Factor-Eliminating Technical Change*

Pietro F. Peretto
Economics Department
Duke University
Durham, NC 27708

John J. Seater
Economics Department
North Carolina State University
Raleigh, NC 27695

March 2007

Abstract

Endogenous growth requires that non-reproducible factors of production be either augmented or eliminated. Attention heretofore has focused almost exclusively on augmentation. In contrast, we study factor elimination. Maximizing agents decide when to reduce the importance of non-reproducible factors. We use a Cobb-Douglas production function with two factors of production, one reproducible ("capital") and one not ("labor"). There is no augmenting progress of any kind, thus excluding the standard engine of growth. What is new is the possibility of changing factor intensities endogenously by spending resources on R&D. The economy starts with no capital and no knowledge of how to use it. By conducting R&D, the economy learns new technologies that use capital, which then is built. There are two possible ultimate outcomes: the economy may achieve perpetual growth, or it may stagnate with no growth. The first outcome is an asymptotic version of the AK model of endogenous growth, and the second outcome is the standard Solow model in the absence of any exogenous sources of growth. Which outcome is achieved depends on parameter values of saving and production, and there always is a feasible saving rate that will give the perpetual growth outcome. The model thus provides a theory of the endogenous emergence of a production technology with constant returns to the reproducible factors, that is, one that is capable of supporting perpetual economic growth. The model also allows derivation of the full transition dynamics, which have interesting properties. One especially notable feature is that the origin is not a steady state. An economy that starts with pure labor production becomes industrialized through its own efforts. The theory thus offers a purely endogenous explanation for the transition from a primitive to a developed economy, in contrast to several well-known theories. Several aspects of the transition paths accord with the evidence, suggesting that the theory is reasonable. In contrast to almost all the existing endogenous growth literature, neither monopoly power nor an externality is a necessary condition for endogenous growth. It is sufficient that firms be able to appropriate the results of their research and development efforts.

JEL Codes: O40, O31, O33

Keywords: Endogenous growth, technical change, factor intensity choice

* We thank the following people for comments that improved the paper: Otilia Boldea, Diego Comin (our discussant at the 2007 AEA meetings); Peter Howitt; Oksana Leukhina, Hernando Zuleta; and participants in the University of North Carolina macroeconomics workshop and the DEGIT XI conference in Jerusalem.

1 Introduction

Perpetual growth of income per capita requires that the marginal products of all reproducible factors of production be bounded away from zero. Virtually all theoretical investigations achieve this necessary condition by augmenting non-reproducible factors, the best-known example undoubtedly being Harrod-neutral, labor-augmenting technical progress. Empirical investigations naturally follow the lead of the theory. There is, however, another way to satisfy the necessary condition for growth: eliminate the non-reproducible factors from the production function. If society can learn to produce without non-reproducible factors, it can grow perpetually. In this paper, we propose an endogenous theory of factor elimination and examine its implications for economic growth and other issues.

Non-reproducible factors act as a drag on the marginal products of reproducible factors. As the ratio of reproducible to non-reproducible factors rises, the marginal products of the reproducible factors fall until they reach sufficiently low values that further accumulation of those factors no longer is economically justified. At that point, growth stops. Augmentation offsets this drag by effectively increasing the amounts of the non-reproducible factors, thereby raising the marginal products of the reproducible factors and thus permitting their accumulation to continue.

It is useful for later intuition to discuss this argument formally. The generic production function is $Y = F(K, L)$, where Y, K, and L are the aggregate amounts of output, capital, and labor, and F satisfies the usual neoclassical assumptions (see, e.g., Barro and Sala-i-Martin 2004, chapter 2, for a list). Suppose population is constant. As K grows, the marginal product of capital shrinks to the point where the marginal benefit of capital accumulation just equals its marginal cost, bringing growth to a halt.[1] The property that guarantees this outcome is the Inada condition that

$$\lim_{K \to +\infty} F_K(K, L) = 0.$$

To generate perpetual growth of income per capita the Solow and Cass models introduce augmentation of labor. The production function has the form $F(K, AL)$, where A is labor-augmenting knowledge that grows at the exogenous rate g. Perpetual growth is feasible because the endowment of "effective" labor AL grows over time and drives up the marginal product of capital, sustaining incentives to accumulation. Specifically, the linear homogeneity of F allows us to write

$$\lim_{K \to +\infty} F_K\left(\frac{K}{A}, L\right)$$

and the ratio K/A remains finite since in steady state K and A grow at the same rate g. This theory is better called a theory *with* growth than a theory *of* growth because growth arises from strictly exogenous forces that the theory makes no attempt to explain. The great advance of endogenous growth theory is precisely to endogenize technical progress. For example, in Romer's (1986) path-breaking model of learning-by-doing, A is proportional to K, so that we have $F(K, KL)$. Variety expansion and quality ladder models have the same property, augmenting the non-reproducible factor and thus enabling perpetual growth.[2]

Elimination of the non-reproducible factors achieves the same end by an entirely different mechanism. The key insight is a straightforward implication of the reason why non-reproducible factors act as a drag on growth. One of the usual Inada conditions is that the marginal product of capital approaches zero as capital approaches infinity. That condition has an important and intuitive basis: it is equivalent to imposing *essentiality* of the non-reproducible factors of production. We can write

$$\lim_{K \to +\infty} F_K(K, L) = \lim_{K \to +\infty} \frac{F(K, L)}{K} = \lim_{K \to +\infty} F\left(1, \frac{L}{K}\right).$$

[1] Of course, if L grows at rate n, then K and Y too grow perpetually at that rate; however, Y/L cannot grow because any attempt to make K grow faster than L leads to a rise in k and thus a drop in the return to K.

[2] See chapters 6 and 7 in Barro and Sala-i-Martin (2004) for a discussion of these models.

This first equality follows from L'Hopital's rule. It shows that the Inada condition is equivalent to the condition that the average product of capital goes to zero, which in turn implies in the Solow model that the growth rate of capital goes to zero. The second equality says that the average product of capital goes to zero because labor is essential, that is, because $F(1,0) = 0$. The Inada condition is equivalent to the essentiality of the non-reproducible factor. It follows that theories of perpetual growth are, in essence, theories of how economic agents overcome scarcity of essential but non-reproducible factors of production. In effect, augmentation overcomes scarcity of a non-reproducible factor by transforming that factor into a reproducible one. In contrast, elimination overcomes the same scarcity by dispensing with the non-reproducible factor altogether.

This growth-through-elimination somewhat resembles growth-through-substitution that arises from a CES production function with a high elasticity of substitution. In the latter type of model, reproducible factors can be substituted for non-reproducible factors fast enough to allow growth to persist, even though there is no augmentation of non-reproducible factors.[3] The main weakness of that theory, however, is that growth is simply a matter of chance, with R&D being totally irrelevant. If mankind has been endowed with a sufficiently high elasticity of substitution, then the technology for producing with less of the non-reproducible factor is available for free. All the economy needs to do is buy reproducible factors to substitute for the non-reproducibles. If the elasticity of substitution is too low, then no amount of R&D can alter the situation, and long-run growth cannot occur. R&D has no role to play. A more satisfying theory would be one in which people could eliminate a non-reproducible factor only by committing resources to the task, which they would do if the benefits of elimination exceed the cost. In what follows, we develop such a theory.

In our theory, maximizing agents decide when it is optimal to reduce the importance of a non-reproducible factor. To keep matters as simple as possible, we use a Cobb-Douglas production function with two factors of production, one reproducible (e.g., physical and/or human capital) and one not (e.g., unskilled labor and/or natural resources). There is no factor-augmenting technical progress of any kind, whether Hicks, Harrod, or Solow neutral, so the standard engine of growth is excluded by construction. What is new is the possibility of changing *factor intensities* (i.e., factor exponents) by devoting resources to R&D. The model is eminently tractable, allowing us to characterize the full transition dynamics. There are two possible ultimate outcomes, depending on a few parameters. One possibility is an economy with perpetual growth and whose aggregate production function asymptotically becomes AK. The other possibility is an economy that settles on a steady state with no growth and a standard aggregate production function with fixed factor intensities bounded away from 0 or 1. If the economy has a sufficiently high saving rate, it will achieve the perpetual growth outcome, and such a saving rate always is feasible. Because the asymptotic AK model is an endogenous outcome of a rational resource allocation process, our theory offers a route to perpetual endogenous growth that avoids most, or perhaps even all, of the criticisms proposed by the skeptics of endogenous growth theory, such as Solow (1994), as discussed below.

The model also offers an interesting theory of economic development and the transition from primitive to advanced technologies. Existing theories suppose that society is endowed from the start with two production functions, one primitive and one advanced, and over time reallocates resources from the former to the latter. For example, Hansen and Prescott (2002) posit primitive and advanced production functions (the "Malthus" and "Solow" technologies) and study a transition driven by the fact that the exogenous rates of technical progress in the two technologies differ. Goodfriend and McDermott (1995) present a theory of development in which the transition from primitive to advanced production is driven by the growth of ideas, which in turn is driven by exogenous population growth through a scale effect. Even Galor and Weil's (2000) model, which has only one technology, relies on an initial episode of costless technical progress to kick the economy out of its initial position of very low productivity, with no growth, and put it on a path of self-sustaining endogenous growth. In our theory, by contrast, there is no exogenous technical progress, no scale effect, and only one initial technology. Alternative technologies appear endogenously, arising only if people use their resources to invent them.

[3] See Chapter 2 of Barro and Sala-i-Martin (2004) for a discussion of this model.

2 A Theory of Factor Elimination

There are three groups of agents: households, producers of final goods, producers of intermediate goods. The number of intermediate goods is fixed and therefore not a source of economic growth. In addition to production, intermediate firms engage in capital accumulation and R&D. The novel feature of our analysis is the nature of the technical change that results from R&D.

2.1 Households

Households own a fixed endowment of a non-reproducible factor. We call this factor "unskilled labor" – just "labor" for short – but it could be any non-reproducible factor, such as land, that limits growth. Households also own firms and receive as dividend income the profits that they generate. We assume that households behave as in the Solow model, supplying labor services inelastically in a competitive market and saving a fixed fraction, s, of their total income (wages plus profits). These assumptions allow us to ignore intertemporal utility maximization and the resulting consumption-saving and labor-leisure choices, greatly simplifying the analysis.

2.2 The Final Good Sector

There is one final good, Y, produced by competitive firms according to the technology

$$Y = \left[\int_0^1 X_i^{\frac{\varepsilon-1}{\varepsilon}} di\right]^{\frac{\varepsilon}{\varepsilon-1}}, \quad \varepsilon > 1, \tag{1}$$

where X_i is the quantity of intermediate good i, ε is the elasticity of substitution between intermediate goods, and we posit a fixed continuum of intermediate goods ranging from 0 to 1 (i.e., there is no variety expansion). The final good is the numeraire, $P_Y \equiv 1$. The final producers maximize profit

$$\pi_Y = Y - \int_0^1 P_i X_i di$$

subject to (1), where P_i is the price of good i. The solution to this problem is the well known demand function

$$X_i = Y \frac{P_i^{-\varepsilon}}{P_X^{1-\varepsilon}}, \tag{2}$$

where

$$P_X = \left[\int_0^1 P_i^{1-\varepsilon} di\right]^{\frac{1}{1-\varepsilon}}$$

is the price index of intermediate goods. Since final producers are competitive and earn zero profit, we have $P_X = P_Y = 1$ and we can drop the index P_X from (2) in the remainder of the analysis.

2.3 The Intermediate Goods Sector

The intermediate goods sector is populated by monopolistically competitive firms that do three things: produce the intermediate goods, invest in capital accumulation, undertake factor-eliminating R&D.

2.3.1 Technologies

We begin with a discussion of the technologies available to the typical firm. To keep the notation simple, we suppress time arguments whenever confusion does not arise.

Production The typical firm hires labor from the households and combines it with its own capital according to the technology

$$X_i = AK_i^{a_i} L_i^{1-a_i}, \quad A > 0. \tag{3}$$

The firm chooses the factor-intensity parameter a_i out of a set of known technologies represented by the interval $a \in [0, \alpha]$, $\alpha < 1$. The upper boundary of this set – the technology frontier – evolves over time as a result of the firm's R&D, discussed momentarily. The total factor productivity (TFP) parameter A, in contrast, is exogenous, constant and common to all firms.[4]

We posit Cobb-Douglas technologies to impose in the simplest possible way *essentiality* for all $a_i \in (0, 1)$.[5] Note, however, that $a_i = 0$ yields $X_i = AL_i$ so that capital is non-essential, whereas $a_i = 1$ yields $X_i = AK_i$ so that labor is non-essential. In our analysis, the fact that the limit case $a_i = 0$ is included in the set of known technologies has profound implications for the economy's dynamics. Whether the economy reaches the limit case $a_i = 1$ is crucial for perpetual growth.

We use the term "capital" in a very broad sense to include all types of reproducible factors of production. Thus we make no distinction between physical and human capital: we include both in our variable K. It would be desirable to extend the theory to distinguish between at least these two broad types of reproducible factors, but doing so is beyond the scope of the present paper.

Note also that referring to a as the "capital share" is not correct because our monopolistic intermediate firms earn excess profit and payments to the two factors of production, K and L, do not exhaust their revenues. The correct term for a is the "elasticity of output with respect to capital", which is rather cumbersome, or the simpler "capital intensity", which is what we use in the remainder of the paper.

Capital Investment The firm's capital stock evolves according to the usual accumulation equation

$$\dot{K}_i = I_i - \delta K_i, \tag{4}$$

where I is gross capital investment in units of the final good and δ is the depreciation rate. To fix terminology, we shall refer to this activity as simply "investment" and to research and development as "R&D".

Research & Development The firm conducts R&D to increase its highest known capital intensity α. We posit the simplest possible research and development technology

$$\dot{\alpha}_i = R_i \tag{5}$$

where R is R&D expenditure in units of the final good.

[4] Sato and Beckmann (1968) provided early evidence that factor intensities in an aggregage Cobb-Douglas production function are not constant and even found that for Japan such function best fit the data among fourteen forms of production function tried. Nonetheless, the only early theoretical investigation of a Cobb-Douglas function with endogenously varying factor intensities was by Kamien and Schwartz (1968). Their model differs substanially from ours in that they restrict attention to an atomistic firm facing fixed factor prices, whereas we study a fully specified general equilibrium model with prices responding to the effects of R&D. Our results are markedly different from theirs, as we discuss below. Seater (2005) studied a model similar in spirit to ours but in which a central planner makes all prodURction and R&D decisions. Seater's results are much more limited than ours because the complicated dynamics emerging from his formulation make characterization of transition path impossible and because imperfect competition is excluded from the analysis. Finally, Zuleta (2006) also examines a model in which R&D alters factor intensities. Zuleta's model incorporates some of the same ingredients as ours but develops along quite different lines because it addresses different issues. Zuleta focuses on the evolution of factors' aggregate income shares whereas we focus on factor elimination and its implications for long-run growth. We discuss the relation of our work to Zuleta's below.

[5] If we posited a CES technology we would need to impose an elasticity of subsitution between capital and labor smaller than 1 to ensure that labor holds down the marginal product of capital. The reason is that with elasticity of subsitution larger than 1 neither capital nor labor is essential and the AK limit can arise even though there is no expenditure on R&D.

2.3.2 The Firm's Decisions

The firm chooses paths of capital intensity, the price of its product, employment, investment and R&D expenditure to maximize the present value of its dividend payments:

$$\max_{\{a_{it},P_{it},L_{it},I_{it},R_{it}\}_{t=0}^{\infty}} \int_0^{\infty} \left(P_{it}X_{it} - w_t L_{it} - I_{it} - R_{it}\right) e^{-\bar{r}_t t} dt,$$

where $\bar{r}_t \equiv \frac{1}{t}\int_0^t r_u du$ is the average interest rate between time 0 and time t, r_u is the instantaneous interest rate at time u, and the optimization is subject to (2), (3), (4), (5) and the restrictions $I_{it} \geq 0$ and $R_{it} \geq 0$. The individual intermediate firm perceives no upper bound on its choices of I or R. In the aggregate, of course, firms' choices must satisfy the constraint $\int_0^1 (I_{it} + R_{it})di = sY_t$ at every time t.

It is convenient to think of the firm as operating two divisions: production and investment. We then rewrite the firm's objective function in a way that reflects that internal structure:

$$\max_{\{a_{it},P_{it},L_{it},I_{it},R_{it}\}_{t=0}^{\infty}} \int_0^{\infty} [(P_{it}X_{it} - w_t L_{it} - r_{Ki}K_{it}) + (r_{Ki}K_{it} - I_{it} - R_{it})]e^{-\bar{r}_t t} dt.$$

The term inside the first set of parentheses is the instantaneous profit of the production division; the term inside the second set of parentheses is the instantaneous profit of the investment division. The production division rents capital from the investment division, paying an internal transfer rate r_K. The investment division receives that rental income and spends resources on investment and R&D. We can then exploit time-separability and solve the firm's maximization problem in two steps. First, the production division chooses the optimal values of P, L, and K taking w, r_K and α as given. Then the investment division chooses I and R.

There is no explicit payment for technology, whose return is included in the total rental income accruing to the investment division, $r_K K$. The investment produces K and α, which affect the total return according to $d\left(r_K K\right) = \left[\partial\left(r_K K\right)/\partial K\right] dK + \left[\partial\left(r_K K\right)/\partial\alpha\right] d\alpha = r_K dK + (\partial r_K/\partial\alpha) K d\alpha$. The second term in this last expression captures the implicit payment to technology. The intuition is that the more capital intensive technology allows the production division to use its capital and labor more efficiently and thus produce more. This extra output makes the production division willing to pay more for capital, and the increased payment for capital compensates the investment division for the R&D that made possible the increase in capital efficiency.

Choice of a, P, L and K We make the usual symmetry assumption that intermediate firms are identical. Henceforth, we omit the i subscript except where clarity requires it. The production division solves

$$\max_{\{a_t,P_t,L_t,K_t\}_{t=0}^{\infty}} \int_0^{\infty} (P_t X_t - w_t L_t - r_{Kt} K_t) e^{-\bar{r}_t t} dt$$

subject to (2) and (3).

Note that by splitting the firm into a production division and an investment division, we have isolated the firm's intratemporal decisions from its intertemporal ones. Note also that price setting and quantity setting are equivalent decisions for a monopolist so that the choice of inputs, which determines the quantity produced, also determines the price at which that output sells. Hence, we can use the demand curve (2) to eliminate P and think of the production division as facing a sequence of independent instantaneous profit maximization problems of the form

$$\max_{a,L,K} \pi = Y^{\frac{1}{\varepsilon}} X^{1-\frac{1}{\varepsilon}} - wL - r_K K$$

subject to (3).

This setup posits a technology choice problem. As α rises in response to R&D, the firm does not forget the lower-a technologies in the interior of the set $[0, \alpha]$ that were previously in use. Given that the firm knows technologies in the interval $a \in [0, \alpha]$, which does it use? All of them? Only that with the highest a? The following proposition due to Zuleta (2006) provides the answer.

Proposition 1 *(Zuleta 2006, Proposition 1). A firm that has available Cobb-Douglas technologies with constant returns to scale and capital intensities in the range $a \in [\alpha_{\min}, \alpha_{\max}]$ uses only one of the following three possible technologies: (1) only that with the lowest capital intensity, (2) only that with the highest capital intensity, or (3) only the two with the lowest and highest capital intensity.*

The intuition behind this result is straightforward. Let k be the firm's capital/labor ratio, and consider two values of a, $a_1 < a_2$. If $k < 1$, then $k^{a_1} > k^{a_2}$ and the firm will choose the lowest a at its disposal. The opposite holds if $k > 1$. It might thus seem that the firm would use only the technology with the lowest or the highest a. That would be true if the firm could use only one of the technologies available to it. If, however, the firm can operate more than one technology simultaneously, it will split its labor force between the two technologies with the highest and lowest values of a. Doing so maximizes the firm's total output. In certain constrained cases that we discuss below, the firm may choose to operate only one of the two extreme technologies rather than both.

In our analysis, we set $\alpha_{\min} = 0$, thus positing that our economy starts with a technology that uses only labor. As time passes, the economy develops – at a cost – increasingly capital intensive technologies, so that $\alpha_{\max} = \alpha > 0$. The firm's production division divides its labor force between a plant that uses the "primitive" technology $a = 0$ and one that uses the most "advanced" technology $a = \alpha$, continuously revising that allocation as new technologies become available, as we now explain.

The firm's total output is the sum of the output from its two plants:

$$X = A\left[(L - l) + \left(K^{\alpha} l^{1-\alpha}\right)\right], \tag{6}$$

where L is the firm's total employment, $l \in [0, L]$ is labor allocated to the advanced plant and $L - l$ labor allocated to the primitive plant. Note that this expression removes essentiality of capital for all $\alpha < 1$, not just for $\alpha = 0$. This provides an interesting ingredient to our development story. An economy that knows only the most primitive technology (i.e., an economy for which $\alpha = 0$) would build no capital and produce with labor only. A more advanced economy with $\alpha > 0$ and some capital K can still use the primitive technology and would produce output even if the capital were to disappear. Thus, thinking of capital intensity as a *choice* – as opposed to an exogenous parameter – changes radically the properties of an otherwise very conventional economy.

The production division's problem now is

$$\max_{l,L,K} \pi = Y^{\frac{1}{\varepsilon}} X^{1-\frac{1}{\varepsilon}} - wL - r_K K$$

subject to (6). The first-order conditions are:

$$Y^{\frac{1}{\varepsilon}} X^{1-\frac{1}{\varepsilon}} \left(1 - \frac{1}{\varepsilon}\right) \frac{1}{X} \frac{\partial X}{\partial K} = r_K; \tag{7}$$

$$Y^{\frac{1}{\varepsilon}} X^{1-\frac{1}{\varepsilon}} \left(1 - \frac{1}{\varepsilon}\right) \frac{1}{X} \frac{\partial X}{\partial L} = w; \tag{8}$$

$$Y^{\frac{1}{\varepsilon}} X^{1-\frac{1}{\varepsilon}} \left(1 - \frac{1}{\varepsilon}\right) \frac{1}{X} \frac{\partial X}{\partial l} = 0. \tag{9}$$

The first condition is the firm's demand for capital, the second is the firm's demand for labor, and the third gives the efficient allocation of labor across the two plants. There is a constraint that we must consider that complicates discussion of these conditions and of the rest of the analysis. Rewrite equation (9) as

$$0 = \frac{\partial X}{\partial l} = A\left[-1 + (1-\alpha)\left(\frac{K}{l}\right)^{\alpha}\right]$$

$$\Rightarrow l = K(1-\alpha)^{\frac{1}{\alpha}} \tag{10}$$

The problem is that the right side of (10) may exceed the firm's total employment, L. In that case, the firm would allocate all its labor to the advanced plant. This possibility becomes relevant in general equilibrium, when the population may act as a binding constraint on the representative firm's total employment. The constraint can be written in several useful ways:

$$K(1-\alpha)^{\frac{1}{\alpha}} \leq L \Leftrightarrow K \leq L\left(\frac{1}{1-\alpha}\right)^{\frac{1}{\alpha}} \Leftrightarrow \frac{K}{L} \leq \left(\frac{1}{1-\alpha}\right)^{\frac{1}{\alpha}} \tag{11}$$

We must write the first-order condition (9) in two parts, one for the unconstrained case and one for the constrained case:

$$l = \begin{cases} K(1-\alpha)^{\frac{1}{\alpha}} & K(1-\alpha)^{\frac{1}{\alpha}} < L \\ L & K(1-\alpha)^{\frac{1}{\alpha}} \geq L \end{cases}. \tag{12}$$

In the same way, equations (7) and (8) must be written in two parts:

$$r_K = \begin{cases} Y^{\frac{1}{\varepsilon}} X^{1-\frac{1}{\varepsilon}} \left(1-\frac{1}{\varepsilon}\right) \frac{1}{X}\left[\alpha A\left(\frac{K}{l}\right)^{\alpha-1}\right] & K(1-\alpha)^{\frac{1}{\alpha}} < L \\ Y^{\frac{1}{\varepsilon}} X^{1-\frac{1}{\varepsilon}} \left(1-\frac{1}{\varepsilon}\right) \frac{1}{X}\left[\alpha A\left(\frac{K}{L}\right)^{\alpha-1}\right] & K(1-\alpha)^{\frac{1}{\alpha}} \geq L \end{cases}; \tag{13}$$

$$w = \begin{cases} A & K(1-\alpha)^{\frac{1}{\alpha}} < L \\ Y^{\frac{1}{\varepsilon}} X^{1-\frac{1}{\varepsilon}} \left(1-\frac{1}{\varepsilon}\right) \frac{1}{X}\left[(1-\alpha)\left(\frac{K}{L}\right)^{\alpha}\right] & K(1-\alpha)^{\frac{1}{\alpha}} \geq L \end{cases}. \tag{14}$$

Note that, in the unconstrained version of these alternatives, the capital/labor ratio is K/l, whose denominator is *not* the firm's total labor employment L (and thus in general equilibrium is *not* the economy's labor endowment) because the firm splits total labor across the two plants. It is important for what follows to understand that (11) constrains the relation between k and α, not the absolute size of the capital stock K. As $\alpha \to 1$, the term $(1-\alpha)^{-1/\alpha} \to \infty$, so that K, K/l, and K/L all can be arbitrarily large.

Figure 1 shows the optimal labor allocation a given α and two different values of K. The horizontal line marked MPL_0 is the marginal product of labor in the primitive plant, the downward sloping line marked MPL_α is the marginal product of labor in the advanced plant. The vertical line is the firm's total employment. Anticipating the properties of the general equilibrium of this model we note that L is also the economy's labor endowment. Figure 1 tells us two things. First, the firm does not shut down the primitive plant unless capital is large. Second, when both plants operate, the firm allocates labor across plants to equalize the marginal products of labor. To see why, suppose that the firm allocates very little labor to the advanced plant. Because l is small, K/l is large and the marginal product of labor in the advanced plant is large. That, however, means the marginal product of capital is small, and the firm is better off increasing l and accepting the reduction in MPL_α in order to have a larger MPK_α. This solution is feasible if desired l does not exceed the firm's total employment. If it does, the constraint $l = L$ binds and the firm shuts down the primitive plant. With sufficiently abundant capital $MPL_\alpha > MPL_0$ for all $l \leq L$.

To see the advantage of keeping the primitive plant operational for small K, observe that the unconstrained part of (12) can be written

$$k = \left(\frac{1}{1-\alpha}\right)^{\frac{1}{\alpha}} > 1. \tag{15}$$

where $k \equiv K/l$. This relation tells us that the firm splits labor across the two plants so that the capital/labor ratio in the advanced plant is (a) independent of factor prices, (b) always larger than 1, (c) increasing in capital intensity α. Property (a) implies that the capital/labor ratio in the advanced plant is different from and independent of the economy's endowment ratio K/L. Property (b) implies that the plant's output is increasing in α. That fact provides the underlying rationale both for using only the $a = \alpha$ technology out of the set $(0, \alpha]$ and for pursuing α-increasing innovations. Property (c) implies that the development of more capital intensive technologies drives up the optimal capital/labor ratio and thus the incentive to accumulate capital. Together, (b) and (c) say that there exists a positive feedback between capital accumulation and the pursuit of higher capital intensity. This feedback explains why the advanced technology becomes AK in the long run, as we explain in detail later.

The intuition for the positive feedback is that, by developing more capital intensive technologies, the firm uses the available capital and labor more efficiently. Specifically, the definition of k allows us to rewrite (6) as

$$X = A\left(L + K\frac{k^{\alpha} - 1}{k}\right).$$

It is then straightforward to show that

$$\left(\frac{1}{1-\alpha}\right)^{\frac{1}{\alpha}} = \arg\max_{k} \frac{k^{\alpha} - 1}{k}.$$

The left side of this equation is the same as the right side of (15), which in turn is a rewriting of the unconstrained part of the first-order condition (12). Thus the firm's (unconstrained) optimal allocation of its total labor L between the primitive and advanced plant is that which maximizes the marginal product of capital in the advanced plant, given the stock of capital available to the firm at the time. Any remaining labor is employed in the primitive technology. As already explained, this maximizing value of k depends on α alone and so does not diminish as K is accumulated. The reason is that the primitive plant provides the firm with a "labor pool" that it can use to keep the maximized marginal product of capital constant even as capital is accumulated. In addition, as we explain later, that same "labor pool" guarantees that it is always desirable to increase α.

It is useful for later discussion to define the maximized value of $(k^{\alpha} - 1)/k$ as

$$m(\alpha) \equiv \max_{k} \frac{k^{\alpha} - 1}{k} = \alpha(1-\alpha)^{\frac{1-\alpha}{\alpha}}. \tag{16}$$

The function $m(\alpha)$ has the following properties: $m' > 0$ and $m'' > 0$ for all $\alpha \in [0, 1]$, $m(0) = 0$, $m(1) = 1$, and $m'(1) = \infty$ (see the Appendix for details). The marginal product of capital in the advanced plant is $A(k^{\alpha} - 1)/k$, and its maximized value is $Am(\alpha)$. We call $m(\alpha)$ the "maximized discretionary marginal product," i.e., the part of the maximized marginal product that the firm can affect.

Choice of I and R We complete our description of the firm's behavior by looking at its intertemporal investment and R&D decisions. The investment division, taking the production division's decision rules as given, chooses I and R to maximize the firm's present value:

$$\max_{\{I_t, R_t\}_{t=0}^{\infty}} \int_0^{\infty} (\pi_t + r_K K_t - I_t - R_t) e^{-\bar{r}_t t} dt,$$

subject to (4), (5) and initial conditions. The current-value Hamiltonian is

$$H = (\pi + r_K K - I - R) + \psi(I - \delta K) + \phi R + \omega_1 I + \omega_2 R, \tag{17}$$

where ψ and ϕ are the costate variables corresponding to K and α, respectively, and ω_1 and ω_2 are Lagrange multipliers satisfying the Kuhn-Tucker conditions:

$$\begin{aligned} \omega_1 &\geq 0, \quad I \geq 0, \quad \omega_1 I = 0; \\ \omega_2 &\geq 0, \quad R \geq 0, \quad \omega_2 R = 0. \end{aligned}$$

We present the full set of necessary conditions for this problem in the Appendix. It is sufficient for the discussion here to state that they yield the following expressions for the returns to capital and R&D

$$r = r_K - \delta,$$

where r_K is given by (13), and

$$r = r_\alpha \equiv \frac{\partial \pi}{\partial \alpha} = Y^{\frac{1}{\varepsilon}} X^{1-\frac{1}{\varepsilon}} \left(1 - \frac{1}{\varepsilon}\right) \frac{1}{X} \frac{\partial X}{\partial \alpha}. \tag{18}$$

The Hamiltonian (17) is linear in the control variables I and R, so we have a bang-bang control problem. That means one of I or R is zero and the other is set at its possible maximum value, determined by the aggregate resource constraint, as explained below. Specifically, whenever $I > 0$ we have $\omega_1 = 0$ and thus $\psi = 1$. Similarly, whenever $R > 0$ we have $\omega_2 = 0$ and $\phi = 1$. The fact that $\psi = 1$ and $\phi = 1$ at all times effectively reduces the dimension of the state space from four to two because instead of having to solve for the paths of K, α, ψ and ϕ, we need to solve only for the paths of K and α.

2.4 Taking Stock: The "hybrid" AK model

Before we discuss the model's general equilibrium solution, it is useful to assess what we have so far. The main innovations of this paper are that:

a. the firm chooses its capital intensity a out of a set of known technologies $[0, \alpha]$;

b. the firm invests in R&D to expand the technology frontier α.

The first innovation changes radically the production structure of the economy and gives rise to what we refer to as the "hybrid AK" model. The second innovation has important implications for long-run growth. In this subsection we focus on the main features of the hybrid AK model. We discuss its dynamic implications in the next section.

The efficiency condition (12) has the striking implication that there exists a region of (α, K)-space where output is *linear* in labor and capital separately. To see this, we use (12) to rewrite (6) as

$$X = \begin{cases} A\left[L + m\left(\alpha\right) K\right] & K < L\left(\frac{1}{1-\alpha}\right)^{\frac{1}{\alpha}} \\ AK^{\alpha} L^{1-\alpha} & K \geq L\left(\frac{1}{1-\alpha}\right)^{\frac{1}{\alpha}} \end{cases}. \tag{19}$$

Recall that the term $m\left(\alpha\right)$ is the maximized (discretionary) marginal product contributed by the advanced plant.

Figure 2 illustrates various aspects of the production technology. The upper panel shows the choice of the optimal labor allocation. By operating simultaneously two plants with $a = 0$ and $a = \alpha$, the firm does not follow the upper envelope of the two technologies (the very heavy dotted curve that first runs along the labor-intensive technology and then along the capital-intensive technology), switching from one to the other at $K/L = 1$. Instead, the firm's operates the two technologies simultaneously, weighted average "total" technology of the two underlying technologies with optimal weights determined by the allocation of labor across plants. The weighted average is shown by the straight line running from the point A on the vertical axis to the tangency point T. To the right of point T the firm no longer can follow the straight line because the constraint (11) binds, so it then follows the curved function by shutting down the labor-intensive technology and devoting all its labor to the capital-intensive plant. This behavior convexifies the overall production function. The profile of the overall production function captures the fundamental idea that the firm, confronted with a menu of technologies having different capital intensities, maximizes output by using both the lowest, $a = 0$, and the highest, $a = \alpha$. Restricting production to just one of the two technologies

would not maximize output. The lower panel show how the choice of factor intensity from the set $a \in [0, a]$ changes the structure of production. As the α-innovations arrive, the production function for the capital-intensive plant rotates and flattens, becoming increasingly linear until it reaches the limit $m(1) = 1$ and is linear everywhere. The tangency point of the "average" technology and the capital-intensive technology moves right as α grows, going to infinity as α goes to 1.

The unconstrained part of the solution to the firm's allocation problem, given by first half of (19), has two characteristics that are important for the economy's dynamics that we discuss in the next section. First, for any given α, capital has *constant* rather than diminishing returns. For given K, the firm uses the labor-intensive technology to absorb any excess labor beyond that necessary to yield the output-maximizing capital/labor ratio in the capital-intensive plant. As a result, the firm always is willing to invest in the marginal unit of capital. Second, the presence of the labor-intensive plant allows the firm to keep the capital/labor ratio in the capital-intensive plant above one, that is, $k > 1$. An increase in α then increases output, thus making the firm always willing to invest in R&D to generate a marginal increase in α. These two aspects of the solution reinforce each other in a way that tends to generate perpetual growth. If the firm had to use all its labor in the capital-intensive plant, neither of these conditions would hold, and the economy would be unable to leave a neighborhood of the origin. These issues are pursued further and explained in detail in the next section.

This reasoning brings to the forefront of our story the notion of "localized elimination" and "separation" (or "segregation" if one finds this word more descriptive) of the factors of production. The firm invests in R&D to learn how to produce without labor ($\alpha = 1$). It thus eliminates labor from the advanced plant and thereby creates endogenously the engine of perpetual growth – a technology that (asymptotically) produces with reproducible inputs only. This elimination is "local" in the sense that it affects only the advanced plant, not the overall production function of the firm. Labor, in fact, is a productive resource that the firm knows how to use, with or without capital. It obviously is not optimal to leave labor idle under these conditions, so the firm completely "separates" labor from capital and thus removes diminishing returns altogether from its technology.

Note that we talk about a "firm" allocating labor to two different "plants" because we found useful to set up the problem with only one decision maker. Is should be obvious that we can decentralize decisions and talk about an "economy" that allocates labor to two different "sectors" linked by the arbitrage condition that wages be equalized. In that case, the reallocation of labor from the primitive to the advanced plant and vice versa is done by the market. The thrust of the story does not change: the economy endogenously creates perpetual growth through R&D investment that brings into existence a technology that eventually uses only capital. As labor demand falls in the capital-using sector, labor is absorbed by the primitive sector. As we shall see, the process ends when capital and labor are fully separated so that neither sector is subject to diminishing returns.

2.5 General Equilibrium Dynamics

To study the general equilibrium dynamics of our economy we need to clear four markets: intermediate goods, final good, labor and assets. The market for intermediate goods is in fact a continuum of monopolistic markets wherein equilibrium follows from the fact that each producer sets its price and thereby chooses a point on the demand curve it faces. Moreover, because the mass of firms is set at 1 (recall the limits of integration in equation (1)), we have that in symmetric equilibrium

$$PX = Y = X \tag{20}$$

and thus $P = 1$. Households sell their labor services inelastically in a competitive market. In equilibrium, therefore, aggregate employment is equal to the economy's labor endowment, L, and the wage rate is the value marginal product of labor

$$w = \begin{cases} \frac{\varepsilon-1}{\varepsilon} A & K < L\left(\frac{1}{1-\alpha}\right)^{\frac{1}{\alpha}} \\ \frac{\varepsilon-1}{\varepsilon} A \left(\frac{K}{L}\right)^{\alpha} & K \geq L\left(\frac{1}{1-\alpha}\right)^{\frac{1}{\alpha}} \end{cases} .$$

Given our assumption that households save a constant fraction, s, of their income, the market-clearing condition for the final goods market is

$$(1-s)Y + R + I = Y.$$

Using (4), (5) and (19), we can write this relation as

$$\dot{\alpha} + \dot{K} + \delta K = \begin{cases} sA\left[L + m\left(\alpha\right)K\right] & K < L\left(\frac{1}{1-\alpha}\right)^{\frac{1}{\alpha}} \\ sAK^{\alpha}L^{1-\alpha} & K \geq L\left(\frac{1}{1-\alpha}\right)^{\frac{1}{\alpha}} \end{cases}. \tag{21}$$

There are two assets in this economy, physical capital and technology, both accumulated by the firm. The resources constraint above ensures that the flow of saving equals the flow of total investment. Therefore, equilibrium of the assets market only requires the additional condition that the firm be indifferent between allocating resources to R&D or to investment.

Our phase diagram uses two loci. The first is the *arbitrage locus*, along which the firm is indifferent between allocating resources to investment or R&D, that is, where $r_{\alpha} = r_K - \delta$. The second is the *stationarity locus*, along which there is no net investment in either K or α so that the resource constraint (21) is satisfied with both $\dot{K} = 0$ and $\dot{\alpha} = 0$.

In general, the constraint (11) may or may not bind. A sufficient condition to guarantee that it never binds along the economy's adjustment path is $L \geq \alpha$ (see the Appendix for the proof). The variable α is restricted to the interval $[0, 1]$. If we take L to be unskilled labor only and use natural units to count population (i.e., use the census count), then $L \geq 1$ because the smallest possible economy comprises one person. With that interpretation of L, the unconstrained case would be the only reasonable case. The situation is less clear when L is taken to be all non-reproducible factors of production. It turns out, however, that, irrespective of the interpretation of L, the constrained case gives results substantially the same as the unconstrained case. We therefore relegate the constrained case to the Appendix and restrict attention in the main text to the "interior," where $l < L$. To discuss the economy's dynamics, we need to determine the arbitrage and stationarity loci. The following propositions, proven in the Appendix, describe the two loci, and Figures 3 and 4 plot them.

Proposition 2 ***Arbitrage Locus.*** *Assume $L \geq 1 \geq \alpha$. Then the arbitrage locus in (α, K) space is*

$$K = \begin{cases} 0 & 0 \leq \alpha \leq \bar{\alpha} \\ \frac{1}{m'(\alpha)}\left[m\left(\alpha\right) - \frac{\delta\varepsilon}{A(\varepsilon-1)}\right] & \bar{\alpha} < \alpha \leq 1 \end{cases} \tag{22}$$

where $\bar{\alpha}$ solves

$$m\left(\alpha\right) = \frac{\delta\varepsilon}{A\left(\varepsilon - 1\right)}.$$

The locus starts at zero and lies on the α-axis from zero to $\bar{\alpha}$. In the interval $(\bar{\alpha}, 1)$, the locus is positive and hump-shaped. The locus crosses the horizontal axis again at exactly $\alpha = 1$. The locus lies everywhere in the region

$$K < L\left(\frac{1}{1-\alpha}\right)^{\frac{1}{\alpha}}$$

where the interior equilibrium $l < L$ holds.

The arbitrage locus is obtained by setting r_{α} equal to $r_K - \delta$, substituting the expressions for r_{α} and r_K obtained earlier, and performing algebraic manipulations. The resulting function has negative values for $\alpha < \bar{\alpha}$, but of course only non-negative values of K are possible. Consequently, the equilibrium locus lies on the horizontal axis for $0 \leq \alpha \leq \bar{\alpha}$, and $K = 0$ for $\alpha \leq \bar{\alpha}$.

Proposition 3 ***Stationarity Locus.*** *The equation for the stationarity locus is*

$$K = \frac{L}{\frac{\delta}{sA} - m(\alpha)}. \tag{23}$$

The locus has an asymptote at $0 < \tilde{\alpha} < 1$*, where* $\tilde{\alpha}$ *solves*

$$\frac{\delta}{sA} = m(\alpha).$$

For $\alpha < \tilde{\alpha}$*,* $K > 0$*, and the locus starts at* $L\frac{sA}{\delta}$ *and goes asymptotically to* $+\infty$ *as* $\alpha \to \tilde{\alpha}$*. For* $\alpha > \tilde{\alpha}$*,* $K < 0$*, and the locus starts at* $-\infty$ *and reaches the value* $-L\left(\frac{\delta}{sA} - 1\right)^{-1}$ *at* $\alpha = 1$*.*

The equation for the stationarity locus is obtained simply by rearranging the resource constraint (21) under the stationarity conditions $\dot{K} = 0$ and $\dot{\alpha} = 0$.

The equilibrium loci are shown in Figures 3 and 4, the phase diagrams for the economy. Figure 3 shows the case where the equilibrium loci do not intersect, and Figure 4 shows the case where they do. In both Figures, points below the arbitrage locus yield $r_\alpha < r_K - \delta$, so that $\dot{\alpha} = 0$ and $\dot{K} > 0$, whereas all points above the locus yield $r_\alpha > r_K - \delta$, so that $\dot{\alpha} > 0$ and $\dot{K} = 0$. Points on the locus yield indifference between investment and R&D so that the economy experiences both $\dot{\alpha} > 0$ and $\dot{K} > 0$. Points below, above, or on the stationarity locus yield $\dot{\alpha} + \dot{K} > 0$, $\dot{\alpha} + \dot{K} < 0$, and $\dot{\alpha} + \dot{K} = 0$, respectively. We refer to the two possible cases as "high saving" and "low saving" because, given all other parameters, a sufficiently high saving rate will guarantee that the two equilibrium loci do not intersect whereas a sufficiently low saving rate will guarantee that they do. The balanced growth paths and transitional dynamics are somewhat different in the two cases. In the Appendix, we prove that at any time there exist saving rates sufficiently high and low to make each case possible.

2.5.1 Equilibrium Dynamics: High Saving

When the saving rate is sufficiently high, the stationarity locus does not intersect the arbitrage locus, and the situation in Figure 3 prevails. The arbitrage and stationarity loci together divide the phase plane into three regions, labelled I, II, and III. In region I, the rate of return to R&D is less than the rate of return to capital, $r_\alpha < r_K - \delta$, gross capital investment I is positive, and R&D is zero. Total asset accumulation is positive ($\dot{\alpha} + \dot{K} > 0$), so I exceeds depreciation δK and K grows with α constant. The dynamic adjustment paths are vertical lines pointing north. In region II, we still have $\dot{\alpha} + \dot{K} > 0$, but now $r_\alpha > r_K - \delta$ so that R&D is positive and gross capital investment is zero. Hence, α grows while K falls because of depreciation. The resulting dynamic adjustment paths point southeast. Region III is like region II except that total asset accumulation is negative ($\dot{K} + \dot{\alpha} < 0$), meaning that gross investment I, though positive, is smaller than depreciation. As in region II, the dynamic adjustment paths point southeast. For the economy's dynamic behavior, regions II and III are essentially the same.

The precise shape of the economy's dynamic adjustment path depends on where the economy starts. Figure 3 shows the possibilities. Irrespective of which path the economy follows, it eventually reaches the $\alpha = 1$ limit and grows forever. In fact, in our simple model, the economy reaches $\alpha = 1$ in finite time (a property we discuss below), and the aggregate production function becomes $Y = X = A(L + K)$. The aggregate production function is never AK because the primitive sector always operates in order to make use of labor L, but it does go asymptotically to AK as K grows without bound.[6]

[6] The result that α goes to 1 is dramatically different from that of Kamien and Schwartz (1968). In their model, α always is bounded away from 1. Kamien and Schwartz study an atomistic firm that takes prices as given. In their model, changes in α never affect any price, resulting in the absence of important feedback channels. In contrast, in our model, all prices are affected by changes in α through the working of general equilibrium, providing a feedback that keeps R&D going until α reaches its upper bound of 1.

Equation (20) implies that the growth rate of Y equals the growth rate of X, which in turn is given by

$$\begin{aligned}\frac{\dot{X}}{X} &= \frac{Am'(\alpha)K\dot{\alpha} + Am(\alpha)\dot{K}}{A[L + m(\alpha)K]} \\ &= \frac{Am(\alpha)K}{A[L + m(\alpha)K]}\left[\frac{\alpha m'(\alpha)}{m(\alpha)}\frac{\dot{\alpha}}{\alpha} + \frac{\dot{K}}{K}\right] \\ &= \frac{Am(\alpha)K}{A[L + m(\alpha)K]}\left[\left\{\ln\left[\left(\frac{1}{1-\alpha}\right)^{\frac{1}{\alpha}}\right]\right\}\frac{\dot{\alpha}}{\alpha} + \frac{\dot{K}}{K}\right]\end{aligned}$$

The path of growth rate is difficult to analyze while $\alpha < 1$ but is straightforward once α reaches its limit of 1. At that point, the growth rate of X simplifies to

$$\left.\frac{\dot{X}}{X}\right|_{\alpha=1} = \frac{K}{L+K}\frac{\dot{K}}{K}$$

The resource constraint (21) also simplifies to

$$\dot{K} + \delta K = sA(L + K)$$

Combining these last two equations yields

$$\begin{aligned}\left.\frac{\dot{X}}{X}\right|_{\alpha=1} &= sA - \delta\frac{K}{L+K} \\ &\downarrow \quad sA - \delta \text{ as } K \to \infty\end{aligned}$$

Thus, after the economy reaches the boundary $\alpha = 1$, the growth rate declines over time and tends to the AK limit as the ratio K grows without bound. The reduced-form aggregate production function (19) no longer features essentiality of labor because the advanced plant is now AK. Labor is still an input in the overall production structure, but it no longer holds down the marginal product of capital as the economy grows. Once $\alpha = 1$, firms can fully separate the factors of production, putting all labor in a plant that uses no capital and putting all capital into another plant that uses no labor. In other words, the economy has *endogenously created* a technology (equivalently, a production sector) that uses reproducible inputs only and thus satisfies the familiar condition for endogenous growth discussed in Rebelo (1991). This result is the most important of our analysis.

Notice that all these results hold for economies that start exactly at the origin. The reason is that these primitive economies do have output – because capital is not essential – and therefore can produce capital according to the accumulation technology (4) and figure out how to use it according to the research technology (5). Interestingly, they start by building up knowledge first, and only when they have developed technologies with sufficient capital intensity do they start building the capital stock. The reason is that with low values of both K and α the advanced sector does not generate enough output to overcome depreciation, making construction of the capital unprofitable.

2.5.2 Equilibrium Dynamics: Low Saving

In Figure 4, the intersection of the equilibrium loci creates the new region IV. The dynamics in regions I-III are as before. In region IV we have $r_\alpha < r_K - \delta$, and thus no R&D, while total asset accumulation is negative, $\dot{K} + \dot{\alpha} = I - \delta K < 0$, meaning that net capital investment is negative. The equilibrium path is vertical with α constant and K falling. The portion of the stationarity locus below the arbitrage locus, shown as the heavy dashed arc between points **x** and **z** in the Figure, is a set of "Solow" steady states in which the economy does not grow. Recall that there is no exogenous growth in population or technology, which is why the "Solow" solution has no growth.[7]

[7] Readers following the details in the Appendix will note that the same two possibilities of perpetual growth or stagnation occur when the labor constraint binds. If saving is high enough, perpetual growth emerges. The reason is that agents look into

2.6 Relation to Other Theories of Factor Substitution

The theory developed here is related to the one version of endogenous growth theory we know that does not rely on factor augmentation — growth through substitution. The clearest case is that in which the economy has a CES production function. If the elasticity of substitution is sufficiently high, the economy achieves growth by building capital at a rapid rate and substituting it for labor. The aspect of this theory that is unsatisfactory is that economic growth is simply a matter of luck: either the economy is endowed with an high elasticity of substitution sufficiently high to support endogenous growth or it isn't. Furthermore, as a practical matter, modern industrial economies apparently do not have elasticities of substitution above the critical level to sustain endogenous growth, yet they have been growing for centuries and show no sign of slowing down.[8] Our theory does not rely on a fortuitous exogenous endowment of a sufficiently high elasticity of substitution. Instead, the necessary elasticity is created through the R&D process: as soon as the economy has at its disposal both the primitive technology with $a = 0$ and any advanced technology with $a = \alpha > 0$. The economy then operates an overall technology with *infinite* elasticity of substitution between capital and labor; see equation (19).

Several authors have proposed interesting models of economic growth that fundamentally are variations on the CES-factor substitution theme. Boldrin and Levine (2002), Givon (2006), and Zeira (1998, 2006) present related models in which the economy successively adopts technologies that are less and less labor-intensive. Although the models are much more elaborate than the simple CES story, the driving mechanism is the same: capital is substituted for labor. There is no R&D sector or any other resource-absorbing activity that produces the technologies to allow the substitution. The economy simply is endowed with the ability to make the substitution. What seems to be an important element of the history of economic growth and transition is missing. Our model addresses that limitation.

3 Economic Implications

Our theory has several interesting implications.

3.1 The Big Three

Three aspects of this economy's dynamics are especially noteworthy: the nature of the origin, the ultimate form of the production function, and the appearance of the economy at any point on its transition path.

In the Solow and Cass models, the origin is an unstable steady state: an economy starting exactly at the origin stays there and one needs a positive initial endowment of capital to get the transition going. In contrast, in our model the origin is not a steady state: an economy starting exactly at the origin moves away from it. What allows this novel behavior is that output is positive, rather than zero, because production of intermediate goods is linear in labor. Consequently, saving is positive and investment occurs. Investing in capital alone would be pointless if α stays at zero, while inventing technologies with positive α would be equally pointless if capital does not accumulate. Interestingly, because of depreciation it is optimal to invest only in α early on and start building up the capital stock only when capital intensity is sufficiently high. This kind of behavior at the origin seems realistic, at least if one accepts the notion that humans arose from other animals that did not build capital or do R&D. In other words, humans started with nothing except their labor and, through their own efforts, moved away from that state. The image we have in mind is the insightful ape in the first act of the movie *2001: A Space Odyssey*, who has no capital but discovers the

the future and see that, even if the current (binding) capital/labor ratio is less than 1 and as a result a current increase in α reduces current output, the future return to increasing α now makes it worthwhile to do so. Eventually, the economy passes out of the constrained region and back into the unconstrained region, where the self-reinforcing effects of increases in K and α drive the economy to perpetual growth.

[8] See Chapter 1 of Barro and Sala-i-Martin (2003) for a discussion of the theory of growth through the CES function and Pereira (2003) for a discussion of the evidence.

use of tools. The difference between the movie and our theory is that in our theory, instead of getting his inspiration exogenously from the Monolith, the ape figures out how to use tools with his own wits.

This behavior at the origin is an interesting result in itself but it also has strong implications for the theory of economic development because the transition from the most primitive technology to industrialization does not require exogenous forces to start the transition. Most existing models of economic transition (e.g., Goodfriend and McDermott 1995, Hansen and Prescott 2002, Lucas 2002) jump-start the process by endowing the economy with two production technologies, one primitive and one advanced. They then characterize the reallocation of resources from one to the other. Galor and Weil (2000) avoid the exogenous endowment of primitive and advanced technologies, constructing a model with only one technology augmented by endogenous technical progress and human capital. However, their model requires an initial period of costless technological advance to kick the economy out of its primitive state and start the transition. These models offer many useful insights, but they leave unanswered the question of where the advanced technology, or the initial advance in technology, comes from. It seems unrealistic to suppose that humans always had at their disposal all the technologies they would ever use, or that the Monolith got the process going. It seems more realistic to treat the production technology itself as an endogenous variable that evolves in response to the effort that people expend on improving it. Indeed, Solow (1994) has remarked:

> "I think that the real value of endogenous growth theory will emerge from its attempt to model the endogenous component of technological progress as an integral part of the theory of economic growth."

The model presented above has that very characteristic. Advanced technologies are generated endogenously by the economy through devotion of resources to research and development. No exogenous progress ever occurs. No exogenous spark is required to ignite the fire of discovery.

The second noteworthy aspect of our economy's dynamics is the behavior at the other end of the transition path, the form of the production function that it ultimately attains. Under some conditions (e.g., a low saving rate) the economy reaches a steady state with $\alpha < 1$. Locally, this is a standard Solow economy with no population growth, no technical progress, and so no economic growth. If conditions differ, however, the economy reaches the $\alpha = 1$ limit in finite time. Thereafter, growth continues perpetually and the economy becomes asymptotically AK as the ratio L/K shrinks to zero. The AK technology is the outcome of an endogenous process of technological change. To put it in more general terms, an economy with constant returns to the reproducible factors and thus perpetual endogenous growth is a possible outcome of the growth process itself. This outcome stands in sharp contrast to Solow's (1994) doubts about endogenous growth theory:

> "The conclusion has to be that [the constant returns] version of the endogenous-growth model is very un-robust. It cannot survive without *exactly* constant returns to capital. But you would have to believe in the tooth fairy to expect that kind of luck."

In fact, no tooth fairy is required for the economy to achieve constant returns to capital and the perpetual endogenous growth that flows from it. Constant returns to capital not only can but inevitably will emerge from an economy that starts with diminishing returns or even no returns at all to capital, provided that the saving rate is sufficiently high. Whether the possibility for perpetual growth is realized thus depends on human choice, not supernatural intervention.

The third noteworthy aspect of our economy's dynamics is the implication it has for what would be seen by an observer with the conventional view that Cobb-Douglas factor intensities are constant. At any transition point on any dynamic adjustment path, a snapshot of the economy would suggest an inability to sustain endogenous growth. The economy would have a Cobb-Douglas technology with constant TFP,

a value of α between 0 and 1, and diminishing returns to the reproducible factor. Even if the economy is guaranteed to attain asymptotically the AK structure supporting perpetual endogenous growth, an observer who imposes the standard hypothesis of constant factor intensities will estimate a production technology that cannot sustain endogenous growth. He would agree with Solow (1994) - to quote him yet again:

> "If [the constant returns version of new growth theory] found strong support in empirical material, one would have to reconsider and perhaps try to find some convincing reason why Nature has no choice but to present us with constant returns to capital. On the whole, however, the empirical evidence appears to be less than not strong; if anything, it goes the other way."

The observer's view would be incorrect because the maintained joint hypothesis is incorrect: capital's intensity α is not constant. In particular, the non-reproducible factor L becomes increasingly inessential as technological progress continues. Increasing emphasis is placed on the reproducible factor K, and as a result endogenous growth is sustainable. Empirical evidence suggests that factor intensities have been changing in ways consistent with our theory. See, for example, Blanchard (1997, 1998), Bound and Johnson (1995), and Krueger (1999), discussed briefly below.

3.2 Other Implications

The theory has obvious implications for the dynamics of the income shares of non-reproducible factors. Our variable L represents all reproducible factors, and that variable's share of national income is given by

$$\begin{aligned} \frac{wL}{Y} &= \frac{AL}{A\left[L + m\left(\alpha\right)K\right]} \\ &= \frac{1}{L + m\left(\alpha\right)\frac{K}{L}} \end{aligned}$$

Because AL is constant and Y rises monotonically, the model predicts that the income shares of non-reproducible factors will fall over time. Good data on income shares are not readily available for most non-reproducible factors, but what data are available suggest that the model's prediction agrees with the facts. Consider three types of non-reproducible factors: unskilled labor, land, and energy. (I) *Unskilled labor.* Bound and Johnson (1995) and Krueger (1999) present evidence that unskilled labor's income share of the US economy has been falling. Krueger reports that the share was down to 6 percent by the mid-1990s. At the same time, the income share of skilled labor has been rising (e.g., Blanchard, 1997, 1998). Recall that our variable K is broadly defined and includes human capital. An increase in α therefore tends to increase skilled labor's income share. (II) *Land.* The return on land is difficult to measure because much of it comes as capital gains, which are not included in national income accounts. Nevertheless, estimates of land's share in aggregate income are available. Land's income share in England was about 25% in 1600 (Clark, 2001), but by 2000 it had dropped to about 0.1% (Bar and Leukhina, 2006).[9] We can get a different, indirect measure of land's income share by looking at agriculture data. Land is a major factor of agricultural production but is of negligible importance in manufacturing and service production. Indeed, land is always included in agricultural production functions but always omitted from an industrial economy's aggregate production function. Consequently, as an economy shifts emphasis from agriculture to manufacturing and service, land's income share falls. Such a shift is evident in the data. In the US, for example, the National Income and Product Accounts show that agriculture accounted for 8.92% of Net National Product in 1929 but only 0.59% in 2005. Similarly, Weil (2005, table 3.1) shows that the share of wealth in the form of agricultural land in the United Kingdom fell from 64% in 1688 to 3% in 1958. (III) *Energy.* Data from the Energy Information Administration (U.S. Department of Energy) and NIPA show that total real expenditures on energy in the US grew over the period 1990-2001 at a rate of about 1.5% per year, whereas real GDP grew at an annual rate of about 2.8%. These two figures indicate an accumulated drop of about 15% in energy expenditures as

[9] The 2000 estimate is based on data from the UK National Statistics, kindly provided to us by Oksana Leukhina.

a share of GDP even over a period marked by unusually large (and likely to be partly reversed) increases in energy prices. Data for longer periods are not readily available, but data for related variables suggest similar drops. For example, energy use relative to GDP in the U.S. fell from 19,566 BTU per real dollar of GDP in 1949 to 9,041 BTU per real dollar in 2005.

The theory has the unusual characteristic that neither imperfect competition in the output market nor externalities are necessary for R&D to take place. The model has been cast in terms of monopolistic competition, but examination of the solution shows that imperfect competition of that type is not necessary for R&D to occur. The important relations are the two equilibrium loci given by equations (22) and (23). The elasticity of substitution ε is the indicator of the firm's price-setting power. Larger values of ε indicate less pricing power, with no price-setting power when $\varepsilon = \infty$. The arbitrage locus is well defined for any admissible value of ε, including infinity, and the stationarity locus is independent of ε. Consequently, everything we have derived is valid for the price-taking case. Note also that there are no externalities in this model. The model thus delivers the possibility of self-sustaining endogenous growth without either of the usual conditions of endogenous growth theory. The important element is that the fruits of R&D are *excludable*. Even in the price-taking case, firms will conduct R&D. The firm's goal is to deliver the maximum possible present value of dividends to its shareholders. The dividends consist of the return to capital. The investment division's part in maximizing present value is to provide the firm with the most efficient production technology possible. It has two tools at its disposal: increasing the capital stock and increasing the capital intensity. Each costs one unit of final goods, so the investment division splits its budget between its two activities in whatever way maximizes productive efficiency. No price-setting power is necessary to make R&D worthwhile.[10]

The characteristics of the transition to the balanced growth path (possibly just a steady state) depend on the economy's starting point, and it is not clear what should be considered the right starting point for humanity. Apparently, initial α was at or near zero. Animals produce income without capital, indicating that capital cannot be essential for them. In terms of our model, their α and K both are zero. Early man would have started with the same production function. (Remember the ape in *2001*.) Nature, however, apparently has endowed the world with a small amount of physical capital. Sticks and stones are available for the taking virtually everywhere. All that is needed is the knowledge to use them. (Again remember the ape in *2001*.) In addition, at the point where children become adults and members of the economic community, they have a fair amount of human capital in the form of language, learned practices, and social skills. Let us consider the possibility, then, that there is a positive minimum amount of total capital $K_{\min}$. In that case, the dynamic adjustment paths in the Figures must be changed to point straight rightward in those parts of regions II and III lying below the minimum K. Instead of initially moving along the α-axis from the origin, the economy would move along the horizontal line $K = K_{\min}$. As in Figure 3, increases in α initially would have no effect on income, which corresponds to the facts of human history, emphasized by Lucas (2002, Chapter 5). Thus even this very simple variation of the theory is capable of producing an adjustment path that resembles the path actually taken by humanity. It would be interesting to see what implications would emerge from a model that extended the theory to include a distinction between physical and human capital and that introduced endogenous demography.

A critical issue for the economy's path is whether the ratio sA/δ is sufficiently high to guarantee that the equilibrium loci do not intersect. That ratio can be increased either by raising sA or by reducing δ. In our model, s, A and δ are exogenous constants, but in reality all three can vary. The saving rate obviously would be endogenous in a complete model of household choice. We have treated it as constant for tractability, not for realism. Similarly, we have taken total factor productivity A to be constant when in fact firms should be able to change it as well as α through R&D. Even the depreciation rate need be constant. Indeed, in reality it seems to fall. Automobiles require less maintenance now than they did two or three decades ago, and solid-state electronics break down far less often than their vacuum tube predecessors. In other words, still another dimension of technical progress is control of the depreciation rate. Making the saving rate, total factor productivity, or the depreciation rate endogenous is well beyond the scope of the present paper, but doing so would be an interesting area for further research.

[10] Excludability means that the market is not contestable and so is not competitive in the usual sense. Boldrin and Levine's (2005) examination of the necessary conditions for innovation depends on the same mechanism and so has the same element of non-competitive behavior.

Our theory offers new perspectives on two aspects of technical change that have been much disucssed in the literature: factor-bias and induction of technical progress. Factor-bias is the change in factors' relative rates of return caused by technical progress, and induced technical progress is progress that arises in response to an increase in the quantity of one of the factors. Virtually all the literature on factor-bias and induction, like the growth literature, is restricted to factor-augmenting technical change, ignoring factor-eliminating progress.[11] It is straightforward to see that factor-eliminating progress is biased toward capital and is induced by increases in capital. Intermediate output is given by equation (19). When the labor constraint is not binding, the marginal products of K and L are $m(a)A$ and A, respectively, and their ratio $m(\alpha)$ is monotonically increasing in α. When the labor constraint binds, the marginal products of K and L are $\alpha A K^{\alpha-1} L^{1-\alpha}$ and $(1-\alpha) A K^{\alpha} L^{-\alpha}$, and their ratio $\alpha(1-\alpha)(K/L)^{-1}$ again is monotonically increasing in α. In both cases, then, technical progress is biased toward capital. Also, an increase in K always makes higher α desirable. The marginal product of α is $Am'(\alpha) K$ when the labor constraint binds and is $AK^{\alpha} L^{(1-\alpha)} \ln(K/L)$. Both are always positive functions of K because $m'(\alpha) > 0$ for all α and $K/L > 1$ whenever the labor constraint binds because the constraint boundary is given by $K/L = (1-\alpha)^{-(1/\alpha)} > 1$ so that $\ln(K/L) > 0$ in the constrained region. Factor-eliminating progress satisfies what Acemoglu (2006) defines as strong relative equilibrium bias, which means that the increase in α induced by an increase in K is sufficiently large to raise the marginal return to K even though the quantity of K has increased, which in itself drives down the marginal return to K.[12]

In our theory, the economy reaches the $\alpha = 1$ limit in finite time. This is because we posited an R&D technology that is independent of the level of α; see (5). If we posited that changing α becomes infinitely costly as α approaches 1 by writing

$$\dot{\alpha} = R(1-\alpha)$$

as in Zuleta (2006), then the economy would reach the $\alpha = 1$ limit only asymptotically. Nothing of substance, however, would change in our story. Our economy *already* reaches the AK structure only asymptotically. Slowing the transition to $\alpha = 1$ (in fact, arbitrarily preventing α from ever getting to 1) complicates the analysis without adding any insight. It is not obvious, moreover, why the productivity of R&D resources should be equal to $1-\alpha$. Interpreting α as knowledge, such a specification implies that knowledge accumulation *reduces* research productivity – a strange idea to say the least! Suppose instead there are positive but diminishing returns to knowledge. Then one could write

$$\dot{\alpha} = R \cdot f(\alpha), \quad f'(\alpha) > 0,\ f''(\alpha) < 0,\ lim_{\alpha \to 1} f'(\alpha) = 0.$$

This specification, too, would complicate the analysis without changing anything of substance. Alternatively, one could write

$$\dot{\alpha} = f(R, \alpha),$$

where $f(\cdot, \cdot)$ is a standard neoclassical production function. Again, nothing of substance would change at the cost of a vastly more complicated model. The reason why playing around with the specific way in which α enters the R&D technology makes no substantive difference is that our story requires that α becomes arbitrarily close to 1, a finite value. We thus do not need assumptions that allow it to grow forever. We chose to work with a specification where α becomes exactly 1 in finite time because it simplifies the analysis drastically.

4 Conclusion

We have proposed a theory of endogenous technical progress that alters factor intensity. As we have seen, the theory can deliver perpetual economic growth without any sort of factor augmenting technical change. In particular, under some parameter settings, the AK model is the endogenous asymptotic limit of the economy. In that case, growth occurs along entire the transition path as well as in the limit. The

[11] See Acemoglu (2002) for a recent review.

[12] Acemoglu's discussion is restricted to a static economy with only factor-augmenting progress, but his concepts carry over directly to our dynamic economy with eliminating progress.

perpetual growth path always is feasible. It always is possible for society to choose a saving rate that will put the economy on it. The theory also offers a totally endogenous explanation for the transition from the primitive state at the dawn of mankind to a modern post-industrial economy. The explanation relies in no way on exogenous technical change, scale effects, or endowments of alternative technologies. Human progress is entirely the result of human activity and human choices. Neither the Monolith nor the Tooth Fairy is required.

Our theory has an important implication for the time path of the price of non-reproducible factors of production. If the economy gets on a path that leads to the AK model, then the non-reproducible factors become asymptotically unimportant, meaning that their prices eventually go to zero. This result is in line with Julian Simon's well known remark:

> "Our supplies of natural resources are not finite in any economic sense. Nor does past experience give reason to expect natural resources to become more scarce. Rather, if history is any guide, natural resources will progressively become less costly, hence less scarce, and will constitute a smaller proportion of our expenses in future years."

If we think of unskilled labor as a non-reproducible factor, our theory has important implications for income distribution dynamics. Assuming that the economy starts at or near the origin, unskilled labor's share drops along the adjustment path. Unskilled labor's share never disappears, but it does go asymptotically to zero. This outcome is similar to what already has happened to land's income share in industrialized economies. It does not mean that unskilled labor will have no income. It is unskilled labor's *share* of income that falls, not its absolute level of income. Unskilled labor always is productive, so it always earns income. Furthermore, *skilled* labor's income share rises because human capital, which is subsumed under our broadly defined "capital," earns an increasing share of the growing national income.

The theory suggests several lines for future research, both theoretical and empirical. An obvious extension would be to include both factor-augmenting and intensity-altering technical change. Empirical evidence suggests that both occur, so a complete theory would accommodate both. We have initial exploration along these lines, and the theory seems tractable and interesting. Another interesting theoretical development would be to make either saving or depreciation endogenous. Endogenous saving would be parallel to moving from the Solow model to the Cass model. It is non-trivial because of the increase in the dimensionality of the system and the fact that saving would behave continuously rather than in a bang-bang manner.[13] Presumably, introducing endogenous depreciation would face similar difficulties.

On the empirical side, our theory suggests that factor intensities should change over time. In general, our theory allows imperfect competition, so factor intensities are not the same as factor shares. If one is willing to assume that factor shares estimated from national income account data provide reasonable estimates of factor intensities, then the time series evidence on the behavior of factor shares is consistent with our theory's predictions. Factor shares change over time, and the shares of non-reproducible factors have been falling. Time variation in factor intensities has an implication for measurement of Total Factor Productivity. Hall and Jones (1999), for example, use a Cobb-Douglas production function to estimate TFP for a large set of countries. They find that cross-country differences in TFP are of paramount importance in explaining cross-country differences in economic performance. In arriving at their estimates, they make the usual assumption that countries have the same capital intensities of 1/3. Gollin (2002), however, has shown that capital shares differ substantially across countries, ranging from 0.17 to 0.66, which suggests that capital intensities similarly differ across countries.[14] Hall and Jones's estimates strongly suggest that technology differences

[13] Zuleta (2006) examines endogenous saving, but he restricts attention to a model in which investments in physical capital K can be instantaneously converted into production knowledge α and vice versa.

[14] Gollin's figures do not distinguish between skilled and unskilled labor, whereas our theory treats skill as a type of capital. For our purposes the correct measure of labor income share would be the share paid to unskilled labor only. We know of no cross-country data on the income share of unskilled labor. Nonetheless, Gollin's figures do suggest considerable cross-country variation in factors' income shares.

are important in explaining cross-country differences, but Gollin's estimates suggest, in conformity with our theory, that it may not be appropriate to summarize all differences in technology by differences in TFP. It would be useful to re-calculate estimates of TFP across countries using Gollin's factor share estimates. It then would be possible to get some idea of how much of the variation in cross-country economic performance depends on differences in TFP and how much depends on differences in factor intensities.

5 Appendix

5.1 The function $m(\alpha)$

The function

$$m(\alpha) = \alpha (1-\alpha)^{\frac{1-\alpha}{\alpha}}$$

has derivatives:

$$\begin{aligned} m'(\alpha) &= m(\alpha) \frac{1}{\alpha^2} \ln \frac{1}{1-\alpha} > 0; \\ m''(\alpha) &= m(\alpha) \frac{1}{\alpha^2} \left[\frac{1}{\alpha} \ln \frac{1}{1-\alpha} \left(\frac{1}{\alpha} \ln \frac{1}{1-\alpha} + \frac{\alpha}{1-\alpha} \right) \right] > 0. \end{aligned}$$

Moreover, we can write:

$$m(\alpha) = \alpha \exp \left[n \frac{\ln(1-\alpha)}{\frac{\alpha}{1-\alpha}} \right];$$

$$m'(\alpha) = \frac{-\ln(1-\alpha)}{\alpha} \exp \left[\frac{\ln(1-\alpha)}{\frac{\alpha}{1-\alpha}} \right].$$

We then have that:

$$\lim_{\alpha \to 0,1} \frac{\ln(1-\alpha)}{\frac{\alpha}{1-\alpha}} = \lim_{\alpha \to 0,1} \frac{-\frac{1}{1-\alpha}}{\frac{1}{(1-\alpha)^2}} = \lim_{\alpha \to 0,1} (\alpha - 1);$$

$$\lim_{\alpha \to 0,1} \frac{-\ln(1-\alpha)}{\alpha} = \lim_{\alpha \to 0,1} \frac{1}{1-\alpha}.$$

Hence:

$$\begin{aligned} m(0) &= 0 \cdot \exp(-1) = 0; \\ m(1) &= 1 \cdot \exp(0) = 1; \\ m'(0) &= 1 \cdot \exp(-1) = e^{-1}; \\ m'(1) &= +\infty \cdot \exp(0) = +\infty. \end{aligned}$$

5.2 The firm's necessary conditions for optimization

The necessary conditions are:

$$\frac{\partial H}{\partial \psi} = I - \delta K; \tag{24}$$

$$\frac{\partial H}{\partial \phi} = R; \tag{25}$$

$$\dot{\psi} = -\frac{\partial H}{\partial K} + r\psi = -\frac{\partial \pi}{\partial K} - r_K + \delta\psi + r\psi; \tag{26}$$

$$\dot{\phi} = -\frac{\partial H}{\partial \alpha} + r\phi = -\frac{\partial \pi}{\partial \alpha} + r\phi; \tag{27}$$

$$\frac{\partial H}{\partial I} = -1 + \psi + \omega_1 = 0; \tag{28}$$

$$\frac{\partial H}{\partial R} = -1 + \phi + \omega_2 = 0; \tag{29}$$

$$\lim_{t \to \infty} \psi_t K_t e^{-\bar{r}_t t} = 0; \tag{30}$$

$$\lim_{t\to\infty} \phi_t \alpha_t e^{-\bar{r}_t t} = 0. \tag{31}$$

To obtain the returns to capital and R&D, we substitute the values $\psi = \phi = 1$ and $\dot{\psi} = \dot{\phi} = 0$ into equations (26) and (27), and note that the investment division takes l, K and L as given in π. This gives us

$$r = r_K - \delta,$$

where r_K is given by (13), and

$$r = r_\alpha \equiv \frac{\partial \pi}{\partial \alpha}.$$

5.3 Labor-Constraint Boundary

The boundary between the region wherethe labor constraint does and does not bind is given by

$$K = L\left(\frac{1}{1-\alpha}\right)^{\frac{1}{\alpha}}$$

The boundary's slope is

$$\frac{dK}{d\alpha} = K\left[\alpha^{-1}(1-\alpha)^{-1} + \alpha^{-2}\ln(1-\alpha)\right]$$

The two terms in brackets are of opposite sign, but plotting the slope shows that it is positive over the entire domain $\alpha \in [0,1]$, indicating that the boundary is increasing everywhere over that domain. Application of L'Hopital's Rule shows that the boundary intersects the K-axis at $K = eL$ and goes to infinity as $\alpha \to 1$. The slope of the boundary is $eL/2$ at $\alpha = 0$.

5.4 Arbitrage Locus and Proof of Proposition 2

We derive the properties of the arbitrage locus under general conditions, obtaining Proposition 2 as part of the derivation. Use (6), (13), (16), and (18) to write:

$$r_K = \begin{cases} \left(1-\frac{1}{\varepsilon}\right) Am(\alpha) & K < L\left(\frac{1}{1-\alpha}\right)^{\frac{1}{\alpha}} \\ \left(1-\frac{1}{\varepsilon}\right) A\alpha\left(\frac{K}{L}\right)^{\alpha-1} & K \geq L\left(\frac{1}{1-\alpha}\right)^{\frac{1}{\alpha}} \end{cases};$$

$$r_\alpha = \begin{cases} \left(1-\frac{1}{\varepsilon}\right) AKm'(\alpha) & K < L\left(\frac{1}{1-\alpha}\right)^{\frac{1}{\alpha}} \\ \left(1-\frac{1}{\varepsilon}\right) AL\left(\frac{K}{L}\right)^{\alpha}\ln\frac{K}{L} & K \geq L\left(\frac{1}{1-\alpha}\right)^{\frac{1}{\alpha}} \end{cases}.$$

Substitute these expressions into the equation for asset market equilibrium, $r_\alpha = r_K - \delta$, and solve for K in terms of α to obtain the equation for the arbitrage locus. Consider the unconstrained and constrained cases in turn.

In the unconstrained region, we have

$$K = \frac{1}{m'(\alpha)}\left[m(\alpha) - \frac{\delta\varepsilon}{A(\varepsilon-1)}\right]$$

The expression in brackets can be positive or negative, depending on parameters magnitudes and the value of α. However, K cannot be negative, so for any α for which the bracketed expression is negative, the arbitrage locus must lie on the horizontal axis (i.e., K is constrained to be zero). Let $\bar{\alpha}$ be the value for which $m(\alpha) = \delta\varepsilon/[A(\varepsilon-1)]$. If $\delta\varepsilon/[A(\varepsilon-1)] \geq 1$, then $\bar{\alpha} \geq 1$, the bracketed expression is non-positive for all feasible values of $\alpha \in [0,1]$, and the entire arbitrage locus lies on the horizontal axis. In that case, all

dynamic adjustment paths lead to $K = 0$, and the economy always degenerates to a primitive state of having no capital and using only the primitive production function $X = AL$. Such a case does not correspond to human history, so we rule it out by assuming that $\delta\varepsilon/[A(\varepsilon-1)] < 1$. Then there exists an $\bar{\alpha} \in (0,1)$ such that $m(\alpha) = \delta\varepsilon/[A(\varepsilon-1)]$. For $\alpha \leq \bar{\alpha}$, the arbitrage locus lies on the horizontal axis, and for $\alpha \leq \bar{\alpha}$, $K = [m(\alpha) - \delta\varepsilon/\{A(\varepsilon-1)\}]/m'(\alpha)$. We thus have the final equation for the unconstrained arbitrage locus:

$$K = \begin{cases} 0 & 0 \leq \alpha \leq \bar{\alpha} \\ \frac{1}{m'(\alpha)}\left[m(\alpha) - \frac{\delta\varepsilon}{A(\varepsilon-1)}\right] & \bar{\alpha} < \alpha \leq 1 \end{cases}$$

which is the equation given in Proposition 2. At $\alpha = 1$, we have $K = 0$ because $m(1) = 1$ and $m'(1) = \infty$.

The slope of the unconstrained arbitrage locus is (after some algebra)

$$\frac{dK}{d\alpha} = 1 - \frac{1}{m(\alpha)}\left[1 + \frac{\alpha^2}{(1-\alpha)\ln(1-\alpha)}\right]\left[m(\alpha) - \frac{\delta\varepsilon}{A(\varepsilon-1)}\right]$$

At $\alpha = \bar{\alpha}$ the slope is 1, and at $\alpha = 1$ it is $-\infty$. The slope therefore changes sign between $\bar{\alpha}$ and 1. We have not been able to derive analytically how many times the slope changes sign (although it clearly must be an odd number), but plots of the arbitrage locus for various parameter values always show a single turning point. The arbitrage locus thus is hump-shaped as stated in Proposition 2 and shown in Figures 3 and 4.

In the constrained region, we obtain an expression for the arbitrage locus that we cannot solve in closed form for K:

$$\left(\frac{K}{L}\right)^{\alpha-1}\left[\alpha - K\ln\frac{K}{L}\right] = \frac{\delta\varepsilon}{A(\varepsilon-1)}$$

The slope is even more complicated:

$$\frac{dK}{d\alpha} = \frac{\alpha(1-\alpha)K^{-1} + (\alpha+L)\ln\frac{K}{L}}{1 + \alpha\ln\frac{K}{L} - K\left(\frac{K}{L}\right)^{1-\alpha}\left(\ln\frac{K}{L}\right)^2}$$

These expressions give little information about the shape of the constrained arbitrage locus, so we turn to a graphical analysis.

Note that at $\alpha = 0$ the equation for the constrained arbitrage locus simplifies to $-L\ln(K/L) = \delta\varepsilon/[A(\varepsilon-1)]$, which can be rearranged as

$$\begin{aligned} K &= L\exp\left[-\frac{\delta\varepsilon}{LA(\varepsilon-1)}\right] \\ &< L \end{aligned}$$

Recall that the constraint boundary starts at $K = eL > L$ and the unconstrained arbitrage locus starts at $K = 0$. Consequently, the constrained arbitrage locus starts below the constraint boundary and above the unconstrained arbitrage locus. At $\alpha = 1$, constrained arbitrage locus equation simplifies to $1 - K\ln(K/L) = \delta\varepsilon/[A(\varepsilon-1)]$, or equivalently

$$\begin{aligned} K\ln\frac{K}{L} &= 1 - \frac{\delta\varepsilon}{A(\varepsilon-1)} \\ &> 0 \end{aligned}$$

$$\begin{aligned} \Rightarrow \quad & 1 < \frac{K}{L} < \infty \\ \iff \quad & L < K < \infty \end{aligned}$$

This last result implies two things: (1) the locus has an overall upward trend because it starts at a value of K less than L and ends at a value greater than L, and (2) the constrained locus intersects the vertical line

$\alpha = 1$ at a finite value. There are three possibilities for the relation among the constrained arbitrage locus, the unconstrained arbitrage locus, and the constraint boundary. First, the constrained arbitrage locus can lie everywhere below the constraint boundary and everywhere above the unconstrained locus as shown in the bottom panel of Figure 5. Second, the three loci could be mutually tangent at a single point, as shown in the middle panel of Figure 5. Third, the loci could intersect as shown in the top panel of Figure 5. That is the case where the labor constraint binds in a meaningful way over part of the $\alpha - K$ space. To the left of the first intersection point, both the unconstrained and constrained arbitrage loci lie below the constraint boundary. That means the constrained locus cannot be effective because it cannot pertain in the unconstrained region (which is the region below the constraint boundary). The same is true to the right of the second intersection point. Between those two points, both loci lie above the constraint boundary, i.e., in the constrained region. In that region, the unconstrained locus cannot be effective. The effective equilibrium locus therefore is the heavy curve that is the union of the relevant segments of the unconstrained and constrained loci: first a segment of the unconstrained locus below the constraint boundary, then a segment of the constrained locus above the boundary, and finally another segment of the unconstrained locus below the boundary. It should be clear from this discussion that the two equilibrium loci must intersect the constraint boundary at common points. Anything else leads to contradictions with the meaning of an equilibrium locus.

From the foregoing, we see that the effective arbitrage locus will consist of only the unconstrained locus if the latter lies everywhere below the constraint boundary, as in the bottom panel of Figure 5. For that to happen, we require that, for all α,

$$L\left(\frac{1}{1-\alpha}\right)^{\frac{1}{\alpha}} > \frac{1}{m'(\alpha)}\left[m(\alpha) - \frac{\delta\varepsilon}{A(\varepsilon-1)}\right] \tag{32}$$

which we can rearrange as

$$\left[\frac{L}{\alpha}\left(\frac{1}{1-\alpha}\right)^{\frac{1}{\alpha}} \ln\left(\frac{1}{1-\alpha}\right)^{\frac{1}{\alpha}} - 1\right] m(\alpha) > -\frac{\delta\varepsilon}{A(\varepsilon-1)}$$

A sufficient condition for this inequality to hold is

$$\frac{L}{\alpha}\left(\frac{1}{1-\alpha}\right)^{\frac{1}{\alpha}} \ln\left(\frac{1}{1-\alpha}\right)^{\frac{1}{\alpha}} > 1.$$

Now observe that

$$\left(\frac{1}{1-\alpha}\right)^{\frac{1}{\alpha}} > 1$$

and

$$\ln\left(\frac{1}{1-\alpha}\right)^{\frac{1}{\alpha}} \geq 1 \text{ because } \left(\frac{1}{1-\alpha}\right)^{\frac{1}{\alpha}} = \exp\left\{\frac{1}{\alpha}\ln\left(\frac{1}{1-\alpha}\right)\right\} \geq e.$$

Then, $L > \alpha$ is sufficient for (32) to hold, as remarked in the main text.

5.5 Stationarity Locus and Proof of Proposition 3

Stationarity requires $\dot{\alpha} = 0$ and $\dot{K} = 0$. Substituting these into the resource constraint, given by equation (21), and solving for K in terms of α yields the following unconstrained and constrained equations for the stationarity locus:

$$K = \begin{cases} L\left[\frac{\delta}{sA} - m(\alpha)\right]^{-1} & K < L\left(\frac{1}{1-\alpha}\right)^{\frac{1}{\alpha}} \\ L\left(\frac{sA}{\delta}\right)^{\frac{1}{1-\alpha}} & K \geq L\left(\frac{1}{1-\alpha}\right)^{\frac{1}{\alpha}} \end{cases}$$

Consider the unconstrained case first. let $\tilde{\alpha}$ be the value of α that solves $m(\alpha) = \delta/(sA)$. The stationarity locus is asymptotic to $\tilde{\alpha}$. The first and second derivatives of the unconstrained locus are

$$\begin{aligned} \frac{dK}{d\alpha} &= L\left[\frac{\delta}{sA} - m(\alpha)\right]^{-2} m'(\alpha) \\ &> 0 \qquad \forall\alpha \end{aligned}$$

$$\begin{aligned} \frac{d^2K}{d\alpha^2} &= L\left[\frac{\delta}{sA} - m(\alpha)\right]^{-2}\left\{2\left[\frac{\delta}{sA} - m(\alpha)\right]^{-1}[m'(\alpha)]^2 + m''(\alpha)\right\} \\ &> 0 \qquad \forall\alpha < \tilde{\alpha} \\ &< 0 \qquad \forall\alpha > \tilde{\alpha} \end{aligned}$$

We are interested in the stationarity locus only over the domain $\alpha \in [0,1]$. If $\delta/(sA) < 1$, then $\tilde{\alpha} < 1$ as well, and the stationarity locus has two branches. The left branch starts at $K = L\frac{sA}{\delta}$ at $\alpha = 0$ and then goes asymptotically to infinity as $\alpha \to \tilde{\alpha}$. The right branch then starts at $-\infty$ immediately to the right of $\tilde{\alpha}$ and rises to $L\left[\frac{\delta}{sA} - 1\right]^{-1} < 0$ at $\alpha = 1$. The lower panel of Figure 6 shows the general shape of the unconstrained stationarity locus. The right branch of the locus is of no interest because it lies in the part of the plane where capital is negative. It is ignored henceforth. The upper panel of Figure 6 shows various possible configurations of the left branch of the unconstrained stationarity locus. If $\delta/(sA) < e$, the locus lies everywhere above the constraint boundary. If $e < \delta/(sA) < 1$, the locus intersects the constraint boundary once. Using the equations for the stationarity locus and the constraint boundary, some algebra shows that value of α at which the intersection occurs is that which solves $(1-\alpha)^{-\frac{1-\alpha}{\alpha}} = sA/\delta$. If $\delta/(sA) = 1$, the locus goes asymptotically to the vertical line $\alpha = 1$. Finally, if $\delta/(sA) > 1$, the locus intersects the vertical line $\alpha = 1$ at a finite value of K.

The constrained stationarity locus starts at the same point as the unconstrained locus. That is, at $\alpha = 0$ the value of K on the two loci are the same, namely, $(sA/\delta)L$. As $\alpha \to 1$, the behavior of the constrained locus depends on the size of sA/δ:

$$K \quad \begin{cases} \to \infty & \text{if } sA/\delta > 1 \\ = L & \text{if } sA/\delta = 1 \\ \to 0 & \text{if } sA/\delta < 1 \end{cases}$$

The first derivative of constrained stationarity locus equation is

$$\begin{aligned} \frac{dK}{d\alpha} &= (1-\alpha)^{-2} L\left(\frac{sA}{\delta}\right)\ln\left(\frac{sA}{\delta}\right) \\ &\gtreqless 0 \quad \text{as } \frac{sA}{\delta} \gtreqless 1 \end{aligned}$$

As $\alpha \to 1$, the slope goes to infinity if $sA/\delta > 1$, equals zero if $sA/\delta = 1$, and requires analysis of the second derivative if $sA/\delta < 1$. The second derivative is

$$\frac{d^2K}{d\alpha^2} = (1-\alpha)^{-3}\left(\frac{sA}{\delta}\right)^{\frac{1}{1-\alpha}} L\ln\left(\frac{sA}{\delta}\right)\left[2 + (1-\alpha)^{-1}\ln\left(\frac{sA}{\delta}\right)\right]$$

This expression is positive if $sA/\delta > 1$, is zero if $sA/\delta = 1$, and changes sign as α increases if $sA/\delta < 1$. If $e^{-2} < sA/\delta < 1$, the second derivative is negative at $\alpha = 0$ and positive at $\alpha = 1$. If $sA/\delta < e^{-2}$, the second derivative is positive for all values of α. Whenever $sA/\delta > 1$, the constrained stationarity locus intersects the constraint boundary once. The point of intersection is found by equating the expressions for the two curves, and the result is

$$\frac{sA}{\delta} = \left(\frac{1}{1-\alpha}\right)^{\frac{1-\alpha}{\alpha}}$$

which gives the value of α at which the intersection occurs. For $\alpha \in (0,1)$, the right side is greater than 1 and less than e. Thus if $sA/\delta \in [1,e]$, there is an intersection. If $sA/\delta > e$, the constrained stationarity locus lies everywhere above the constraint boundary. If $sA/\delta < 1$, the constrained stationarity locus lies everywhere below the constraint boundary. The lower panel of Figure 7 shows various possible configurations of the constrained stationarity locus.

Whenever $\delta/(sA) < 1$, the unconstrained stationarity locus crosses the constraint boundary at some $\hat{\alpha} \in (0,1)$ and lies in the constrained region for values of $\alpha > \hat{\alpha}$. Therefore, to the right of $\hat{\alpha}$ the unconstrained locus is irrelevant. Some algebra shows that the constrained stationarity locus crosses the constraint boundary at the same point that the unconstrained locus crosses, namely, $\hat{\alpha}$. However, the two stationarity loci have at least one other crossing at a value of α between 0 and $\hat{\alpha}$. One way to see this result is to note that the slope of the constrained locus is less than the slope of the unconstrained locus at both $\alpha = 0$ and $\alpha = \hat{\alpha}$, implying at least one crossing between those two values of α. The upper panel of Figure 7 shows the relation between the three curves. In that panel, the effective stationarity locus is the heavy curve consisting of the part of the unconstrained stationarity locus for $\alpha < \hat{\alpha}$ and the constrained stationarity locus for $\alpha > \hat{\alpha}$.

Collecting these results gives us the following equation for the effective stationarity locus:

$$K = \begin{cases} \dfrac{L}{\left(\frac{\delta}{sA}\right)^{\frac{1}{1-\alpha}}} & \frac{sA}{\delta} \geq e \\ \left.\begin{array}{ll} \dfrac{L}{\left(\frac{\delta}{sA}\right)^{\frac{1}{1-\alpha}}} & 0 \leq \alpha \leq \hat{\alpha} \\ \dfrac{L}{\frac{\delta}{sA} - m(\alpha)} & \hat{\alpha} \leq \alpha \leq 1 \end{array}\right\} & e > \frac{sA}{\delta} \geq 1 \\ \dfrac{L}{\frac{\delta}{sA} - m(\alpha)} & 1 > \frac{sA}{\delta} \end{cases}$$

In the main text, we have restricted attention to the case where the stationarity locus lies everywhere above the arbitrage locus, so we have not bothered to show there the constrained part of the stationarity locus.

Finally, we note that there always exists a value of the saving rate s sufficiently high to guarantee that the stationarity locus lies above the arbitrage locus and a value of s sufficiently low that the two loci intersect. Recall that $\bar{\alpha}$ is the value of α at which the unconstrained arbitrage locus first crosses into the positive quadrant, and $\tilde{\alpha}$ is the asymptote of the unconstrained stationarity locus. If the asymptote is to the left of $\bar{\alpha}$, the two loci cannot intersect. The value of $\tilde{\alpha}$ latter depends on the saving rate s, so we want to see if there is a saving rate such that $\tilde{\alpha} < \bar{\alpha}$. From earlier definitions we have that $\bar{\alpha}$ is the value of α for which $m(\alpha) = \varepsilon\delta/[(\varepsilon-1)A]$ and $\tilde{\alpha}$ is the value for which $m(\alpha) = \delta/(sA)$. Then note that

$$\begin{aligned} m(\tilde{\alpha}) &= \frac{\delta}{sA} \\ &= \frac{1}{s}\frac{\varepsilon-1}{\varepsilon}\frac{\varepsilon}{\varepsilon-1}\frac{\delta}{A} \\ &= \frac{1}{s}\frac{\varepsilon-1}{\varepsilon}m(\bar{\alpha}) \end{aligned}$$

Choosing $s > (\varepsilon-1)/\varepsilon$ then guarantees that $\tilde{\alpha} < \bar{\alpha}$ because $m(\alpha)$ is increasing in α. Such a choice of s always is possible because $(\varepsilon-1)/\varepsilon < 1$.

5.6 Transversality conditions: investment division

After the economy reaches the $\alpha = 1$ limit, the interest rate is constant and equal to the net marginal product of capital, $r = \frac{\varepsilon-1}{\varepsilon}A - \delta$. We assume $\frac{\varepsilon-1}{\varepsilon}A - \delta > 0$ to guarantee $r > 0$. Asymptotically, capital grows at rate $sA - \delta$. Also, $\psi = 1$ for all t. We thus have

$$\lim_{t\to\infty} \psi_t K_t e^{-\bar{r}_t t} = \lim_{t\to\infty} K_0 e^{(sA-\delta)t} e^{-\left(\frac{\varepsilon-1}{\varepsilon}A-\delta\right)t} = 0.$$

This surely holds for $sA \leq \delta$. For $sA > \delta$, it holds if $1/\varepsilon < 1 - s$ and fails to hold if $1/\varepsilon \geq 1 - s$. We thus must impose an upper bound on the monopoly profit ratio $1/\varepsilon$. Also, $\alpha = 1$ and $\phi = 1$ yield

$$\lim_{t\to\infty} \phi_t \alpha_t e^{-\bar{r}_t t} = \lim_{t\to\infty} e^{-\left(\frac{\varepsilon-1}{\varepsilon}A-\delta\right)t} = 0.$$

If the economy reaches a "Solow" steady state with constant capital and $\alpha < 1$, the transversality conditions (30) and (31) are surely satisfied.

References

[1] Acemoglu, Daron K. "Directed Technical Change," *Review of Economic Studies* 69, October 2002, pp. 781-809.

[2] Acemoglu, Daron K. "Equilibrium Bias of Technology," working paper, Massachusetts Institute of Technology, June 2006.

[3] Bar, Michael, and Oksana Leukhina. "Demographic Transition and Industrial Revolution: A Coincidence?" working paper, March 2006, University of North Carolina at Chapel Hill.

[4] Barro, Robert J., and X. Sala-i-Martin. *Economic growth, 2nd ed.* MIT Press, Cambridge MA., 2004.

[5] Blanchard, Olivier J. "The Medium Run," *Brookings Papers on Economic Activity* 2, 1997, pp. 89-158.

[6] Blanchard, Olivier J. "Revisiting European Unemployment: Unemployment, Capital Accumulation, and Factor Prices," Geary Lecture, Economic and Social Research Institute 1998.

[7] Boldrin, Michele, and David Levine. "Factor-Saving Innovation," *Journal of Economic Theory* 105, July 2002, pp. 18-41.

[8] Bound, John, and George Johnson. "What Are the Causes of Rising Wage Inequality in the United States?" *Economic Policy Review*, Federal Reserve Bank of New York, January 1995, pp. 9-17.

[9] Clark, Gregory. "The Secret History of the Industrial Revolution," working paper, University of California at Davis, October 2001.

[10] Givon, Danny. "Factor Replacement versus Factor Substitution, Mechanization and Asymptotic Harrod Neutrality," working paper, Hebrew University, March 2006.

[11] Gollin, Douglas. "Getting Income Shares Right," *Journal of Political Economy* 110, April 2002, pp. 458-474.

[12] Hall, Robert A., and Charles I. Jones. "Why Do Some Countries Produce So Much More Output per Worker than Others?" *Quarterly Journal of Economics* 114, February 1999, pp. 83-116.

[13] Hansen, Gary D., and Edward C. Prescott. "Malthus to Solow," *American Economic Review* 92, September 2002, pp. 1205-1217.

[14] Kamien, Morton I., and Nancy L. Schwartz. "Optimal 'Induced' Technical Change," *Econometrica* 36, January 1968, pp. 1-17.

[15] Krueger, Alan B. "Measuring Labor's Share," *American Economic Review* 89, May 1999, pp. 45-51.

[16] Lucas, Robert E., Jr. *Lectures on Economic Growth*, Harvard University Press, Cambridge MA, 2002.

[17] Pereira, Claudiney M. *Empirical Essays on the Elasticity of Substitution, Technical Change, and Economic Growth.* Doctoral dissertation, North Carolina State University, 2003.

[18] Rebelo, Sergio. "Long Run Policy Analysis and Long Run Growth," *Journal of Political Economy* 99, June1991, pp. 500-521.

[19] Sato, Ryuzo, and Martin J. Beckmann. "Neutral Innovations and Production Functions," *Review of Economic Studies* 35, January 1968, pp. 57-66.

[20] Seater, John J. "Share-Altering Technical Progress," in *Economic Growth and Productivity*, L. A. Finley, ed., Nova Science Publishers, Hauppage NY, 2005, pp. 59-84.

[21] Weil, David N. *Economic Growth.* Pearson Addison-Wesley, Boston MA., 2005.

[22] Solow, Robert M. "Perspectives on Growth Theory," *Journal of Economic Perspectives* 8, Winter 1994, pp. 45-54.

[23] Zeira, Joseph. "Workers, Machines, and Economic Growth," *Quarterly Journal of Economics* 113, November 1998, pp. 1091-1117.

[24] Zeira, Joseph. "Machines as Engines of Growth," working paper, Hebrew University, 2006.

[25] Zuleta, Hernando. "Factor Saving Innovations and Factors Income Shares," working paper, Instituto Tecnologico Autonomo de Mexico (ITAM)., February 2006.

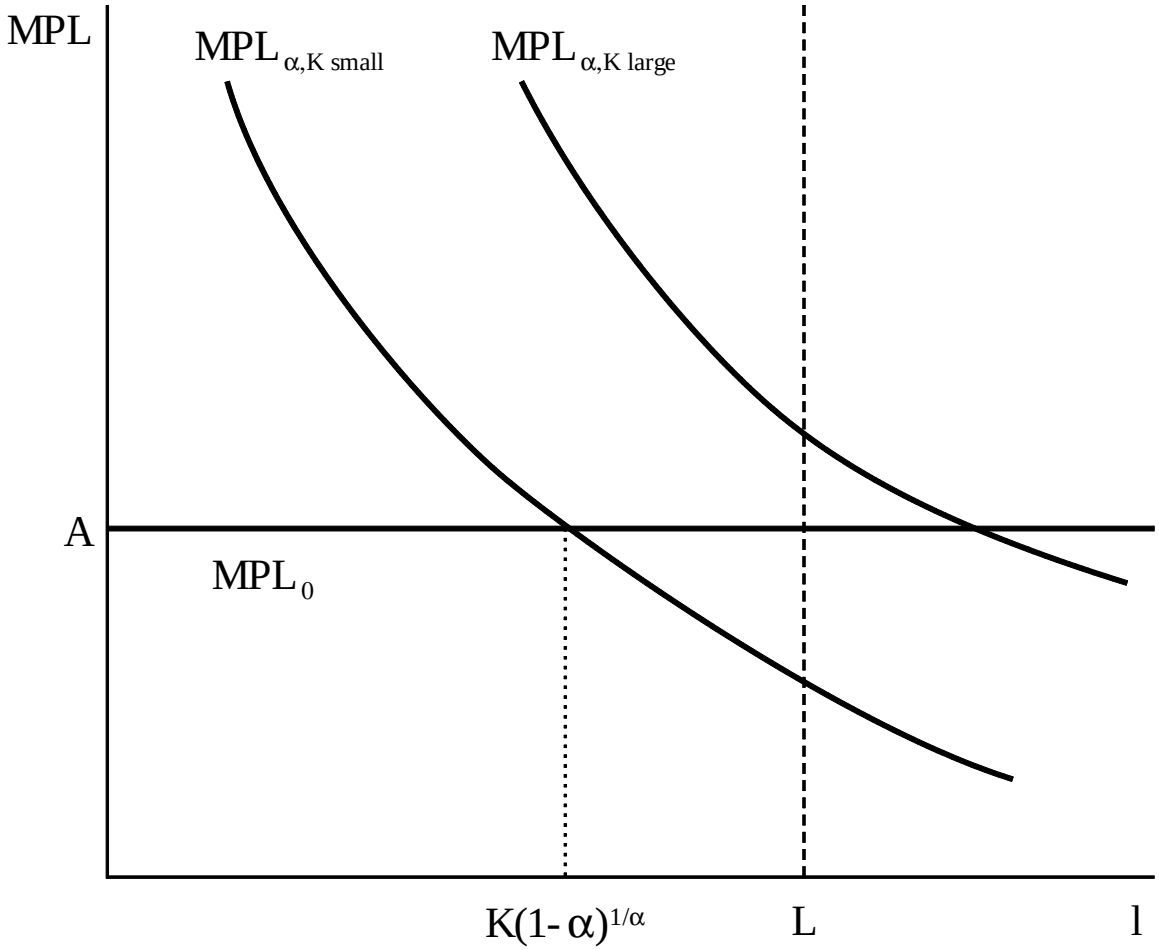

Figure 1: Optimal labor allocation

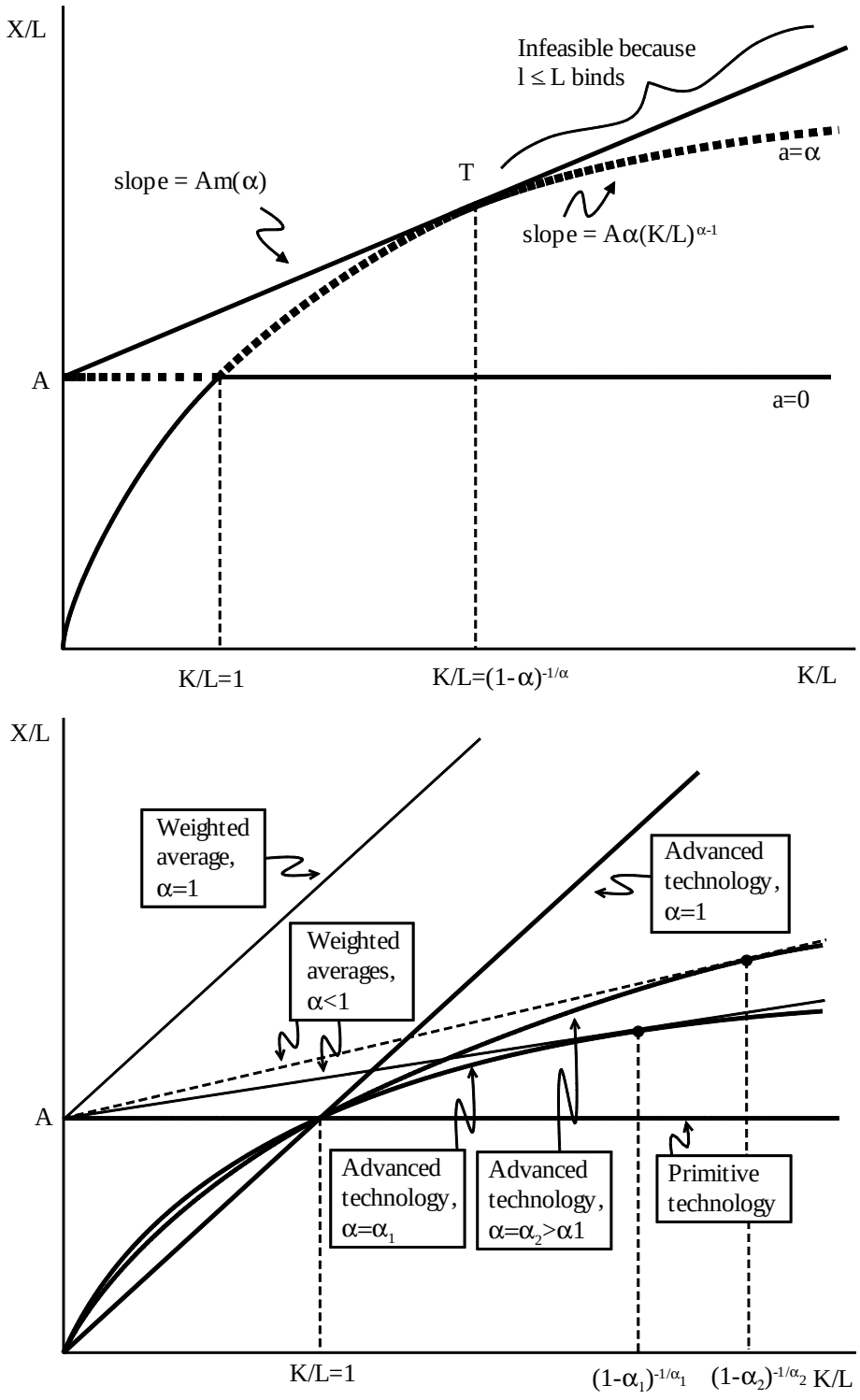

Figure 2: The composite production function

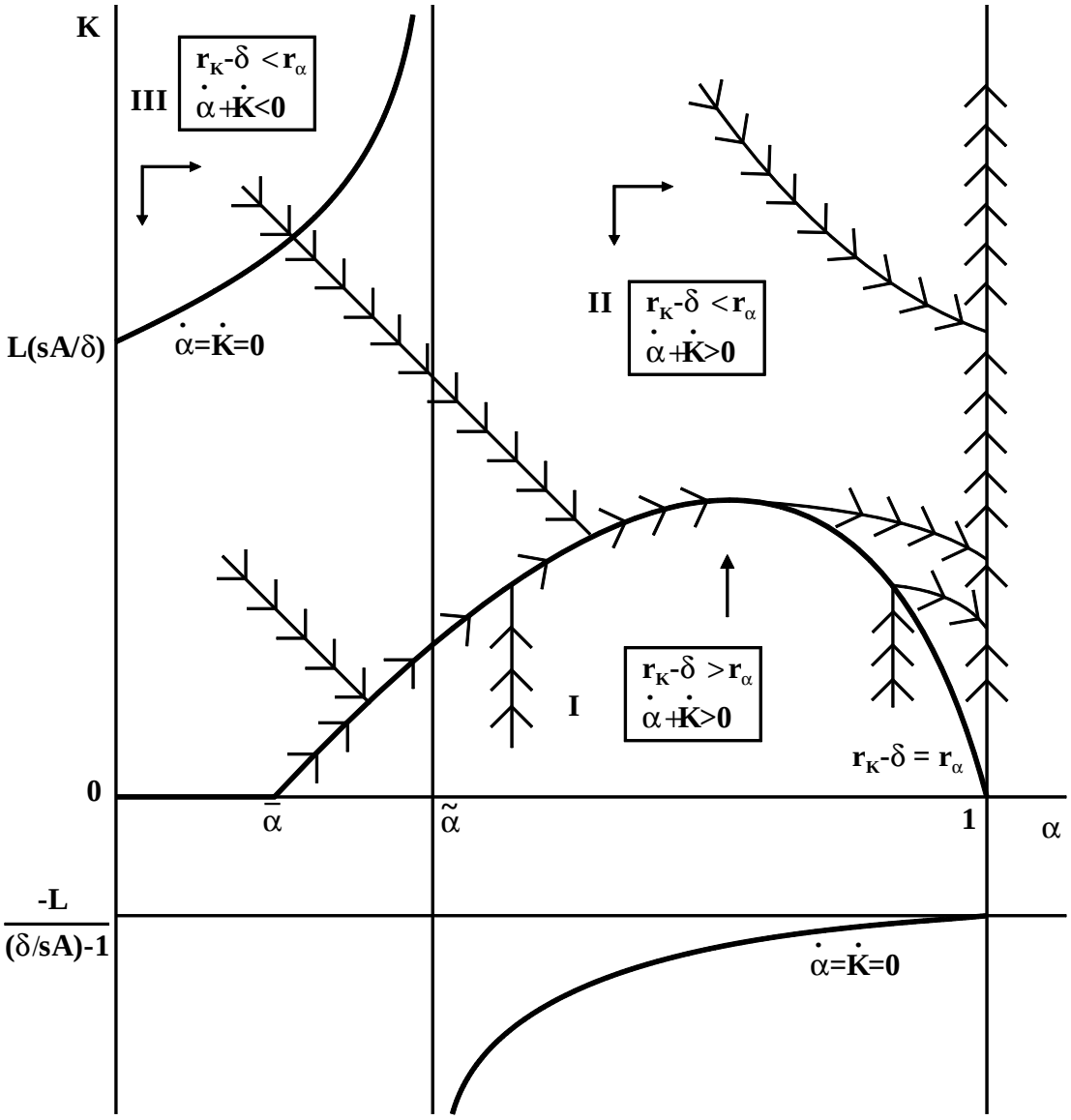

Figure 3: Phase diagram, high saving

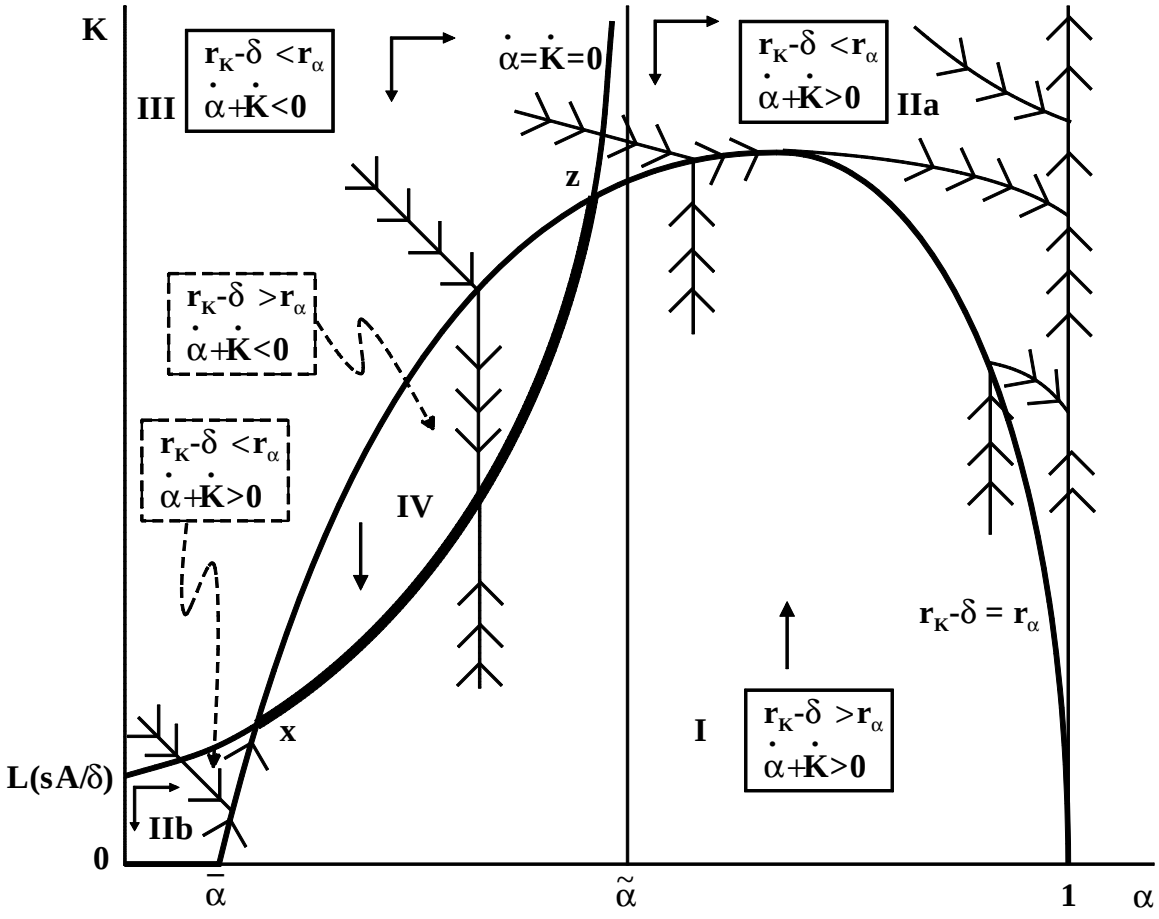

Figure 4: Phase diagram, low saving

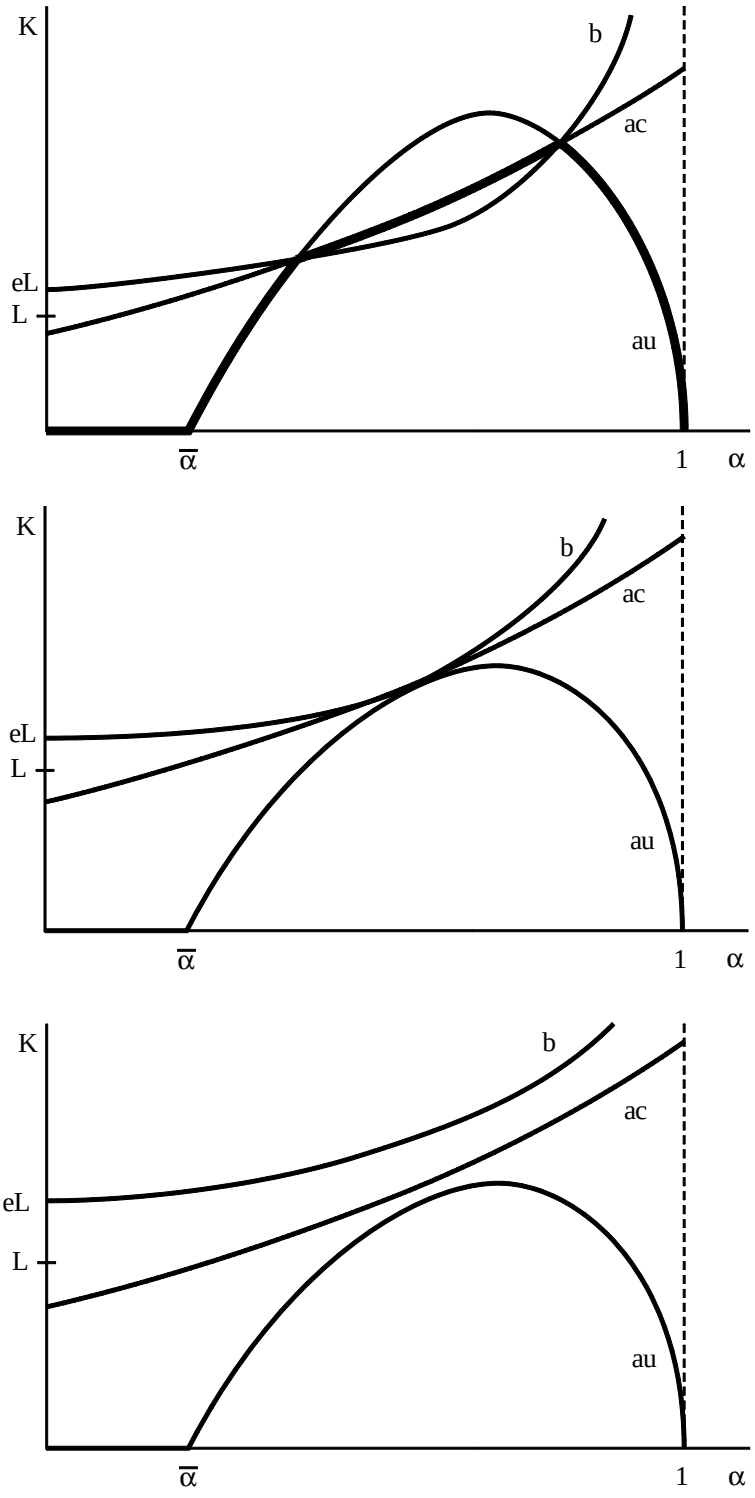

Figure 5: Unconstrained and constrained arbitrage loci

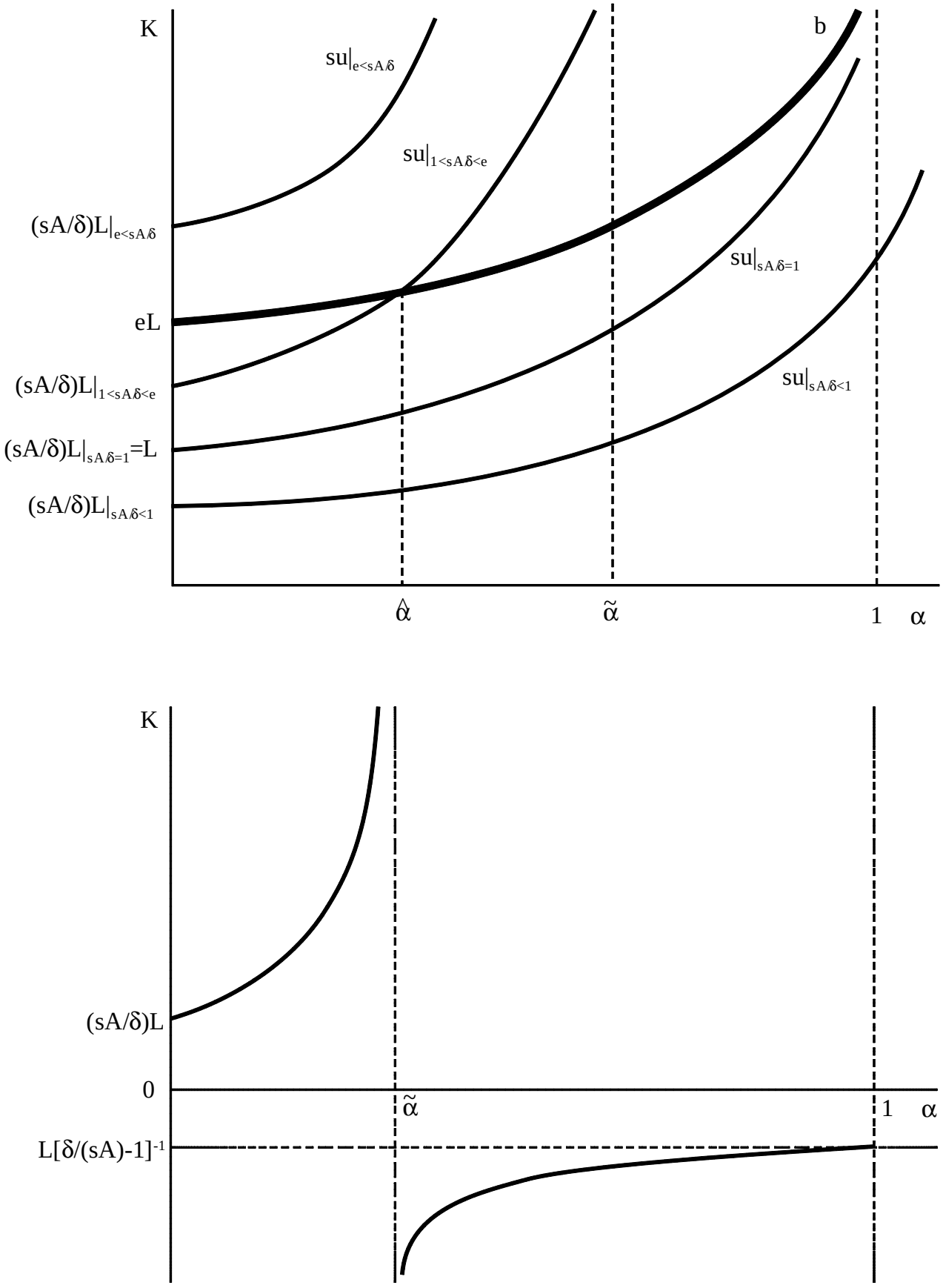

Figure 6: Unconstrained stationarity locus

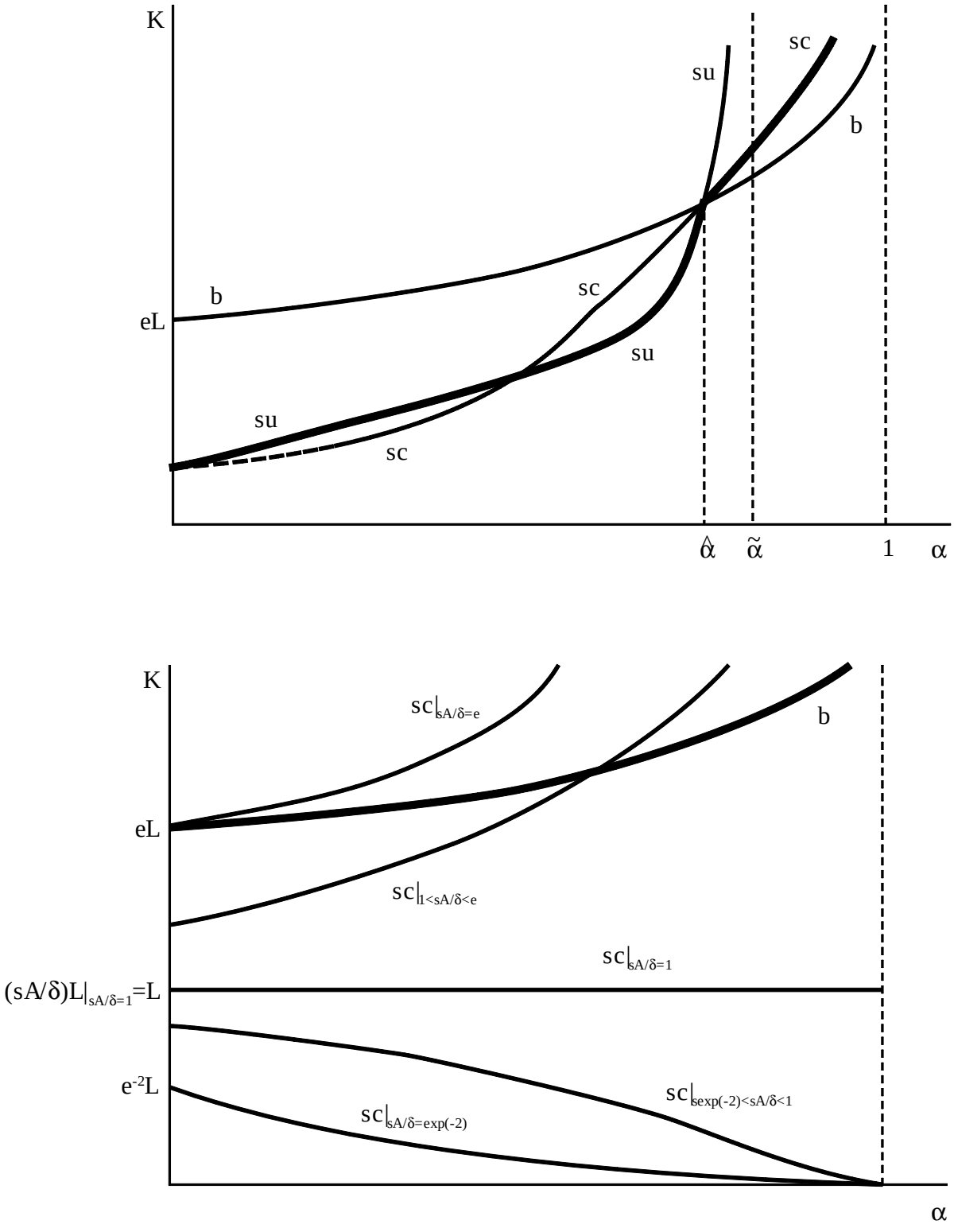

Figure 7: Unconstrained and constrained stationarity loci