Markups, Firm Scale, and Distorted Economic Growth*

Mohamad Adhami[1] Jean-Felix Brouillette[2] Emma Rockall[3]

January 13, 2024

Please click here for the latest version.

Abstract

We study the consequences of markups for long-run economic growth in a model of firm-driven endogenous technological change. In this framework, differentiated firms engage in monopolistic competition, charge heterogeneous markups, and make forward-looking investments in R&D to improve their process efficiency. Markups distort the scale at which these firms operate and, therefore, affect their incentives to invest in R&D. With dispersion in markups, both the aggregate and cross-firm allocations of such investments are distorted. Using firm-level administrative data from France to discipline our model, we find that correcting the product market distortions induced by markups increases the long-run growth rate of productivity by 1.2 percentage points per year. Nearly 75% of this faster productivity growth can be achieved by simply reallocating R&D resources across firms, revealing that the dispersion in markups, rather than their average level, is more detrimental to economic growth.

*We are grateful to our advisors, Pete Klenow, Chad Jones, and Chris Tonetti, for their guidance and support. We would also like to thank Adrien Auclert, Luigi Bocola, Huiyu Li, Ernest Liu, Monika Piazzesi, Martin Schneider, Kurt Sweat, and seminar participants at Stanford University and the Federal Reserve Board of San Francisco for helpful comments. This project was supported by the George P. Shultz Dissertation Support Fund and the B.F. Haley and E.S. Shaw Fellowship for Economics at SIEPR. Access to confidential data, on which this work is based, has been made possible within a secure environment offered by CASD – Centre d'accès sécurisé aux donneées (Ref. 10.34724/CASD). We are indebted to Philippe Aghion for generously hosting us at the Collège de France and to Antonin Bergeaud and Maxime Gravoueille for their help with the data. All errors are our own.

[1]Stanford University. E-mail: adhami@stanford.edu

[2]Stanford University. E-mail: jfbrou@stanford.edu

[3]Stanford University. E-mail: erockall@stanford.edu

1 Introduction

The widespread rise in market concentration in recent decades has raised concerns about the aggregate consequences of product market power.[1] Concurrently, a series of studies have concluded that the macroeconomic costs of markups can be substantial, providing grounds for these concerns.[2] These costs ensue from markups distorting the scale at which firms operate. But how does this distortion affect their incentives to bring better or cheaper products to the market?

This paper quantifies the impact on economic growth of rectifying the product market distortions induced by markups. The assumption that forms the basis of our analysis is that economic growth is sustained by the research and development (R&D) investments of imperfectly competitive firms. While markups are required to recoup the fixed cost of an investment in R&D, they also distort the scale at which firms produce, affecting their incentives to invest in R&D through *partial* and *general* equilibrium channels.

In partial equilibrium, imperfectly competitive firms curtail their own production. However, since the idea behind a technological improvement is nonrival, the return on an investment in R&D increases with the scale at which a firm operates. Therefore, by restricting their output, firms limit the scale at which their improved technology can be deployed, lowering the return on their investment in R&D. Moreover, if these firms differ in the markups they command and differentially hold back their production, both the aggregate and cross-firm allocations of such investments are distorted.

However, in general equilibrium, the consequences of markups for economic growth are a priori ambiguous. Indeed, if firms limit the scale at which they produce, aggregate demand for factors of production will be lower. In environments where those factors (e.g. labor) are shared between the purposes of production and R&D, the low demand for production might increase the availability of resources for R&D and, in turn, speed up the pace of economic growth. To quantitatively resolve this ambiguity, we put forth a general equilibrium model of firm-driven endogenous technological change.

In the model we propose, differentiated firms engage in monopolistic competition and face non-isoelastic demand curves from a final sector that uses their varieties as inputs to produce a final good. These demand curves satisfy Marshall (1890)'s second law of demand, which states that lower prices are met with less elastic demand. This property of demand implies that, in equilibrium, more productive firms are larger and

[1] See Grullon, Larkin and Michaely (2019), Autor, Dorn, Katz, Patterson and Van Reenen (2020), De Loecker, Eeckhout and Unger (2020), and Kehrig and Vincent (2021).

[2] See Baqaee and Farhi (2019), Bilbiie, Ghironi and Melitz (2019), Behrens, Mion, Murata and Suedekum (2020), Edmond, Midrigan and Xu (2023) and Afrouzi, Drenik and Kim (2023).

command higher markups–a fact supported by cumulating empirical evidence.[3]

Over time, firms improve their process efficiency by hiring labor to engage in R&D[4] However, this decision is risky since firms receive independent Brownian productivity shocks in each period. As these firms become more productive and lower their price to attract more demand, they are met with a progressively less elastic demand schedule, enabling them to charge higher markups. This pursuit of profit opportunities through R&D investment stands as one of the two engines of economic growth.

The second one unfolds through the selective replacement of unsuccessful firms by more productive newcomers, sustained by the endogenous process of entry and exit. New entrants incur a labor-denominated entry cost to imperfectly imitate the existing technology of a randomly selected incumbent. Meanwhile, incumbent firms must bear a fixed overhead labor cost per unit of time to remain in business, and failure to do so results in endogenous exit. To meet these costs, and demand from production and R&D, labor is elastically supplied by a representative household.

Through a generalized method of moments (GMM) strategy, we estimate the model's structural parameters from a comprehensive administrative panel dataset of French firms from 2009 to 2019. A central challenge of this exercise is to discipline the degree of markup dispersion in the model and its two sources of productivity growth. In our counterfactual analysis, the extent of R&D reallocation is contingent upon the former, and its consequences for economic growth depend on the latter. Since our model's sole source of markup dispersion derives from size differences across firms, we replicate the empirical relationship between firm-level markups and market shares and the extent of firm size heterogeneity in the data. We further take advantage of the panel structure of this dataset to replicate firm-level and aggregate growth moments.

With the quantified model at hand, we conduct several counterfactual exercises. We first quantify the impact on economic growth of correcting product market distortions from markups. In our model, this can be achieved through size-dependent production subsidies to firms, inducing each to price at marginal cost.[5,6] This intervention increases the growth rate of total-factor productivity (TFP) by 1.2 percentage points per year in the long run, which is attributable to three factors: (1) an increase in aggregate R&D, (2) a reallocation of R&D across firms, and (3) a higher rate of entry and exit.

[3]See De Loecker and Warzynski (2012), Amiti, Itskhoki and Konings (2014), De Loecker, Goldberg, Khandelwal and Pavcnik (2016), Amiti, Itskhoki and Konings (2019), De Loecker et al. (2020), Autor et al. (2020) and De Ridder, Grassi and Morzenti (2023).

[4]In Appendix A.3, we show that this theoretical framework is isomorphic to one in which firms improve the quality of their product when quality and quantity are perfect substitutes.

[5]This could alternatively be achieved by enabling firms to engage in perfect price discrimination.

[6]This intervention alone does not decentralize the optimal allocation of resources, which is derived and presented in Appendix A.1.8.

As the intervention takes hold, firms scale up, further invest in R&D, and achieve faster productivity growth. This dynamic is most pronounced for the largest firms, which initially commanded the highest markups and thus restricted their output to the greatest extent. Correcting product market distortions induces these firms to disproportionately expand their production, yielding a differentially higher return on their investments in R&D. Consequently, their more frequent improvements in process efficiency are rolled out over a larger volume of output, further contributing to productivity growth.

In contrast, the smallest, least productive firms face intensified competitive pressures and are thus edged out of the market at a higher rate. As R&D employment is reallocated towards their larger competitors, these small firms trail behind and exit at a higher rate, unable to cover the overhead cost, which grows faster due to more rapid wage growth. Since more productive new entrants replace these unsuccessful firms, the productivity pool undergoes more frequent improvements, thereby accelerating TFP growth.

We conduct alternative exercises to gain insight into the catalysts behind this growth acceleration. To assess the role of a greater *aggregate* allocation of labor to R&D relative to its *reallocation* across firms, we consider a "constrained" intervention in which we hold the former fixed. Here, firm-level production subsidies are upheld, but a uniform tax is concomitantly levied on their R&D expenditures to leave the aggregate allocation of labor to R&D at a level commensurate with the initial equilibrium. This intervention achieves an increase in productivity growth of almost 75% of the increment recorded under its "unconstrained" counterpart, revealing a prominent role for the reallocation, rather than expansion, of R&D employment.

Finally, we explore more flexible tax and subsidy schemes to disentangle the role of the aggregate markup from that of markup dispersion. A uniform subsidy that rectifies the level of markups while leaving their dispersion unchanged, *reduces* the long-run growth rate of TFP by a muted 4 basis points. On one hand, as firms expand in scale and distribute their R&D expenditures over more units sold, the return on those investments increases. On the other, these firms demand more production labor and inadvertently bid up the cost of R&D through a higher wage. These offsetting channels indicate that while the level of markups significantly restricts the scale of the economy (Edmond et al., 2023), it has nearly no bearing on the rate at which it grows.

Conversely, a size-dependent tax and subsidy scheme that addresses the dispersion in markups while holding their average level fixed increases long-run productivity growth by 1.3 percentage points, a slight uptick from the baseline subsidy scheme. Notably, this scheme is nearly budget-neutral, even raising revenue equivalent to 0.6% of output, which is rebated to the household. The ensuing reallocation of R&D employment from small, unproductive firms towards their larger, further expanding competitors ensures that (1) process efficiency improvements are rolled out over more units sold, and (2) more productive new entrants more frequently replace inefficient exiting firms.

These counterfactual exercises indicate that rectifying product market distortions induced by markups has a substantial positive effect on economic growth. This is mostly due to a reallocation of R&D employment from small to large firms. Our findings suggest that the welfare consequences of markups may surpass our previous approximations, providing further grounds for concern about recent trends in market concentration.

The rest of the paper is outlined as follows. In the remainder of this section, we discuss the relevant literature. Section 2 provides partial equilibrium intuition on the dynamic consequences of markups. Section 3 presents our general equilibrium model. Section 4 describes the quantification of this model. Section 5 presents the results of our counterfactual analysis and Section 6 concludes.

Related Literature

Our paper is primarily related to a longstanding literature on the macroeconomic costs of product market power. Classic analyses can be traced back to Smith (1776), Lerner (1934), Harberger (1954) and Dixit and Stiglitz (1977). Quantitatively, Baqaee and Farhi (2019), Bilbiie et al. (2019), Behrens et al. (2020), Edmond et al. (2023) and Afrouzi et al. (2023) are recent examples of papers concluding that the aggregate efficiency losses from markups can be large. The mechanisms stressed in this literature revolve around markups restricting the economy's scale of production, distorting the allocation of factors of production across firms, and inducing inefficient entry (Dhingra and Morrow, 2019).

Yet, by distorting the scale at which firms operate, the markups they command alter their incentives to invest in R&D. This is akin to a firm-specific "market size" effect. The notion that incentives for R&D are dictated by the extent of the market is extensively discussed in Schmookler (1966) and succinctly captured by a quote from Matthew Boulton, a mechanical engineer, and business partner of James Watt:

> "It would not be worth my while to make [steam engines] for three countries only; but I find it very well worth my while to make [them] for all the world."
> – Matthew Boulton (Scherer, 1965)

This phenomenon is explored in the analyses of Arrow (1962) and Dasgupta and Stiglitz (1980), and appears in the lab equipment model of Rivera-Batiz and Romer (1991).[7] Our contribution is to quantify its consequences for economic growth in a framework

[7] Unlike the canonical Romer (1990) model, the lab equipment model features no technology spillovers. Nevertheless, economic growth is inefficiently low in the decentralized equilibrium as private incentives

that features heterogeneity in markups. This is motivated by the mounting evidence of substantial markup dispersion across firms (De Loecker and Warzynski, 2012; Amiti et al., 2014; De Loecker et al., 2016; Amiti et al., 2019; De Loecker et al., 2020; De Ridder et al., 2023). To tractably yet endogenously replicate such dispersion, our model features a non-isoelastic demand system à la Kimball (1995), with the functional form proposed by Klenow and Willis (2016).

Under such a demand system, markup dispersion derives from price differences across firms, which, in our model, reflect heterogeneity in their process efficiency. To endogenously deliver such heterogeneity, our model borrows from the firm dynamics literature and extends the frameworks of Hopenhayn (1992) and Luttmer (2007) to allow for forward-looking and risky investments in R&D.[8] Following Ericson and Pakes (1995), Benhabib, Perla and Tonetti (2021) and Lashkari (2023), the firm's investment choice is formulated as a stochastic optimal control problem, and its endogenous exit decision takes the form of an optimal stopping time problem.

The counterfactual analysis we conduct in this model differs from those considered in Peters (2020), Cavenaile, Celik and Tian (2021) and Voronina (2021), which are otherwise closely related to our paper. The former two propose models of growth through creative destruction in which heterogeneous markups arise endogenously as the outcome of firms' investments in R&D. Peters (2020) quantifies the aggregate static efficiency losses from markups whereas our focus is on their consequences for long-run economic growth. Cavenaile et al. (2021) study an economy whose structural parameters change over time to replicate the observed trend in markups and quantify the extent to which the resulting static efficiency costs are mitigated or amplified by the endogenous response of firms' investments in R&D.

Our counterfactual analysis is closest to Voronina (2021), who puts forth a theory of firm-driven endogenous growth in which markups are heterogeneous but *exogenous*. Through the lens of this model, she quantifies the improvement in welfare that a social planner can achieve by choosing flexible transfers to firms. However, the planner can design these flexible transfers to fix other market failures (e.g., technology spillovers) such that this counterfactual exercise does not *isolate* the costs of markups. In contrast, the subsidy schedule we consider is constrained to a structure that induces all firms to price at marginal cost, thus directly addressing the product market distortions induced by markups.

to create new intermediate varieties are weakened by final good producers who substitute marked-up intermediates for labor.

[8]This is in line with the findings of Foster, Haltiwanger and Krizan (2001) and Garcia-Macia, Hsieh and Klenow (2019) who infer large contributions to productivity growth from entry and exit, and technology improvements carried by incumbent firms on their existing products.

The cross-firm reallocation of R&D employment instigated by the intervention we consider relates our paper to a growing literature on the misallocation R&D resources, including Akcigit, Celik and Greenwood (2016), Acemoglu, Akcigit, Alp, Bloom and Kerr (2018), Akcigit, Hanley and Serrano-Velarde (2020), Liu and Ma (2021), Hopenhayn and Squintani (2021), Chen, Liu, Suárez Serrato and Xu (2021), Akcigit, Hanley and Stantcheva (2022), König, Storesletten, Song and Zilibotti (2022), De Ridder (2023), Ayerst (2023) and Lehr (2023). We contribute to this literature by quantifying a novel mechanism that can give rise to such misallocation, which results from the dispersion of markups across firms within an industry.

Acemoglu (2023) and Aghion, Bergeaud, Boppart, Klenow and Li (2023) are closely related to our study, as they investigate the consequences of markups on the allocation of R&D resources. Acemoglu (2023) studies their allocation across sectors (rather than firms) in a setting with heterogeneous (yet exogenous) markups. Aghion et al. (2023) distinguish between "good" and "bad" markups. The former reflects a firm's quality advantage over its competitors, which confers positive technology spillovers onto other firms, while the latter reflects its higher process efficiency, with no associated spillovers. Hence, they find that the allocation of R&D resources is inefficiently distorted away from high markup firms in the former but not the latter case.

2 Partial Equilibrium Intuition

We consider a simple two-period partial equilibrium model to form intuition on how markups alter private incentives for productivity-enhancing investments. In this setting, either a profit-maximizing monopolist or a welfare-maximizing agent operates a firm and allocates resources to achieve an endogenously chosen reduction in its marginal cost (Arrow, 1962; Dasgupta and Stiglitz, 1980; Tirole, 1988; Garella, 2012).

The setup is as follows. In both periods, a household inelastically supplies a factor whose price is exogenous and normalized to unity. This household has preferences over the consumption of a commodity whose price is denoted by p. Assume that these preferences imply a twice differentiable demand function $y(p)$ that satisfies:

$$\frac{\partial y(p)}{\partial p} < 0, \quad \vartheta(p) \equiv -\frac{\partial \ln(y(p))}{\partial \ln(p)} > 1 \quad \text{and} \quad \varepsilon(p) \equiv \frac{\partial \ln(\vartheta(p))}{\partial \ln(p)} \in \mathbb{R}$$

where $\vartheta(p)$ denotes the price elasticity of demand at price p, and $\varepsilon(p)$ denotes the "super-elasticity" of demand at that price. In the post-period, the commodity is produced by a firm using the factor supplied by the household according to a technology with constant returns to scale described by the marginal cost function:

$$c(z') = \exp(-z').$$

Here, z' denotes the firm's productivity in the post-period, which can be controlled by the agent operating the firm in the pre-period. More specifically, $i(g)$ units of the factor can be invested in the pre-period to achieve a g% improvement in the firm's post-period process efficiency:

$$z' = g + z$$

where z denotes the firm's pre-period process efficiency. The twice differentiable function i is assumed to be strictly increasing and strictly convex and satisfies $i(0) = 0$ and $\lim_{g\to\infty} i(g) = \infty$.

In this environment, we now compare the decision problems of a profit-maximizing monopolist and a welfare-maximizing agent operating the firm. In the post-period, both agents face a *static* problem. Given the household's demand function, they must choose a unit price at which to sell the commodity to maximize profits (producer surplus) or social surplus (the sum of producer and consumer surplus). The two objectives are respectively denoted by $\pi(z', p)$ and $S(z', p)$:[9]

$$\pi(z', p) \equiv [p - \exp(-z')]y(p) \quad \text{and} \quad S(z', p) \equiv \pi(z', p) + \int_p^{\overline{p}} y(p')\mathrm{d}p'.$$

Assuming demand is positive at optimally chosen prices, it is straightforward to show that maximized producer and social surpluses are given by:

$$\pi(z') \equiv \frac{p(z')y(p(z'))}{\vartheta(p(z'))} \quad \text{and} \quad S(z') \equiv \int_{c(z')}^{\overline{p}} y(p)\mathrm{d}p$$

where $p(z')$ denotes the usual profit-maximizing price implicitly defined as:

$$p(z') \equiv \frac{\vartheta(p(z'))}{\vartheta(p(z')) - 1} \times c(z').$$

Let us now consider the agents' *dynamic* problem. In the pre-period, they must choose a factor allocation to investments in R&D to maximize post-period producer or social surplus. Up to a first-order approximation of these objectives and assuming no time

[9] Here, $\overline{p}$ denotes the choke price.

discounting, these dynamic problems are described by:

$$\max_{g}\{\pi(z)+\pi'(z)g-i(g)\} \quad \text{and} \quad \max_{g}\{S(z)+S'(z)g-i(g)\}$$

where $\pi'(z)$ and $S'(z)$ denote the partial derivatives of producer and social surplus with respect to the firm's initial productivity. The first-order conditions of each problem are:

$$\pi'(z)=i'(g) \quad \text{and} \quad S'(z)=i'(g).$$

Therefore, private and social incentives for marginal cost reductions may not coincide if *marginal* producer and social surpluses differ. The proposition that follows characterizes the ratio $R(z)\equiv \pi'(z)/S'(z)$ of these objects.

Proposition 1. *The ratio $R(z)$ of marginal producer to social surplus from an infinitesimal reduction in marginal cost is characterized by:*

$$R(z)\equiv\frac{\pi'(z)}{S'(z)}=\frac{y(p(z))}{y(\exp(-z))}<1.$$

When the welfare-maximizing agent is instead constrained to produce at the same scale as the monopolist, the ratio $R^c(z)$ of marginal producer surplus to "constrained" marginal social surplus is characterized by:

$$R^c(z)\equiv\frac{\pi'(z)}{\pi'(z)+C'(z)}=\frac{\vartheta(p(z))+\varepsilon(p(z))-1}{2\vartheta(p(z))+\varepsilon(p(z))-1}<1$$

where $C(z)\equiv\int_{p(z)}^{\overline{p}}y(p)dp$ denotes consumer surplus.

Proposition 1 shows that in this setting, the welfare-maximizing agent always faces stronger incentives to achieve a marginal cost reduction than the monopolist, even when the two are constrained to operate at the same scale. The intuition behind this proposition is twofold. First, since the welfare-maximizing agent optimally operates at a larger scale than the monopolist, the same reduction in marginal cost applies to more units produced, thus begetting larger total cost savings. This is evident from the expression of the ratio $R(z)$, which is equal to the ratio of the quantities produced by the monopolist and the welfare-maximizing agent.

Second, at a given scale of operation, the welfare-maximizing agent internalizes that a reduction in marginal cost may achieve additional consumer surplus, whereas the monopolist does not. Intuitively, since the monopolist does not price discriminate (by assumption), it faces a trade-off as it becomes more productive. It can lower its price to sell additional units, but doing so comes at the cost of reaping lower producer surplus per unit sold. In contrast, the split between consumer and producer surplus is irrelevant to the welfare-maximizing agent. This agent internalizes that an increase in consumer surplus exactly balances any loss in producer surplus from a price reduction. This is illustrated in the second part of the proposition, where it is clear that the monopolist only "appropriates" a fraction of the marginal surplus achieved by the marginal cost reduction. We now take this first proposition further to characterize how the distance between private and social incentives for R&D depends on the firm's initial productivity.

Proposition 2. *The elasticity of the ratio $R(z)$ with respect to the firm's initial productivity is characterized by:*

$$\frac{\partial \ln(R(z))}{\partial z} = \vartheta(p(z))\varrho(z) - \vartheta(\exp(-z)) \quad \textit{where} \quad \varrho(z) \equiv -\frac{\partial \ln(p(z))}{\partial z}.$$

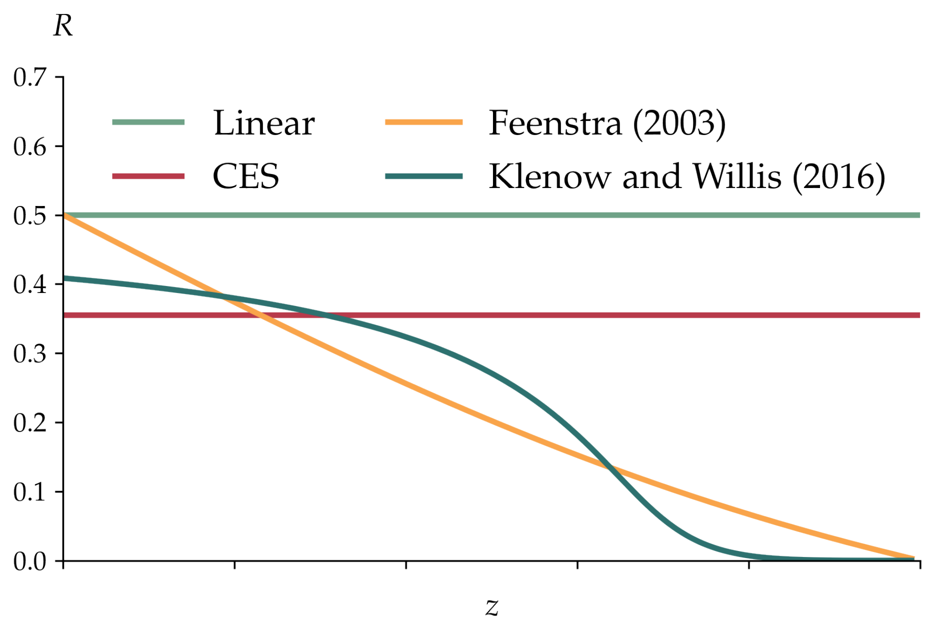
Here, $\varrho(z)$ denotes the monopolist's productivity "pass-through" (i.e., the percent change in the monopolist's price following a one percent improvement in its productivity), which is a function of the price elasticity and super-elasticity of demand:

$$\varrho(z) = \frac{\vartheta(p(z)) - 1}{\vartheta(p(z)) + \varepsilon(p(z)) - 1}.$$

Proposition 2 shows that the distance between private and social incentives depends on the price elasticity and super-elasticity of demand. To illustrate this, Figure 1 plots the ratio $R(z)$ implied by four common demand functions: the isoelastic (CES) demand function, the linear demand function, the Klenow and Willis (2016) specification of the Kimball (1995) demand function and the Translog demand function proposed by Feenstra (2003). Under the CES demand function, social incentives for marginal cost reductions exceed private incentives by the same proportion, regardless of the firm's productivity. In contrast, for the non-isoelastic Kimball and Translog demand functions, the distance between social and private incentives increases with firm productivity. However, this property does not generically hold for all non-isoelastic demand functions, as the linear demand function features a constant ratio $R(z) = 1/2$.

Since this section presented a stylized partial equilibrium environment, we have abstracted from the possibility that the private and social costs of investments in R&D may not coincide. This motivates the following section, which presents a more involved theory where market prices may fail to reflect the social value of resources. As discussed in Section 1, this possibility must be crucially accounted for when quantifying the consequences of markups for economic growth in general equilibrium.

Figure 1: Private vs. Social Incentives for Productivity Improvements



Note: The vertical axis measures the ratio $R(z)$ of marginal producer to social surplus from an infinitesimal reduction in marginal cost.

3 Theory

In this section, we propose a general equilibrium theory of endogenous economic growth that builds on the partial equilibrium intuition presented in the previous section. We extend the model of Luttmer (2007) to allow for endogenous productivity improvements by incumbent firms who face non-isoelastic demand curves.

3.1 Economic Environment

Preferences

Consider an economy populated by an infinitely-lived representative household of unit measure with separable preferences over consumption C_t and hours worked H_t such that lifetime utility is defined as:

$$U_0 = \int_0^\infty e^{-\rho t}[\ln(C_t) - v(H_t)]\mathrm{d}t. \tag{1}$$

Here, $\rho > 0$ is the household's rate of time preference, the function v is strictly increasing and convex, and time is continuous and indexed by $t \in \mathbb{R}_0^+$.

Technology

The economy is composed of two sectors: the final and intermediate sectors. The final sector produces a final good using a continuum of differentiated varieties indexed by j from the intermediate sector. The final sector's production technology has constant returns to scale and is defined implicitly by the following Kimball (1995) aggregator:

$$\int_{j \in \mathcal{J}_t} \Upsilon(\hat{y}_{jt}) \mathrm{d}j = 1 \quad \text{where} \quad \hat{y}_{jt} \equiv \frac{y_{jt}}{Y_t}. \tag{2}$$

Here, Y_t denotes aggregate output and y_{jt} is the quantity of variety j used in production. The function Υ is strictly increasing, strictly concave, and satisfies $\Upsilon(1) = 1$. In what will follow, we denote the measure of varieties at time t by $M_t \equiv |\mathcal{J}_t|$. This production function belongs to the family of homothetic aggregators with direct implicit additivity (HDIA) as defined in Matsuyama and Ushchev (2017). In particular, it nests the Dixit and Stiglitz (1977) aggregator when $\Upsilon(\hat{y}) = \hat{y}^{\frac{\theta-1}{\theta}}$, where $\theta > 1$ would denote the constant elasticity of substitution across varieties.

Each variety is produced by a single firm from the intermediate sector using physical capital and production labor with Hicks-neutral productivity z_{jt} according to a Cobb-Douglas production technology:

$$y_{jt} = \exp(z_{jt}) k_{jt}^{\alpha} l_{jt}^{1-\alpha}. \tag{3}$$

Here, k_{jt} and l_{jt} respectively denote the quantities of capital and labor used in production, and $\alpha \in [0, 1]$ denotes the output elasticity of capital. As in Hopenhayn (1992) and Luttmer (2007), firms must pay an overhead of $c_O > 0$ units of labor per unit of time to remain active. If this cost is unpaid, a firm must irreversibly exit. Firms may also exit exogenously at Poisson rate $\chi > 0$.

At any point in time, a firm is fully described by its productivity $z_t \in \mathbb{R}$ such that, from this point on, we abandon the j-index notation. Over time, firms can improve their process efficiency by allocating labor to R&D. More precisely, productivity evolves according to a controlled diffusion process of the form:

$$\mathrm{d}z_t = \gamma_t \mathrm{d}t + \sigma \mathrm{d}B_t$$

where $\gamma_t > 0$ is the controlled drift, $\mathrm{d}B_t$ is the standard normal increment of a Brownian motion and $\sigma > 0$ is its standard deviation.[10] Defining a firm's productivity relative to the least productive firm in the economy as $\hat{z}_t \equiv z_t - \underline{z}_t$, we obtain the following law of motion by Itô's lemma:

$$\mathrm{d}\hat{z}_t = (\gamma_t - g_t)\mathrm{d}t + \sigma \mathrm{d}B_t. \tag{4}$$

Here, $\underline{z}_t$ and g_t respectively denote the productivity lower bound and its instantaneous rate of change. The labor requirement to achieve a drift of γ for a firm with relative productivity $\hat{z}$ is $i(\gamma, \hat{z}) : \mathbb{R}_0^+ \times \mathbb{R}_0^+ \to \mathbb{R}_0^+$, where the function i is strictly increasing and convex in its first argument and satisfies $i(0, \hat{z}) = 0$ and $\gamma(\hat{z}) < \infty$ for all $\hat{z} \in [0, \infty)$.

In every period, a measure of potential entrants can allocate $c_E > 0$ units of labor to achieve a unit flow of entry and start producing with a relative productivity draw from the cumulative density function (CDF) $F_t^E(\hat{z})$. This function is a transformation of the relative productivity CDF of incumbent firms $F_t(\hat{z})$, as defined by the non-decreasing function $T : [0, 1] \to [0, 1]$ such that $T(0) = 0$ and $T(1) = 1$:[11]

$$F_t^E(\hat{z}) = 1 - T[1 - F_t(\hat{z})]. \tag{5}$$

In particular, we assume that this transformation is such that the right tail of the relative productivity distribution decays faster for entrants than incumbents:

$$\lim_{\hat{z} \to \infty} \frac{1 - F_t^E(\hat{z})}{1 - F_t(\hat{z})} = 0.$$

As discussed in Section 3.3, we impose this condition to achieve a unique stationary distribution of relative productivity on a balanced growth path (e.g., a simple power function for T satisfies this condition if the exponent is greater than one).

Resource Constraints

The final good can either be consumed or invested in physical capital:

$$\dot{K}_t + \delta K_t + C_t \leq Y_t \quad \text{where} \quad K_t \equiv \int_0^\infty k_t(\hat{z}) \mathrm{d}M_t(\hat{z}). \tag{6}$$

Here, $\delta > 0$ is the rate at which capital depreciates and $M_t(\hat{z}) = F_t(\hat{z}) M_t$ denotes the cumulative density of firms with relative productivity below $\hat{z}$ at time t. Labor supplied

[10] Productivity shocks are independent and identically distributed across firms.

[11] This type of transformation is often referred to as a dual distortion function.

by the household can be allocated to either production, innovation, entry, or overhead:

$$L_t + I_t + c_E E_t + c_O M_t \leq H_t \tag{7}$$

where E_t denotes the aggregate flow of entry, and the aggregate allocations of labor to production and innovation are defined as:

$$L_t \equiv \int_0^\infty l_t(\hat{z}) \mathrm{d}M_t(\hat{z}) \quad \text{and} \quad I_t \equiv \int_0^\infty i(\gamma_t(\hat{z}), \hat{z}) \mathrm{d}M_t(\hat{z}).$$

Laws of Motion

Denoting the density of firms with relative productivity equal to $\hat{z}$ at time t by $m_t(\hat{z})$, the Kolmogorov forward (KF) equation describing its evolution over time is:

$$\dot{m}_t(\hat{z}) = \mathcal{A}_t^* m_t(\hat{z}) + E_t f_t^E(\hat{z}) \quad \forall \hat{z} > 0. \tag{8}$$

Here, $f_t^E(\hat{z})$ is the probability density function from which entrants draw their relative productivity, and $\mathcal{A}_t^*$ is the adjoint of the operator $\mathcal{A}_t$, defined as:

$$\mathcal{A}_t \equiv [\gamma_t(\hat{z}) - g_t]\partial_{\hat{z}} + (\sigma^2/2)\partial_{\hat{z}\hat{z}} - \chi \tag{9}$$

where $\partial_{\hat{z}}$ and $\partial_{\hat{z}\hat{z}}$ denote the first and second partial derivative operators with respect to $\hat{z}$. The measure of varieties then follows the law of motion:

$$\dot{M}_t = E_t - \chi M_t - (\sigma^2/2) M_t''(0).$$

The economic environment is summarized in Table 1.

3.2 Decision Problems

We now define the decision problems of economic agents, which determine equilibrium prices and quantities on the final good, varieties, labor, and asset markets. In terms of market structure, we assume that all agents partake in perfect competition in all markets besides intermediate firms who engage in monopolistic competition and choose the price at which to sell their variety.

Table 1: The economic environment

(1)	$U_0 = \int_0^\infty e^{-\rho t}[\ln(C_t) - v(H_t)]\mathrm{d}t$	Preferences
(2)	$\int_0^\infty \Upsilon(\hat{y}_t(\hat{z}))\mathrm{d}M_t(\hat{z}) = 1$	Final good production technology
(3)	$y_t(\hat{z}) = \exp(\hat{z} + \underline{z}_t)k_t(\hat{z})^\alpha l_t(\hat{z})^{1-\alpha}$	Variety production technology
(4)	$\mathrm{d}\hat{z}_t = (\gamma_t - g_t)\mathrm{d}t + \sigma \mathrm{d}B_t$	Innovation technology
(5)	$F_t^E(\hat{z}) = 1 - T[1 - F_t(\hat{z})]$	Entrants' productivity distribution
(6)	$\dot{K}_t + \delta K_t + C_t \leq Y_t$	Final good resource constraint
(7)	$L_t + I_t + c_E E_t + c_O M_t \leq H_t$	Labor resource constraint
(8)	$\dot{m}_t(\hat{z}) = \mathcal{A}_t^* m_t(\hat{z}) + E_t f_t^E(\hat{z})$	Incumbents' productivity density

The Household's Problem

Taking prices as given, the household's problem is to choose its consumption and hours worked to maximize lifetime utility subject to a flow budget constraint:

$$\max_{\{C_t, H_t\}_{t\geq 0}} \int_0^\infty e^{-\rho t}[\ln(C_t) - v(H_t)]\mathrm{d}t \quad \text{s.t.} \quad \dot{A}_t = r_t A_t + w_t H_t - C_t$$

where w_t denotes the wage rate, A_t is the value of physical capital and corporate assets, and r_t is the rate of return on those assets:

$$A_t = K_t + \int_0^\infty V_t(\hat{z})\mathrm{d}M_t(\hat{z}) \quad \text{where} \quad \lim_{t\to\infty} e^{-\int_0^t r_{t'}\mathrm{d}t'} A_t = 0.$$

Here, $V_t(\hat{z})$ denotes the value of a firm with relative productivity $\hat{z}$, which is yet to be defined. The household's problem thus delivers the usual intertemporal Euler equation and static first-order condition:

$$\frac{\dot{C}_t}{C_t} = r_t - \rho \quad \text{and} \quad v'(H_t)C_t = w_t.$$

The Final Sector's Problem

Taking prices as given, the final sector's problem is to choose its relative demand for each variety to maximize profits in each period:

$$\max_{\{\hat{y}_t(z)\}_{\hat{z}=0}^{\infty}} \left\{ P_t - \int_0^{\infty} p_t(\hat{z})\hat{y}_t(\hat{z})\mathrm{d}M_t(\hat{z}) \right\} Y_t \quad \text{s.t.} \quad \int_0^{\infty} \Upsilon(\hat{y}_t(\hat{z}))\mathrm{d}M_t(\hat{z}) = 1$$

where P_t and $p_t(\hat{z})$ respectively denote the price of the final good and the price charged by a firm with relative productivity $\hat{z}$. Therefore, this problem delivers the following inverse demand functions:

$$p_t(\hat{z}) = \Upsilon'(\hat{y}_t(\hat{z}))P_t D_t \quad \text{where} \quad P_t \equiv \int_0^{\infty} p_t(\hat{z})\hat{y}_t(\hat{z})\mathrm{d}M_t(\hat{z}).$$

Here, the final good is chosen as the numéraire such that $P_t = 1$ for all t and D_t is a demand index defined as:

$$D_t \equiv \left(\int_0^{\infty} \Upsilon'(\hat{y}_t(\hat{z}))\hat{y}_t(\hat{z})\mathrm{d}M_t(\hat{z}) \right)^{-1}.$$

The Firm's Static Problem

Firms engage in monopolistic competition in the product market but perfect competition in the input markets. A firm chooses the price at which to sell its variety and its demand for physical capital and production labor to maximize profits in each period. The firm takes as given the demand for its variety, the rental rate of capital r_t and the wage rate w_t, which delivers the following problem:

$$\begin{aligned} \pi_t(\hat{z}) = & \max_{p_t(\hat{z}),k_t(\hat{z}),l_t(\hat{z})} \{p_t(\hat{z})y_t(\hat{z}) - (r_t + \delta)k_t(\hat{z}) - w_t l_t(\hat{z})\} - w_t c_O \\ & \text{s.t.} \quad p_t(\hat{z}) = \Upsilon'(\hat{y}_t(\hat{z}))D_t. \end{aligned}$$

The firm's optimal choices of physical capital and production labor imply that we can rewrite the problem as:

$$\pi_t(\hat{z}) = \max_{p_t(\hat{z})} \{[p_t(\hat{z}) - \varsigma_t \exp(-\hat{z} - \underline{z}_t)]\hat{y}_t(\hat{z})\} Y_t - w_t c_O$$

where ς_t denotes the producer price index of inputs:

$$\varsigma_t \equiv \left(\frac{r_t+\delta}{\alpha}\right)^{\alpha}\left(\frac{w_t}{1-\alpha}\right)^{1-\alpha}.$$

This, in turn, implies that the firm sets its price to a markup $\mu_t(\hat{z})$ above marginal cost:

$$p_t(\hat{z}) = \mu_t(\hat{z}) \times \frac{\varsigma_t}{\exp(\hat{z}+\underline{z}_t)} \quad \text{where} \quad \mu_t(\hat{z}) \equiv \frac{\vartheta_t(\hat{z})}{\vartheta_t(\hat{z})-1}$$

and where $\vartheta_t(\hat{z})$ denotes the price elasticity of demand:

$$\vartheta_t(\hat{z}) \equiv -\frac{\Upsilon'(\hat{y}_t(\hat{z}))}{\Upsilon''(\hat{y}_t(\hat{z}))\hat{y}_t(\hat{z})} \in (1,\infty).$$

Since the markup function is monotonic in relative demand, which is itself monotonic in the firm's price, a unique solution exists for the latter as an implicit function of relative productivity and calendar time. As such, firm-level profits can be expressed as:

$$\pi_t(\hat{z}) = \frac{p_t(\hat{z})\hat{y}_t(\hat{z})Y_t}{\vartheta_t(\hat{z})} - w_t c_O.$$

The Firm's Dynamic Problem

Given the above static profit function and taking the wage rate as given, firms control the drift of their productivity and choose an optimal exit time τ at which to shut down operations:

$$V_t(\hat{z}) = \max_{\tau,\{\gamma_s\}_{s\geq t}} \mathbb{E}_{\hat{z}} \left\{ \int_t^{t+\tau} e^{-\int_t^s (r_{t'}+\chi)\mathrm{d}t'} [\pi_s(\hat{z}_s) - w_s i(\gamma_s,\hat{z}_s)]\mathrm{d}s \right\}$$

where $\mathbb{E}_{\hat{z}}$ denotes the expectation operator with respect to the diffusion process $\{\hat{z}_s\}_{s\geq t}$ when its initial value is $\hat{z}_t = \hat{z}$. Within the continuation region of productivity (i.e., where it is not optimal to exit), the firm's value function satisfies the standard Hamilton-Jacobi-Bellman (HJB) equation:

$$r_t V_t(\hat{z}) = \pi_t(\hat{z}) + \max_{\gamma}\{\mathcal{A}_t V_t'(\hat{z}) - w_t i(\gamma,\hat{z})\} + \dot{V}_t(\hat{z})$$

with value matching, smooth pasting, and first-order conditions:

$$V_t(0) = V_t'(0) = 0 \quad \text{and} \quad V_t'(\hat{z}) = w_t \times \frac{\partial i(\gamma,\hat{z})}{\partial \gamma}.$$

The Entrant's Problem

Entrants engage in perfect competition on the labor market and, therefore, choose a flow of entry to maximize future expected profits while taking the wage rate as given:

$$V_t^E = \max_{E_t} \left\{ E_t \int_0^\infty V_t(\hat{z}) \mathrm{d}F_t^E(\hat{z}) - w_t c_E E_t \right\}.$$

The first-order condition of the entrant's problem delivers what will be referred to as the free-entry condition, which is here written in complementary-slackness form:

$$\left(\int_0^\infty V_t(\hat{z}) \mathrm{d}F_t^E(\hat{z}) - w_t c_E \right) E_t = 0.$$

The derivations of the optimality conditions are presented in Appendix A.1.

3.3 Equilibrium Allocation

Now that all decision problems have been described, we can define the concept of an equilibrium allocation.

Definition 1. *Given initial conditions $\{\underline{z}_0, K_0, m_0(\hat{z})\}$, an equilibrium allocation consists of time paths for quantities, prices, and policy functions such that the following conditions hold:*

1. $\{C_t, H_t\}_{t\geq 0}$ *solve the household's problem.*
2. $\{\hat{y}_t(\hat{z})\}_{t\geq 0}$ *solve the final sector's problem.*
3. $\{p_t(\hat{z}), k_t(\hat{z}), l_t(\hat{z})\}_{t\geq 0}$ *solve the firm's static problem.*
4. $\{\gamma_t(\hat{z}), \underline{z}_t\}_{t\geq 0}$ *solve the firm's dynamic problem.*
5. $\{E_t\}_{t\geq 0}$ *solves the entrant's problem.*
6. $\{Y_t\}_{t\geq 0}$ *satisfies the Kimball (1995) aggregator.*
7. $\{p_t(\hat{z})\}_{t\geq 0}$ *clear the variety markets.*
8. $\{w_t\}_{t\geq 0}$ *clears the labor market.*
9. $\{r_t\}_{t\geq 0}$ *clears the asset market.*
10. *The capital stock evolves according to equation (6).*
11. *The density of firms evolves according to equation (8).*

Aggregation

Despite its complex structure, our theory admits tractable aggregation of the equilibrium allocation. In particular, aggregate output can be expressed as:

$$Y_t = Z_t K_t^{\alpha} L_t^{1-\alpha} \quad \text{where} \quad Z_t \equiv \left(\int_0^{\infty} \hat{y}_t(\hat{z}) \exp(-\hat{z} - \underline{z}_t) \mathrm{d}M_t(\hat{z}) \right)^{-1}.$$

Here, Z_t denotes the economy's TFP, which is a quantity-weighted harmonic aggregate of firm-level productivity. The aggregate demand for physical capital and production labor is given by:

$$K_t = \frac{\alpha Y_t}{(r_t + \delta)\mathcal{M}_t} \quad \text{and} \quad L_t = \frac{(1-\alpha)Y_t}{w_t \mathcal{M}_t}$$

where $\mathcal{M}_t$ denotes the cost-weighted average of firm-level markups:

$$\mathcal{M}_t \equiv \frac{\int_0^{\infty} \mu_t(\hat{z}) \hat{y}_t(\hat{z}) \exp(-\hat{z}) \mathrm{d}F_t(\hat{z})}{\int_0^{\infty} \hat{y}_t(\hat{z}) \exp(-\hat{z}) \mathrm{d}F_t(\hat{z})}. \tag{10}$$

Therefore, we recover the result from Edmond et al. (2023) that the "aggregate" markup reduces the quantity of variable inputs used in production.

Balanced Growth Path

With these aggregation results, we now define the concept of a balanced growth path (BGP) equilibrium allocation. The following propositions characterize the growth rate of TFP and the asymptotic behavior of the relative productivity distribution on this BGP.

Definition 2. *A BGP equilibrium allocation is an equilibrium allocation as defined in Definition 1 such that all quantities, prices, and policy functions are either stationary or grow at a constant rate, and the distribution of relative productivity is stationary.*

Since the economy is growing over time, the distribution of firm-level productivity behaves as a "traveling wave". Hence, for this distribution to be stationary, it must be normalized by a variable that travels at the same speed on a BGP. Here, we choose this variable to be the endogenous exit threshold such that the scale of the productivity distribution (and of the economy more generally) is determined by the initial condition for that threshold.[12] On a BGP, the growth rate of TFP is characterized by the following proposition.

[12] In all theories of exponential economic growth, the scale of the economy is determined by the initial

Proposition 3. *Letting the price elasticity of demand $\vartheta(\hat{z})$ as well as the firm's productivity pass-through $\varrho(\hat{z})$ be defined as in Section 2, the stationary growth rate of TFP can be decomposed into the contribution of (1) incumbent firms' productivity drift, (2) incumbent firms' productivity volatility, (3) endogenous exit and (4) exogenous exit:*

$$
\begin{aligned}
g = & \frac{\int_0^\infty [\vartheta(\hat{z})\varrho(\hat{z}) - 1]\hat{y}(\hat{z}) \exp(-\hat{z})\gamma(\hat{z}) dF(\hat{z})}{\int_0^\infty [\vartheta(\hat{z})\varrho(\hat{z}) - 1]\hat{y}(\hat{z}) \exp(-\hat{z}) dF(\hat{z})} && \textit{Incumbents' drift} \\
& - \frac{(\sigma^2/2) \int_0^\infty [\vartheta(\hat{z})\varrho(\hat{z}) - 1]\hat{y}(\hat{z}) \exp(-\hat{z}) dF'(\hat{z})}{\int_0^\infty [\vartheta(\hat{z})\varrho(\hat{z}) - 1]\hat{y}(\hat{z}) \exp(-\hat{z}) dF(\hat{z})} && \textit{Incumbents' volatility} \\
& + \frac{(\sigma^2/2) F''(0) [\int_0^\infty \hat{y}(\hat{z}) \exp(-\hat{z}) dF^E(\hat{z}) - \hat{y}(0)]}{\int_0^\infty [\vartheta(\hat{z})\varrho(\hat{z}) - 1]\hat{y}(\hat{z}) \exp(-\hat{z}) dF(\hat{z})} && \textit{Endogenous exit} \\
& - \frac{\chi [\int_0^\infty \hat{y}(\hat{z}) \exp(-\hat{z}) dF(\hat{z}) - \int_0^\infty \hat{y}(\hat{z}) \exp(-\hat{z}) dF^E(\hat{z})]}{\int_0^\infty [\vartheta(\hat{z})\varrho(\hat{z}) - 1]\hat{y}(\hat{z}) \exp(-\hat{z}) dF(\hat{z})} && \textit{Exogenous exit}.
\end{aligned}
$$

Here, the growth rate of TFP is also denoted by g since it must be equated to the growth rate of the endogenous exit threshold on a BGP.

Despite its complexity, the growth rate derived in Proposition 3 is similar to those obtained in other prominent endogenous growth models.[13] To see this, suppose for simplicity that the price elasticity of demand is constant $\vartheta(\hat{z}) = \theta > 1$, the productivity pass-through is constant and complete $\varrho(\hat{z}) = 1$ and the productivity drifts are constant $\gamma(\hat{z}) = \gamma > 0$.[14] Suppose further that entrants draw their productivity as to capture a fraction $s_E < 1$ of the average market share of incumbent firms while endogenously exiting firms' market share is equal to a fraction $s_X < s_E$ of that average. Under those assumptions, the above formula boils down to:

$$g = \gamma + \frac{(\theta - 1)\sigma^2}{2} + \frac{\sigma^2 F''(0)(s_E - s_X)}{2(\theta - 1)} - \frac{\chi(1 - s_E)}{\theta - 1}.$$

The first term reflects the positive contribution of incumbent firms' productivity drift to economic growth. The second term reflects how the volatility of incumbent firms' productivity contributes positively to growth. Indeed, since varieties are substitutes ($\theta > 1$), independent productivity shocks allow the final sector to reallocate expenditures

condition of the underlying linear ordinary differential equation (e.g., the initial capital stock in the "AK" model, the initial level of technology in the neoclassical growth model, or the initial population level in the Jones (1995) model).

[13] See Lashkari (2023) for a discussion of the growth rate that arises in different theories of endogenous technological change.

[14] The assumptions of a constant price elasticity of demand and a complete pass-through are obtained with a CES demand function.

from varieties that receive bad productivity shocks to those that receive good ones. The third and fourth terms reflect the contribution of entry and exit, by which entrants replace two types of firms: (1) the least productive firms who are swept below the endogenous exit threshold at rate $(\sigma^2/2)F''(0)$ and (2) randomly selected firms who exit exogenously at rate χ. Since the measure of varieties is constant on a BGP, entry does not contribute positively to growth through a "love for variety".[15]

The general formula provided in Proposition 3 further accounts for heterogeneity in demand elasticities, pass-throughs, and productivity drifts. This heterogeneity matters in that productivity improvements can be translated into *further* economic growth if (1) they are rolled out over more units produced, (2) they are passed on through lower prices, and (3) these lower prices are rewarded with additional demand. To see this, notice that the term reflecting the growth contribution of incumbents' investments in R&D is a weighted average of their productivity drifts:

$$\int_0^\infty \omega(\hat{z})\gamma(\hat{z})\mathrm{d}\hat{z} \quad \text{where} \quad \omega(\hat{z}) \equiv \frac{[\vartheta(\hat{z})\varrho(\hat{z})-1]\hat{y}(\hat{z})\exp(-\hat{z})F'(\hat{z})}{\int_0^\infty [\vartheta(\hat{z})\varrho(\hat{z})-1]\hat{y}(\hat{z})\exp(-\hat{z})F'(\hat{z})\mathrm{d}\hat{z}}.$$

In particular, the term $\hat{y}(\hat{z})\exp(-\hat{z})$ in the "weights" $\omega(\hat{z})$ is proportional to a firm's total expenditures on inputs, whereas the term $\vartheta(\hat{z})\varrho(\hat{z})-1$ denotes the elasticity of those expenditures with respect to the firm's productivity. Hence, these weights reflect the extent to which a firm *differentially* expands in scale following an improvement in its process efficiency.

Intuitively, this implies that the rate of TFP growth will be higher if the firms that achieve the largest drift in productivity are not only larger and more numerous but also the firms towards which demand is mostly reallocated as a result. The magnitude of this correlation depends on the stationary distribution of relative productivity, whose asymptotic behavior is characterized by the following proposition.

Proposition 4. *On a BGP, if* $\chi > 0$, $\lim_{\hat{z}\to\infty}\gamma(\hat{z}) = \overline{\gamma} < \infty$ *and the transformation function* T *satisfies the assumptions described in Section 3, the distribution of relative productivity asymptotes to an exponential distribution as* $\hat{z}\to\infty$*:*

$$\lim_{\hat{z}\to\infty} F(\hat{z}) = 1-\exp(-\lambda\hat{z}) \quad \textit{where} \quad \lambda \equiv \frac{g-\overline{\gamma}+\sqrt{(g-\overline{\gamma})^2+2\chi\sigma^2}}{\sigma^2}$$

and where g *is the stationary growth rate of TFP. If* $g > \overline{\gamma}+\sigma^2/2-\chi$*, the stationary distribution of* $\exp(\hat{z})$ *is Pareto with shape parameter* $\lambda > 1$ *and has a finite mean.*

The rate parameter λ of the exponential tail is inversely related to the dispersion in

[15] This is a result of our assumption of a constant population.

relative productivity. Hence, more churning from a higher growth rate g or a higher exit rate χ implies a thinner right tail. In contrast, a higher instantaneous productivity volatility σ implies a fatter right tail. Since TFP growth is endogenous, Proposition 4 illustrates how firm-level heterogeneity determines aggregate economic growth, which determines the extent of this heterogeneity.

3.4 Characterization

For our theory to deliver quantifiable predictions, we impose additional parametric functional form assumptions on preferences and technologies. This subsection describes those choices, which are both standard in the literature and consistent with relevant empirical regularities.

In terms of preferences, we assume that the household has standard MaCurdy (1981) flow disutility from hours worked:

$$v(H) = \beta \times \frac{H^{1+\eta}}{1+\eta}$$

where $\eta > 0$ is the inverse of the Frisch elasticity of labor supply and $\beta > 0$ is the utility weight on hours worked.

The final sector's Kimball (1995) production technology is defined according to the functional form introduced by Klenow and Willis (2016):[16]

$$\Upsilon(\hat{y}) = 1 + (\theta - 1)\exp(1/\epsilon)\epsilon^{\theta/\epsilon - 1}\left[\Gamma\left(\frac{\theta}{\epsilon}, \frac{1}{\epsilon}\right) - \Gamma\left(\frac{\theta}{\epsilon}, \frac{\hat{y}^{\epsilon/\theta}}{\epsilon}\right)\right]$$

where $\theta > 1$ and $\epsilon > 0$. Note that as $\epsilon \to 0$, this functional form converges to the Dixit and Stiglitz (1977) aggregator with constant elasticity of substitution across varieties. This functional form is chosen because it is flexible enough to capture two important empirical regularities, sometimes referred to as Marshall (1890)'s second and third laws of demand (Matsuyama and Ushchev, 2022). These properties of demand state that the price elasticity of demand increases in the price charged, whereas its rate of change (the "super-elasticity" of demand) decreases therein. This implies that, in equilibrium, more productive firms are larger and command both higher markups and lower pass-throughs, which is empirically documented in Amiti et al. (2014) and Amiti et al. (2019). Specifically, these markups and pass-throughs take the following form as functions of

[16]Here, $\Gamma(s, x) = \int_x^\infty t^{s-1}e^{-t}\mathrm{d}t$ denotes the upper incomplete gamma function.

relative demand:

$$\mu(\hat{y}) = \frac{\theta}{\theta - \hat{y}^{\epsilon/\theta}} \quad \text{and} \quad \varrho(\hat{y}) = \frac{\theta - \hat{y}^{\epsilon/\theta}}{\theta + \epsilon - \hat{y}^{\epsilon/\theta}}.$$

As in Acemoglu et al. (2018) and Akcigit and Kerr (2018), the firm's innovation technology is characterized by an isoelastic cost function:

$$i(\gamma, \hat{z}) = \frac{\exp[c_I + (1+\zeta)\hat{z}]\gamma^{1+\zeta}}{1+\zeta}$$

where $c_I > 0$ measures the scale of that cost function and $\zeta > 0$ disciplines its elasticity. In particular, this functional form implies that all firms must allocate the same quantity of labor to achieve a given *absolute* drift of relative productivity. Therefore, achieving a *proportional* drift becomes more costly as a firm becomes more productive.

Finally, we follow Benhabib et al. (2021) and choose the transformation function $T(x) = x^{\xi}$ where $\xi > 1$ such that:

$$F_t^E(\hat{z}) = 1 - [1 - F_t(\hat{z})]^{\xi}.$$

This functional form is parsimonious and can satisfy the required assumptions to achieve a stationary distribution of relative productivity on a BGP. Specifically, for $\xi > 1$, entrants start producing with lower relative productivity than incumbents on average. Table 2 summarizes the functional form assumptions.

Table 2: Functional forms

Function	Source
$v(H) = \beta \times \frac{H^{1+\eta}}{1+\eta}$	MaCurdy (1981)
$\Upsilon(\hat{y}) = 1 + (\theta - 1)\exp(1/\epsilon)\epsilon^{\theta/\epsilon - 1}\left[\Gamma\left(\frac{\theta}{\epsilon}, \frac{1}{\epsilon}\right) - \Gamma\left(\frac{\theta}{\epsilon}, \frac{\hat{y}^{\epsilon/\theta}}{\epsilon}\right)\right]$	Klenow and Willis (2016)
$i(\gamma, \hat{z}) = \exp[c_I + (1+\zeta)\hat{z}]\gamma^{1+\zeta}/(1+\zeta)$	Acemoglu et al. (2018)
$F_t^E(\hat{z}) = 1 - [1 - F_t(\hat{z})]^{\xi}$	Benhabib et al. (2021)

4 Quantification

In this section, we present the estimation of our theory's structural parameters, which is performed via a GMM strategy, targeting aggregate and firm-level moments from France. We first describe the data from which these moments are calculated, after which we discuss the identification of each parameter.

4.1 Data

Our primary source of data is the *Fichier Approché des Résultats d'Esane* (FARE), which is an annual panel dataset with the balance sheet and income statements of all firms in France that are subject to the standard corporate tax (excluding the financial and farming sectors). Our sample consists of 5.4 million (firm-year) observations between 2009 and 2019, with around 830 thousand unique firms overall and 460 thousand firms each year.[17] The variables of interest are the firm's main industry of operation, value-added, wage bill, and its stock of capital.

From this data, we measure the markup of firm j from industry i in year t as:[18]

$$\mu_{jit} = \frac{p_{jit}y_{jit}}{[(r_t+\delta)k_{jit}]^{\alpha_{it}}(w_t l_{jit})^{1-\alpha_{it}}}$$

where $p_{jit}y_{jit}$ is its value-added, k_{jit} is its stock of capital (in current value), $w_t l_{jt}$ is its total expenditures on labor and α_{it} is the output elasticity of physical capital, which we assume is common to all firms in the same 2-digit NACE industry.[19] Given the Cobb-Douglas production function, we calculate this elasticity as the cost-weighted average of each firm's capital cost share in industry i and year t:[20]

$$\alpha_{it} = \sum_{j\in i} \frac{\omega_{jit}(r_t+\delta)k_{jit}}{[(r_t+\delta)k_{jit}+w_t l_{jit}]} \quad \text{where} \quad \omega_{jit} \equiv \frac{(r_t+\delta)k_{jit}+w_t l_{jit}}{\sum_{j\in i}[(r_t+\delta)k_{jit}+w_t l_{jit}]}.$$

Finally, we define a firm's market share as the share of value-added it captures in a given year within its 5-digit NACE industry. The empirical relationship between the logarithm of markups and that of market shares is plotted in Figure 2(a) as a binned scatter plot.

[17] Refer to Appendix C for details on criteria we set for inclusion of an observation in our sample.

[18] This uses Euler's theorem and abstracts from the proportionality term $\alpha^{\alpha}(1-\alpha)^{1-\alpha}$. We verified that including that term affects the level of the implied markups but not their correlation with market shares.

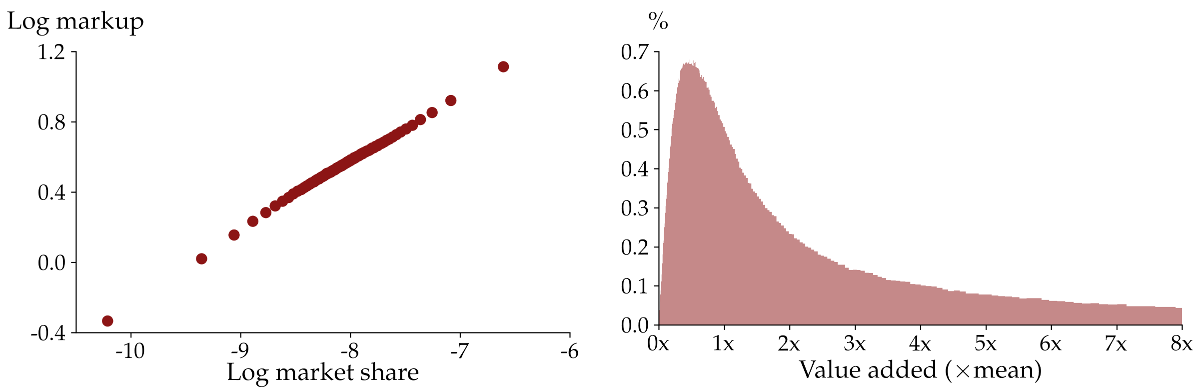
[19] See Appendix C for further details on the construction of each variable used in the measurement of firm-level markups.

[20] The chosen value for $r_t+\delta$ is consistent with the equilibrium interest rate of our model and the rate of physical capital depreciation we assumed.

Not only do we recover evidence for Marshall (1890)'s second law of demand, which has been documented in many other settings, but the relationship between markups and market shares is remarkably log-linear. In Figure, 2(b), we plot the empirical histogram of firm-level value-added (in proportion to the industry mean) from which the extent of size dispersion in the data is evident.

Figure 2: Markups and Firm Size

(a) Markups and Market Shares (b) Value-Added Histogram



Note: The underlying data for these two plots is from the FARE dataset between 2009 and 2019. Markups and market shares are calculated as described in Section 4.1. In Panel 2(a), year, industry, industry-year, firm, and age fixed effects are removed from each variable before plotting the 100 bin-scatter points. In Panel 2(b), the 2-digit industry mean is removed from firm-level value-added such that the histogram is expressed in proportion to the industry mean.

4.2 Structural Estimation

Our theory features 14 parameters to be determined, collected in the set Ω:

$$\Omega = \{\rho, \beta, \eta, \theta, \epsilon, \alpha, \delta, c_O, \sigma, c_I, \zeta, c_E, \xi, \chi\}.$$

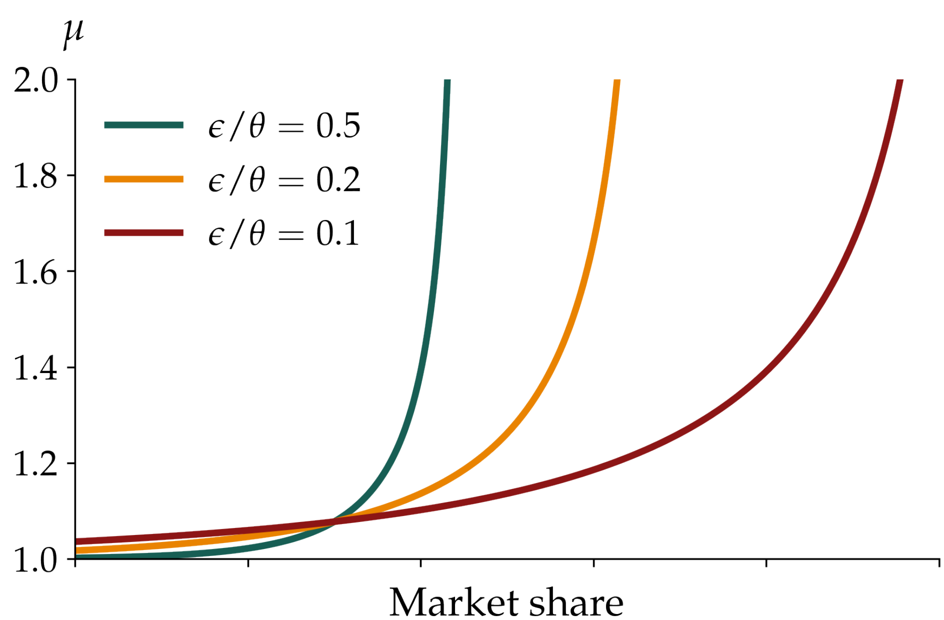
We assign conventional values to $\{\rho, \eta, \alpha, \delta\}$, normalize c_O, and estimate or normalize the remaining parameters. We set the representative household's rate of time preference ρ to 0.04, the parameter η to one to deliver a unit Frisch elasticity of labor supply, the capital share to $1/3$, and the depreciation rate of physical capital to 0.06.[21] The overhead

[21]The cross-year average of the industry capital shares α_{it} we measure in the FARE data is equal to 0.22. To deal with this discrepancy, we consider a robustness check in Appendix C.3 where we inflate capital shares by a constant such that they aggregate to our model's capital share of $1/3$.

cost parameter c_O is normalized without loss of generality.[22]

We then independently identify the following three parameters $\{\beta, \chi, \epsilon/\theta\}$ with three moments. First, the utility weight on labor supply β is chosen to match average hours worked per year per person in France between 1995 and 2019.[23] Second, the exogenous exit rate χ is set equal to the average exit rate of 1.34% for firms with more than ten employees between 2009 and 2019 in France.[24] Third, the ratio ϵ/θ of the Klenow and Willis (2016) elasticity and super-elasticity parameters is estimated from the relationship between firm-level markups and market shares in our data. Figure 3 illustrates how that relationship depends on the value ascribed to the ratio ϵ/θ.

Figure 3: Markups and Market Shares



Note: This figure plots the relationship between firm-level markups and market shares, where the horizontal axis is defined on a logarithmic scale. Units are omitted for that axis since they provide no information.

Following Edmond et al. (2023), we show in Appendix C that in a generalization of our theory (with time-varying industry-level demand shifters and time-invariant

[22]Given an initial condition for the exit threshold, we choose its value such that aggregate output is normalized to unity in our model's initial stationary equilibrium.

[23]For a time endowment of 16 hours per day and 365 days per year, the average of 1540.3 hours worked per person per year (calculated from EU-KLEMS) implies a value of $1540.3/(16 \times 365) \approx 0.26$ for H_t.

[24]This moment is calculated from Eurostat's Business Demography Statistics, which goes back to 2008.

firm-level demand shifters), this relationship is nonlinear and given by:

$$\mu_{jit}^{-1} + \ln(1 - \mu_{jit}^{-1}) = b + b_t + b_i + b_{it} + b_j + (\epsilon/\theta)\ln(s_{jit}) \quad \text{where} \quad s_{jit} \equiv \frac{p_{jit}y_{jit}}{P_{it}Y_{it}}.$$

Here, s_{jit} denotes the market share of firm j operating within industry i in year t, b is a constant, b_t is a time fixed effect, b_i is an industry fixed effect, b_{it} is a time-industry fixed effect and b_j is a firm fixed effect. We estimate this relationship in the FARE data using our firm-level markup and market share measurements described in Section 4.1. The results of this estimation exercise are reported in Table 3 for the entire sample and for the manufacturing sector only. In our preferred specification with all levels of fixed effects, we obtain an estimate of $\epsilon/\theta = 0.243$ with a standard error of 0.001 clustered at the firm level. This is in line with the estimates reported in Edmond et al. (2023) and Amiti et al. (2019) who find values of 0.16 and 0.32, respectively. The latter identify this ratio by matching the variability in markups and resulting pass-throughs among Belgian manufacturing firms.

Table 3: Markups and Market Shares

	Dependent variable: $\mu_{jit}^{-1} + \ln(1 - \mu_{jit}^{-1})$					
	Full sample			Manufacturing		
$\ln(s_{jit})$	0.047	0.234	0.243	0.049	0.321	0.331
	(0.000)	(0.001)	(0.001)	(0.001)	(0.004)	(0.004)
Firm fixed effects		Y	Y		Y	Y
Industry × year fixed effects	Y	Y	Y	Y	Y	Y
Industry fixed effects			Y			Y
Year fixed effects			Y			Y
Age group fixed effects	Y		Y	Y		Y
R^2	0.090	0.505	0.507	0.056	0.489	0.490
Observations	4.9M	4.9M	4.9M	0.5M	0.5M	0.5M

Note: Firm-level markups and market shares are constructed from the FARE dataset as described in Section 4.1. This table presents different regression specifications with firm fixed effects, 5-digit NACE industry fixed effects, and age group fixed effects (for 20 evenly-spaced age groups). Standard errors (in parentheses) are clustered at the firm level. The total number of observations is below the total sample size of 5.4M because negative markups were estimated for some firms.

Finally, the remaining six parameters $\{\sigma, c_I, \zeta, c_E, \xi, \theta\}$ are jointly identified by the following six moments via a GMM estimation strategy:

1. An aggregate (cost-weighted average) markup of 1.3, averaging the range of 1.1 to 1.5 estimated by De Ridder et al. (2023) using the FARE (manufacturing) data.
2. The average annual growth rate of 1.16% of real GDP per hour worked in France between 1995 and 2019 calculated from EU-KLEMS's national growth accounts.
3. The average annual growth rate of (deflated) firm-level value added of 1.24% calculated from the FARE data.
4. The average annual exit rate of 5.61% among all French firms between 2009 and 2019 calculated from Eurostat's Business Demography Statistics.
5. The average size (value added) of entrants relative to incumbents of 31% calculated from the FARE data.
6. The within-industry standard deviation of log value added of 1.54 calculated from the FARE data.

The objective we minimize is the squared percent deviation between these moments and their counterpart in our theory's stationary equilibrium allocation. We pose this estimation exercise as a mathematical program with equilibrium constraints (MPEC) (Su and Judd, 2012; Dubé, Fox and Su, 2012). Doing so allows us to perform the parameter search without repeatedly solving the model's equilibrium conditions at each guess of parameters.[25] Table 4 reports the resulting values of our structural parameters.

Table 5 compares the empirical and theoretical moments listed above, evaluated at the estimated parameter values. All of our targeted moments are matched with relatively high accuracy. Our model also replicates several untargeted moments in the FARE data, such as the Gini coefficient of value-added, the share of total value-added captured by the largest firms, the relative size of entrants by employment, and the average and median age of a firm. It is also consistent with the empirically evidenced decreasing relationships between (1) the rate of exit and firm age (Caves, 1998), and (2) firm-level sales volatility and firm size (Yeh, 2021).

Figure 4 plots several static firm-level outcomes against relative productivity, where the stationary distribution of the latter is plotted in transparency to emphasize the "relevant" domain of each of those functions. Panel 4(a) plots the downward-sloping elasticity of the demand schedule faced by the firm, illustrating Marshall (1890)'s second law of demand. Panel 4(b) plots the price chosen by the firm (relative to the choke price),

[25] Appendix B.2 describes this exercise in detail and provides a formal discussion on identification.

Table 4: Structural Parameters

Parameter	Symbol	Value
Household preferences:		
Rate of time preference	ρ	0.04
Labor supply utility weight	β	10.3
Frisch elasticity of labor supply reciprocal	η	1
Final sector technology:		
Klenow and Willis (2016) elasticity parameter	θ	14.9
Klenow and Willis (2016) super-elasticity parameter	ϵ	3.62
Firm production technology:		
Output elasticity of physical capital	α	0.33
Depreciation rate of physical capital	δ	0.06
Overhead cost parameter	c_O	0.03
Firm innovation technology:		
Brownian motion standard deviation	σ	0.03
Innovation cost scale parameter	c_I	9.23
Innovation cost elasticity parameter	ζ	0.99
Entry and exit:		
Entry cost parameter	c_E	6.61
Entry distribution parameter	ξ	1.71
Exogenous exit rate	χ	1.34%

Note: This table presents the assigned/estimated structural parameters of our theory.

while Panel 4(c) plots the implied markup over marginal cost, which is increasing in size. Finally, Panel 4(d) plots the market share captured by more or less productive firms.

In Figure 5, we present the probability distribution function of firm-level markups, including various percentiles. Although our target is a cost-weighted average markup of 1.3, the *unweighted* median markup stands notably lower at 1.1. The implied dispersion in markups also appears more modest than its empirical counterpart. While we infer an interquartile range of 0.05 for the logarithm of markups, De Ridder et al. (2023) estimate values of 0.48 and 0.2, respectively, using information on quantities or revenues from the FARE data (for the manufacturing sector). This illustrates that our model strictly captures the markup dispersion systematically related to firm size. In that sense, our

Table 5: Moments

Moment	Source	Model	Data
Targeted:			
Aggregate markup	FARE	1.30	1.30
GDP per hour worked growth rate	EU-KLEMS	1.23%	1.16%
Incumbent value added growth rate	FARE	1.18%	1.24%
Exit rate of all firms	Eurostat	5.32%	5.61%
Relative size of entrants by value added	FARE	0.30	0.31
Standard deviation of log value added	FARE	1.51	1.54
Untargeted:			
Gini coefficient of value added	FARE	0.78	0.73
Top 5% value added share	FARE	52%	51%
Top 10% value added share	FARE	70%	64%
Relative size of entrants by employment	FARE	0.33	0.32
Average firm age	FARE	27.8	19.3
Median firm age	FARE	15.1	15.7

Note: This table presents moments (targeted or not in our GMM estimation exercise) and their resulting value in our model. Moments measured in the FARE data are first calculated within 2-digit NACE industries and then aggregated with each industry's share of total value added.

estimation strategy leans somewhat conservatively on the extent of such dispersion.

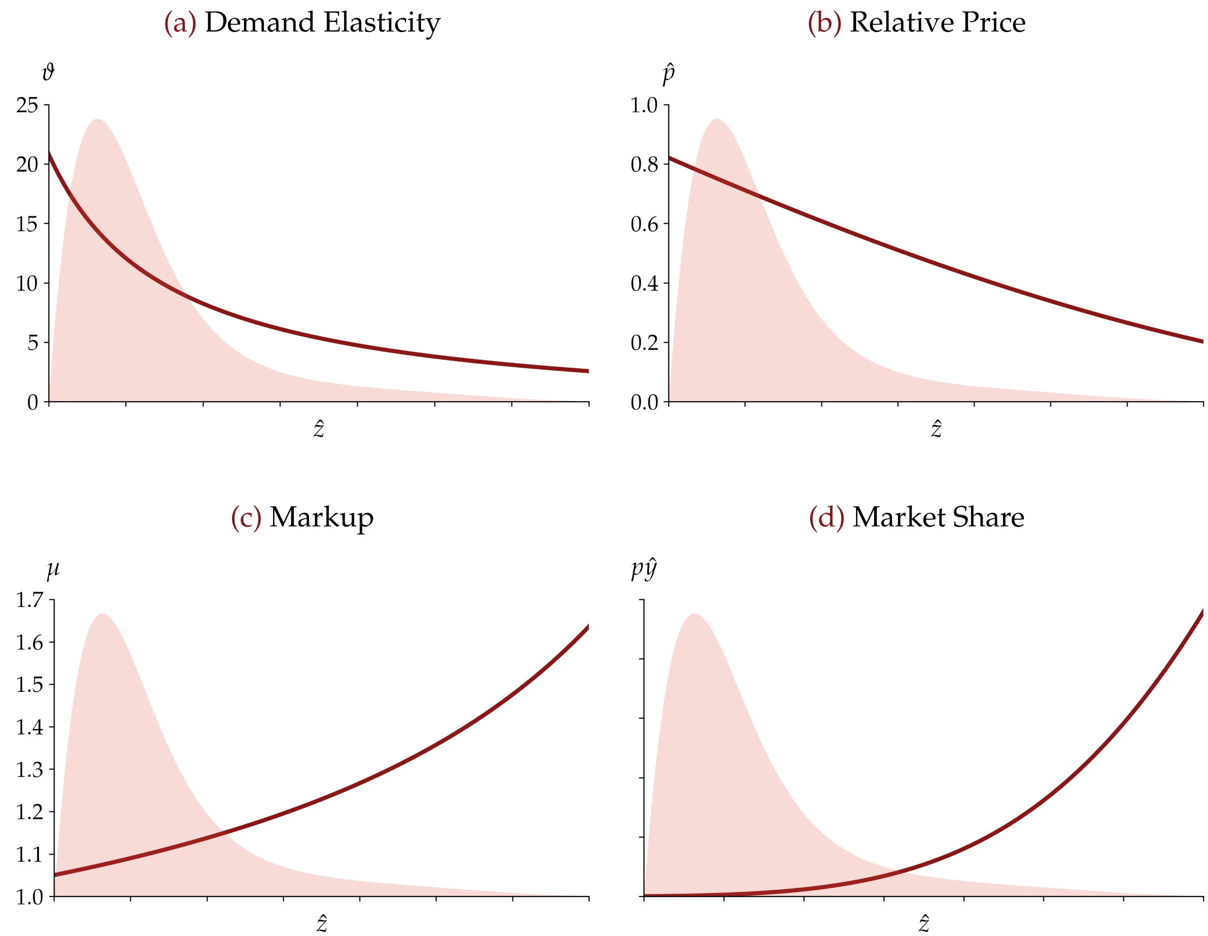
Parameter Identification

While the six parameters $\{\sigma, c_I, \zeta, c_E, \xi, \theta\}$ are jointly identified by the six moment conditions above, we see it as informative to intuitively discuss the reasoning behind our selection of these six moments. Appendix B.2 provides a more formal analysis of the identification of these parameters.

We include the aggregate markup as a target in our estimation exercise since it directly affects how much firms limit the scale at which they operate in partial equilibrium.[26] While none of the aforementioned parameters explicitly appear in the expression for the aggregate markup, $\{\sigma, c_I, \zeta, \xi, \theta\}$ directly affect the shape of the productivity distribution over which firm-level markups are aggregated and are, in that sense, identified by that

[26] Edmond et al. (2023) find that the "static" welfare losses from markups exhibit significant convexity in the aggregate markup target.

Figure 4: Static Firm-Level Outcomes

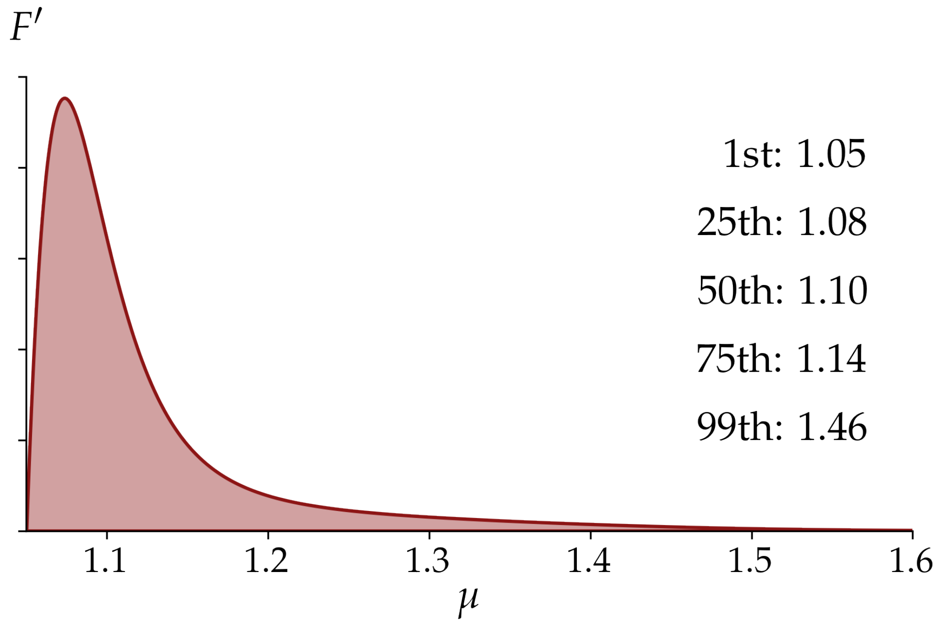


Note: Here, all variables are defined according to the estimated values for θ and ϵ, which are reported in Table 4. The distribution plotted in transparency is the stationary distribution of relative productivity. Units are omitted on axes where they are not informative.

aggregate. This is evident from the KFE in equation (8), in which these parameters appear either through the firm-level productivity process, the Kimball (1995) aggregation condition, or the dynamics of entry and exit.

We aim for a value of 1.3 for the aggregate markup, which aligns with the range of values estimated by De Ridder et al. (2023) using the FARE data for the manufacturing sector. More precisely, they calculate the sales-weighted harmonic average markup to be approximately equal to 1.1 or 1.5 with revenue and quantity data, respectively. Our choice of targeting the average of these two estimates is motivated by recent evidence that markups are challenging to identify with either revenue or production data (Bond, Hashemi, Kaplan and Zoch, 2021; De Ridder et al., 2023; Flynn, Traina and Gandhi, 2019;

Figure 5: Distribution of Firm-Level Markups



Note: This figure plots the unweighted distribution of firm-level markups with its 1st, 25th, 50th, 75th and 99th percentiles. The interquartile range of the logarithm of markups is equal to 0.05, which is more modest than the values of 0.48 and 0.2 estimated by De Ridder et al. (2023) on quantity and revenue data, respectively.

Raval, 2023).[27] In Section 5.5, we consider how our results change as we aim for lower or higher values.

We target the *aggregate* and *firm-level* value-added growth rate moments with two objectives in mind. First, it appears important that our parameterization be consistent with the growth rate of the French economy. Second, comparing this growth rate to that of continuing firms in the FARE data is indirectly informative about their contribution to the latter.[28] As such, the aggregate and firm-level growth rate moments are intended to identify the parameters of the firm's innovation cost function, c_I and ζ, thereby delineating the pace of economic growth and the contribution of incumbent firms to this progress.

As accentuated in the technology diffusion literature, yet another contribution to long-run productivity growth comes from the selective survival of successful firms and the adoption of existing technologies by new entrants. The scope of this contribution

[27]For different empirical counterparts to the markup in our theory, we have estimated cost-weighted average markups that ranged from 5% to almost 80%.

[28]Despite the absence of a clear, direct empirical counterpart, there have been attempts to infer this contribution indirectly, as evidenced by Luttmer (2007) and Garcia-Macia et al. (2019).

hinges on two factors: (1) the frequency at which underperforming firms are supplanted by more efficient newcomers and (2) the productivity differential between these two groups of firms. To discipline these two factors, we target the rate at which firms exit–which must be equated to the entry rate on a balanced growth path since our model features a constant population–and the initial value-added of new entrants relative to incumbents, which positively correlates with productivity in our model.[29] Hence, these two moments partly identify the parameters ξ and σ, where the former regulates the transformation of the incumbent distribution from which entrants draw their relative productivity, and the latter directly influences the rate at which unsuccessful firms are swept below the endogenous exit threshold.

Finally, we include the standard deviation of the logarithm of value added as a target in our estimation exercise, enabling us to regulate the extent of markup dispersion. Our model posits that the sole source of dispersion in markups across firms comes from their endogenous differences in process efficiency, which cause them to charge varying prices at which demand is more or less elastic. That is, the extent of dispersion in markups is intrinsically tied to the degree of heterogeneity in firm size. As contended in Edmond et al. (2023), we regard this approach as more conservative than the alternative of directly targeting the empirically observed dispersion in markups, which could instead reflect dispersion in other types of distortions unrelated to markups.

5 Counterfactual Analysis

To quantify how heterogeneous markups differentially distort firms' incentives to invest in R&D, we implement size-dependent subsidies to production, inducing each firm to price at marginal cost. We analyze how firm-level decisions, their aggregation, and economic growth endogenously respond to this intervention.

5.1 Policy Intervention

Why do we consider such a particular policy intervention rather than simply comparing the economy's laissez-faire and optimal resource allocations? The reason is that, in order to achieve a stationary distribution of relative productivity, our economic environment features technological spillovers across firms, which introduces externalities that are

[29] We target the exit rate rather than the entry rate, since the latter may reflect long-run growth in the number of firms, which is not a feature of our model.

not intrinsically tied to markups.[30] Therefore, a comparison of the laissez-faire and optimal allocations would not *isolate* the consequences of markups for economic growth. Instead, the implementation of subsidies devised to compel firms to price at marginal cost directly addresses the product market distortions induced by markups, which is the focus of this paper. Following Edmond et al. (2023), we show in Appendix A.1.8 that the size-dependent subsidy scheme $T_t(\hat{y})$ that achieves this is:

$$T_t(\hat{y}) = [\Upsilon(\hat{y}) - \Upsilon'(\hat{y})\hat{y}]Y_t D_t \tag{11}$$

which we assume is financed by lump-sum taxes levied on the household.[31] Under this policy, firms optimally price at marginal cost as the subsidy schedule outlined in equation (11) is such that they consider the final sector's output rather than their revenue as part of their objective. That is, these subsidies enable firms to capture the entire consumer surplus. However, since the representative household holds a diversified portfolio of all firms in the economy, this surplus is returned to it in the form of dividends.

As the intervention is rolled out, firms operate at a larger scale. This is depicted in Panel 6(a), which plots firms' production employment in both the pre- and post-policy stationary equilibria. In most of the figures that follow, the horizontal axis is represented by a firm's initial markup before the intervention, and its range covers more than 99.9% of the measure of firms. Notably, production employment increases for all firms but is also reallocated towards larger, more productive firms that initially commanded a higher markup. While firm-level production employment roughly doubles on average, it almost triples for the largest firms. The reallocation of innovation employment is even starker, as illustrated in Panel 6(b). On average, firm-level innovation employment more than triples. Yet, it shrinks by about 75% for the least productive firms, while their most productive counterparts see a more than 5-fold increase in their allocation.

Table 6 presents the aggregation of these firm-level outcomes. Aggregate demand for production, innovation, and entry labor increases considerably, matched by a 23% increase in labor supply.[32] However, the aggregate labor allocation to overhead contracts by 41.8% due to a proportional decrease in the measure of varieties. Although more resources are allocated to entry due to the greater scale and convexity of profits, the endogenous exit rate increases disproportionately from 4% to 15.5%, thus depleting the stock of varieties. As R&D resources are redirected towards more productive firms, the smallest firms fail to keep up with the competition, trail behind, and eventually exit

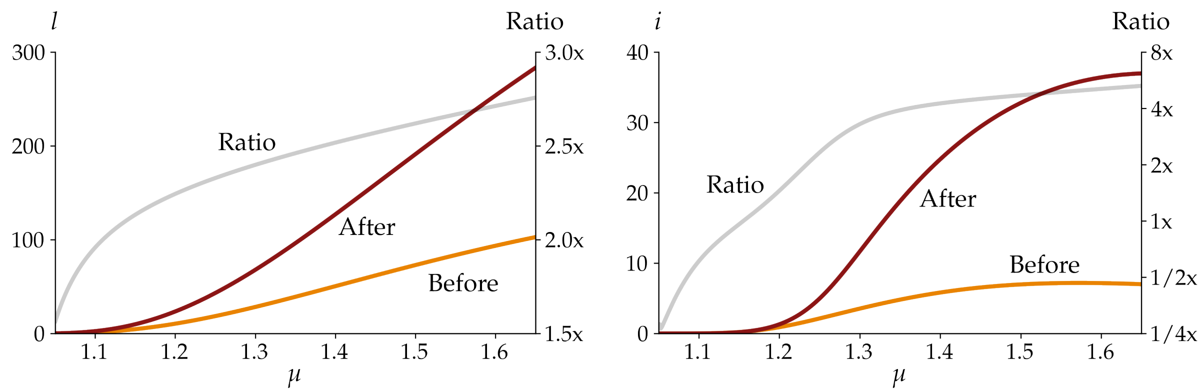
[30] These externalities are further discussed in Section 5.4 and appear in Appendix A.1.8, where we derive the optimal allocation of resources.

[31] This policy intervention is large in magnitude, equivalent to about 29% of aggregate output.

[32] As a reference, Edmond et al. (2023) find, under targeted aggregate markups of 1.25 and 1.35, that the same policy intervention achieves a 30.1% and 42.1% increase in labor supply, respectively.

Figure 6: Firm-Level Labor Allocations

(a) Production Employment (b) Innovation Employment



Note: Both functions are plotted over the same support of initial markups, covering more than 99.9% of the measure of firms both before and after the intervention. The red and orange lines (left axes) respectively plot labor allocations before and after the policy intervention. The gray line (right axis) plots their ratio (post- relative to pre-policy).

(endogenously) at a higher rate.

This dynamic is most clearly illustrated in Figure 7. In particular, Panel 7(a) plots the firm-level productivity drifts achieved pre- and post-policy. Consistent with Figure 6, the growth trajectory of the smaller, less efficient firms decelerates post-policy, and, as a result, they congregate near the exit threshold. This is depicted in Panel 7(b), where it is shown that the post-policy stationary distribution of relative productivity admits a higher density of small firms. However, it also features a higher density of fast-growing large firms, such that the distribution becomes slightly bimodal after the intervention. Market concentration rises only slightly after the intervention, as presented in the last two rows of Table 6, which report the share of total value added captured by the top 5% and 10% largest firms.

A closer look at the firm's dynamic first-order condition sheds light on the disparity in growth trajectories depicted in Panel 7(a):

$$\frac{V_t'(\hat{z})}{w_t} = \frac{\partial i(\gamma, \hat{z})}{\partial \gamma}.$$

This condition implies that a firm will achieve a larger productivity drift if the resulting change in its value is large relative to the prevailing wage rate. Despite the absence of an analytical solution for the former, one can gain insight into what determines the extent of the firm's marginal value through the following asymptotic proposition:

Table 6: Economic Aggregates

Aggregate	Before	After	Change
Labor allocations:			
Labor supply	0.264	0.325	+23.0%
Production labor	0.227	0.260	+14.5%
Innovation labor	0.020	0.036	+76.9%
Entry labor	0.015	0.028	+84.7%
Overhead labor	0.002	0.001	-41.8%
Firms, entry and exit:			
Measure of varieties	0.044	0.025	-41.8%
Entry rate	5.32%	16.87%	+11.6p.p.
Endogenous exit rate	3.97%	15.52%	+11.6p.p.
Market concentration:			
Top 5% value added share	51.6%	54.6%	+3.0p.p.
Top 10% value added share	69.8%	77.2%	+7.4p.p.

Note: This table presents the pre- and post-policy level of various economic aggregates as well as the corresponding percentage change.

Proposition 5. *On a balanced growth path, the firm's value function asymptotes to the present discounted value of asymptotic profits:*

$$\lim_{\hat{z}\to\infty} V_t(\hat{z}) = \overline{V}_t \quad \textit{where} \quad \overline{V}_t \equiv \frac{\overline{\pi} Y_t D_t - w_t c_O}{\rho + \chi}$$

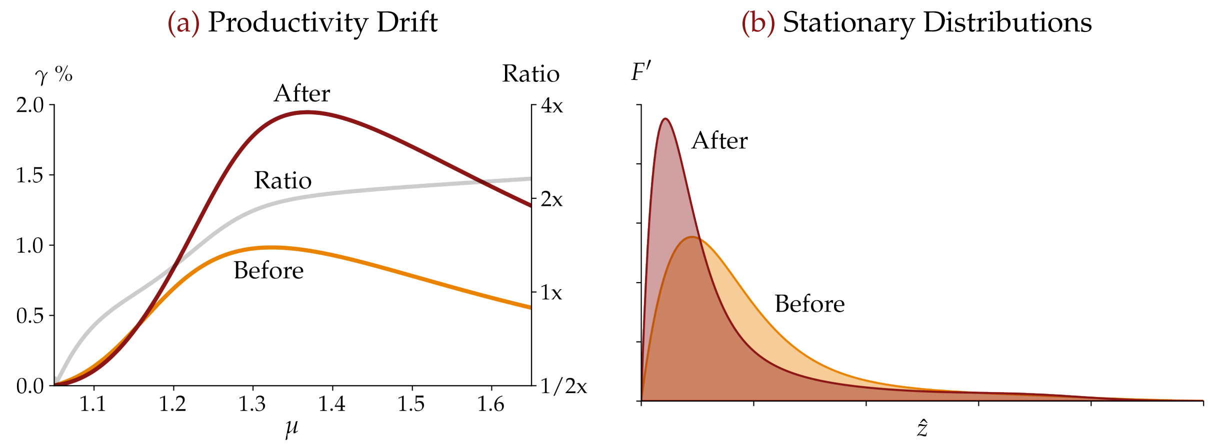
and where the constant $\overline{\pi}$ is given by:

$$\overline{\pi} = \begin{cases} (\theta - 1)\exp[(1-\theta)/\epsilon]\theta^{\theta/\epsilon - 1} & \textit{Pre-policy,} \\ 1 + (\theta-1)\exp(1/\epsilon)\epsilon^{\theta/\epsilon-1}\Gamma\left(\frac{\theta}{\epsilon}, \frac{1}{\epsilon}\right) & \textit{Post-policy.} \end{cases}$$

Using Proposition 5 together with the expression for aggregate labor demand, let us denote the firm's value function relative to its asymptote by $\hat{V}_t(\hat{z}) \equiv V_t(\hat{z})/\overline{V}_t$. This allows us to recast the firm's dynamic first-order condition as:

$$\frac{\overline{\pi}\mathcal{M}_t L_t D_t - (1-\alpha)c_O}{(1-\alpha)(\rho+\chi)} \times \hat{V}_t'(\hat{z}) = \frac{\partial i(\gamma, \hat{z})}{\partial \gamma}.$$

Figure 7: Productivity Drifts and Stationary Distributions



Note: For consistency, both functions are plotted over the same relative productivity support as in Figure 6. In Panel 7(a), the red and orange lines (left axes) respectively plot productivity drifts before and after the policy intervention. The gray line (right axis) plots their ratio (post- relative to pre-policy). It is worth noting that the distributions' support in Panel 7(b) are defined *relative* to the least productive firm in each allocation. The *level* of the exit threshold may change due to the policy intervention.

Given that the term $\overline{\pi}\mathcal{M}_t L_t D_t$ increases by a factor of almost five after the intervention, all firms would be set on a trajectory of accelerated growth were it not for changes in the curvature of the firm's value function. Consequently, the curvature in the subsidy scheme effectively reallocates innovation employment away from smaller firms towards their larger counterparts.

5.2 Long-Run Economic Growth

What are the implications of these findings for long-run economic growth? Our analysis reveals that the stationary growth rate of TFP increases by 1.2 percentage points, from 0.82% to 2.05%. Due to physical capital accumulation, this leads to a corresponding increase in the growth rate of aggregate output, from an initially targeted value of 1.23% to a counterfactual value of 3.08%.

Table 7 provides a breakdown of the pre- and post-policy growth rates, along with the corresponding change, according to Proposition 3. In particular, this decomposition presents the contributions of incumbent firms' productivity drift and volatility, as well as that of entry and exit. Of noteworthy importance is the role of incumbents' productivity drift, which constitutes the largest share of both growth levels (before and after the intervention): it respectively accounts for 65.5% and 50% of TFP growth in the pre- and post-policy stationary equilibria. Nonetheless, it contributes slightly less to the growth rate differential between the two, comprising 39.8% of it. The large remaining share of the growth rate change is attributed to the contribution from entry and exit.

Table 7: Growth Rate Decomposition

Contribution	Before	After	Change (p.p.)
Incumbent drift	0.54%	1.02%	+0.49%
Incumbent volatility	0.18%	0.19%	+0.01%
Entry and exit	0.10%	0.84%	+0.74%
Total	0.82%	2.05%	+1.23%

Note: This table presents the contributions to the level and change of TFP growth.

This latter contribution is an indirect consequence of the R&D reallocation prompted by the policy. As R&D expenditures are redirected from small, unproductive firms to their larger, more productive competitors, the former grapple with dwindling resources, trail behind, and exit at a higher rate. However, these departing firms are replaced by new entrants who seize the opportunity to replicate the existing technologies of more productive incumbents. Consequently, the higher churn rate brought about by the R&D reallocation leads to more frequent improvements in the pool of productivity.

As discussed in Section 3.3, the growth contribution from incumbent innovation takes the form of a weighted average of firm-level productivity drifts. Denoting changes between the pre- and post-policy stationary equilibria by the operator Δ, the change in this weighted average can be decomposed as:

$$\begin{aligned} \Delta \int_0^\infty \omega(\hat{z})\gamma(\hat{z})\mathrm{d}\hat{z} = & \int_0^\infty \Delta\gamma(\hat{z}) \times \omega(\hat{z})\mathrm{d}\hat{z} && \text{Productivity growth} && (62.7\%) \\ & + \int_0^\infty \Delta\omega(\hat{z}) \times \gamma(\hat{z})\mathrm{d}\hat{z} && \text{Market expansion} && (14.9\%) \\ & + \int_0^\infty \Delta\omega(\hat{z}) \times \Delta\gamma(\hat{z})\mathrm{d}\hat{z} && \text{Covariance} && (22.4\%). \end{aligned} \tag{12}$$

The first term reflects the average change in firm-level productivity drifts, evaluated over the initial composition of firms. The second term instead reflects the change in firms' market expansion responses, keeping their productivity drifts constant. Finally, the third term captures the covariance between these changes. The "productivity growth" term accounts for 62.7% of the total, compared to 14.9% and 22.4% for the "market expansion" and "covariance" terms, respectively.

Therefore, the larger contribution from incumbent innovation is mostly due to (1) firms simply growing faster on average and (2) initially larger firms achieving the largest increases in their productivity drift. However, a noteworthy contribution comes from the heightened correlation between firms' productivity growth, size, and density. Larger firms, who initially commanded higher markups and whose scale was consequently most restricted, achieve disproportionately faster productivity growth, differentially expand in scale, and become more numerous following the intervention. Accordingly, process efficiency improvements are rolled out over more units sold.

5.3 Aggregate and Cross-Firm Allocations of R&D

In light of the preceding analysis, a question arises: is the accelerated growth in TFP the result of a greater aggregate allocation of labor to innovation, or is it attributable to its reallocation across firms? To elucidate this, we consider an alternative intervention wherein the baseline subsidies are upheld, but concomitantly, a uniform tax is levied on firms' R&D expenditures. This tax is precisely chosen to fix the aggregate allocation of labor to innovation at a level commensurate with the initial stationary equilibrium, which isolates the role of R&D reallocation.

Table 8: Growth Rate Decomposition for Fixed Innovation Labor

Contribution	Before	Baseline		Fixed R&D	
		After	Change	After	Change
Incumbent drift	0.54%	1.02%	+0.49%	0.71%	+0.17%
Incumbent volatility	0.18%	0.19%	+0.01%	0.22%	+0.04%
Entry and exit	0.10%	0.84%	+0.74%	0.82%	+0.72%
Total	0.82%	2.05%	+1.23%	1.74%	+0.92%

Note: This table presents the contributions to the level and change of TFP growth when fixing or not the aggregate allocation of labor to innovation to its initial level before the implementation of the policy intervention. Doing so requires imposing a uniform tax of 44.5% on firms' expenditures on R&D.

The outcomes of this alternate policy are detailed in Table 8. It achieves an increase in

TFP growth equal to 74.8% of the increment observed under the baseline policy, where variations in the aggregate allocation of labor to innovation were admissible. A little over three-quarters of this accelerated growth results from an intensified selection of firms. The redistribution of R&D resources from less productive, smaller firms to their larger, more efficient competitors induces the exit of the former, which are replaced by more productive newcomers. Furthermore, nearly one-fifth of the uplift in TFP growth is derived from an expanded contribution from the productivity drift of incumbent firms. Although the aggregate allocation of labor to innovation is kept fixed, the per-firm average allocation escalates by 49.7% owing to the diminishing number of firms.

5.4 Aggregate Markup and Markup Dispersion

To get a deeper sense of the driving forces behind the acceleration in TFP growth, we disentangle the role of the *aggregate* markup from that of markup *dispersion*. To do so, we consider a slightly more general tax and subsidy schedule. As described in Edmond et al. (2023), the transfers of equation (11) can be generalized with the following parameterization:

$$T_t(\hat{y}) = [\tau_0 \Upsilon(\hat{y}) + \tau_1 \Upsilon'(\hat{y})\hat{y}] Y_t D_t \tag{13}$$

where τ_0 and τ_1 can be appropriately chosen to either mitigate the level or dispersion in markups. In particular, we show in Appendix A.1.9 that setting $\tau_0 = 0$ and $\tau_1 = \mathcal{M}_t - 1$ delivers a uniform subsidy scheme, which leaves the dispersion in markups unchanged from the initial equilibrium, but eliminates the aggregate markup from equation (10). Instead, setting $\tau_0 = \mathcal{M}_t^{-1}$ and $\tau_1 = -1$ delivers a size-dependent scheme of taxes and subsidies, which eliminates markup dispersion, holding the aggregate markup fixed.

Table 9 replicates Table 7 but for those alternative schemes. In particular, the two columns labeled "Level fix" and "Dispersion fix", respectively refer to the scheme that either rectifies the level or dispersion in markups. The "Level fix" has a comparatively muted impact on long-run TFP growth, decreasing it by a slight 4 basis points. This subdued response is largely due to (1) a nearly unchanged growth contribution from incumbent firms' investments in R&D, and (2) a weak reallocation of R&D across firms. The former reflects a pecuniary externality: as firms expand in scale and demand more production labor, they bid up the cost of R&D through a higher wage. This conclusion contrasts with the findings of Edmond et al. (2023), who infer an important role for the aggregate markup in distorting the scale of the economy at a point in time. Instead, we find the level of markups to have little to no bearing on the rate at which the economy grows in the long run.

Table 9: Growth Rate Decomposition for Alternative Transfer Schedules

Contribution	Before	Baseline		Level fix		Dispersion fix	
		After	Change	After	Change	After	Change
Incumbent drift	0.54%	1.02%	+0.49%	0.51%	-0.03%	1.06%	+0.53%
Incumbent vol.	0.18%	0.19%	+0.01%	0.18%	-0.00%	0.19%	+0.01%
Entry and exit	0.10%	0.84%	+0.74%	0.09%	-0.01%	0.89%	+0.79%
Total	0.82%	2.05%	+1.23%	0.78%	-0.04%	2.15%	+1.33%

Note: This table presents the contributions to the level and change (p.p.) of TFP growth under alternative policy interventions. Specifically, the columns labeled "Baseline", "Level fix" and "Dispersion fix" refer to the transfers that rectify both, and either the level or dispersion in markups, respectively.

Conversely, the tax and subsidy schedule that eliminates markup dispersion while leaving their average level unchanged achieves an increase in long-run TFP growth that is even larger than that from the baseline intervention.[33] The largest contributors are, here again, the terms reflecting incumbents' productivity drift and entry and exit. Nonetheless, it is worth noting that this faster productivity growth is not necessarily indicative of an improvement in welfare. On the one hand, too few resources may be directed towards production under this allocation, thus reducing aggregate output. On the other hand, this scheme might strike a more optimal balance between rectifying product market distortions and addressing other market failures.

As mentioned earlier, these market failures take the form of technological externalities across firms. As entering firms draw their relative productivity from a transformation of the incumbent distribution, the latter do not internalize that their R&D investments benefit future cohorts of firms. Further, as emphasized in Lashkari (2023), since the lower bound of the productivity support is endogenous, an unproductive firm may choose to stay in business to extract ex-post rents, but in doing so, it "pollutes" the pool from which entrants draw their relative productivity.

All else equal, these inefficiencies imply that the market would (1) allocate too few resources to R&D and (2) harbor an inefficiently large density of small unproductive firms. The "Dispersion fix" tax and subsidy scheme inadvertently addresses both of

[33] This intervention is nearly budget-neutral. It raises revenue amounting to 0.6% of output, which is rebated to the household.

these market failures, albeit imperfectly. In preserving the level of markups, production labor demand remains low, thus "freeing up" resources for R&D. Further, this scheme takes the form of a tax for the smallest firms, who initially charged below-average markups. To induce these firms to increase their markup to the pre-policy average, their output must be taxed. Since this tax is passed on through higher prices, it reduces the demand they face and edges them out of the market, thereby improving the productivity pool. Yet, without a comprehensive welfare analysis, it remains unclear which scheme achieves the largest improvement in welfare.

5.5 Robustness

An important assumption entertained in the quantification of our model is to target an aggregate markup of 30%. De Ridder et al. (2023) measure a sales-weighted *harmonic* average markup of 10% and 50% using French firm-level revenue and quantity data, respectively.[34] To assess the implications of this assumption, Table 10 replicates Table 7 with a targeted aggregate markup of 10% or 50%.[35] Structural parameters are re-estimated under these alternative targets. With a lower target of 10%, the implications of the intervention are more tempered. The increase in the long-run growth rate of TFP is significantly muted at 29 basis points. However, we see a nontrivial contribution of 13 basis points from the faster productivity growth achieved by incumbent firms.

Table 10: Growth Rate Decomposition for Alternative Aggregate Markup Targets

Contribution	$\mathcal{M}$=1.1			$\mathcal{M}$=1.5		
	Before	After	Change	Before	After	Change
Incumbent drift	0.15%	0.28%	+0.13%	0.83%	2.12%	+1.29%
Incumbent vol.	0.33%	0.47%	+0.14%	0.14%	0.15%	+0.01%
Entry and exit	0.28%	0.30%	+0.02%	-0.20%	0.77%	+0.97%
Total	0.76%	1.05%	+0.29%	0.77%	3.03%	+2.26%

Note: This table presents the contributions to the level and change of TFP growth under alternative aggregate markup targets. Specifically, the columns labeled "$\mathcal{M}$=1.1" and "$\mathcal{M}$=1.5" respectively refer to parameterizations that target a cost-weighted average markup of 1.1 and 1.5.

[34] As a reference point, Aghion et al. (2023) entertain an aggregate markup of 50%.

[35] Appendix C.3 further replicates Tables 4, 5 and 6 for these cases.

With a target of 50% for the aggregate markup, the repercussions of the intervention are magnified. TFP growth increases by 2.26 percentage points, with a considerably larger contribution from incumbent firms' productivity drift. Decomposing the change in this term according to equation (12) reveals a larger contribution (46.1% of the total) from a greater correlation between firm-level productivity drifts and their market expansion responses. Meanwhile, the contribution of larger productivity drifts, holding fixed the composition of firms, is slightly subdued relative to baseline (48.2% of the total).

5.6 Discussion

In this subsection, we delve into the nuances of our theoretical framework by exploring a range of possible extensions and alternative assumptions. Two forthcoming extensions are to (1) conduct a comprehensive welfare analysis incorporating transition dynamics and (2) characterize the optimal allocation of resources. First, as emphasized in Atkeson, Burstein and Chatzikonstantinou (2019), transition dynamics tend to be slow in models of endogenous economic growth. Such inertia might curtail the intervention's welfare implications if the accelerated pace of productivity growth mostly materializes in the distant future. Second, detailing the optimal allocation of resources will provide insight into the efficiency properties of this intervention.

Process vs. Product Innovation

In our framework, firms invest in R&D to achieve improvements in their *process efficiency*. An alternative assumption is to consider *product quality* improvements as the objective for those investments. We show in Appendix A.3 that when the quality and quantity of a product are perfect substitutes, this alternative is isomorphic to our model. While the assumption of perfect substitutability between quantity and quality is undoubtedly stylized, the economic modeling of the latter lacks a cohesive consensus. Hence, further study of product quality improvements under non-isoelastic demand systems presents a promising frontier for exploration.

Production vs. Innovation Resource Substitution

Our analysis yields a perhaps unexpected result: a uniform subsidy that rectifies the aggregate markup–while preserving the dispersion in markups–mildly diminishes the long-run growth rate of TFP. It is worth noting that this result is partly attributable to our theoretical choices. As posited in Section 3.1, labor can be seamlessly reallocated between production and innovation. Consequently, by inducing firms to employ more labor in production, a uniform output subsidy can potentially reduce the availability of labor for R&D. This reveals that the degree of substitutability between production and innovation resources determines the extent to which the aggregate markup can alter economic growth.

To elucidate, consider an alternative economy wherein an elastic supply of 'skilled' and 'unskilled' labor can only be allocated to innovation and production, *respectively*. Under these circumstances, an output subsidy would leave the availability of 'skilled' labor unchanged, as it is non-transferable to production. A more contrasting alternative is that of an economy in which the final sector (instead of intermediate firms) uses labor and intermediate inputs to produce the final good. In this environment, rectifying the level of markups would induce the final sector to reallocate expenditures toward (initially marked-up) intermediates and away from production labor, thus freeing up resources for R&D rather than restricting them. In that sense, we consider our choices to be conservative.

Monopolistic vs. Oligopolistic Competition

Another assumption of our model is the tractable premise of *monopolistic* competition. Notwithstanding, Edmond et al. (2023) find that the efficiency losses from markups are greater under *oligopolistic* competition, through which they infer greater dispersion therein. It remains to be shown whether this conclusion extends to the dynamic costs of markups, but one might conjecture that it could, based on the following argument. The Klenow and Willis (2016) functional form makes two counterfactual predictions under monopolistic competition. First, in order to accommodate a reasonable distribution of markups, it requires unreasonably little heterogeneity in pass-throughs, which grates against the evidence documented in Amiti et al. (2019). Second, the sharp concavity in demand at lower prices appears empirically inconsistent when compared to the heavy observed tails of firm-level employment or sales.

What are the implications of these counterfactual predictions for economic growth? With little dispersion in pass-throughs prior to the intervention, the transition to marginal cost pricing (complete pass-throughs for all firms) barely improves the reallocation of demand towards the most productive firms. Further, the pronounced decline in the elasticity of demand implies that lower prices are not met with substantially higher demand. These two properties of this functional form limit the extent to which marginal cost pricing can increase the correlation between firms' productivity growth and their resulting expansion responses. Amiti et al. (2019) argue that models of oligopolistic competition can replicate the joint distribution of markups and pass-throughs more closely. This suggests that the consequences of markups might be amplified under a market structure that (1) aligns with lower pass-throughs for large firms and (2) obviates the requirement for a steeply declining demand elasticity at lower prices.

Customer Acquisition

Neglecting the role of customer acquisition might also understate such consequences. Afrouzi et al. (2023) find that in a model parallel to Edmond et al. (2023), considering the endogenous accumulation of a customer base unveils greater dispersion in markups, thereby intensifying the implied efficiency losses from markups.[36] In addition, Einav, Klenow, Levin and Murciano-Goroff (2021) document that, while new entrants accrue lower total sales than incumbents, their average sales per customer are nearly equivalent. Since a firm's process efficiency is related to that intensive (rather than extensive) margin of demand, these findings suggest that new entrants might be more productive than inferred in models omitting a customer margin, such as ours. Hence, accounting for this extensive margin of demand could amplify the role played by the selective displacement of unsuccessful firms by more productive newcomers.

Firm Ownership, Nonlinear Pricing, and Semi-Endogenous Growth

However, other forces not encompassed in our analysis–such as concentrated firm ownership, nonlinear pricing, and semi-endogenous growth–might instead temper our conclusions. Boar and Midrigan (2019) find that when firm ownership is concentrated, tensions arise between concerns for efficiency and inequality. Whether these trade-offs are mitigated or amplified in models of endogenous economic growth remains an open question. When firms engage in second-degree price discrimination, as in Bornstein and Peter (2023), their ability to appropriate a larger fraction of the consumer surplus might narrow the gap between private and social incentives for innovation. The work of Jones (1995) adds yet another layer of nuance. Should economic growth be intrinsically tied to demographic dynamics, the acceleration in TFP growth we identified might only be transient. Combining the insights from semi-endogenous growth theory with models of firm-led technological change, such as those proposed by Peretto (1998), Dinopoulos and Thompson (1998), Young (1998) and Howitt (1999), offers a promising route for future exploration.

[36] These findings may extend to frictions in accumulating factors of production, such as in Bilal, Engbom, Mongey and Violante (2021).

6 Conclusion

We studied the consequences of markups for long-run economic growth in a general equilibrium model of firm-led endogenous technological change. In our model, firms engage in monopolistic competition, charge heterogeneous markups, and their profit-seeking investments in R&D propel economic growth.

Two economic insights have formed the basis of our argument in this paper. Since ideas are nonrival, investments in R&D have increasing returns to scale. Yet, markups distort the scale at which firms operate and, therefore, affect their incentives to invest in R&D. In general equilibrium, however, markups also distort the aggregate demand for labor, thereby affecting its availability for R&D.

To quantify our model, we estimated its parameters using French macroeconomic and firm-level administrative data. We found that an intervention correcting the product market distortions induced by markups raises productivity growth by 1.2 percentage points in the long run. This accelerated growth results from an increase in aggregate R&D employment, its reallocation towards firms with a broader market reach, and a higher rate of churn from entry and exit.

We conducted two alternative exercises to elucidate these findings further. We first explored a "constrained" intervention, which kept the aggregate allocation of labor to innovation fixed, to determine whether the uptick in TFP growth is predominantly due to the allocation of additional resources to R&D or a reallocation of these resources across firms. Our findings suggest that nearly 75% of the baseline acceleration in growth can be attained by simply reallocating a fixed quantity of innovation labor from small, unproductive firms to their larger, more productive competitors.

Our final counterfactual exercise disentangled the role of the aggregate markup from that of markup dispersion. We found that rectifying the level of markups has a relatively muted impact on long-run TFP growth, as firms bid up the cost of R&D investments via a higher wage. In contrast, neutralizing the dispersion in markups achieves slightly faster TFP growth than the baseline intervention. This exercise revealed that the dispersion in markups, rather than their average level, is more detrimental to economic growth.

To conclude, we emphasize that heterogeneous markups serve as just one illustration of a distortion that differentially affects firms' production scale and, thus, their incentives to improve their technology. Hsieh and Klenow (2009) document that such distortions (e.g., size-dependent taxes and regulations, financial frictions, or political connections) are plausibly large and pervasive, and Baqaee and Farhi (2019) show that input-output linkages substantially amplify their consequences on allocative efficiency. In light of this, one cannot help but surmise that the real world might be riddled with impediments to long-run economic growth awaiting further study.

References

Acemoglu, Daron, "Distorted Innovation: Does the Market Get the Direction of Technology Right?," *AEA Papers and Proceedings*, May 2023, *113*, 1–28.

__ **, Ufuk Akcigit, Harun Alp, Nicholas Bloom, and William Kerr**, "Innovation, Reallocation, and Growth," *American Economic Review*, November 2018, *108* (11), 3450–91.

Afrouzi, Hassan, Andres Drenik, and Ryan Kim, "Concentration, Market Power, and Misallocation: The Role of Endogenous Customer Acquisition," Working Paper 31415, National Bureau of Economic Research June 2023.

Aghion, Philippe, Antonin Bergeaud, Timo Boppart, Peter J. Klenow, and Huiyu Li, "Good Rents versus Bad Rents: R&D Misallocation and Growth," Working Paper 2023.

Akcigit, Ufuk and William R. Kerr, "Growth through Heterogeneous Innovations," *Journal of Political Economy*, 2018, *126* (4), 1374–1443.

__ **, Douglas Hanley, and Nicolas Serrano-Velarde**, "Back to Basics: Basic Research Spillovers, Innovation Policy, and Growth," *The Review of Economic Studies*, 10 2020, *88* (1), 1–43.

__ **,** __ **, and Stefanie Stantcheva**, "Optimal Taxation and R&D Policies," *Econometrica*, 2022, *90* (2), 645–684.

__ **, Murat Alp Celik, and Jeremy Greenwood**, "Buy, Keep, or Sell: Economic Growth and the Market for Ideas," *Econometrica*, 2016, *84* (3), 943–984.

Amiti, Mary, Oleg Itskhoki, and Jozef Konings, "Importers, Exporters, and Exchange Rate Disconnect," *American Economic Review*, July 2014, *104* (7), 1942–78.

__ **,** __ **, and** __ , "International Shocks, Variable Markups, and Domestic Prices," *The Review of Economic Studies*, 02 2019, *86* (6), 2356–2402.

Andrews, Isaiah, Matthew Gentzkow, and Jesse M. Shapiro, "Measuring the Sensitivity of Parameter Estimates to Estimation Moments," *The Quarterly Journal of Economics*, 06 2017, *132* (4), 1553–1592.

Arrow, Kenneth J., "Economic Welfare and the Allocation of Resources for Invention," in "The Rate and Direction of Inventive Activity: Economic and Social Factors," Princeton University Press, 1962, pp. 609–626.

Atkeson, Andrew, Ariel T. Burstein, and Manolis Chatzikonstantinou, "Transitional Dynamics in Aggregate Models of Innovative Investment," *Annual Review of Economics*, 2019, *11* (1), 273–301.

Autor, David, David Dorn, Lawrence F Katz, Christina Patterson, and John Van Reenen, "The Fall of the Labor Share and the Rise of Superstar Firms," *The Quarterly Journal of Economics*, 02 2020, *135* (2), 645–709.

Ayerst, Stephen B., "Innovator Heterogeneity, R&D Misallocation and the Productivity Growth Slowdown," Working Paper 2023.

Baqaee, David Rezza and Emmanuel Farhi, "Productivity and Misallocation in General Equilibrium," *The Quarterly Journal of Economics*, 09 2019, *135* (1), 105–163.

_ , _ , and Kunal Sangani, "The Darwinian Returns to Scale," *The Review of Economic Studies*, 06 2023, p. rdad061.

Behrens, Kristian, Giordano Mion, Yasusada Murata, and Jens Suedekum, "Quantifying the Gap Between Equilibrium and Optimum under Monopolistic Competition," *The Quarterly Journal of Economics*, 05 2020, *135* (4), 2299–2360.

Benhabib, Jess, Jesse Perla, and Christopher Tonetti, "Reconciling Models of Diffusion and Innovation: A Theory of the Productivity Distribution and Technology Frontier," *Econometrica*, 2021, *89* (5), 2261–2301.

Bilal, Adrien, Niklas Engbom, Simon Mongey, and Giovanni L Violante, "Labor Market Dynamics When Ideas are Harder to Find," Working Paper 29479, National Bureau of Economic Research November 2021.

Bilbiie, Florin O., Fabio Ghironi, and Marc J. Melitz, "Monopoly Power and Endogenous Product Variety: Distortions and Remedies," *American Economic Journal: Macroeconomics*, October 2019, *11* (4), 140–74.

Boar, Corina and Virgiliu Midrigan, "Markups and Inequality," Working Paper 25952, National Bureau of Economic Research June 2019.

Bond, Steve, Arshia Hashemi, Greg Kaplan, and Piotr Zoch, "Some Unpleasant Markup Arithmetic: Production Function Elasticities and Their Estimation From Production Data," *Journal of Monetary Economics*, 2021, *121*, 1–14.

Bornstein, Gideon and Alessandra Peter, "Nonlinear Pricing and Misallocation," Working Paper 2023.

Byrd, Richard H., Jorge Nocedal, and Richard A. Waltz, "KNITRO: An Integrated Package for Nonlinear Optimization," *Large-Scale Nonlinear Optimization*, 2006, pp. 35–59.

Cavenaile, Laurent, Murat Alp Celik, and Xu Tian, "Are Markups Too High? Competition, Strategic Innovation, and Industry Dynamics," Working Paper 2021.

Caves, Richard E., "Industrial Organization and New Findings on the Turnover and Mobility of Firms," *Journal of Economic Literature*, 1998, *36* (4), 1947–1982.

Chen, Zhao, Zhikuo Liu, Juan Carlos Suárez Serrato, and Daniel Yi Xu, "Notching R&D Investment with Corporate Income Tax Cuts in China," *American Economic Review*, July 2021, *111* (7), 2065–2100.

Dasgupta, Partha and Joseph Stiglitz, "Industrial Structure and the Nature of Innovative Activity," *The Economic Journal*, 06 1980, *90* (358), 266–293.

Dhingra, Swati and John Morrow, "Monopolistic Competition and Optimum Product Diversity under Firm Heterogeneity," *Journal of Political Economy*, 2019, *127* (1), 196–232.

Dinopoulos, Elias and Peter Thompson, "Schumpeterian Growth without Scale Effects," *Journal of Economic Growth*, 1998, *3* (4), 313–335.

Dixit, Avinash K. and Joseph E. Stiglitz, "Monopolistic Competition and Optimum Product Diversity," *The American Economic Review*, 1977, *67* (3), 297–308.

_ **and Robert S. Pindyck**, *Investment under Uncertainty*, Princeton University Press, 1994.

Dubé, Jean-Pierre, Jeremy T. Fox, and Che-Lin Su, "Improving the Numerical Performance of Static and Dynamic Aggregate Discrete Choice Random Coefficients Demand Estimation," *Econometrica*, 2012, *80* (5), 2231–2267.

Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu, "How Costly Are Markups?," *Journal of Political Economy*, 2023, *131* (7), 1619–1675.

Einav, Liran, Peter J Klenow, Jonathan D Levin, and Raviv Murciano-Goroff, "Customers and Retail Growth," Working Paper 29561, National Bureau of Economic Research December 2021.

Ericson, Richard and Ariel Pakes, "Markov-Perfect Industry Dynamics: A Framework for Empirical Work," *The Review of Economic Studies*, 01 1995, *62* (1), 53–82.

Feenstra, Robert C., "A Homothetic Utility Function for Monopolistic Competition Models, Without Constant Price Elasticity," *Economics Letters*, 2003, *78* (1), 79–86.

Flynn, Zach, James Traina, and Amit Gandhi, "Measuring Markups with Production Data," Working Paper 2019.

Foster, Lucia, John C. Haltiwanger, and Cornell J. Krizan, "Aggregate Productivity Growth: Lessons From Microeconomic Evidence," in "New Developments in Productivity Analysis," University of Chicago Press, 2001, pp. 303–372.

Garcia-Macia, Daniel, Chang-Tai Hsieh, and Peter J. Klenow, "How Destructive Is Innovation?," *Econometrica*, 2019, *87* (5), 1507–1541.

Garella, Paolo G., "Monopoly incentives for cost-reducing R&D," *Economics Letters*, 2012, *117* (1), 21–24.

Grullon, Gustavo, Yelena Larkin, and Roni Michaely, "Are US Industries Becoming More Concentrated?," *Review of Finance*, 04 2019, *23* (4), 697–743.

Harberger, Arnold C., "Monopoly and Resource Allocation," *American Economic Review*, May 1954, *44* (2), 77–87.

Hopenhayn, Hugo A., "Entry, Exit, and firm Dynamics in Long Run Equilibrium," *Econometrica*, 1992, *60* (5), 1127–1150.

Hopenhayn, Hugo and Francesco Squintani, "On the Direction of Innovation," *Journal of Political Economy*, 2021, *129* (7), 1991–2022.

Howitt, Peter, "Steady Endogenous Growth with Population and R & D Inputs Growing," *Journal of Political Economy*, 1999, *107* (4), 715–730.

Hsieh, Chang-Tai and Peter J. Klenow, "Misallocation and Manufacturing TFP in China and India," *The Quarterly Journal of Economics*, 11 2009, *124* (4), 1403–1448.

Jones, Charles I., "R&D-Based Models of Economic Growth," *Journal of Political Economy*, 1995, *103* (4), 759–784.

Kehrig, Matthias and Nicolas Vincent, "The Micro-Level Anatomy of the Labor Share Decline," *The Quarterly Journal of Economics*, 03 2021, *136* (2), 1031–1087.

Kimball, Miles S., "The Quantitative Analytics of the Basic Neomonetarist Model," *Journal of Money, Credit and Banking*, 1995, *27* (4), 1241–1277.

Klenow, Peter J. and Jonathan L. Willis, "Real Rigidities and Nominal Price Changes," *Economica*, 2016, *83* (331), 443–472.

König, Michael, Kjetil Storesletten, Zheng Song, and Fabrizio Zilibotti, "From Imitation to Innovation: Where Is All That Chinese R&D Going?," *Econometrica*, 2022, *90* (4), 1615–1654.

Lashkari, Danial, "Innovation Policy in Theories of Knowledge Diffusion and Selection," Working Paper 2023.

Lehr, Nils H., "R&D Return Dispersion and Economic Growth – The Case of Inventor Market Power," Working Paper 2023.

Lerner, Abba P., "The Concept of Monopoly and the Measurement of Monopoly Power," *The Review of Economic Studies*, 06 1934, *1* (3), 157–175.

Liu, Ernest and Song Ma, "Innovation Networks and R&D Allocation," Working Paper 29607, National Bureau of Economic Research December 2021.

Loecker, Jan De and Frederic Warzynski, "Markups and Firm-Level Export status," *American Economic Review*, May 2012, *102* (6), 2437–71.

_ , Jan Eeckhout, and Gabriel Unger, "The Rise of Market Power and the Macroeconomic Implications," *The Quarterly Journal of Economics*, 01 2020, *135* (2), 561–644.

_ , Pinelopi K. Goldberg, Amit K. Khandelwal, and Nina Pavcnik, "Prices, Markups, and Trade Reform," *Econometrica*, 2016, *84* (2), 445–510.

Lubin, Miles, Oscar Dowson, Joaquim Dias Garcia, Joey Huchette, Benoît Legat, and Juan Pablo Vielma, "JuMP 1.0: Recent improvements to a modeling language for mathematical optimization," *Mathematical Programming Computation*, 2023.

Luttmer, Erzo G. J., "Selection, Growth, and the Size Distribution of Firms," *The Quarterly Journal of Economics*, 08 2007, *122* (3), 1103–1144.

MaCurdy, Thomas E., "An Empirical Model of Labor Supply in a Life-Cycle Setting," *Journal of Political Economy*, 1981, *89* (6), 1059–1085.

Marshall, Alfred, *Principles of Economics*, Macmillan for the Royal Economic Society, London, 1961, 1890.

Matsuyama, Kiminori and Philip Ushchev, "Beyond CES: Three Alternative Classes of Flexible Homothetic Demand Systems," *Global Poverty Research Lab Working Paper*, 2017, (17-109).

__ **and** __ , "Selection and Sorting of Heterogeneous Firms Through Competitive Pressures," Working Paper 2022.

Miranda, Mario J. and Paul L. Fackler, *Applied Computational Economics and Finance*, MIT press, 2004.

Peretto, Pietro F, "Technological change and population growth," *Journal of Economic Growth*, 1998, *3* (4), 283–311.

Peters, Michael, "Heterogeneous Markups, Growth, and Endogenous Misallocation," *Econometrica*, 2020, *88* (5), 2037–2073.

Raval, Devesh, "Testing the Production Approach to Markup Estimation," *The Review of Economic Studies*, 01 2023, p. rdad002.

Ridder, Maarten De, "Market Power and Innovation in the Intangible Economy," Working Paper 2023.

__ **, Basile Grassi, and Giovanni Morzenti**, "The Hitchhiker's Guide to Markup Estimation," Working Paper 2023.

Rivera-Batiz, Luis A. and Paul M. Romer, "Economic Integration and Endogenous Growth," *The Quarterly Journal of Economics*, 05 1991, *106* (2), 531–555.

Romer, Paul M., "Endogenous Technological Change," *Journal of Political Economy*, 1990, *98* (5, Part 2), S71–S102.

Scherer, Frederic M., "Invention and Innovation in the Watt-Boulton Steam-Engine Venture," *Technology and Culture*, 1965, *6* (2), 165–187.

Schmookler, Jacob, *Invention and Economic Growth*, Cambridge, MA and London, England: Harvard University Press, 1966.

Smith, Adam, *An Inquiry Into the Nature and Causes of the Wealth of Nations*, W. Strahan and T. Cadell, London, 1776.

Stokey, Nancy L., *The Economics of Inaction: Stochastic Control Models with Fixed Costs*, Princeton University Press, 2009.

Su, Che-Lin and Kenneth L. Judd, "Constrained Optimization Approaches to Estimation of Structural Models," *Econometrica*, 2012, *80* (5), 2213–2230.

Tirole, Jean, *The Theory of Industrial Organization*, Vol. 1 of *MIT Press Books*, The MIT Press, December 1988.

Voronina, Maria, "Endogenous Growth and Optimal Market Power," Working Paper 2021.

Wächter, Andreas and Lorenz T. Biegler, "On the Implementation of a Primal-Dual Interior Point Filter Line Search Algorithm for Large-Scale Nonlinear Programming," *Mathematical programming*, 2006, *106*, 25–57.

Yeh, Chen, "Revisiting the Origins of Business Cycles with the Size-Variance Relationship," *Federal Reserve Bank of Richmond mimeo*, 2021.

Young, Alwyn, "Growth without Scale Effects," *Journal of Political Economy*, 1998, *106* (1), 41–63.

A Theoretical Appendix

This section of the Appendix provides derivations, proofs and extensions for the results presented in Section 3. Appendix A.1 provides derivations, Appendix A.2 provides proofs and Appendix A.3 discusses extensions to our framework.

A.1 Derivations

A.1.1 Discretization

In some derivations and proofs presented in this appendix, working with a discrete-time approximation of the model is convenient. As such, we follow the approach in Dixit and Pindyck (1994) and consider two discrete spaces $\tau \in \mathbb{N} \cup \{0\}$ and $j \in \mathbb{Z}$ for time and productivity, respectively. We define time intervals of length $\Delta > 0$ and productivity intervals of length $\Delta_z = \sigma\sqrt{\Delta}$ such that $t = \tau\Delta$ and $z = j\Delta_z$. With these definitions, the firm's productivity diffusion process can be approximated by a random walk. That is, between τ and $\tau + 1$, the probability of going from j to $j \pm 1$ is given by:

$$p_\tau^+(\gamma_\tau(j)) = [1 + \gamma_\tau(j)\sqrt{\Delta}/\sigma]/2 \quad \text{and} \quad p_\tau^-(\gamma_\tau(j)) = [1 - \gamma_\tau(j)\sqrt{\Delta}/\sigma]/2.$$

A.1.2 The Household's Problem

Taking prices as given, the household's problem is to choose its consumption and hours worked to maximize lifetime utility subject to a flow budget constraint:

$$\max_{\{C_t, H_t\}_{t \geq 0}} \int_0^\infty e^{-\rho t}[\ln(C_t) - v(H_t)]\mathrm{d}t \quad \text{s.t.} \quad \dot{A}_t = r_t A_t + w_t H_t - C_t.$$

Reformulating the Household's problem using the current-value Hamiltonian, we have:

$$\mathcal{H}_t(C_t, H_t, A_t, \nu_t) = \ln(C_t) - v(H_t) + \nu_t(r_t A_t + w_t H_t - C_t)$$

where ν_t denotes the costate variable. The first-order conditions are:

$$\frac{\partial \mathcal{H}_t}{\partial C_t} = 1/C_t - \nu_t = 0,$$
$$\frac{\partial \mathcal{H}_t}{\partial H_t} = -v'(H_t) + \nu_t w_t = 0,$$
$$\frac{\partial \mathcal{H}_t}{\partial A_t} = \nu_t r_t = \nu_t \rho - \dot{\nu}_t$$

together with the No-Ponzi and transversality conditions, which jointly imply:

$$\lim_{t\to\infty} e^{-\int_0^t r_{t'}\mathrm{d}t'} A_t = 0$$

Combining these equations, we obtain the usual intertemporal Euler equation and static first-order condition:

$$\frac{\dot{C}_t}{C_t} = r_t - \rho \quad \text{and} \quad v'(H_t)C_t = w_t.$$

A.1.3 The Final Sector's Problem

Taking prices as given, the final sector's problem is to choose its relative demand for each variety to maximize profits in each period:

$$\max_{\{\hat{y}_t(\hat{z})\}_{\hat{z}=0}^{\infty}} \left\{ P_t - M_t \int_0^{\infty} p_t(\hat{z})\hat{y}_t(\hat{z})\mathrm{d}F_t(\hat{z}) \right\} Y_t \quad \text{s.t.} \quad M_t \int_0^{\infty} \Upsilon(\hat{y}_t(\hat{z}))\mathrm{d}F_t(\hat{z}) = 1.$$

Reformulating the final sector's problem as a cost-minimization problem subject to the Kimball (1995) aggregator constraint using the Lagrangian, we have:

$$\mathcal{L}_t(\{\hat{y}_t(\hat{z})\}_{\hat{z}=0}^{\infty}, \nu_t) = M_t Y_t \int_0^{\infty} p_t(\hat{z})\hat{y}_t(\hat{z})\mathrm{d}F_t(\hat{z}) + \nu_t \left(M_t \int_0^{\infty} \Upsilon(\hat{y}_t(\hat{z}))\mathrm{d}F_t(\hat{z}) - 1 \right)$$

where ν_t now denotes the Lagrange multiplier. The first-order conditions are:

$$p_t(\hat{z}) = \nu_t \Upsilon'(\hat{y}_t(\hat{z}))/Y_t \quad \text{and} \quad M_t \int_0^{\infty} \Upsilon(\hat{y}_t(\hat{z}))\mathrm{d}F_t(\hat{z}) = 1.$$

Since the final sector is perfectly competitive and makes no profit, we have:

$$P_t = M_t \int_0^{\infty} p_t(\hat{z})\hat{y}_t(\hat{z})\mathrm{d}F_t(\hat{z}).$$

Substituting in the first-order conditions, we obtain a solution for ν_t:

$$\nu_t = P_t Y_t D_t \quad \text{where} \quad D_t \equiv \left(M_t \int_0^{\infty} \Upsilon'(\hat{y}_t(\hat{z}))\hat{y}_t(\hat{z})\mathrm{d}F_t(\hat{z}) \right)^{-1}.$$

This delivers the following inverse demand functions:

$$p_t(\hat{z}) = \Upsilon'(\hat{y}_t(\hat{z}))P_t D_t.$$

A.1.4 The Firm's Static Problem

Firms engage in monopolistic competition in the product market but perfect competition in the input markets. That is, a firm chooses the price at which to sell its variety as well as its demand for physical capital and production labor to maximize profits in each period. The firm takes as given the demand for its variety, the rental rate of capital r_t, the wage rate w_t and any transfer $T_t(\hat{y}_t(\hat{z}))$, which delivers the following problem:

$$\pi_t(\hat{z}) = \max_{p_t(\hat{z}),k_t(\hat{z}),l_t(\hat{z})} \{p_t(\hat{z})y_t(\hat{z}) - (r_t+\delta)k_t(\hat{z}) - w_t l_t(\hat{z}) + T_t(\hat{y}_t(\hat{z}))\} - w_t c_O$$

subject to the inverse demand function $p_t(\hat{z}) = \Upsilon'(\hat{y}_t(\hat{z}))D_t$. Let us first consider the sub-problem of optimally choosing the demand for capital and labor, which can be reformulated as a cost-minimization problem. Using the Lagrangian, we have:

$$\mathcal{L}_t(k_t(\hat{z}), l_t(\hat{z}), \nu_t(\hat{z})) = (r_t+\delta)k_t(\hat{z}) + w_t l_t(\hat{z}) + \nu_t(\hat{z})[y_t(\hat{z}) - \exp(\hat{z}+\underline{z}_t)k_t(\hat{z})^\alpha l_t(\hat{z})^{1-\alpha}]$$

where $\nu_t(\hat{z})$ denotes the Lagrange multiplier. The first-order conditions are:

$$\begin{aligned} k_t(\hat{z}) &= \frac{\alpha \nu_t(\hat{z}) y_t(\hat{z})}{r_t+\delta}, \\ l_t(\hat{z}) &= \frac{(1-\alpha)\nu_t(\hat{z}) y_t(\hat{z})}{w_t}, \\ y_t(\hat{z}) &= \exp(\hat{z}+\underline{z}_t)k_t(\hat{z})^\alpha l_t(\hat{z})^{1-\alpha}. \end{aligned}$$

Solving for the Lagrange multiplier, we have:

$$\nu_t(\hat{z}) = \varsigma_t \exp(-\hat{z}-\underline{z}_t) \quad \text{where} \quad \varsigma_t \equiv \left(\frac{r_t+\delta}{\alpha}\right)^\alpha \left(\frac{w_t}{1-\alpha}\right)^{1-\alpha}.$$

Therefore, we can rewrite the firm's static problem as:

$$\begin{aligned} \pi_t(\hat{z}) = \max_{p_t(\hat{z})} & \{[p_t(\hat{z}) - \varsigma_t \exp(-\hat{z}-\underline{z}_t)]\hat{y}_t(\hat{z})Y_t + T_t(\hat{y}_t(\hat{z}))\} - w_t c_O \\ \text{s.t.} \quad & p_t(\hat{z}) = \Upsilon'(\hat{y}_t(\hat{z}))D_t. \end{aligned}$$

Reformulating it as a choice of $\hat{y}_t(\hat{z})$ given the inverse demand function $p_t(\hat{z})$, we have:

$$\pi_t(\hat{z}) = \max_{\hat{y}_t(\hat{z})} \{[\Upsilon'(\hat{y}_t(\hat{z}))D_t - \varsigma_t \exp(-\hat{z}-\underline{z}_t)]\hat{y}_t(\hat{z})Y_t + T_t(\hat{y}_t(\hat{z}))\} - w_t c_O.$$

The first-order condition is:

$$[\Upsilon''(\hat{y}_t(\hat{z}))\hat{y}_t(\hat{z}) + \Upsilon'(\hat{y}_t(\hat{z}))]D_t + T_t'(\hat{y}_t(\hat{z}))/Y_t = \varsigma_t \exp(-\hat{z} - \underline{z}_t).$$

With $T_t(\hat{y}_t(\hat{z})) = 0$ for all $\hat{z}$, we can rearrange this expression as:

$$p_t(\hat{z}) = \frac{\mu_t(\hat{z})\varsigma_t}{\exp(\hat{z} + \underline{z}_t)} \quad \text{where} \quad \mu_t(\hat{z}) \equiv \frac{\vartheta_t(\hat{z})}{\vartheta_t(\hat{z}) - 1}$$

and where $\vartheta_t(\hat{z})$ denotes the price elasticity of demand:

$$\vartheta_t(\hat{z}) \equiv -\frac{\Upsilon'(\hat{y}_t(\hat{z}))}{\Upsilon''(\hat{y}_t(\hat{z}))\hat{y}_t(\hat{z})} \in (1, \infty).$$

Substituting the monopoly pricing function in the profit function, we have:

$$\pi_t(\hat{z}) = \frac{p_t(\hat{z})\hat{y}_t(\hat{z})Y_t}{\vartheta_t(\hat{z})} - w_t c_O.$$

Denoting the firm's price relative to the choke price as $\hat{p}_t(\hat{z}) \equiv p_t(\hat{z})/\overline{p}_t$, we can rewrite:

$$\hat{p}(\hat{z}) = \frac{\mu(\hat{z})\varsigma_t/\overline{p}_t}{\exp(\hat{z} + \underline{z}_t)} \quad \text{and} \quad \pi_t(\hat{z}) = \frac{\hat{p}(\hat{z})\hat{y}(\hat{z})\overline{p}_t Y_t}{\vartheta(\hat{z})} - w_t c_O.$$

With the transfers from equation (11), the first-order condition is instead:

$$\hat{p}(\hat{z}) = \frac{\varsigma_t/\overline{p}_t}{\exp(\hat{z} + \underline{z}_t)}$$

such that all firms price at marginal cost. This implies that profits are simply given by:

$$\pi_t(\hat{z}) = [\Upsilon(\hat{y}(\hat{z})) - \Upsilon'(\hat{y}(\hat{z}))\hat{y}(\hat{z})]Y_t D_t - w_t c_O.$$

A.1.5 The Firm's Dynamic Problem

Given the above static profit function and taking the wage rate as given, firms control the drift of their productivity and choose an exit stopping time τ:

$$V_t(z) = \max_{\tau, \{\gamma_s\}_{s \geq t}} \mathbb{E}_z \left\{ \int_t^{t+\tau} e^{-\int_t^s (r_{t'} + \chi)\mathrm{d}t'} [\pi_s(z_s) - w_s i(\gamma_s, z_s - \underline{z}_t)]\mathrm{d}s \right\}$$

where $\mathbb{E}_z$ denotes the expectation operator with respect to the diffusion process $\{z_s\}_{s \geq t}$ when its initial value is $z_t = z$. With the definitions presented in Appendix A.1.1, we consider the discrete-time recursive formulation of the firm's dynamic problem:

$$V_\tau(j) = \pi_\tau(j)\Delta + \max_\gamma\{(1-\chi\Delta)(1-r_\tau\Delta)[p_\tau^+(\gamma)\max\{V_{\tau+i}(j+1),0\} + p_\tau^-(\gamma)\max\{V_{\tau+1}(j-1),0\}] - w_\tau i(\gamma, j\Delta_z - \underline{z}_\tau)\Delta\}.$$

At a productivity state j such that $V_{\tau+1}(j-1) > 0$, the firm's value function satisfies:

$$V_\tau(j) = \pi_\tau(j)\Delta + \max_\gamma\{(1-\chi\Delta)(1-r_\tau\Delta)[p_\tau^+(\gamma)V_{\tau+1}(j+1) + p_\tau^-(\gamma)V_{\tau+1}(j-1)] - w_\tau i(\gamma, j\Delta_z - \underline{z}_\tau)\Delta\}.$$

Up to a second-order approximation of $V_{\tau+1}(j\pm 1)$ around $V_\tau(j) = V_t(z)$, we find:

$$V_t(z) = \pi_t(z)\Delta + \max_\gamma\{(1-\chi\Delta)(1-r_t\Delta)[V_t(z) + \dot{V}_t(z)\Delta + \ddot{V}_t(z)\Delta^2/2 + \sigma^2 V_t''(z)\Delta/2 + (2p_t^+(\gamma)-1)V_t'(z)\sigma\sqrt{\Delta} + o(\Delta)] - w_t i(\gamma, z - \underline{z}_\tau)\Delta\}$$

where a single and double dot above a function respectively denote its first and second partial derivatives with respect to time. Subtracting $(1-\chi\Delta)(1-r_t\Delta)V_t(z)$ from both sides and substituting in the expression for $p_t^+(\gamma)$, we find:

$$(r_t\Delta + \chi\Delta + r_t\chi\Delta^2)V_t(z) = \pi_t(z)\Delta + \max_\gamma\{(1-\chi\Delta)(1-r_t\Delta)[\dot{V}_t(z)\Delta + \ddot{V}_t(z)\Delta^2/2 + \gamma V_t'(z)\Delta + V_t''(z)\sigma^2\Delta/2 + o(\Delta)] - w_t i(\gamma, z - \underline{z}_t)\Delta\}.$$

If we divide both sides of this equation by Δ and then take the limit as $\Delta \to 0$, we obtain the following HJBE:

$$(r_t + \chi)V_t(z) = \pi_t(z) + \max_\gamma\{\gamma V_t'(z) - w_t i(\gamma, z - \underline{z}_t)\} + \sigma^2 V_t''(z)/2 + \dot{V}_t(z).$$

With the change of variable $\hat{z} \equiv z - \underline{z}_t$ where $\dot{\underline{z}}_t = g_t$, we can rewrite:

$$(r_t + \chi)V_t(\hat{z}) = \pi_t(\hat{z}) + \max_\gamma\{(\gamma - g_t)V_t'(\hat{z}) - w_t i(\gamma, \hat{z})\} + \sigma^2 V_t''(\hat{z})/2 + \dot{V}_t(\hat{z}).$$

Let us examine the case where the firm's productivity nears the exit threshold. We have assumed that the value of exiting the market is equal to zero, which implies that

$V_t(0) = 0$. Evaluating the firm's value function at the exit threshold then delivers:

$$\begin{aligned}
0 &= \pi_t(0)\Delta \\
&+ \max_{\gamma}\{(1-\chi\Delta)(1-r_t\Delta)[p_t^+(\gamma)\max\{V_t'(0)\sigma\sqrt{\Delta} + V_t''(0)\sigma^2\Delta/2 + o(\Delta), 0\} \\
&+ p_t^-(\gamma)\max\{-V_t'(0)\sigma\sqrt{\Delta} + V_t''(0)\sigma^2\Delta/2 + o(\Delta), 0\}] - w_t i(\gamma, z - \underline{z}_t)\Delta\}.
\end{aligned}$$

If we divide both sides by $\sqrt{\Delta}$ and then take the limit as $\Delta \to 0$, we obtain:

$$\max\{V_t'(0), 0\} + \max\{-V_t'(0), 0\} = 0,$$

which implies the smooth pasting condition $V_t'(0) = 0$. Therefore, as in Stokey (2009), the optimality conditions of the firm's dynamic problem are the value matching, smooth pasting, and first-order conditions:

$$V_t(0) = 0, \quad V_t'(0) = 0 \quad \text{and} \quad V_t'(\hat{z}) = w_t \times \frac{\partial i(\gamma, \hat{z})}{\partial \gamma}$$

together with the "no bubble" condition:

$$\lim_{\hat{z}\to\infty} V_t(\hat{z}) = \lim_{\hat{z}\to\infty} \max_{\{\gamma_s\}_{s\geq t}} \mathbb{E}_{\hat{z}} \left\{ \int_t^\infty e^{-\int_t^s (r_{t'}+\chi)\mathrm{d}t'} [\pi_s(\hat{z}_s) - w_s i(\gamma_s, \hat{z}_s)]\mathrm{d}s \right\}.$$

A.1.6 The Kolmogorov Forward Equation

Here again, we will derive the KFE from its discrete-time analog. But in addition, we also keep track of a firm's age a through the definition $a = i\Delta$ for $i \in \mathbb{N} \cup \{0\}$. Denoting the measure of firms with productivity j and age i at time τ by $m_\tau(j, i)$, we have the following law of motion for all $i > 0$:

$$m_{\tau+1}(j, i) = (1-\chi\Delta)[m_\tau(j+1, i-1)p_\tau^-(\gamma_\tau(j+1)) + m_\tau(j-1, i-1)p_\tau^+(\gamma_\tau(j-1))]$$

together with the "boundary condition" $m_\tau(j, 0) = E_t f_\tau^E(j)$ for all j. Taking a second-order approximation of $m_\tau(j \pm 1, i-1)$ around $m_\tau(j, i) = m_t(z, a)$, we find:

$$\begin{aligned}
m_\tau(j \pm 1, i-1) &= m_t(z, a) \pm \partial_z m_t(z, a)\Delta_z - \partial_a m_t(z, a)\Delta \\
&\quad + [\partial_{zz} m_t(z)\Delta_z^2 + \partial_{aa} m_t(z, a)\Delta^2 + \partial_{za} m_t(z, a)\Delta_z\Delta]/2 + o(\Delta).
\end{aligned}$$

Taking a second-order approximation of $\gamma_\tau(j \pm 1)$ around $\gamma_\tau(j) = \gamma_t(z)$, we find:

$$\gamma_\tau(j \pm 1) = \gamma_t(z) \pm \partial_z \gamma_t(z)\Delta_z + \partial_{zz}\gamma_t(z)\Delta_z^2/2 + o(\Delta).$$

Hence, we can rewrite:

$$\begin{aligned}
&m_\tau(j+1,i-1)p_\tau^-(\gamma_\tau(j+1)) \\
&+m_\tau(j-1,i-1)p_\tau^+(\gamma_\tau(j-1)) = [1-\partial_z\gamma_t(z)\Delta]\{m_t(z,a)-\partial_a m_t(z,a)\Delta \\
&+[\partial_{zz}m_t(z)\Delta_z^2+\partial_{aa}m_t(z,a)\Delta^2+\partial_{za}m_t(z,a)\Delta_z\Delta]/2\}+o(\Delta) \\
&-\partial_z m_t(z,a)\Delta[\gamma_t(z)+\partial_{zz}m_t(z,a)\Delta_z^2/2+o(\Delta)].
\end{aligned}$$

Substituting this in the law of motion for $m_\tau(j,i)$ and taking the limit as $\Delta \to 0$:

$$\dot{m}_t(z,a) = -\partial_z[\gamma_t(z)m_t(z,a)] - \partial_a m_t(z,a) + (\sigma^2/2)\partial_{zz}m_t(z,a) - \chi m_t(z,a).$$

With the change of variable $\hat{z} \equiv z - \underline{z}_t$ where $\dot{\underline{z}}_t = g_t$, we can rewrite:

$$\dot{m}_t(\hat{z},a) = -\partial_{\hat{z}}[(\gamma_t(\hat{z})-g_t)m_t(\hat{z},a)] - \partial_a m_t(\hat{z},a) + (\sigma^2/2)\partial_{\hat{z}\hat{z}}m_t(\hat{z},a) - \chi m_t(\hat{z},a).$$

The boundary conditions in terms of relative productivity are:

$$\lim_{\hat{z}\to 0} m_t(\hat{z},a) = \lim_{\hat{z}\to\infty} m_t(\hat{z},a) = 0 \ \ \forall \hat{a}.$$

The boundary conditions in terms of age are:

$$\lim_{a\to 0} m_t(\hat{z},a) = E_t f_t^E(\hat{z}) \quad \text{and} \quad \lim_{a\to\infty} m_t(\hat{z},a) = 0 \ \ \forall \hat{z}.$$

The total measure of firms is defined as:

$$M_t \equiv \int_0^\infty \int_0^\infty m_t(\hat{z},a)\mathrm{d}a\mathrm{d}\hat{z}.$$

Differentiating this equation with respect to time and substituting in the KFE, we find the law of motion for the measure of firms:

$$\dot{M}_t = [e_t - \chi - (\sigma^2/2)\lim_{\hat{z}\to 0} f_t'(\hat{z})]M_t \quad \text{where} \quad e_t \equiv E_t/M_t$$

and where the joint and marginal probability density functions of relative productivity and age are defined as:

$$\begin{aligned}
f_t(\hat{z},a) &\equiv m_t(\hat{z},a)/M_t, \\
f_t(\hat{z}) &\equiv \lim_{a\to\infty} m_t(\hat{z},a)/M_t, \\
f_t(a) &\equiv \lim_{\hat{z}\to\infty} m_t(\hat{z},a)/M_t.
\end{aligned}$$

Similarly, the joint and marginal cumulative density functions of relative productivity and age are defined as:

$$\begin{aligned}
F_t(\hat{z}, a) &\equiv \textstyle\int_0^{\hat{z}} \int_0^{a} f_t(\hat{z}', a')\mathrm{d}a'\mathrm{d}\hat{z}', \\
F_t(\hat{z}) &\equiv \textstyle\int_0^{\hat{z}} f_t(\hat{z}')\mathrm{d}\hat{z}', \\
F_t(a) &\equiv \textstyle\int_0^{a} f_t(a')\mathrm{d}a'.
\end{aligned}$$

A.1.7 Aggregation

By the definition of the final good's price index, we have:

$$1 = M_t \int_0^\infty p_t(\hat{z})\hat{y}_t(\hat{z})\mathrm{d}F_t(\hat{z}).$$

Substituting in the monopoly pricing condition, we obtain:

$$1 = \varsigma_t M_t \int_0^\infty \mu_t(\hat{z})\hat{y}_t(\hat{z}) \exp(-\hat{z} - \underline{z}_t)\mathrm{d}F_t(\hat{z}).$$

Multiplying both sides by the definition of TFP Z_t:

$$Z_t = \varsigma_t \mathcal{M}_t \quad \text{where} \quad \mathcal{M}_t \equiv \frac{\int_0^\infty \mu_t(\hat{z})\hat{y}_t(\hat{z}) \exp(-\hat{z})\mathrm{d}F_t(\hat{z})}{\int_0^\infty \hat{y}_t(\hat{z}) \exp(-\hat{z})\mathrm{d}F_t(\hat{z})}.$$

The definitions of aggregate physical capital and production labor demand are:

$$K_t = M_t \int_0^\infty k_t(\hat{z})\mathrm{d}F_t(\hat{z}) \quad \text{and} \quad L_t = M_t \int_0^\infty l_t(\hat{z})\mathrm{d}F_t(\hat{z}).$$

Substituting in the firm-level demand functions for physical capital and production labor derived in Appendix A.1.4, we obtain:

$$K_t = \frac{\alpha\varsigma_t Y_t}{(r_t + \delta)Z_t} = \frac{\alpha Y_t}{(r_t + \delta)\mathcal{M}_t} \quad \text{and} \quad L_t = \frac{(1-\alpha)\varsigma_t Y_t}{w_t Z_t} = \frac{(1-\alpha)Y_t}{w_t \mathcal{M}_t}.$$

Solving for aggregate output delivers:

$$Y_t = Z_t K_t^\alpha L_t^{1-\alpha}.$$

Under the transfers of equation (11), we simply have that $\mathcal{M}_t = 1$.

A.1.8 The Optimal Allocation of Resources

In this section, we characterize the optimal allocation of resources. To do so, we consider the problem of a benevolent social planner maximizing social welfare subject to the economy's resource constraints and the law of motion for the density of firms. More specifically, this planner chooses:

$$\{\{k_t(z), l_t(z), \gamma_t(z)\}_{z=\underline{z}_t}^{\infty}, \underline{z}_t, C_t, H_t, E_t\}_{t=0}^{\infty}$$

to maximize the following objective:

$$U_0 = \int_0^{\infty} e^{-\rho t}[\ln(C_t) - v(H_t)]\mathrm{d}t$$

subject to the following constraints for all t and z:

$$\begin{aligned}
K_t &\geq \textstyle\int_{\underline{z}_t}^{\infty} k_t(z)m_t(z)\mathrm{d}z, \\
H_t &\geq \textstyle\int_{\underline{z}_t}^{\infty}[l_t(z) + i(\gamma_t(z), z - \underline{z}_t) + c_O]m_t(z)\mathrm{d}z + c_E E_t, \\
Y_t &\geq \dot{K}_t + \delta K_t + C_t, \\
\dot{m}_t(z) &= -\partial_z[\gamma_t(z)m_t(z)] + (\sigma^2/2)m_t''(z) + E_t f_t^E(z) - \chi m_t(z)
\end{aligned}$$

where $m_t(z)$ denotes the density of firms with productivity equal to z and $f_t^E(z)$ denotes the probability density function from which entrants draw their productivity:[37]

$$f_t^E(z) = T'[1 - F_t(z)]f_t(z).$$

In addition, aggregate output Y_t is implicitly defined as:

$$\int_{\underline{z}_t}^{\infty} \Upsilon(y_t(z)/Y_t)m_t(z)\mathrm{d}z = 1 \quad \text{where} \quad y_t(z) = \exp(z)k_t(z)^{\alpha}l_t(z)^{1-\alpha}.$$

As is, this problem is challenging because the density function of productivity $m_t(z)$ is part of the aggregate state of the economy, and it is an infinite-dimensional object. To circumvent this challenge, we follow the approach in Lashkari (2023) and discretize the time dimension of the problem. With the discretization scheme presented in Appendix A.1, we can restate the planner's problem as controlling:

$$\mathbf{u}_\tau = \{\{k_\tau(j), l_\tau(j), \gamma_\tau(j)\}_{j=\underline{j}_\tau}^{\infty}, \underline{j}_\tau, C_\tau, H_\tau, E_\tau\}.$$

[37] With a slight abuse of notation, we use the same notation to identify the (probability) density function of both absolute and relative productivity.

to maximize the following objective:

$$U_0 = \sum_{\tau=0}^{\infty} e^{-\rho\tau\Delta}[\ln(C_\tau) - v(H_\tau)]\Delta$$

subject to the following constraints for all $\tau \in \mathbb{N} \cup \{0\}$ and $j \geq \underline{j}_\tau \in \mathbb{Z}$:

$$\begin{aligned}
K_\tau &\geq \textstyle\sum_{j=\underline{j}_\tau}^{\infty} k_\tau(j) m_\tau(j)\Delta_z, \\
H_\tau &\geq \textstyle\sum_{j=\underline{j}_\tau}^{\infty} [l_\tau(j) + i(\gamma_\tau(j), (j - \underline{j}_\tau)\Delta_z) + c_O] m_\tau(j)\Delta_z + c_E E_\tau, \\
K_{\tau+1} &\leq (Y_\tau - C_\tau)\Delta + (1 - \delta\Delta)K_\tau, \\
m_{\tau+1}(j) &= (1 - \chi\Delta)[m_\tau(j+1)p_\tau^-(\gamma_\tau(j+1)) + m_\tau(j-1)p_\tau^+(\gamma_\tau(j-1))] + E_\tau f_\tau^E(j)\Delta.
\end{aligned}$$

Collecting the state variables in the vector $\mathbf{x}_\tau = \{K_\tau, \{m_\tau(j)\}_{j=\underline{j}_\tau}^{\infty}\}$, we can express the dynamic and static constraints in vector form as:

$$\mathbf{x}_{\tau+1} = \mathbf{f}_\tau^D(\mathbf{u}_\tau, \mathbf{x}_\tau) \quad \text{and} \quad \mathbf{f}_\tau^S(\mathbf{u}_\tau, \mathbf{x}_\tau) = \mathbf{0}, \quad \forall \tau \in \mathbb{N} \cup \{0\}.$$

We denote the Lagrange multipliers for these dynamic and static constraints by:

$$\boldsymbol{\nu}_{\tau+1}^D = \{1, \{\nu_{\tau+1}^m(j)\Delta_z\}_{j=\underline{j}_{\tau+1}}^{\infty}\} \quad \text{and} \quad \boldsymbol{\nu}_\tau^S = \{\nu_\tau^K, \nu_\tau^H\}, \quad \forall \tau \in \mathbb{N} \cup \{0\}.$$

Here again, aggregate output Y_τ is implicitly defined as:

$$\sum_{j=\underline{j}_\tau}^{\infty} \Upsilon(y_\tau(j)/Y_\tau) m_\tau(j)\Delta_z = 1 \quad \text{where} \quad y_\tau(j) = \exp(j\Delta_z) k_\tau(j)^\alpha l_\tau(j)^{1-\alpha}.$$

Hence, we can define the planner's current-value Hamiltonian as:

$$\mathcal{H}_\tau(\mathbf{u}_\tau, \mathbf{x}_\tau, \boldsymbol{\nu}_\tau^D, \boldsymbol{\nu}_\tau^S) = [\ln(C_\tau) - v(H_\tau)]\Delta + \nu_{\tau+1}{\boldsymbol{\nu}_{\tau+1}^D}^\top \mathbf{f}_t^D(\mathbf{u}_\tau, \mathbf{x}_\tau) + \Delta\nu_{\tau+1}{\boldsymbol{\nu}_\tau^S}^\top \mathbf{f}_t^S(\mathbf{u}_\tau, \mathbf{x}_\tau).$$

The first-order conditions with respect to $k_\tau(j)$ and $l_\tau(j)$ are given by:

$$\begin{aligned}
\Delta\nu_{\tau+1} \times \frac{\partial Y_\tau}{\partial y_\tau(j)} \times \frac{\partial y_\tau(j)}{\partial k_\tau(j)} - \Delta\nu_\tau^K \nu_{\tau+1} m_\tau(j)\Delta_z = 0, \\
\Delta\nu_{\tau+1} \times \frac{\partial Y_\tau}{\partial y_\tau(j)} \times \frac{\partial y_\tau(j)}{\partial l_\tau(j)} - \Delta\nu_\tau^H \nu_{\tau+1} m_\tau(j)\Delta_z = 0.
\end{aligned}$$

To obtain an expression for $\frac{\partial Y_\tau}{\partial y_\tau(j)}$, we can differentiate the implicit definition of aggregate output and use the (discrete) definition of the demand index to find:

$$\frac{\partial Y_\tau}{\partial y_\tau(j)} = \Upsilon'(y_\tau(j)/Y_\tau)D_\tau m_\tau(j)\Delta_z.$$

Substituting this in the previous two equations and simplifying, we find expressions for the optimal allocations of production labor and physical capital across firms:

$$\begin{aligned} k_\tau(j) &= \alpha\Upsilon'(y_\tau(j)/Y_\tau)D_\tau y_\tau(j)/\nu^K_\tau, \\ l_\tau(j) &= (1-\alpha)\Upsilon'(y_\tau(j)/Y_\tau)D_\tau y_\tau(j)/\nu^H_\tau. \end{aligned}$$

The first-order conditions with respect to $\gamma_\tau(j)$ imply:

$$i'(\gamma_\tau(j), j\Delta_z) = (1-\chi\Delta)[\nu^m_{\tau+1}(j+1) - \nu^m_{\tau+1}(j-1)]/(2\Delta_z\nu^H_\tau).$$

The first-order conditions with respect to C_τ, H_τ, E_τ and K_τ imply:

$$\begin{aligned} C_\tau &= \nu^{-1}_{\tau+1}, \\ H_\tau &= v'^{-1}(\nu^H_\tau \nu_{\tau+1}), \\ \nu^H_\tau c_E &= \textstyle\sum_{j=\underline{j}_\tau}^{\infty} \nu^m_{\tau+1}(j) f^E_\tau(j)\Delta_z, \\ e^{\rho\Delta}\nu_\tau &= \nu_{\tau+1}[(\nu^K_\tau - \delta)\Delta + 1]. \end{aligned}$$

Finally, the first-order conditions with respect to $m_\tau(j)$ for $\nu^m_{\tau+1}(j-1) > 0$ imply:

$$\begin{aligned} e^{\rho\Delta}\nu^m_\tau(j)\nu_\tau/(\Delta\nu_{\tau+1}) = \Delta_z^{-1} \times \frac{\partial Y_\tau}{\partial m_\tau(j)} &- \nu^K_\tau k_\tau(j) - \nu^H_\tau[l_\tau(j) + i(\gamma_\tau(j), (j-\underline{j}_\tau)\Delta_z) + c_O] \\ &+ (1-\chi\Delta)[\nu^m_{\tau+1}(j+1)p^+_\tau(\gamma_\tau(j)) + \nu^m_{\tau+1}(j-1)p^-_\tau(\gamma_\tau(j))]/\Delta \\ &+ E_\tau \textstyle\sum_{j'=\underline{j}_\tau}^{\infty} \nu^m_{\tau+1}(j') \times \frac{\partial f^E_\tau(j')}{\partial m_\tau(j)}. \end{aligned}$$

Once again, to obtain an expression for $\frac{\partial Y_\tau}{\partial m_\tau(j)}$, we can differentiate the implicit definition of aggregate output to find:

$$\frac{\partial Y_\tau}{\partial m_\tau(j)} = \Upsilon(y_\tau(j)/Y_\tau)Y_\tau D_\tau \Delta_z.$$

To obtain an expression for $\frac{\partial f^E_\tau(j')}{\partial m_\tau(j)}$, let us first state the discrete definition of $f^E_\tau(j')$:

$$f^E_\tau(j') = T'[1 - M_\tau^{-1} \textstyle\sum_{k=\underline{j}_\tau}^{j'} m_\tau(k)\Delta_z] m_\tau(j') M_\tau^{-1} \quad \text{where} \quad M_\tau = \textstyle\sum_{k=\underline{j}_\tau}^{\infty} m_\tau(k)\Delta_z.$$

Hence, we find that:

$$\begin{aligned} \frac{\partial f^E_\tau(j')}{\partial m_\tau(j)} &= T''[1 - F_\tau(j')][F_\tau(j') - \mathbb{1}_{\{j' \geq j\}}] f_\tau(j') M_\tau^{-1} \Delta_z \\ &\quad + T'[1 - F_\tau(j')][\mathbb{1}_{\{j'=j\}} - f_\tau(j')\Delta_z] M_\tau^{-1}. \end{aligned}$$

Substituting these results in the first-order conditions for $m_\tau(j)$, we have:

$$\begin{aligned} (\nu^K_\tau - \delta + 1/\Delta)\nu^m_\tau(j) &= [\Upsilon(\hat{y}_\tau(j)) - \Upsilon'(\hat{y}_\tau(j))\hat{y}_\tau(j)] Y_\tau D_\tau - \nu^H_\tau [i(\gamma_\tau(j), (j - \underline{j}_\tau)\Delta_z) + c_O] \\ &\quad + (1 - \chi\Delta)[\nu^m_{\tau+1}(j+1) p^+_\tau(\gamma_\tau(j)) + \nu^m_{\tau+1}(j-1) p^-_\tau(\gamma_\tau(j))]/\Delta \\ &\quad + e_\tau \textstyle\sum_{j'=\underline{j}_\tau}^{\infty} \nu^m_{\tau+1}(j') T''[1 - F_\tau(j')] F_\tau(j') f_\tau(j') \Delta_z \\ &\quad - e_\tau \textstyle\sum_{j'=j}^{\infty} \nu^m_{\tau+1}(j') T''[1 - F_\tau(j')] f_\tau(j') \Delta_z \\ &\quad + e_\tau [\nu^m_{\tau+1}(j) f^E_\tau(j)/f_\tau(j) - \nu^H_\tau c_E]. \end{aligned}$$

To obtain an expression for ν^K_τ, we can use the market clearing condition for physical capital, from which we find:

$$\nu^K_\tau = \alpha Y_\tau / K_\tau.$$

Similarly, defining L_τ as aggregate production labor, we find that:

$$\nu^H_\tau = (1 - \alpha) Y_\tau / L_\tau.$$

Substituting the optimality conditions for production labor and physical capital in the firm's production function and integrating over all firms, we also find that:

$$Y_\tau = Z_\tau K^\alpha_\tau L^{1-\alpha}_\tau \quad \text{where} \quad \left(\textstyle\sum_{j=\underline{j}_\tau}^{\infty} \hat{y}_\tau(j) \exp(-j\Delta_z) m_\tau(j) \Delta_z \right)^{-1}.$$

From this point on, we take the limit of the previous equations as $\Delta \to 0$. Let us begin with the first-order conditions and law of motion for K_τ, from which we find:

$$\dot{C}_t = (\alpha Y_t / K_t - \delta - \rho) C_t \quad \text{and} \quad \dot{K}_t = Y_t - \delta K_t - C_t.$$

Let us now define the firm's social value function $V_t(z) \equiv \nu^m_\tau(j)$. We can take a second-order approximation of $\nu^m_{\tau+1}(j \pm 1)$ around z and t to find:

$$\nu^m_{\tau+1}(j \pm 1) = V_t(z) \pm V'_t(z)\Delta_z + \dot{V}_t(z)\Delta + [V''_t(z)\Delta_z^2 + \ddot{V}_t(z)\Delta^2 \pm \dot{V}'_t(z)\Delta_z\Delta]/2 + o(\Delta).$$

Hence, we can rewrite:

$$\nu^m_{\tau+1}(j+1) - \nu^m_{\tau+1}(j-1) = 2V'_t(z)\Delta_z + \dot{V}'_t(z)\Delta_z\Delta + o(\Delta).$$

Substituting this in the optimality condition for the controlled drift $\gamma_t(z) \equiv \gamma_\tau(j)$ and taking the limit as $\Delta \to 0$:

$$i'(\gamma_t(z), z - \underline{z}_t) = V'_t(z)/\nu^H_t.$$

Similarly, we can rewrite:

$$\begin{aligned}\nu^m_{\tau+1}(j-1)p^-_\tau(\gamma_\tau(j)) + \nu^m_{\tau+1}(j+1)p^+_\tau(\gamma_\tau(j)) = V_t(z) + \dot{V}_t(z)\Delta \\ + [V''_t(z)\sigma^2\Delta + \ddot{V}_t(z)\Delta^2]/2 + o(\Delta) \\ + [V'_t(z)\Delta + \dot{V}'_t(z)\Delta^2/2 + o(\Delta)/(2\sigma)]\gamma_t(z).\end{aligned}$$

Substituting this in the first-order condition for $m_\tau(j)$ and taking the limit as $\Delta \to 0$:

$$\begin{aligned}(\nu^K_t - \delta + \chi)V_t(z) = {} & [\Upsilon(\hat{y}_t(z)) - \Upsilon'(\hat{y}_t(z))\hat{y}_t(z)]Y_t D_t - \nu^H_t[i(\gamma_t(z), z - \underline{z}_t) + c_O] \\ & + \gamma_t(z)V'_t(z) + (\sigma^2/2)V''_t(z) + \dot{V}_t(z) \\ & + e_t \textstyle\int_{\underline{z}_t}^\infty V_t(x)T''[1 - F_t(x)]F_t(x)\mathrm{d}F_t(x) \\ & - e_t \textstyle\int_z^\infty V_t(x)T''[1 - F_t(x)]\mathrm{d}F_t(x) \\ & + e_t[V_t(z)f^E_t(z)/f_t(z) - \nu^H_t c_E].\end{aligned}$$

Integrating by parts and rearranging, we can rewrite:

$$\begin{aligned}(\nu^K_t - \delta + \chi)V_t(z) = {} & [\Upsilon(\hat{y}_t(z)) - \Upsilon'(\hat{y}_t(z))\hat{y}_t(z)]Y_t D_t - \nu^H_t[i(\gamma_t(z), z - \underline{z}_t) + c_O] \\ & + \gamma_t(z)V'_t(z) + (\sigma^2/2)V''_t(z) + \dot{V}_t(z) \\ & + e_t \textstyle\int_{\underline{z}_t}^z V'_t(x)F_t(x)f_t(x)^{-1}\mathrm{d}F^E_t(x) \\ & - e_t \textstyle\int_z^\infty V'_t(x)[1 - F_t(x)]f_t(x)^{-1}\mathrm{d}F^E_t(x).\end{aligned}$$

The remaining optimality conditions are given by:

$$H_t = v'^{-1}(\nu^H_t/C_t) \quad \text{and} \quad \nu^H_t c_E = \int_{\underline{z}_t}^\infty V_t(z)\mathrm{d}F^E_t(z)$$

together with the value matching and smooth pasting conditions, which determine the socially optimal exit threshold $\underline{z}_t$:

$$V_t(\underline{z}_t) = V_t'(\underline{z}_t) = 0.$$

Here, since firms have no scrap value, it must be that the optimal social value of a firm that exits is equal to zero. Otherwise, the planner would either increase or decrease the exit threshold to eliminate firms with a negative social value or take in those with a positive one. The smooth pasting condition for the firm's social value can be derived as it was derived for the firm's private value in Appendix A.1.5. Let us now define the socially optimal productivity density function $m_t(z) \equiv m_\tau(j)$. We can take a second-order approximation of $m_\tau(j \pm 1)$ around z to find:

$$m_\tau(j \pm 1) = m_t(z) \pm m_t'(z)\Delta_z + m_t''(z)\Delta_z^2/2 + o(\Delta).$$

Similarly, we can take a second-order approximation of $\gamma_\tau(j \pm 1)$ around z to find:

$$\gamma_\tau(j \pm 1) = \gamma_t(z) \pm \gamma_t'(z)\Delta_z + \gamma_t''(z)\Delta_z^2/2 + o(\Delta).$$

Substituting these in the law of motion for $m_\tau(j)$ and taking the limit as $\Delta \to 0$:

$$\dot{m}_t(z) = -\partial_z[\gamma_t(z)m_t(z)] + (\sigma^2/2)m_t''(z) + E_t f_t^E(z) - \chi m_t(z).$$

Integrating over all firms, we find:

$$\dot{M}_t = [e_t - \chi - (\sigma^2/2)f_t'(\underline{z}_t)]M_t.$$

Decentralizing the Optimal Allocation of Resources

To decentralize the optimal allocation of resources, a government may use two firm-specific fiscal instruments: production and overhead labor taxes and subsidies. The optimal relative quantity produced by the firm is given by:

$$\hat{y}_t(\hat{z}) = \Upsilon'^{-1}[\exp(-\hat{z} - \underline{z}_t)Z_t D_t^{-1}].$$

We want to solve for the transfer schedule $T_t(\hat{y}_t(\hat{z}))$ that induces each firm to produce at that optimal level by maximizing profits:

$$\pi_t(\hat{z}) = \max_{\hat{y}_t(\hat{z})}\{[\Upsilon'(\hat{y}_t(\hat{z}))D_t - \varsigma_t \exp(-\hat{z} - \underline{z}_t)]\hat{y}_t(\hat{z})Y_t + T_t(\hat{y}_t(\hat{z}))\} - w_t c_O.$$

The private first-order condition of this problem is:

$$[\Upsilon''(\hat{y}_t(\hat{z}))\hat{y}_t(\hat{z}) + \Upsilon'(\hat{y}_t(\hat{z}))]D_t + T_t'(\hat{y}_t(\hat{z}))/Y_t = \varsigma_t \exp(-\hat{z} - \underline{z}_t).$$

Substituting in the optimal level of production, we find:

$$T_t'(\hat{y}_t(\hat{z})) = -\Upsilon''(\hat{y}_t(\hat{z}))\hat{y}_t(\hat{z})Y_t D_t.$$

This is an ordinary differential equation, which can be integrated by parts with the initial condition $T_t(0) = 0$:

$$T_t(\hat{y}_t(\hat{z})) = -Y_t D_t \int \Upsilon''(\hat{y}')\hat{y}'\mathrm{d}\hat{y}' + \mathcal{C}_0 = [\Upsilon(\hat{y}_t(\hat{z})) - \Upsilon'(\hat{y}_t(\hat{z}))\hat{y}_t(\hat{z})]Y_t D_t.$$

Here, $\mathcal{C}_0$ is the constant of integration, equal to zero given the initial condition. Then, for the firm's social and value functions to coincide, the overhead labor expenditures of a firm with relative productivity $\hat{z}$ must be "taxed" at rate:

$$\tau_t^O(\hat{z}) = e_t[\textstyle\int_{\hat{z}}^{\infty} V_t'(x)[1 - F_t(x)]f_t(x)^{-1}\mathrm{d}F_t^E(x) - \int_0^{\hat{z}} V_t'(x)F_t(x)f_t(x)^{-1}\mathrm{d}F_t^E(x)]/(w_t c_O).$$

With these instruments, it is straightforward to show that the decentralized equilibrium allocation must coincide with the optimal allocation of resources.

A.1.9 Alternative Transfer Schedules

In this subsection, we derive the transfers presented in Section 5.4. Let us first rewrite the firm's static problem of choosing a scale at which to produce in order to maximize profits given a demand schedule:

$$\pi_t(\hat{z}) = \max_{\hat{y}_t(\hat{z})}\{[\Upsilon'(\hat{y}_t(\hat{z}))D_t - \varsigma_t \exp(-\hat{z} - \underline{z}_t)]\hat{y}_t(\hat{z})Y_t + T_t(\hat{y}_t(\hat{z}))\} - w_t c_O.$$

The first-order condition of this problem is:

$$[\Upsilon''(\hat{y}_t(\hat{z}))\hat{y}_t(\hat{z}) + \Upsilon'(\hat{y}_t(\hat{z}))]D_t + T_t'(\hat{y}_t(\hat{z}))/Y_t = \varsigma_t \exp(-\hat{z} - \underline{z}_t).$$

Let us now momentarily abandon the $\hat{z}$ notation for clarity. If the transfers are intended to induce firms to charge a markup of $\mathcal{M}_t$ above marginal cost, we can substitute the constraint $\Upsilon'(\hat{y})D_t = \mathcal{M}_t\varsigma_t \exp(-\hat{z} - \underline{z}_t)$ in the first-order condition to find:

$$T_t'(\hat{y}) = [(\mathcal{M}_t^{-1} - 1)\Upsilon'(\hat{y}) - \Upsilon''(\hat{y})\hat{y}]Y_t D_t.$$

Integrating by parts with the initial condition $T_t(0) = 0$, we find:

$$T_t(\hat{y}) = [\mathcal{M}_t^{-1}\Upsilon(\hat{y}) - \Upsilon'(\hat{y})\hat{y}]Y_t D_t.$$

If the transfers are instead intended to induce firms to charge a markup of $\mu(\hat{y})/\mathcal{M}_t$ above marginal cost, we substitute the constraint $\Upsilon'(\hat{y})D_t = \mu(\hat{y})\varsigma_t \exp(-\hat{z} - \underline{z}_t)/\mathcal{M}_t$ in the first-order condition to find:[38]

$$T_t'(\hat{y}) = (\mathcal{M}_t - 1)[\Upsilon'(\hat{y}) + \Upsilon''(\hat{y})\hat{y}]Y_t D_t.$$

Integrating by parts with the same initial condition, we find:

$$T_t(\hat{y}) = (\mathcal{M}_t - 1)\Upsilon'(\hat{y})\hat{y}Y_t D_t.$$

A.1.10 Relative Prices and Demand Under Klenow and Willis (2016)

From the final sector's problem and the firm's static problem, we have the following relative demand function and monopoly pricing condition for varieties:

$$\hat{y}_t(\hat{z}) = \Upsilon'^{-1}[p_t(\hat{z})/D_t] \quad \text{and} \quad p_t(\hat{z}) = \frac{\mu_t(\hat{z})\varsigma_t}{\exp(\hat{z} + \overline{z}_t)}$$

where we have the following two definitions:

$$\mu_t(\hat{z}) \equiv \frac{\vartheta_t(\hat{z})}{\vartheta_t(\hat{z}) - 1} \quad \text{and} \quad \vartheta_t(\hat{z}) \equiv -\frac{\Upsilon'(\hat{y}_t(\hat{z}))}{\Upsilon''(\hat{y}_t(\hat{z}))\hat{y}_t(\hat{z})}.$$

Denoting the choke price by $\overline{p}_t \equiv \Upsilon'(0)D_t$, we can define a variety's relative price as $\hat{p}_t(\hat{z}) \equiv p_t(\hat{z})/\overline{p}_t$. This allows us to rewrite the relative demand function as:

$$\hat{y}(\hat{p}) = \Upsilon'^{-1}[\hat{p}\Upsilon'(0)]$$

which is now a stationary function of the corresponding variety's relative price. Using the Klenow and Willis (2016) specification of the Kimball (1995) aggregator, we can rewrite the monopoly pricing condition (relative to the choke price) as:

$$\hat{p}_t(\hat{z}) = \frac{\varsigma_t/\overline{p}_t}{[1 + (\epsilon/\theta)\ln(\hat{p}_t(\hat{z}))]\exp(\hat{z} + \underline{z}_t)}$$

[38] Here, $\mu(\hat{y}) = \Upsilon'(\hat{y})/[\Upsilon'(\hat{y}) + \Upsilon''(\hat{y})\hat{y}]$.

Note that we can rearrange this equation to obtain:

$$\exp\{(\theta/\epsilon)[\varsigma_t[\hat{p}_t(\hat{z})\overline{p}_t \exp(\hat{z}+\underline{z}_t)]^{-1} - 1]\}\hat{p}_t(\hat{z})^{-1} = 1$$

Multiplying both sides by $(\theta/\epsilon)\exp(\theta/\epsilon - \hat{z} - \underline{z}_t)\varsigma_t/\overline{p}_t$, we have:

$$W^{-1}\{(\theta/\epsilon)\varsigma_t[\hat{p}_t(\hat{z})\overline{p}_t \exp(\hat{z}+\underline{z}_t)]^{-1}\} = (\theta/\epsilon)\exp(\theta/\epsilon - \hat{z} - \underline{z}_t)\varsigma_t/\overline{p}_t$$

where W is the Lambert W-function defined by the inverse mapping $W^{-1}(x) = xe^x$. Finally, solving for $\hat{p}_t(\hat{z})$ delivers:

$$\hat{p}_t(\hat{z}) = \frac{(\theta/\epsilon)\exp(-\hat{z} - \underline{z}_t)\varsigma_t/\overline{p}_t}{W[(\theta/\epsilon)\exp(\theta/\epsilon - \hat{z} - \underline{z}_t)\varsigma_t/\overline{p}_t]}.$$

The Lambert W-function has two useful properties:

1. $W(x) > 0$ for all $x > 0$ and $\lim_{x\to 0} W(x) = 0$.
2. $W'(x) = [x + e^{W(x)}]^{-1}$ for all $x > 0$ and $\lim_{x\to 0} W'(x) = 1$.

We can use those properties to study the limiting behavior of a variety's relative price and demand for the most productive firms:

Proposition 6. *At any given point in time, a variety's relative price and demand asymptote to constants as $\hat{z} \to \infty$:*

$$\lim_{\hat{z}\to\infty} \hat{p}_t(\hat{z}) = \exp(-\theta/\epsilon) \quad \textit{and} \quad \lim_{\hat{z}\to\infty} \hat{y}(\hat{z}) = \theta^{\theta/\epsilon}.$$

Let us now consider the transfer schedule of equation (13):

$$T_t(\hat{y}) = [\tau_0\Upsilon(\hat{y}) + \tau_1\Upsilon'(\hat{y})\hat{y}]Y_t D_t.$$

In particular, we will explore the three cases derived in the previous subsection of this appendix: $(\tau_0, \tau_1) \in \{(1, -1), (0, \mathcal{M}_t - 1), (\mathcal{M}_t^{-1}, -1)\}$. In the first case, the firm's optimally chosen relative price is:

$$\hat{p}_t(\hat{z}) = \frac{\varsigma_t/\overline{p}_t}{\exp(\hat{z}+\underline{z}_t)}$$

which asymptotes to zero as $\hat{z} \to \infty$. Therefore, the firm's relative demand asymptotes to infinity as $\hat{z} \to \infty$. However, firm-level profits remain finite at any given point in time as $\hat{z} \to \infty$. Indeed, we have that:

$$\lim_{\hat{z}\to\infty} \pi_t(\hat{z}) = \left[1 + (\theta - 1)\exp(1/\epsilon)\epsilon^{\theta/\epsilon - 1}\Gamma\left(\frac{\theta}{\epsilon}, \frac{1}{\epsilon}\right)\right] Y_t D_t - w_t c_O.$$

In the second case, the firm's optimally chosen relative price is:

$$\hat{p}_t(\hat{z}) = \frac{\mu(\hat{z})\varsigma_t/\overline{p}_t}{\mathcal{M}_t \exp(\hat{z} + \underline{z}_t)}.$$

With the same derivation as above, we can express that relative price function as:

$$\hat{p}_t(\hat{z}) = \frac{(\theta/\epsilon)\exp(-\hat{z} - \underline{z}_t)\varsigma_t/(\overline{p}_t\mathcal{M}_t)}{W[(\theta/\epsilon)\exp(\theta/\epsilon - \hat{z} - \underline{z}_t)\varsigma_t/(\overline{p}_t\mathcal{M}_t)]}$$

which holds the same asymptotic properties as in Proposition 6. In the third case, the firm's optimally chosen relative price is:

$$\hat{p}_t(\hat{z}) = \frac{\mathcal{M}_t\varsigma_t/\overline{p}_t}{\exp(\hat{z} + \underline{z}_t)}$$

which asymptotes to zero as $\hat{z} \to \infty$. Therefore, the firm's relative demand asymptotes to infinity as $\hat{z} \to \infty$. However, firm-level profits remain finite at any given point in time as $\hat{z} \to \infty$. Indeed, we have that:

$$\lim_{\hat{z}\to\infty} \pi_t(\hat{z}) = \left[1 + (\theta - 1)\exp(1/\epsilon)\epsilon^{\theta/\epsilon - 1}\Gamma\left(\frac{\theta}{\epsilon}, \frac{1}{\epsilon}\right)\right] Y_t D_t/\mathcal{M}_t - w_t c_O.$$

A.1.11 Computing a Balanced Growth Path Equilibrium Allocation

Consider a balanced growth path equilibrium allocation as defined in Definition 2 with initial condition $\underline{z}_0 = 0$. Using the household's Euler equation together with the restriction that the value of a firm must grow at the same rate as aggregate consumption, we have the following expression for the firm's HJBE in the continuation region:

$$(\rho + \chi)V_t(\hat{z}) = \pi_t(\hat{z}) + \max_{\gamma}\{(\gamma - g)V_t'(\hat{z}) - w_t i(\gamma, \hat{z})\} + \sigma^2 V_t''(\hat{z})/2.$$

Here, g still denotes the instantaneous growth rate of TFP, which is constant on a balanced growth path. Profits and the wage rate both grow at constant rate $g/(1-\alpha)$ such that we can define the stationary function $V(\hat{z}) \equiv V_t(\hat{z})\exp[-gt/(1-\alpha)]$ and rewrite the firm's HJBE as:

$$(\rho+\chi)V(\hat{z}) = \pi_0(\hat{z}) + \max_{\gamma}\{(\gamma-g)V'(\hat{z}) - w_0 i(\gamma,\hat{z})\} + \sigma^2 V''(\hat{z})/2$$

where X_0 denotes the detrended value of a variable X_t that grows at a constant rate on a balanced growth path. As in Appendix A.1.5, the firm's dynamic problem delivers the value matching, smooth pasting and first-order conditions:

$$V(0)=0, \quad V'(0)=0 \quad \text{and} \quad \gamma(\hat{z}) = \left[\frac{V'(\hat{z})}{w_0\exp(c_I + (1+\zeta)\hat{z})}\right]^{\frac{1}{\zeta}}.$$

Substituting this first-order condition in the firm's stationary HJBE, we obtain a second-order nonlinear ordinary differential equation:

$$(\rho+\chi)V(\hat{z}) = \pi_0(\hat{z}) + \frac{\zeta\gamma(\hat{z})V'(\hat{z})}{1+\zeta} - gV'(\hat{z}) + \frac{\sigma^2 V''(\hat{z})}{2}$$

in which the stationary profit function is given by:

$$\pi_0(\hat{z}) = \begin{cases} \hat{p}(\hat{z})\hat{y}(\hat{z})^{1+\epsilon/\theta}\Upsilon'(0)Y_0 D/\theta - w_0 c_O & \text{Pre-policy,} \\ [\Upsilon(\hat{y}(\hat{z})) - \Upsilon'(\hat{y}(\hat{z}))\hat{y}(\hat{z})]Y_0 D - w_0 c_O & \text{Post-policy,} \end{cases}$$

and the relative demand and relative price functions are in turn given by:

$$\hat{p}(\hat{z}) = \begin{cases} \frac{(\theta/\epsilon)\exp(-\hat{z}-\underline{z}_0)\varsigma_0/\overline{p}}{W[(\theta/\epsilon)\exp(\theta/\epsilon-\hat{z}-\underline{z}_0)\varsigma_0/\overline{p}]} & \text{Pre-policy,} \\ \exp(-\hat{z}-\underline{z}_0)\varsigma_0/\overline{p} & \text{Post-policy,} \end{cases}$$
$$\hat{y}(\hat{z}) = [-\epsilon\max\{\ln(\hat{p}(\hat{z})),0\}]^{\theta/\epsilon}.$$

It is straightforward to verify that Y_t must grow at the same rate as w_t, and ς_t must grow at the same rate as $\underline{z}_t$ on a balanced growth path. From Proposition 5, we know that the stationary value function asymptotes to an endogenous constant $\overline{V}$:

$$\overline{V} = \frac{w_0 c_O(1-x)}{(\rho+\chi)x} \quad \text{where} \quad x \equiv \frac{w_0 c_O}{\overline{\pi} Y_0 D} \in (0,1)$$

where the constant $\overline{\pi}$ is given by:

$$\overline{\pi} = \begin{cases} (\theta-1)\exp[(1-\theta)/\epsilon]\theta^{\theta/\epsilon-1} & \text{Pre-policy,} \\ 1+(\theta-1)\exp(1/\epsilon)\epsilon^{\theta/\epsilon-1}\Gamma\left(\frac{\theta}{\epsilon},\frac{1}{\epsilon}\right) & \text{Post-policy.} \end{cases}$$

Therefore, let us respectively define the firm's *normalized* value and profit functions as $\hat{V}(\hat{z}) \equiv V(\hat{z})/\overline{V}$ and $\hat{\pi}_0(\hat{z}) \equiv \pi_0(\hat{z})/\overline{V}$ such that we can rewrite the firm's HJBE as:

$$(\rho + \chi)\hat{V}(\hat{z}) = \hat{\pi}_0(\hat{z}) + \frac{\zeta\gamma(\hat{z})\hat{V}'(\hat{z})}{1+\zeta} - g\hat{V}'(\hat{z}) + \frac{\sigma^2\hat{V}''(\hat{z})}{2}$$

with boundary conditions $\hat{V}(0) = 0$ and $\lim_{\hat{z}\to\infty} \hat{V}(\hat{z}) = 1$, and where the first-order condition of the firm's dynamic problem becomes:

$$\gamma(\hat{z}) = \left[\frac{c_O(1-x)\hat{V}'(\hat{z})}{(\rho+\chi)\exp(c_I + (1+\zeta)\hat{z})x}\right]^{\frac{1}{\zeta}}.$$

Since the measure of varieties is constant on a balanced growth path, the entry rate must be equated to the sum of the exogenous and endogenous exit rates:

$$e = \chi + (\sigma^2/2)F''(0).$$

The stationary Kolmogorov forward equation of the CDF $F(\hat{z})$ therefore delivers the following second-order nonlinear ordinary differential equation:

$$0 = -[\gamma(\hat{z}) - g]F'(\hat{z}) + (\sigma^2/2)\{F''(\hat{z}) - F''(0)[1-F(\hat{z})]\} + e[F^E(\hat{z}) - F(\hat{z})]$$

with boundary conditions $F(0) = 0$ and $\lim_{\hat{z}\to\infty} F(\hat{z}) = 1$.

To solve for the two second-order nonlinear ordinary differential equations above (the firm's HJBE and the KFE), we need equilibrium conditions that pin down the aggregate variables they depend on. More precisely, given an initial condition $\underline{z}_0$, we are looking for eleven equations to identify the unknowns $\{Y_0, C_0, Z_0, w_0, \varsigma_0, r, \overline{p}, D, M, H, g\}$:

1. The Kimball (1995) aggregation condition:

$$\int_0^\infty \Upsilon(\hat{y}(\hat{z}))\mathrm{d}M(\hat{z}) = 1.$$

2. The demand index:

$$D = \left(\int_0^\infty \Upsilon'(\hat{y}(\hat{z}))\hat{y}(\hat{z})\mathrm{d}M(\hat{z})\right)^{-1}.$$

3. The free-entry condition:

$$(1-x)\int_0^\infty \hat{V}(\hat{z})\mathrm{d}F^E(\hat{z}) = (c_E/c_O)(\rho+\chi)x.$$

4. The producer price index:

$$\varsigma_0 = \left(\frac{r+\delta}{\alpha}\right)^{\alpha} \left(\frac{w_0}{1-\alpha}\right)^{1-\alpha}.$$

5. The household's Euler equation:

$$\frac{g}{1-\alpha} = r - \rho.$$

6. The household's static first-order condition:

$$\beta H^{\eta} C_0 = w_0.$$

7. The final good market clearing condition:

$$C_0 = Y_0 \left[1 - \frac{\alpha\varsigma_0(r+\delta-\rho)}{Z_0(r+\delta)}\right].$$

8. The labor market clearing condition:

$$\frac{(1-\alpha)\varsigma_0 Y_0}{w_0 Z_0} + \frac{c_O(1-x)\int_0^{\infty}\gamma(\hat{z})\hat{V}'(\hat{z})\mathrm{d}M(\hat{z})}{(1+\zeta)(\rho+\chi)x} + c_E e M + c_O M = H.$$

9. Total-factor productivity:

$$Z_0 = \exp(\underline{z}_0)\left(\int_0^{\infty}\hat{y}(\hat{z})\exp(-\hat{z})\mathrm{d}M(\hat{z})\right)^{-1}.$$

10. The choke price:

$$\overline{p} = \Upsilon'(0)D.$$

11. The value matching condition.

A.2 Proofs

Proof of Propositions 1 and 2. The monopolist's profits are given by:

$$\pi(z) = \frac{p(z)y(p(z))}{\vartheta(p(z))}.$$

The elasticity of profits with respect to productivity is given by:

$$\frac{\partial \ln(\pi(z))}{\partial z} = [\vartheta(p(z)) + \varepsilon(p(z)) - 1]\varrho(z) \quad \text{where} \quad \varrho(z) \equiv -\frac{\partial \ln(p(z))}{\partial z}$$

where $\varrho(z)$ denotes the productivity "pass-through". Using the expression provided in Section 2 for the profit-maximizing price, we can rewrite $\varrho(z)$ as:

$$\varrho(z) = \frac{\vartheta(p(z)) - 1}{\vartheta(p(z)) + \varepsilon(p(z)) - 1}.$$

Substituting this result in the previous expression, we have:

$$\frac{\partial \ln(\pi(z))}{\partial z} = \vartheta(p(z)) - 1$$

such that the partial derivative of profits with respect to productivity is given by:

$$\pi'(z) = y(p(z)) \exp(-z).$$

The social surplus is instead given by:

$$S(z) = \int_{c(z)}^{\overline{p}} y(p)\mathrm{d}p.$$

Therefore, its partial derivative with respect to productivity is given by:

$$S'(z) = y(\exp(-z)) \exp(-z).$$

This shows that the ratio $R(z) \equiv \pi'(z)/S'(z)$ is simply given by the output ratio of the monopolist and the welfare-maximizing agent, which completes the proof for the first part of Proposition 1. Taking the elasticity of this ratio with respect to productivity and using the expression derived above for the monopolist's productivity pass-through completes the proof of Proposition 2. To prove the second part of Proposition 1, define consumer surplus as:

$$C(z) \equiv \int_{p(z)}^{\overline{p}} y(p)\mathrm{d}p.$$

Its partial derivative with respect to productivity is given by:

$$\begin{aligned} C'(z) &= -\frac{\partial p(z)}{\partial z} \times y(p(z)) \\ &= \varrho(z)p(z)y(p(z)) \\ &= \frac{\vartheta(p(z))}{\vartheta(p(z)) + \varepsilon(p(z)) - 1} \times \pi'(z). \end{aligned}$$

Therefore, the ratio of marginal producer surplus to the sum of marginal consumer and producer surplus is:

$$\frac{\pi'(z)}{C'(z) + \pi'(z)} = \frac{\vartheta(p(z)) + \varepsilon(p(z)) - 1}{2\vartheta(p(z)) + \varepsilon(p(z)) - 1}$$

which completes the proof for the second part of Proposition 1. □

Proof of Proposition 3. Let us start with the expression derived in Appendix A.1.7 for aggregate output:

$$Y_t = Z_t K_t^{\alpha} L_t^{1-\alpha} \quad \text{where} \quad Z_t \equiv \left(M_t \int_0^{\infty} \hat{y}_t(\hat{z}) \exp(-\hat{z} - \underline{z}_t) \mathrm{d}F_t(\hat{z}) \right)^{-1}.$$

On a balanced growth path, L_t and M_t are constant, and Y_t and K_t grow at the same rate. This implies that aggregate output grows at rate $g/(1-\alpha)$ where g is the constant growth rate of the endogenous exit threshold, which must grow at the same rate as TFP on a balanced growth path. To derive an expression for g, let us differentiate the logarithm of Z_t with respect to time, to obtain:

$$\frac{\dot{Z}_t}{Z_t} = g_t - \frac{\dot{M}_t}{M_t} + \frac{(\dot{\varsigma}_t/\varsigma_t - g_t) \int_0^{\infty} \vartheta_t(\hat{z}) \varrho_t(\hat{z}) \hat{y}_t(\hat{z}) \exp(-\hat{z}) \mathrm{d}F_t(\hat{z})}{\int_0^{\infty} \hat{y}_t(\hat{z}) \exp(-\hat{z}) \mathrm{d}F_t(\hat{z})} - \frac{\int_0^{\infty} \hat{y}_t(\hat{z}) \exp(-\hat{z}) \dot{F}_t'(\hat{z}) \mathrm{d}\hat{z}}{\int_0^{\infty} \hat{y}_t(\hat{z}) \exp(-\hat{z}) \mathrm{d}F_t(\hat{z})}.$$

Here g_t and ς_t still denote the rate of change of the endogenous exit threshold and the producer price index, respectively. On a balanced growth path, since the measure of varieties is stationary and the producer price index grows at the same constant rate as the endogenous exit threshold, we obtain:

$$\frac{\int_0^{\infty} \hat{y}_t(\hat{z}) \exp(-\hat{z}) \dot{F}_t'(\hat{z}) \mathrm{d}\hat{z}}{\int_0^{\infty} \hat{y}_t(\hat{z}) \exp(-\hat{z}) \mathrm{d}F_t(\hat{z})} = 0.$$

Substituting in the Kolmogorov forward equation of $F_t(\hat{z})$ and abandoning the time subscripts, we have:

$$0 = -\chi \int_0^\infty \hat{y}(\hat{z}) \exp(-\hat{z}) \mathrm{d}F(\hat{z}) + e \int_0^\infty \hat{y}(\hat{z}) \exp(-\hat{z}) \mathrm{d}F^E(\hat{z})$$

$$- \int_0^\infty \hat{y}(\hat{z}) \exp(-\hat{z}) \frac{\partial[\gamma(\hat{z})F'(\hat{z})]}{\partial \hat{z}} \mathrm{d}\hat{z} \tag{A.1}$$

$$+ g \int_0^\infty \hat{y}(\hat{z}) \exp(-\hat{z}) F''(\hat{z}) \mathrm{d}\hat{z} \tag{A.2}$$

$$+ (\sigma^2/2) \int_0^\infty \hat{y}(\hat{z}) \exp(-\hat{z}) F'''(\hat{z}) \mathrm{d}\hat{z} \tag{A.3}$$

Now, let us consider each of the three last terms above separately. First, using integration by parts, the term (A.1) can be rewritten as:

$$(\text{A.1}) = -[\hat{y}(\hat{z}) \exp(-\hat{z}) \gamma(\hat{z}) F'(\hat{z})]_{\hat{z}=0}^\infty + \int_0^\infty [\vartheta(\hat{z})\varrho(\hat{z}) - 1] \hat{y}(\hat{z}) \exp(-\hat{z}) \gamma(\hat{z}) \mathrm{d}F(\hat{z})$$

where $\varrho(\hat{z})$ denotes the productivity "pass-through":

$$\varrho(\hat{z}) \equiv -\frac{\partial \ln(\hat{p}(\hat{z}))}{\partial \hat{z}}.$$

Similarly, the term (A.2) can be rewritten as:

$$(\text{A.2}) = g[\hat{y}(\hat{z}) \exp(-\hat{z}) F'(\hat{z})]_{\hat{z}=0}^\infty - g \int_0^\infty [\vartheta(\hat{z})\varrho(\hat{z}) - 1] \hat{y}(\hat{z}) \exp(-\hat{z}) \mathrm{d}F(\hat{z}).$$

Notice that the boundary conditions $\lim_{\hat{z}\to\infty} F'(\hat{z}) = F'(0) = 0$, the smooth pasting condition (implying that $\gamma(0) = 0$ by the firm's dynamic first-order condition), the limit $\lim_{\hat{z}\to\infty} \hat{y}(\hat{z}) = \theta^{\theta/\epsilon}$ (proved in Proposition 6) and the limit $\lim_{\hat{z}\to\infty} \gamma(\hat{z}) = 0$ (derived in Appendix A.1.11) imply that:

$$[\hat{y}(\hat{z}) \exp(-\hat{z}) F'(\hat{z})]_{\hat{z}=0}^\infty = [\hat{y}(\hat{z}) \exp(-\hat{z}) \gamma(\hat{z}) F'(\hat{z})]_{\hat{z}=0}^\infty = 0$$

Using integration by parts once again, the term (A.3) can be rewritten as:

$$(\text{A.3}) = (\sigma^2/2)[\hat{y}(\hat{z}) \exp(-\hat{z}) F''(\hat{z})]_{\hat{z}=0}^\infty - (\sigma^2/2) \int_0^\infty [\vartheta(\hat{z})\varrho(\hat{z}) - 1] \hat{y}(\hat{z}) \exp(-\hat{z}) \mathrm{d}F'(\hat{z}).$$

Notice here again that the boundary condition $\lim_{\hat{z}\to\infty} F''(\hat{z}) = 0$ together with the limit $\lim_{\hat{z}\to\infty} \hat{y}(\hat{z}) = \theta^{\theta/\epsilon}$ (proved in Proposition 6) imply that:

$$[\hat{y}(\hat{z}) \exp(-\hat{z}) F''(\hat{z})]_{\hat{z}=0}^\infty = -\hat{y}(0) F''(0).$$

Collecting all terms above and using the fact that the measure of varieties is constant on a balanced growth path (implying that the entry rate is equal to the sum of the endogenous and exogenous exit rates), we can solve for g to obtain the expression:

$$
\begin{aligned}
g = & \frac{\int_0^\infty [\vartheta(\hat{z})\varrho(\hat{z}) - 1]\hat{y}(\hat{z})\exp(-\hat{z})\gamma(\hat{z})\mathrm{d}F(\hat{z})}{\int_0^\infty [\vartheta(\hat{z})\varrho(\hat{z}) - 1]\hat{y}(\hat{z})\exp(-\hat{z})\mathrm{d}F(\hat{z})} \\
& - \frac{(\sigma^2/2)\int_0^\infty [\vartheta(\hat{z})\varrho(\hat{z}) - 1]\hat{y}(\hat{z})\exp(-\hat{z})\mathrm{d}F'(\hat{z})}{\int_0^\infty [\vartheta(\hat{z})\varrho(\hat{z}) - 1]\hat{y}(\hat{z})\exp(-\hat{z})\mathrm{d}F(\hat{z})} \\
& + \frac{(\sigma^2/2)F''(0)[\int_0^\infty \hat{y}(\hat{z})\exp(-\hat{z})\mathrm{d}F^E(\hat{z}) - \hat{y}(0)]}{\int_0^\infty [\vartheta(\hat{z})\varrho(\hat{z}) - 1]\hat{y}(\hat{z})\exp(-\hat{z})\mathrm{d}F(\hat{z})} \\
& - \frac{\chi[\int_0^\infty \hat{y}(\hat{z})\exp(-\hat{z})\mathrm{d}F(\hat{z}) - \int_0^\infty \hat{y}(\hat{z})\exp(-\hat{z})\mathrm{d}F^E(\hat{z})]}{\int_0^\infty [\vartheta(\hat{z})\varrho(\hat{z}) - 1]\hat{y}(\hat{z})\exp(-\hat{z})\mathrm{d}F(\hat{z})}.
\end{aligned}
$$

□

Proof of Proposition 4. We want to prove that on a balanced growth path equilibrium allocation:

$$\lim_{\hat{z}\to\infty} F(\hat{z}) = 1 - \exp(-\lambda\hat{z}).$$

To do so, let us guess that as $\hat{z} \to \infty$:

$$
\begin{aligned}
F(\hat{z}) &= 1 - \exp(-\lambda\hat{z}), \\
F'(\hat{z}) &= \lambda\exp(-\lambda\hat{z}), \\
F''(\hat{z}) &= -\lambda^2\exp(-\lambda\hat{z})
\end{aligned}
$$

for $\lambda > 0$. Substituting these guesses in the stationary KFE derived in Appendix A.1.11 and cancelling terms:

$$0 = [g - \gamma(\hat{z})]\lambda + (\sigma^2\lambda^2/2 - \chi)\left\{\frac{T[\exp(-\lambda\hat{z})]}{\exp(-\lambda\hat{z})} - 1\right\}.$$

Taking the limit of this equation and using the assumptions that $\lim_{\hat{z}\to\infty}\gamma(\hat{z}) = \overline{\gamma} < \infty$ and $\lim_{\hat{z}\to\infty}\frac{1-F_t^E(\hat{z})}{1-F_t(\hat{z})} = 0$, we obtain the quadratic equation:

$$0 = \sigma^2\lambda^2/2 - (g - \overline{\gamma})\lambda - \chi.$$

The positive root of this quadratic equation is:

$$\lambda = \frac{g - \overline{\gamma} + \sqrt{(g-\overline{\gamma})^2 + 2\chi\sigma^2}}{\sigma^2}.$$

For that root to be greater than one, we have the additional restriction:

$$g > \overline{\gamma} + \sigma^2/2 - \chi$$

which implies that the distribution of $\exp(\hat{z})$ is Pareto with a finite mean. $\square$

Proof of Proposition 5. Using the household's Euler equation together with the restriction that the value of a firm must grow at the same rate as aggregate output on a balanced growth path, we have the following expression for the firm's HJBE:

$$(\rho+\chi)V_t(\hat{z}) = \pi_t(\hat{z}) + \max_{\gamma}\{(\gamma - g)V_t'(\hat{z}) - w_t i(\gamma,\hat{z})\} + \sigma^2 V_t''(\hat{z})/2$$

with value matching, smooth pasting and first-order conditions:

$$V_t(0) = 0, \quad V_t'(0) = 0 \quad \text{and} \quad \gamma_t(\hat{z}) = \left[\frac{V_t'(\hat{z})}{w_t \exp(c_I + (1+\zeta)\hat{z})}\right]^{1/\zeta}.$$

Substituting the first-order condition in the firm's HJBE, we obtain the second-order nonlinear ordinary differential equation:

$$(\rho+\chi)V_t(\hat{z}) = \pi_t(\hat{z}) + \frac{\zeta\gamma_t(\hat{z})V_t'(\hat{z})}{1+\zeta} - gV_t'(\hat{z}) + \frac{\sigma^2 V_t''(\hat{z})}{2}$$

in which the profit function is given by:

$$\pi_t(\hat{z}) = \begin{cases} \hat{p}_t(\hat{z})\hat{y}_t(\hat{z})^{1+\epsilon/\theta}\Upsilon'(0)Y_t D_t/\theta - w_t c_O & \text{Pre-policy,} \\ [\Upsilon(\hat{y}_t(\hat{z})) - \Upsilon'(\hat{y}_t(\hat{z}))\hat{y}_t(\hat{z})]Y_t D_t - w_t c_O & \text{Post-policy} \end{cases}$$

and the relative demand and relative price functions are in turn given by:

$$\hat{p}_t(\hat{z}) = \begin{cases} \frac{(\theta/\epsilon)\exp(-\hat{z}-\underline{z}_t)\varsigma_t/\overline{p}_t}{W[(\theta/\epsilon)\exp(\theta/\epsilon-\hat{z}-\underline{z}_t)\varsigma_t/\overline{p}_t]} & \text{Pre-policy,} \\ \exp(-\hat{z}-\underline{z}_t)\varsigma_t/\overline{p}_t & \text{Post-policy,} \end{cases}$$
$$\hat{y}_t(\hat{z}) = [-\epsilon\max\{\ln(\hat{p}_t(\hat{z})), 0\}]^{\theta/\epsilon}.$$

Let us now guess and verify that on a balanced growth path, the firm's value function asymptotes to an endogenous constant as $\hat{z} \to \infty$:

$$\lim_{\hat{z}\to\infty} V_t(\hat{z}) = \overline{V}_t.$$

Taking the limit of the firm's HJBE and using Proposition 6, we obtain:

$$(\rho + \chi)\overline{V}_t = \lim_{\hat{z}\to\infty} \pi_t(\hat{z}) = \overline{\pi} Y_t D_t - w_t c_O$$

where the constant $\overline{\pi}$ is given by:

$$\overline{\pi} = \begin{cases} (\theta - 1)\exp[(1-\theta)/\epsilon]\theta^{\theta/\epsilon - 1} & \text{Pre-policy,} \\ 1 + (\theta - 1)\exp(1/\epsilon)\epsilon^{\theta/\epsilon - 1}\Gamma\left(\frac{\theta}{\epsilon}, \frac{1}{\epsilon}\right) & \text{Post-policy} \end{cases}$$

which completes the proof. □

Proof of Proposition 6. We want to prove that:

$$\lim_{\hat{z}\to\infty} \hat{p}_t(\hat{z}) = \exp(-\theta/\epsilon) \quad \text{and} \quad \lim_{\hat{z}\to\infty} \hat{y}_t(\hat{z}) = \theta^{\theta/\epsilon}.$$

Let us remember that a variety's relative price is given by:

$$\hat{p}_t(\hat{z}) = \frac{(\theta/\epsilon)\exp(-\hat{z} - \underline{z}_t)\varsigma_t/\overline{p}_t}{W[(\theta/\epsilon)\exp(\theta/\epsilon - \hat{z} - \underline{z}_t)\varsigma_t/\overline{p}_t]}.$$

Defining the change of variable $x \equiv (\theta/\epsilon)\exp(\theta/\epsilon - \hat{z} - \underline{z}_t)\varsigma_t/\overline{p}_t$ and using l'Hôpital's rule together with the second property of the Lambert W-function, we find:

$$\lim_{\hat{z}\to\infty} \hat{p}_t(\hat{z}) = \lim_{x\to 0} \frac{x}{W(x)\exp(\theta/\epsilon)} = \exp(-\theta/\epsilon).$$

Clearly, since the relative demand function is given by:

$$\hat{y}(\hat{p}) = \begin{cases} [-\epsilon \ln(\hat{p})]^{\theta/\epsilon} & \text{if } \hat{p} < 1, \\ 0 & \text{otherwise} \end{cases}$$

we immediately obtain the result that $\lim_{\hat{z}\to\infty} \hat{y}_t(\hat{z}) = \theta^{\theta/\epsilon}$. □

A.3 Extensions

Product Quality Improvements

In this section, we consider both productivity and quality improvements as the nature of an innovation. In particular, we denote a firm's quality relative to the lowest quality $\underline{q}_t$ in the economy by $\hat{q} \in [0, \infty)$ and adopt the theoretical definition of a variety's quality proposed by Baqaee, Farhi and Sangani (2023). That is, the quality and quantity of a product are assumed to be perfect substitutes.

Taking prices as given, the final sector's problem is to choose its relative demand for each variety to maximize profits in each period:

$$\max_{\{\hat{y}_t(\hat{z},\hat{q})\}_{\hat{z},\hat{q}=0}^{\infty}} \left\{ P_t - M_t \int p_t(\hat{z},\hat{q}) \hat{y}_t(\hat{z},\hat{q}) \mathrm{d}F_t(\hat{z},\hat{q}) \right\} Y_t$$
$$\text{s.t.} \quad M_t \int \Upsilon[\exp(\hat{q} + \underline{q}_t) \hat{y}_t(\hat{z},\hat{q})] \mathrm{d}F_t(\hat{z},\hat{q}) = 1.$$

Reformulating the final sector's problem as a cost-minimization problem subject to the Kimball (1995) aggregator constraint using the Lagrangian, we have:

$$\mathcal{L}_t(\{\hat{y}_t(\hat{q})\}_{\hat{z},\hat{q}=0}^{\infty}, \nu_t) = M_t \int p_t(\hat{z},\hat{q}) \hat{y}_t(\hat{z},\hat{q}) \mathrm{d}F_t(\hat{z},\hat{q}) + \nu_t \left(M_t \int \Upsilon[\exp(\hat{q} + \underline{q}_t) \hat{y}_t(\hat{z},\hat{q})] \mathrm{d}F_t(\hat{z},\hat{q}) - 1 \right)$$

where ν_t now denotes the Lagrange multiplier. The first-order conditions are:

$$p_t(\hat{z},\hat{q}) = \nu_t \exp(\hat{q} + \underline{q}_t) \Upsilon'[\exp(\hat{q} + \underline{q}_t) \hat{y}_t(\hat{z},\hat{q})] \quad \text{and} \quad M_t \int \Upsilon[\exp(\hat{q} + \underline{q}_t) \hat{y}_t(\hat{z},\hat{q})] \mathrm{d}F_t(\hat{z},\hat{q}) = 1.$$

Since the final sector is perfectly competitive and makes no profit, we have:

$$P_t = M_t \int p_t(\hat{z},\hat{q}) \hat{y}_t(\hat{z},\hat{q}) \mathrm{d}F_t(\hat{z},\hat{q}).$$

Substituting in the first-order conditions, we obtain a solution for ν_t:

$$\nu_t = P_t D_t \quad \text{where} \quad D_t \equiv \left(M_t \int \Upsilon'[\exp(\hat{q} + \underline{q}_t) \hat{y}_t(\hat{z},\hat{q})] \exp(\hat{q} + \underline{q}_t) \hat{y}_t(\hat{z},\hat{q}) \mathrm{d}F_t(\hat{z},\hat{q}) \right)^{-1}.$$

This delivers the following inverse demand functions:

$$p_t(\hat{z},\hat{q}) = \exp(\hat{q} + \underline{q}_t) \Upsilon'[\exp(\hat{q} + \underline{q}_t) \hat{y}_t(\hat{z},\hat{q})] P_t D_t.$$

Firms engage in monopolistic competition in the product market but perfect competition in the input markets. That is, a firm chooses the price at which to sell its variety as well as its demand for physical capital and production labor to maximize profits in each period. The firm takes as given the demand for its variety, the rental rate of capital r_t and the wage rate w_t, which delivers the following problem:

$$\pi_t(\hat{z},\hat{q}) = \max_{p_t(\hat{z},\hat{q}),k_t(\hat{z},\hat{q}),l_t(\hat{z},\hat{q})} \{p_t(\hat{z},\hat{q})y_t(\hat{z},\hat{q}) - (r_t+\delta)k_t(\hat{z},\hat{q}) - w_t l_t(\hat{z},\hat{q})\} - w_t c_O$$

subject to the inverse demand function $p_t(\hat{z},\hat{q}) = \exp(\hat{q}+\underline{q}_t)\Upsilon'[\exp(\hat{q}+\underline{q}_t)\hat{y}_t(\hat{z},\hat{q})]D_t$. Let us first consider the sub-problem of optimally choosing the demand for capital and labor, which can be reformulated as a cost-minimization problem. Using the Lagrangian:

$$\mathcal{L}_t(k_t(\hat{z},\hat{q}), l_t(\hat{z},\hat{q}), \nu_t) = (r_t+\delta)k_t(\hat{z},\hat{q}) + w_t l_t(\hat{z},\hat{q}) + \nu_t[y_t(\hat{z},\hat{q}) - \exp(\hat{z}+\underline{z}_t)k_t(\hat{z},\hat{q})^{\alpha} l_t(\hat{z},\hat{q})^{1-\alpha}]$$

where ν_t denotes the Lagrange multiplier. The first-order conditions are:

$$\begin{aligned} k_t(\hat{z},\hat{q}) &= \frac{\alpha\nu_t y_t(\hat{z},\hat{q})}{r_t+\delta}, \\ l_t(\hat{z},\hat{q}) &= \frac{(1-\alpha)\nu_t y_t(\hat{z},\hat{q})}{w_t}, \\ y_t(\hat{z},\hat{q}) &= \exp(\hat{z}+\underline{z}_t)k_t(\hat{z},\hat{q})^{\alpha} l_t(\hat{z},\hat{q})^{1-\alpha}. \end{aligned}$$

Solving for the Lagrange multiplier, we have:

$$\nu_t = \varsigma_t \exp(-\hat{z}-\underline{z}_t) \quad \text{where} \quad \varsigma_t \equiv \left(\frac{r_t+\delta}{\alpha}\right)^{\alpha}\left(\frac{w_t}{1-\alpha}\right)^{1-\alpha}.$$

Therefore, we can rewrite the firm's static problem as:

$$\begin{aligned} \pi_t(\hat{z},\hat{q}) = \max_{p_t(\hat{z},\hat{q})} &\{[p_t(\hat{z},\hat{q}) - \varsigma_t \exp(-\hat{z}-\underline{z}_t)]\hat{y}_t(\hat{z},\hat{q})\}Y_t - w_t c_O \\ \text{s.t.} \quad & p_t(\hat{z},\hat{q}) = \exp(\hat{q}+\underline{q}_t)\Upsilon'[\exp(\hat{q}+\underline{q}_t)\hat{y}_t(\hat{z},\hat{q})]D_t. \end{aligned}$$

Reformulating it as a choice of $\hat{y}_t(\hat{z},\hat{q})$ given the inverse demand function $p_t(\hat{z},\hat{q})$:

$$\pi_t(\hat{z},\hat{q}) = \max_{\hat{y}_t(\hat{z},\hat{q})}\{[\exp(\hat{q}+\underline{q}_t)\Upsilon'[\exp(\hat{q}+\underline{q}_t)\hat{y}_t(\hat{z},\hat{q})]D_t - \varsigma_t \exp(-\hat{z}-\underline{z}_t)]\hat{y}_t(\hat{z},\hat{q})\}Y_t - w_t c_O.$$

The first-order condition is:

$$\{\Upsilon''[\exp(\hat{q}+\underline{q}_t)\hat{y}_t(\hat{z},\hat{q})]\exp(\hat{q}+\underline{q}_t)\hat{y}_t(\hat{z},\hat{q}) + \Upsilon'[\exp(\hat{q}+\underline{q}_t)\hat{y}_t(\hat{z},\hat{q})]\}\exp(\hat{q}+\underline{q}_t)D_t = \varsigma_t \exp(-\hat{z}-\underline{z}_t).$$

We can rearrange this expression as:

$$p_t(\hat{z},\hat{q}) = \frac{\mu_t(\hat{z},\hat{q})\varsigma_t}{\exp(\hat{z}+\underline{z}_t)} \quad \text{where} \quad \mu_t(\hat{z},\hat{q}) \equiv \frac{\vartheta_t(\hat{z},\hat{q})}{\vartheta_t(\hat{z},\hat{q})-1}$$

and where $\vartheta_t(\hat{z},\hat{q})$ denotes the price elasticity of demand:

$$\vartheta_t(\hat{z},\hat{q}) \equiv -\frac{\Upsilon'[\exp(\hat{q}+\underline{q}_t)\hat{y}_t(\hat{z},\hat{q})]}{\Upsilon''[\exp(\hat{q}+\underline{q}_t)\hat{y}_t(\hat{z},\hat{q})]\exp(\hat{q}+\underline{q}_t)\hat{y}_t(\hat{z},\hat{q})} \in (1,\infty).$$

Substituting the monopoly pricing function in the profit function, we have:

$$\pi_t(\hat{z},\hat{q}) = \frac{p_t(\hat{z},\hat{q})\hat{y}_t(\hat{z},\hat{q})Y_t}{\vartheta_t(\hat{z},\hat{q})} - w_t c_O.$$

Denoting the quality-specific choke price by $\overline{p}_t(\hat{q}) \equiv \exp(\hat{q}+\underline{q}_t)\Upsilon'(0)D_t$, we can define a variety's relative price as $\hat{p}_t(\hat{z},\hat{q}) \equiv p_t(\hat{z},\hat{q})/\overline{p}_t(\hat{q})$. This allows us to rewrite the relative demand function as:

$$\hat{y}_t(\hat{z},\hat{q}) = \Upsilon'^{-1}[\hat{p}_t(\hat{z},\hat{q})\Upsilon'(0)]/\exp(\hat{q}+\underline{q}_t).$$

Using the Klenow and Willis (2016) specification of the Kimball (1995) aggregator, we can rewrite the monopoly pricing condition (relative to the choke price) as:

$$\hat{p}_t(\hat{z},\hat{q}) = \frac{\varsigma_t}{[1+(\epsilon/\theta)\ln(\hat{p}_t(\hat{z},\hat{q}))]\exp(\hat{z}+\underline{z}_t)\overline{p}_t(\hat{q})}$$

Note that we can rearrange this equation to obtain:

$$\exp\{(\theta/\epsilon)[\varsigma_t[\hat{p}_t(\hat{z},\hat{q})\exp(\hat{z}+\underline{z}_t)\overline{p}_t(\hat{q})]^{-1}-1]\}\hat{p}_t(\hat{z},\hat{q})^{-1} = 1$$

Multiplying both sides by $(\theta/\epsilon)\exp(\theta/\epsilon)\varsigma_t[\exp(\hat{z}+\underline{z}_t)\overline{p}_t(\hat{q})]^{-1}$, we have:

$$W^{-1}\{(\theta/\epsilon)\varsigma_t[\hat{p}_t(\hat{z},\hat{q})\exp(\hat{z}+\underline{z}_t)\overline{p}_t(\hat{q})]^{-1}\} = (\theta/\epsilon)\exp(\theta/\epsilon)\varsigma_t[\exp(\hat{z}+\underline{z}_t)\overline{p}_t(\hat{q})]^{-1}.$$

Finally, solving for $\hat{p}_t(\hat{z},\hat{q})$ delivers:

$$\hat{p}_t(\hat{z},\hat{q}) = \frac{(\theta/\epsilon)\exp(-\hat{z}-\underline{z}_t)\varsigma_t/\overline{p}_t(\hat{q})}{W[(\theta/\epsilon)\exp(\theta/\epsilon)\exp(-\hat{z}-\underline{z}_t)\varsigma_t/\overline{p}_t(\hat{q})]}.$$

Substituting in the expression for the quality-specific choke price and defining the firm's composite state variable as $x_t \equiv z_t + q_t$, we have:

$$\hat{p}_t(\hat{x}) = \frac{(\theta/\epsilon)\exp(-\hat{x} - \underline{x}_t)\varsigma_t/\overline{p}_t}{W[(\theta/\epsilon)\exp(\theta/\epsilon)\exp(-\hat{x} - \underline{x}_t)\varsigma_t/\overline{p_t}]} \quad \text{where} \quad \overline{p}_t \equiv \Upsilon'(0)D_t$$

which is isomorphic to our framework. Therefore, profits are given by:

$$\pi_t(\hat{x}) = \frac{\hat{p}_t(\hat{x})\Upsilon'^{-1}[\hat{p}_t(\hat{x})\Upsilon'(0)]\overline{p}_t Y_t}{\vartheta_t(\hat{x})} - w_t c_O$$

which is also isomorphic to our framework. Using the above expressions, the firm's markup is given by the following function:

$$\mu_t(\hat{x}) = \frac{\theta/\epsilon}{W[(\theta/\epsilon)\exp(\theta/\epsilon)\exp(-\hat{x} - \underline{x}_t)\varsigma_t/\overline{p}_t]}.$$

Inverting this function, we have:

$$\hat{x}_t(\mu) = \ln(\mu) + (\theta/\epsilon)(1 - \mu^{-1}) + \ln(\varsigma_t/\overline{p}_t) - \underline{x}_t.$$

B Numerical Appendix

This section of the Appendix provides details on the numerical strategies we use to solve and quantify the model.

B.1 Spectral Collocation and Quadrature

Following Miranda and Fackler (2004), we approximate the solutions of the HJBE and KFE using spectral collocation in the spatial dimension on the interval $\hat{z} \in [0, \infty)$. But first, we consider a change of variable to approximate the solutions of the HJBE and KFE on the unit line:

$$V_t(\hat{z}) = \mathcal{V}_t(\tilde{z}) \quad \text{and} \quad F_t(\hat{z}) = \mathcal{F}_t(\tilde{z}) \quad \text{where} \quad \tilde{z} \equiv \frac{\nu\hat{z}}{\nu\hat{z} + 1}.$$

Here, $\nu > 0$ is a parameter governing the curvature of the change of variable. With this change of variable, we can use the chain rule to obtain:

$$\begin{aligned}
V_t'(\hat{z}) &= \nu(1-\tilde{z})^2\mathcal{V}'(\tilde{z}), \\
V_t''(\hat{z}) &= \nu^2(1-\tilde{z})^3[(1-\tilde{z})\mathcal{V}_t''(\tilde{z}) - 2\mathcal{V}_t'(\tilde{z})], \\
F_t'(\hat{z}) &= \nu(1-\tilde{z})^2\mathcal{F}_t'(\tilde{z}), \\
F_t''(\hat{z}) &= \nu^2(1-\tilde{z})^3[(1-\tilde{z})\mathcal{F}_t''(\tilde{z}) - 2\mathcal{F}_t'(\tilde{z})].
\end{aligned}$$

Under this reformulation, the boundary conditions of $\mathcal{V}_t(\tilde{z})$ and $\mathcal{F}_t(\tilde{z})$ are:

$$\mathcal{V}_t(0) = \mathcal{V}_t'(0) = \mathcal{F}_t(0) = 0 \quad \text{and} \quad \mathcal{F}_t(1) = 1.$$

We approximate the functions $\mathcal{V}_t(\hat{z})$ and $\mathcal{F}_t(\hat{z})$ over $n-2$ Chebyshev nodes $\{\tilde{z}_i\}_{i=2}^{n-1}$ on the unit line to which we append the boundaries $\tilde{z}_1 = 0$ and $\tilde{z}_n = 1$. The approximation is a linear combination of Chebyshev basis functions $\{b_j(\tilde{z})\}_{j=1}^{n}$ of degree n whose time-varying coefficients $\{c_j^V(t), c_j^F(t)\}_{j=1}^{n}$ are to be determined:

$$\mathcal{V}_t(\tilde{z}_i) \approx \sum_{j=1}^{n} c_j^V(t) b_j(\tilde{z}_i) \quad \text{and} \quad \mathcal{F}_t(\tilde{z}_i) \approx \sum_{j=1}^{n} c_j^F(t) b_j(\tilde{z}_i).$$

In particular, the coefficients $\{c_j^V(t), c_j^F(t)\}_{j=1}^{n}$ are chosen to satisfy the HJBE and KFE over the nodes $\{\tilde{z}_i\}_{i=2}^{n-1}$ as well as the boundary conditions of $\mathcal{V}_t(\hat{z})$ and $\mathcal{F}_t(\hat{z})$ at $\tilde{z}_1 = 0$ and $\tilde{z}_n = 1$. In addition, the value of the free boundary $\underline{z}_t$ is chosen to satisfy the smooth pasting condition $\mathcal{V}_t'(0) = 0$. However, the solutions of the HJBE and KFE depend on unknown endogenous economic variables that are constrained by equilibrium conditions. These equilibrium conditions involve integrals which we approximate using Chebyshev–Gauss quadrature over the same set of nodes $\{\tilde{z}_i\}_{i=2}^{n-1}$ as above. That is, for an arbitrary function f, we approximate the following integral:

$$\int_0^\infty f_t(\hat{z})\mathrm{d}\hat{z} = \int_0^1 \frac{f_t(\hat{z}(\tilde{z}))}{\nu(1-\tilde{z})^2}\mathrm{d}\tilde{z} \approx \sum_{i=2}^{n-1} \frac{f_t(\hat{z}(\tilde{z}_i))}{\nu(1-\tilde{z}_i)^2}\omega_i \quad \text{where} \quad \hat{z}(\tilde{z}) \equiv \frac{\tilde{z}}{\nu(1-\tilde{z})}$$

and where $\{\omega_i\}_{i=2}^{n-1}$ denote the Chebyshev–Gauss quadrature weights.

B.2 MPEC-BGP Estimation Strategy

To estimate the structural parameters of our theory, solving for a balanced growth path equilibrium at multiple parameter guesses is computationally expensive. An alternative approach discussed in Su and Judd (2012) and Dubé et al. (2012) is to reformulate the estimation problem as a mathematical program with equilibrium constraints (MPEC).

Instead of solving for a balanced growth path equilibrium allocation at multiple guesses of these parameters, our approach is to treat the model's equilibrium conditions as constraints in the optimization problem. That is, we search for parameters as well as endogenous economic variables to minimize a GMM objective subject to the constraints that the model's equilibrium conditions are met.

More formally, let us describe the balanced growth path equilibrium allocation of our model as an $N_x \times 1$ vector of endogenous variables $\mathbf{x}$ that depend on an $N_\Omega \times 1$ vector of parameters $\boldsymbol{\Omega}$ through the $N_x \times 1$ vector of equilibrium conditions:

$$h(\mathbf{x}, \boldsymbol{\Omega}) = \mathbf{0}.$$

We let $X(\boldsymbol{\Omega})$ denote the set of all $\mathbf{x}$ such that $h(\mathbf{x}, \boldsymbol{\Omega}) = \mathbf{0}$:

$$X(\boldsymbol{\Omega}) := \{\mathbf{x} : h(\mathbf{x}, \boldsymbol{\Omega}) = \mathbf{0}\}.$$

For some weighting $N_m \times N_m$ matrix $\mathbf{W}$ and an $N_m \times 1$ vector of moments $m(\mathbf{x}, \boldsymbol{\Omega})$, define the GMM estimator as the vector $\boldsymbol{\Omega}^*$ that solves the problem:

$$\boldsymbol{\Omega}^* = \arg\min_{\boldsymbol{\Omega}} \{ \min_{\mathbf{x} \in X(\boldsymbol{\Omega})} m(\mathbf{x}, \boldsymbol{\Omega})^\top \mathbf{W} m(\mathbf{x}, \boldsymbol{\Omega}) \}.$$

Denote by $\mathbf{x}^*(\boldsymbol{\Omega}^*)$ the optimal solution of endogenous variables for this problem. This approach is particularly powerful when the functions $h(\mathbf{x}, \boldsymbol{\Omega})$ and $m(\mathbf{x}, \boldsymbol{\Omega})$ are both twice differentiable in their arguments since we can exploit their Jacobian and Hessian to efficiently find a solution. We implement this using the commercial nonlinear solver KNITRO (Byrd, Nocedal and Waltz, 2006) as well as the open source nonlinear solver IPOPT (Wächter and Biegler, 2006) through the interface of JuMP (Lubin, Dowson, Dias Garcia, Huchette, Legat and Vielma, 2023), a modeling language for mathematical optimization embedded in Julia.

B.2.1 Identification

Andrews, Gentzkow and Shapiro (2017) suggest that researchers report the sensitivity of their parameter estimates to moment conditions. This sensitivity matrix is denoted by $\boldsymbol{\Lambda}$ and defined as:

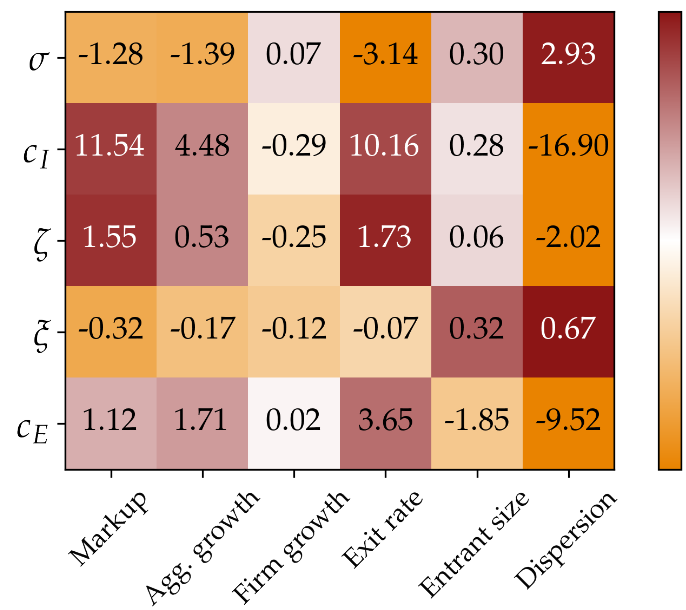
$$\boldsymbol{\Lambda} = -[\mathbf{G}(\mathbf{x}^*(\boldsymbol{\Omega}^*), \boldsymbol{\Omega}^*)^\top \mathbf{W} \mathbf{G}(\mathbf{x}^*(\boldsymbol{\Omega}^*), \boldsymbol{\Omega}^*)]^{-1} \mathbf{G}(\mathbf{x}^*(\boldsymbol{\Omega}^*), \boldsymbol{\Omega}^*)^\top \mathbf{W}$$

where $\mathbf{G}(\mathbf{x}^*(\mathbf{\Omega}^*), \mathbf{\Omega}^*)$ is the $N_m \times N_\Omega$ Jacobian of $m(\mathbf{x}, \mathbf{\Omega})$ with respect to parameters, evaluated at $\mathbf{x}^*(\mathbf{\Omega}^*)$ and $\mathbf{\Omega}^*$. However, to calculate this Jacobian, we must account for the dependency between endogenous variables and parameters through the equilibrium conditions $h(\mathbf{x}^*(\mathbf{\Omega}^*), \mathbf{\Omega}^*) = \mathbf{0}$. As such, we can define that Jacobian using the implicit function theorem:

$$\mathbf{G}(\mathbf{x}^*(\mathbf{\Omega}^*), \mathbf{\Omega}^*) := \nabla_{\mathbf{\Omega}} m(\mathbf{x}^*, \mathbf{\Omega}^*) - \nabla_{\mathbf{x}} m(\mathbf{x}^*, \mathbf{\Omega}^*)[\nabla_{\mathbf{x}} h(\mathbf{x}^*, \mathbf{\Omega}^*)]^{-1} \nabla_{\mathbf{\Omega}} h(\mathbf{x}^*, \mathbf{\Omega}^*).$$

Figures B.8 and B.9 respectively plot the matrices $\mathbf{\Lambda}$ and $\mathbf{G}(\mathbf{x}^*(\mathbf{\Omega}^*), \mathbf{\Omega}^*)$ (in elasticity form) for the five jointly estimated parameters and the six moment conditions presented in Section 4.2.

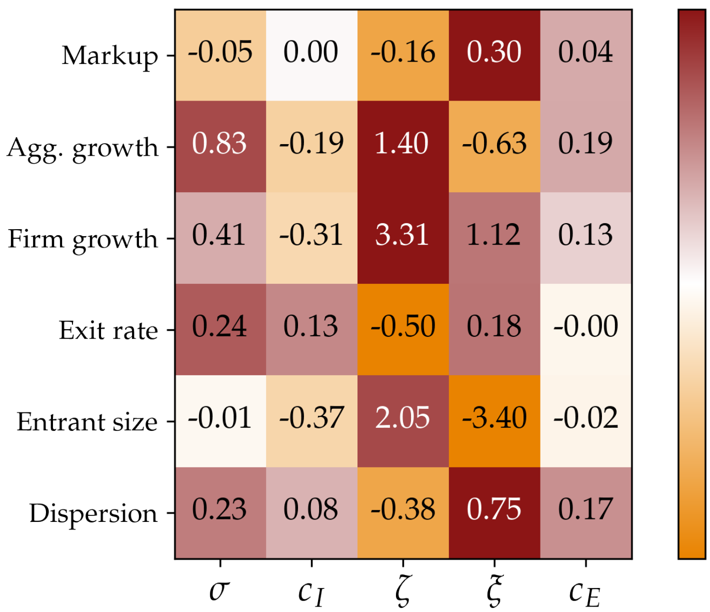
Figure B.8: Sensitivity Matrix



Note: This heat map plots the sensitivity matrix $\mathbf{\Lambda}$ of parameters with respect to moment conditions. The color palette is normalized for each row separately to ease visualization and the matrix is presented in elasticity form.

A first observation is that the standard deviation of the Brownian motion σ is mostly identified by the dispersion of value added and the overall exit rate. It is perhaps no surprise that this parameter is sensitive to the dispersion in value-added, since it determines the volatility of the productivity process. But this volatility also influences the rate at which unproductive firms are swept below the exit threshold, explaining why σ is sensitive to the exit rate. Further, this parameter is also sensitive to the aggregate markup–since it determines the shape of the relative productivity distribution over which firm-level markups are aggregated–and to the aggregate growth rate, which is directly reliant on this parameter as noted in Proposition 3.

Figure B.9: Jacobian Matrix



Note: This heat map plots the Jacobian matrix $\mathbf{G}(\mathbf{x}^*(\mathbf{\Omega}^*), \mathbf{\Omega}^*)$ of moment conditions with respect to parameters. The color palette is normalized for each row separately to ease visualization and the matrix is presented in elasticity form.

Similarly, the innovation cost scale and elasticity parameters c_I and ζ are mainly identified by these same four moments. The intuition underlying these relationships is as follows: to decide how much to invest in R&D, the firm compares the marginal value to the marginal cost of such investments. The parameters c_I and ζ influence the scale and shape of that marginal cost function, which thus dictate the firm's productivity drift function and in turn the shape of the relative productivity distribution. Since these four moments are explicitly dependent on this stationary distribution, it is consistent with our understanding that they identify c_I and ζ.

The parameter ξ determines the transformation of the incumbent distribution from which entrants draw their relative productivity. Therefore, it is not surprising that this parameter is sensitive to the relative size of entrants. However, it is also sensitive to the dispersion of value-added since it acts as a compressing force on the variance of the relative productivity distribution. Therefore, and by the same intuition mentioned above, this parameter is also sensitive to the aggregate markup.

Finally, the entry cost parameter c_E is primarily identified by the dispersion in value-added, which is intricately tied to the shape of the relative productivity distribution. But the shape of that distribution is in large part determined by firms' investments in R&D who compete against one another for labor. The intensity of that competition, in turn, is dictated by the endogenous measure of varieties, which depends on the flow of entrants and, consequently, on the entry cost parameter itself.

C Empirical Appendix

This section of the Appendix provides details on the construction of variables of interest from our main source of data and on the estimation of our theory's structural parameters.

C.1 Data

Our main source of data is the *Fichier Approché des Résultats d'Esane* (FARE). This is an annual panel dataset, covering the period 2009–2019, with the balance sheet and income statements for the universe of firms in France that are subject to the standard corporate tax (excluding the financial and farming sectors).

Given that this dataset is compiled from tax declarations, the unit of observation is a legal entity (*unité légale*), each identified by a unique Siren number. Recognizing that this does not correspond to what users of the data would call a firm, the National Institute of Statistics and Economic Studies (INSEE) developed definitions of consolidated firms (*entreprises profilées*): a collection of legal entities that are part of the same group identified by a unique Sirus number. We use the *Contour des Entreprises Profilées* from 2019 to define the boundaries of the firms in our sample.

The main variables of interest are the firm's industry of operation, value-added, wage bill, and capital stock.

- The main industry of operation for the firm is a 5-digit industry from the NAF classification. For consolidated firms, this is provided in the Contours files.
- Annual value added excludes VAT and is calculated as gross output (sum of sales and the gross value of stored production) net of expenditures on intermediate inputs, materials, and other external expenses, as well as changes in the stock of intermediate inputs and materials.
- The annual wage bill includes all labor costs for the firm and is obtained by summing expenditures on salaries (including bonuses) and social security payments.

- Capital is the sum of tangible capital, inventories, and rental and lease payments. We calculate the book value of tangible capital as the gross acquisition value net of accumulated depreciation. To convert to current prices, we multiply this book value by the ratio of the aggregate price index for gross fixed capital formation in the current year to its value in the acquisition year.[39] For every firm-by-year observation, we have an estimate of the acquisition year of capital since, assuming a constant depreciation rate, the ratio of accumulated depreciation to gross acquisition value can be used to recover an average age for the firm's capital stock.[40] Formally, denoting by k^{T}_{jit} the stock of tangible capital of firm j in industry i in year t, k^{I}_{jit} its inventories and R_{jit} its lease and rental payments, we define:

$$k_{jit} = k^{\mathrm{T}}_{jit} + k^{\mathrm{I}}_{jit} + \frac{R_{jit}}{r_t + \delta} \quad \text{s.t.} \quad (r_t + \delta)k_{jit} = (r_t + \delta)(k^{\mathrm{T}}_{jit} + k^{\mathrm{I}}_{jit}) + R_{jit}.$$

We obtain firm-level value-added, wage bill and capital stock by summing each of these variables over the different legal units that constitute a firm.

We keep in our sample private businesses with a regular taxation scheme and drop any firm with a negative value for value-added, wage bill, stock of tangible capital, inventories, or rental and lease payments. Additionally, we winsorize each of these variables at the 1% level at the 2-digit industry by year level. With these selection criteria, we end up with a sample of 5,423,743 (firm-year) observations between 2009 and 2019, with 831,297 unique firms overall. Table C.11 presents summary statistics for the main variables of interest in our sample.

Table C.11: Summary Statistics

Variable	Mean	5th %ile	25th %ile	Median	75th %ile	95th %ile
Value Added	938	26	116	265	584	2790
Wage bill	700	18	94	214	467	2149
Tangible capital + inventories	823	1	19	77	256	1987

Note: Units are in thousands of current Euros.

[39] This price index is provided by the INSEE.
[40] We use a depreciation rate of 10% for the tangible capital stock, consistent with French accounting standards.

C.2 Structural Estimation

The Klenow and Willis (2016) Parameters

A key parameter to discipline in our theory is the ratio of ϵ and θ. This ratio measures how steeply markups increase in firms' productivity and therefore partly determines the degree of dispersion in markups. If both markups and market shares are observed at the firm level, one can use the demand functions for varieties to estimate this ratio.[41] In particular, let us consider a generalization of the Kimball (1995) aggregator:

$$\sum_{i=1}^{I} x_{it} \int_{j \in \mathcal{J}_t} q_{ji} \Upsilon(\hat{y}_{jit}) \mathrm{d}j = 1$$

where x_{it} is an industry-time-specific demand shifter and q_{ji} denotes the unobserved time-invariant "quality" of variety j. This demand system implies the following inverse demand functions in logarithms:

$$\ln(p_{jit}) = \ln(x_{it}) + \ln(q_{ji}) + \ln(D_t) + \ln(\Upsilon'(\hat{y}_{jit}))$$

where D_t is the time-varying demand index and the function $\Upsilon'(\hat{y})$ is given by:

$$\Upsilon'(\hat{y}) = \left(\frac{\theta - 1}{\theta}\right) \exp\left(\frac{1 - \hat{y}^{\epsilon/\theta}}{\epsilon}\right).$$

Adding $\ln(\hat{y}_{jit})$ to both sides of this equation, and denoting the market share of firm j from industry i at time t by s_{jit}, we obtain:

$$\ln(s_{jit}) = \ln(x_{it}) + \ln(q_{ji}) + \ln(D_t) + \ln(\Upsilon'(\hat{y}_{jit})) + \ln(\hat{y}_{jit}).$$

Since markups are related to relative demand as:

$$\mu_{jit}^{-1} = 1 - \hat{y}_{jit}^{\epsilon/\theta}/\theta$$

we can rewrite the previous equation as:

$$\mu_{jit}^{-1} + \ln(1 - \mu_{jit}^{-1}) = \psi + (\epsilon/\theta)[\ln(s_{jit}) - \ln(x_{it}) - \ln(q_{ji}) - \ln(D_t)]$$

where the constant is given by $\psi \equiv \frac{\theta-1}{\theta} - (\epsilon/\theta)\ln\left(\frac{\theta-1}{\theta}\right)$. Hence, regressing the nonlinear transformation $\mu_{jit}^{-1} + \ln(1 - \mu_{jit}^{-1})$ of firm-level markups on firm-level market shares,

[41] We calculate μ_{jit} and market share as described in Section 4.1

controlling for industry, industry-time, time and firm fixed effects delivers a consistent estimate of θ/ϵ.

Calculating the Targeted Moments

Our GMM estimation strategy uses three moments calculated from the FARE data:

- Within-industry standard deviation of log value added.
- Average annual growth of firm-level value-added, deflated with the GDP deflator.
- Relative size of entrants, calculated as the ratio of the average value added of an incumbent to the average value added of an entrant. In calculating this moment, we treat any firm as an entrant if it has been created within the past five years, and an incumbent otherwise.

We calculate each of these moments at the 2-digit industry-by-year level. We then aggregate by year, weighting each industry by its share of value added in that year. Finally, we take a simple average of each of these moments across the different years in our sample.

C.3 Additional Tables

In this section of the Appendix, we present tables for the alternative assumptions and policy counterfactuals discussed and considered in the main text. In particular, we replicate Tables 3, 4, 5 and 6.

Table C.12: Markups and Market Shares

	Dependent variable: $\mu_{jit}^{-1} + \ln(1 - \mu_{jit}^{-1})$					
	Unadjusted capital share			Adjusted capital share		
$\ln(s_{jit})$	0.047	0.234	0.243	0.048	0.224	0.231
	(0.000)	(0.001)	(0.001)	(0.000)	(0.001)	(0.001)
Firm fixed effects		Y	Y		Y	Y
Industry $\times$ year fixed effects	Y	Y	Y	Y	Y	Y
Industry fixed effects			Y			Y
Year fixed effects			Y			Y
Age group fixed effects	Y		Y	Y		Y
R^2	0.090	0.505	0.507	0.087	0.523	0.524
Observations	4.9M	4.9M	4.9M	5M	5M	5M

Note: Firm-level markups and market shares are constructed from the FARE dataset as described in Section 4.1. This table presents different regression specifications with firm fixed effects, 5-digit NACE industry fixed effects and age group fixed effects (for a total of 20 evenly-spaced age groups). Standard errors (in parentheses) are clustered at the firm level. The total number of observations is below the total sample size of 5.4M because negative markups were estimated for some firms. In the columns labeled "adjusted capital share", the industry-specific capital cost shares measured in the data are inflated by a constant such that the aggregate capital cost share is equal to 1/3, consistent with our calibrated model.

Table C.13: Economic Aggregates for Fixed Innovation Labor

Aggregate	Before	After	Change
Labor allocations:			
Labor supply	0.264	0.319	+20.8%
Production labor	0.227	0.263	+16.1%
Innovation labor	0.020	0.020	0.0%
Entry labor	0.015	0.034	+122.2%
Overhead labor	0.002	0.001	-33.2%
Firms, entry and exit:			
Measure of varieties	0.044	0.029	-33.2%
Entry rate	5.32%	17.68%	+12.4p.p.
Endogenous exit rate	3.97%	16.33%	+12.4p.p.

Note: This table presents the pre- and post-policy level of various economic aggregates as well as the corresponding percentage change when fixing the aggregate allocation of labor to innovation to its initial level before the implementation of the policy intervention. Doing so requires imposing a uniform tax of 44.5% on firms' expenditures on R&D.

Table C.14: Economic Aggregates for Alternative Transfer Schedules

Aggregate	Before	Level fix		Dispersion fix	
		After	Change	After	Change
Labor allocations:					
Labor supply	0.264	0.304	+15.3%	0.284	+7.8%
Production labor	0.227	0.271	+19.5%	0.214	-5.7%
Innovation labor	0.020	0.017	-12.9%	0.039	+93.4%
Entry labor	0.015	0.014	-8.9%	0.031	+100.2%
Overhead labor	0.002	0.001	-1.7%	0.001	-40.5%
Firms, entry and exit:					
Measure of varieties	0.044	0.043	-1.7%	0.026	-40.5%
Entry rate	5.32%	4.93%	-0.4p.p.	17.89%	+12.6p.p.
Endogenous exit rate	3.97%	3.59%	-0.4p.p.	16.54%	+12.6p.p.

Note: This table presents the pre- and post-policy level of various economic aggregates as well as the corresponding percentage change under alternative policy interventions. Specifically, the columns labeled "Baseline", "Level fix" and "Dispersion fix" refer to the transfers that rectify both, and either the level or dispersion in markups, respectively.

Table C.15: Structural Parameters for Alternative Aggregate Markup Targets

Parameter	Symbol	Value $\mathcal{M}$=1.1	Value $\mathcal{M}$=1.5
Household preferences:			
Rate of time preference	ρ	0.04	0.04
Labor supply utility weight	β	11.6	9.3
Frisch elasticity of labor supply reciprocal	η	1	1
Final sector technology:			
Klenow and Willis (2016) elasticity param.	θ	54.5	6.8
Klenow and Willis (2016) super-elasticity param.	ϵ	26.78	1.53
Firm production technology:			
Output elasticity of physical capital	α	0.33	0.33
Depreciation rate of physical capital	δ	0.06	0.06
Overhead cost parameter	c_O	0.05	0.03
Firm innovation technology:			
Brownian motion standard deviation	σ	0.03	0.05
Innovation cost scale parameter	c_I	11.04	4.83
Innovation cost elasticity parameter	ζ	1.10	0.48
Entry and exit:			
Entry cost parameter	c_E	1.15	3.49
Entry distribution parameter	ξ	2.03	2.39
Exogenous exit rate	χ	1.34%	1.34%

Note: This table presents the assigned/estimated structural parameters of our theory under alternative aggregate markup targets. Specifically, the columns labeled "$\mathcal{M}$=1.1" and "$\mathcal{M}$=1.5" respectively refer to parameterizations that target a cost-weighted average markup of 1.1 and 1.5.

Table C.16: Moments for Alternative Aggregate Markup Targets

Moment	Model $\mathcal{M}$=1.1	Model $\mathcal{M}$=1.5	Data
Cost-weighted average markup	1.10	1.50	
GDP per hour worked growth rate	1.15%	1.16%	1.16%
Incumbent value added growth rate	1.24%	1.24%	1.24%
Exit rate of all firms	10.24%	3.62%	5.61%
Relative size of entrants by value added	0.37	0.26	0.31
Standard deviation of log value added	1.54	1.42	1.54

Note: This table presents targeted moments and their resulting value in our model under alternative aggregate markup targets. Specifically, the columns labeled "$\mathcal{M}$=1.1" and "$\mathcal{M}$=1.5" respectively refer to parameterizations that target a cost-weighted average markup of 1.1 and 1.5.

Table C.17: Economic Aggregates for Alternative Aggregate Markup Targets

Aggregate	$\mathcal{M}$=1.1			$\mathcal{M}$=1.5		
	Before	After	Change	Before	After	Change
Labor allocations:						
Labor supply	0.264	0.285	+7.9%	0.264	0.368	+39.6%
Production labor	0.245	0.258	+5.3%	0.211	0.256	+21.6%
Innovation labor	0.005	0.016	+191.3%	0.034	0.079	+134.2%
Entry labor	0.053	0.047	-10.3%	0.029	0.058	+101.4%
Overhead labor	0.003	0.002	-43.5%	0.004	0.002	-45.7%
Firms, entry and exit:						
Measure of varieties	0.078	0.044	-43.5%	0.119	0.065	-45.7%
Entry rate	10.24%	16.27%	+6.0p.p.	3.62%	13.43%	+9.8p.p.
Endogenous exit rate	8.90%	14.93%	+6.0p.p.	2.28%	12.09%	+9.8p.p.

Note: This table presents the pre- and post-policy level of various economic aggregates as well as the corresponding percentage change under alternative aggregate markup targets. Specifically, the columns labeled "$\mathcal{M}$=1.1" and "$\mathcal{M}$=1.5" respectively refer to parameterizations that target a cost-weighted average markup of 1.1 and 1.5.