Capping Profits for Efficiency

Gabriele Patete*

January 31, 2025

Latest version

Abstract

In light of recent concerns over firms' market power, I study the impact of a novel policy—a cap on firms' profit-to-cost ratios—on economic efficiency. I show that this tool, or the equivalent excess-profits tax, mitigates the misallocation of resources across firms due to heterogeneous markups by increasing production, particularly for high-markup firms. Unlike traditional firm-specific interventions that require firm-specific information, this policy reduces markups progressively despite being uniform for all firms. In a general equilibrium model with oligopolistic competition and CES demand, I show that the optimal cap replicates the effects of firm-specific price controls, restoring allocative efficiency. In a more general framework, I show that a mix of uniform tax rates—excess-profits, profit, and sales tax—can implement the social optimum. The quantitative estimates of the optimal mix prescribe a positive excess-profits tax, a negative sales tax, and a zero (or even negative) profit tax.

Keywords: concentration, misallocation, optimal taxation.

JEL codes: L1, H2, D4.

*University of Zurich, Department of Economics: gabriele.patete@econ.uzh.ch. I thank Alan Auerbach, Alessandro Ferrari, David Hémous, and Florian Scheuer for their precious and irreplaceable guidance. I am grateful to all the people who, over time, have provided their unique perspectives on the paper, contributing significantly to its development: Carlo Cusumano, Elie Gerschel, Nick Netzer, Claudia Marangon, Virgiliu Midrigan, Massimo Morelli, Lorenzo Pesaresi, Marek Pycia, Steven Raphael, Matteo Saccarola, Emmanuel Saez, Felix Samy Soliman, Armin Schmutzler, Guido Tabellini, Joachim Voth, Roberto Weber, and Danny Yagan. I thank the seminar audiences at the Department of Economics of the University of Zurich, the Department of Economics of the University of California, Berkeley, and the Institute for Research on Labor and Employment in Berkeley for comments that substantially improved the paper. All errors are my own.

1 Introduction

The recent inflationary period substantially raised attention toward firms' profit margins. In many countries, this led to discussions or reforms of the taxes levied on businesses, including standard corporate taxes, mandatory profit-sharing, or excess-profits taxation.[1] At the same time, recent empirical evidence shows that the price markups firms charge above their marginal production costs—especially those of large firms—have grown in the last decades (e.g., Autor et al., 2020; De Loecker et al., 2020). Price markups could be inefficient for various reasons, even when new firms can freely enter the market. First, markups imply that firms produce less than consumers would be willing to buy. Second, they provide distorted incentives for firm or product creation. Third, when firms charge heterogeneous markups, they distort the allocation of production factors across firms, reducing the economy's productivity.

The last problem is particularly daunting for regulators because, so far, previous research has advocated for firm-specific policies—namely, size-dependent output subsidies and cost-dependent price controls—to match the heterogeneity of firms' characteristics (e.g., Baron and Myerson, 1982; Edmond et al., 2023). However, these tools are not easily implementable because of the complexity of the design and the demanding information requirements.[2] In general, public interventions suffer from the trade-off between the need to fit the characteristics of a specific context and the need to be easy to implement. Recent work (Melitz et al., 2024; Nocco et al., 2024) shows that there is no easy solution to this trade-off in the case of misallocation induced by heterogeneous markups: traditional taxes and subsidies cannot address misallocation if they are uniform across firms unless imposing restrictive, non-standard assumptions on demand. The key challenge, therefore, is providing policy-makers with a set of tools that are effective in theory against the *heterogeneous* market power of firms but that also have practical relevance thanks to the simplicity of their design.

In this paper, I propose a *uniform* policy that can, under standard macroeconomic assumptions, address this misallocation of resources: a cap on the profit-to-cost ratio of firms, i.e., a regulation that prevents firms from having too high profits relative to their costs. In partial equilibrium, I show that this regulation induces firms to reduce the markup over their marginal cost and increase production. In addition, the impact of

[1] In addition to proposals to raise standard corporate taxes ("Kamala Harris backs plan to raise US corporate tax rate to 28%"), several countries have adopted mandatory profit-sharing systems (e.g., France in 2023) or excess-profits taxes, mainly on firms in the energy or banking sector, as a reaction to the increase in prices and margins (e.g., all EU Member States after 2022). See Appendix B for a summary of the recent policy debate on these policies.

[2] For example, firm-specific output subsidies require knowledge of the demand structure (Edmond et al., 2023), while firm-specific price controls require knowledge of marginal costs (Baron and Myerson, 1982).

the policy on firm markups is progressive: high-markup firms are more affected than low-markup firms, meaning that this policy reduces the dispersion of markups in the economy. Both these effects increase allocative efficiency. The former decreases the aggregate distortion in the supply of factors of production; the latter improves the allocation of resources across heterogeneous firms. The policy obtains these gains at the cost of smaller firm profits, which are necessary for business creation. Therefore, this regulation is most beneficial when the economy features too many small firms or when the additional entry distortion is more than offset by the efficiency gains. However, I show later that this policy can be coupled with uniform sales and profit subsidies that offset the potential adverse effects on firm entry.

I also show that a planner implementing a uniform cap on the profit-to-cost ratio of firms needs different information relative to the traditional policies that reduce markups progressively. While firm-specific output subsidies and price controls require the observability of firm output or demand, the planner's information set needed to implement the cap on the profit-to-cost ratio consists of the information reported in a firm's financial statement (i.e., *values* of revenues and costs).[3] Therefore, an informationally-constrained planner who does not have access to data on firm-level output or demand prefers to enforce a cap on firms' profit-to-cost ratio. Intuitively, the desirable effects this policy achieves rely on the fact that it does not target revenues and costs separately but their ratio. As long as this ratio is reliably linked to firms' markups, under the regulation, firms increase their production in a way that offsets the firm-specific distortion.

To study the general equilibrium effects of this new policy, I derive the closed-form, optimal profit-to-cost cap in a model of oligopolistic competition with CES demand featuring firms that differ in their productivity and markups. I show that the optimal *uniform* level of the policy restores the social optimum, as characterized by the choices of a planner constrained only by the resource constraint, the technology of production, and the technology of entry. Most importantly, in such a setting, one uniform cap mandated for all firms can replicate the effects of optimal firm-specific price controls. In other words, employing a profit-to-cost cap allows regulators to induce a response equivalent to enforcing a different price cap for each firm without the limitations of doing so directly. The optimal level of the policy induces firms to charge the same constant markup. This is more efficient than the status quo for three reasons. First, it reduces the aggregate markup in the economy. Second, it reduces the dispersion in firm markups. Third, because the decentralized market economy features too many firms compared to the social optimum, the reduction in the level and dispersion of markups comes with no additional distortion in entry. All inefficiencies are targeted at the same time with a single policy tool. In particular, the

[3]In practice, the policy needs to be implemented on a measure of costs that includes the wage bill, intermediaries expenses, materials expenses, and an estimate of the rental cost of capital. Profits are, therefore, given by sales minus the sum of these costs.

optimal cap enforces monopolistic competition pricing in an oligopolistically competitive economy, restoring efficiency as in Dixit and Stiglitz (1977).

As a next step, I relax the assumptions on the market structure (allowing for both monopolistic and oligopolistic competition) and the demand structure (allowing for heterogeneous firm-specific demand), and I show that there exists a closed-form, optimal mix of *uniform* policy rates (profit-to-cost cap, sales tax, profit tax) that enforces the social optimum. The result that the optimal policy rates, despite the heterogeneity of firms and the generality of the demand structure, are uniform is most striking, given that prior research has relied on firm-specific policies to implement the first best. Each tool addresses one of the three types of allocative inefficiency: the implicit output tax due to the aggregate markup, the misallocation of production across firms, and the distorted entry incentives. Still, this result relies on the fact that entry technology depends only on the aggregate economic environment and not on individual firm characteristics. The policy mix, therefore, cannot also address, for example, distortions in firm selection, except in oligopolistic competition with CES demand.[4] The framework I consider nests the one by Edmond et al. (2023), on which I rely to derive estimates of the optimal cap on the profit-to-cost ratio in the context of oligopolistic competition with CES demand and monopolistic competition with Kimball demand. The optimal cap on profits varies between 0.02 and 0.23 of total costs, depending on the market structure.

Finally, I show that regulators can cap the profit-to-cost ratio through a tax on the gap between profits and a given fraction of costs—a specific form of excess-profits taxation. In particular, suppose that regulators do not want firms to have profits higher than 10% of costs (i.e., the cap on the profit-to-cost ratio is 0.1). Then, the equivalent tax is levied on the gap between profits and 10% of costs, where this fraction of costs defines the level of profits with respect to which the excess is computed.[5] If the tax rate is sufficiently high, firms always prefer to increase production instead of paying the tax, adapting their behavior as they would under a profit-to-cost cap. Consequently, the optimal policy mix that I design provides an efficiency-based argument for a comprehensive reform of corporate taxation built around three uniform tools: a profit tax, a sales tax, and an excess-profits tax. When the objective is to offset market power distortions, the main policy recommendation of this paper, as implied by the quantitative estimates, is to rely on positive excess-profits taxes combined with a negative sales tax and a zero (or even

[4]For example, Melitz (2003) models both the ex-ante choice of firms to enter the market, which depends on expected profits, and the ex-post choice of firms to stay on the market after entering, which depends on individual productivity. The second choice is usually referred to as firm selection.

[5]In the public discourse, the term excess-profits tax often means a tax on firm profits net of capital costs to emphasize that taxes on accounting profits do not deduct the implicit cost of capital. In economic models, this would be simply called a profit tax because profits are net of capital costs. As a result, the policy studied in this paper is an excess-profits tax in the economic sense, in that it taxes the excess between economic profits (net of costs including capital costs) and a given fraction of costs. Costs are, therefore, deducted twice from revenues to compute the tax base.

negative) profit tax. Contrary to the traditional, extraordinary implementation of windfall taxes to raise additional revenues, this project highlights a structural role excess-profits taxes can play in correcting allocative inefficiencies, providing a conceptual framework to evaluate recent policy proposals in this direction.

The recent French reform of their mandatory profit-sharing requirement is particularly interesting as an example of profits capping. According to a new regulation introduced on November 23, 2023, a class of firms (with more than 11 employees, not previously subject to mandatory profit-sharing) will be subject to profit-sharing obligations whenever they have after-tax net profits of *at least 1% of turnover* for three consecutive years. Because this requirement is strictly linked to a requirement on the profit-to-cost ratio, implementing this new profit-sharing scheme can be expected to produce desirable, targeted effects on firms. In particular, the higher the penalty for violating the requirement (i.e., in this case, the amount of profits to share), the stronger the incentives to adapt production accordingly. This reform can, therefore, constitute a promising setting for future studies that reinforce the validity of the theoretical results and the practicality of the proposal.

1.1 Relation to literature

This paper is linked to several active strands of the literature. First is the positive and normative research on markups. The studies by De Loecker et al. (2020) and Autor et al. (2020) have documented trends of increasing firm markups and decreasing labor shares in the last decades, as well as an increase in the markup dispersion, driven mainly by larger firms, and a reallocation of production from firms with higher labor share to firms with lower labor share. This evidence has spurred many follow-up analyses that have corroborated such evidence or highlighted methodological issues. In addition, many scholars have explored the welfare costs induced by product market power. Starting from Dixit and Stiglitz (1977) and Mankiw and Whinston (1986), recent studies such as Dhingra and Morrow (2019), Zhelobodko et al. (2012), and Edmond et al. (2023) explore theoretically and quantitatively the role firm heterogeneity plays in generating allocative inefficiencies. For example, Edmond et al. (2023) highlight that, when firms are heterogeneous, subsidizing entry into the market is not effective in decreasing the welfare costs of markups (as opposed to a context with a representative firm, as in Bilbiie et al., 2019). Compared to this existing literature, my paper studies a new policy that effectively constrains firms' market power, i.e., a cap on the profit-to-cost ratio of firms. In addition, it highlights the close link between this regulation and an excess-profits tax, providing an efficiency argument to employ this tool as a structural component of the taxes levied on businesses and offering a framework to evaluate recent policy proposals in this direction.

Second is the recent research on factor misallocation across firms (e.g., Restuccia and

Rogerson, 2008; Hsieh and Klenow, 2009) showing that the dispersion in firm-level revenue TFP generates a misallocation of production factors across firms, decreasing aggregate productivity. Regarding product market power, Baqaee and Farhi (2020) suggest that eliminating the misallocation resulting from dispersed markups would increase value-added aggregate productivity by 20%. Some subsequent analyses have downsized the welfare costs of misallocation. For example, Edmond et al. (2023) suggest that eliminating the misallocation due only to the markup variation systematically correlated with firm size leads to value-added productivity gains not higher than 2%-6%. Melitz et al. (2024) and Nocco et al. (2024) look for uniform policy tools (taxes and subsidies on firms) that can also address the misallocation induced by heterogeneous markups, concluding this is feasible only under restrictive assumptions on demand. In particular, they find that a demand system inducing constant absolute pass-through from marginal costs to prices is both necessary and sufficient for the existence of effective non-discriminatory policies (profit taxes, sales taxes, cost taxes). Compared to this existing literature, my paper studies a policy that is, under standard macroeconomic assumptions, at the same time, non-discriminatory across firms and able to reduce the dispersion of markups, increasing allocative efficiency. In particular, with respect to Melitz et al. (2024) and Nocco et al. (2024), I highlight that, instead of using profit taxes, sales taxes, or cost taxes separately, a tax on the gap between profits and a given share of costs can achieve the objective with a uniform rate.

In addition, my paper complements other studies concerning public policy and market power that instead tackle the redistributional concern, often with an optimal screening problem under asymmetric information. For example, Boar and Midrigan (2023) find that, in a Mirrleesian setting with a utilitarian regulator, the optimal policy reduces wealth and income inequality by redistributing market share to bigger firms. This is because smaller firms are owned privately by rich individuals, while large firms have dispersed ownership. Similarly, Eeckhout et al. (2021) study the optimal, constrained-efficient income taxation in a context where profits reward entrepreneurs' labor effort instead of covering entry costs. Similarly, my paper also complements the classic work of Baron and Myerson (1982). In particular, as opposed to what I do, they solve an optimal screening problem by assuming that regulators cannot observe costs but can observe the demand structure and prices and quantities separately.

The paper is structured as follows: Section 2 analyzes the effects of capping the profit-to-cost ratio of firms on their behavior in partial equilibrium; Section 3 derives the optimal cap on the profit-to-cost ratio of firms in a general equilibrium model with oligopolistic competition and CES demand; Section 4 derives the optimal policy mix in a more general economy, without restrictions imposed on the market structure or the demand structure; Section 5 presents extensions and discussions; and Section 6 concludes.

2 A cap on the profit-to-cost ratio

In this section, I analyze, in a general setting, the effect on firms' production decisions of capping their profit-to-cost ratio.[6] I also compare this regulation to alternative policies, namely output subsidies and price controls, and characterize its comparative advantage as a tool for a planner with information constraints. Finally, I show that a cap on the profit-to-cost ratio can also be implemented with a tax on the gap between profits and a corrected measure of costs.

2.1 Capping the profit-to-cost ratio

Firms.—Consider a price-setting firm indexed by its unit cost of production $c > 0$.[7] The firm chooses quantities $y(c)$ and price $p(c)$ to maximize profits $\pi(c) = p(c)y(c) - cy(c)$. Suppose firms' price choice is constrained by a twice continuously differentiable inverse demand function $\mathrm{p}(y)$ for $y \geq 0$, with $\mathrm{p}(.)$ strictly decreasing in y. At the optimum, it holds $p(c) = \mathrm{p}(y(c))$. The price set by the firm is then given by $p(c) = c/[1 - \epsilon_p(y(c))]$, where $\epsilon_p(y(c)) = -\mathrm{p}'(y(c))y(c)/\mathrm{p}(y(c))$ denoting the elasticity of inverse demand. The markup charged by a firm c is given by $\mu(y(c)) = 1/(1 - \epsilon_p(y(c)))$.

ASSUMPTION 1. (Firm regularity conditions.)

1. Revenues $\mathrm{p}(y(c))y(c)$ are continuous, strictly concave in quantity and satisfy Inada conditions, i.e., $\lim_{y \to 0} [\mathrm{p}(y(c))y(c)]' = +\infty$ and $\lim_{y \to +\infty} [\mathrm{p}(y(c))y(c)]' = 0$.
2. The inverse demand elasticity $\epsilon_p(y(c))$ is bounded between $m > 0$ and $1 - m < 1$.

Assumption 1.1 ensures that there exists a (unique) quantity $y(c)$ that equates the marginal revenues of the firm to its marginal costs and that this is a sufficient condition for optimality. Assumption 1.2 ensures that markups are well-behaved. (Dhingra and Morrow, 2019.)

A cap on the profit-to-cost ratio.—I study the effects of an upper bound $\rho \geq 0$ mandated on the profit-to-cost ratio of firm c. In particular, the following restriction is imposed on the firm profit-maximization:

$$\pi(c, \rho)/cy(c, \rho) = [p(c, \rho)y(c, \rho)/cy(c, \rho)] - 1 \leq \rho. \tag{1}$$

If the restriction binds, Assumption 1.1 ensures that, at the optimum, condition (1) holds with equality: quasi-concavity of profits implies that whenever $\pi(c, \rho)/cy(c, \rho) <$

[6] While Section 2 is devoted to the positive analysis of the effects of the policy on production decisions in a partial equilibrium framework, Section 3 provides a normative analysis of the impact of the policy in a general equilibrium, taking into account entry decisions as well.

[7] All results can be extended under any cost function that has constant elasticity to output, as shown in Appendix A (Extension of Proposition 1b).

ρ, it is always profitable for the firm to bring production closer to the unconstrained optimum. Similarly, it also ensures that the optimal price is given by market inverse demand $p(y(c,\rho))$: whenever the price is lower than the market willingness to pay for it, it is always profitable for the firm to increase production. Therefore, the optimal production level of the firm is characterized by

$$\pi(c,\rho) = \rho c y(c,\rho).$$

The following proposition describes the impact of the policy on firm choices:

PROPOSITION 1. Under Assumption 1, a binding cap on the profit-to-cost ratio implies:

1. An increase in production: $y(c,\rho) > y(c,\infty)$.
2. A progressive reduction in markups: the implied percentage change in markups $\tau(c,\rho)$ is increasing in $\mu(c,\infty)$, with $\mu(c,\rho) = (1-\tau(c,\rho))\mu(c,\infty)$.

Proof. See Appendix A. □

The policy has two main consequences on firm behavior. First, it induces the firm to increase production. Even though the firm could meet the mandate just by lowering its price instead of increasing production, this is never profitable: increasing production allows the firm to expand the maximum level of profits it can achieve when subject to the mandate, compensating partially the loss in earnings due to regulation. In other words, this policy improves allocative efficiency in any economy that is inefficiently small as a result of product market power.

Second, the effect of a cap on the profit-to-cost ratio is larger the higher the markup. Let $\tau(c,\rho)$ be the percentage reduction in markups defined by $\mu(y(c,\rho)) = [1-\tau(c,\rho)]\mu(y(c))$. Proposition 1 shows that $\tau(c,\rho)$ is increasing in $\mu(y(c))$. In an economy populated by firms that are heterogeneous in markups, the policy creates an incentive for the firms for which it binds to uniformize their markups. In the absence of fixed costs of production (or if fixed costs are observable), markups are homogeneous after the implementation of the mandate, i.e., $\mu(y(c,\rho)) = \mu(\rho) = 1+\rho$.[8] This effect improves allocative efficiency by correcting the misallocation induced by heterogeneous markups.[9] Intuitively, given two firms producing two differentiated products at marginal costs c and c', standard Pareto efficiency requires equating the marginal rate of substitution (MRS) to the marginal rate

[8] I show in Proposition 1b how the results in Proposition 1 extend to an environment with fixed production costs.

[9] A misallocation induced by the dispersion of revenue total factor productivity across firms, as described by Restuccia and Rogerson (2008), or Hsieh and Klenow (2009).

of transformation (MRT) for any two given products. Because, in a simple context, the MRS is equal to the ratio of prices, while the MRT is equal to the ratio of marginal costs, there is inefficiency whenever firms charge heterogeneous markups: MRS $= p(c)/p(c') = \mu(c)c/\mu(c')c' \neq c/c' =$ MRT. The introduction of the policy delivers $p(c,\rho)/p(c',\rho) = c/c'$ for any two affected firms c and c', ensuring that a consumer's rate of substitution between the two products equals the rate of transformation of the production system.

Proposition 1 shows how a mandate on the profit-to-cost ratio of firms is not only an effective policy tool to counteract the inefficiencies induced by firms' market power but also a finely tuned policy tool. One policy (one single upper bound ρ) applied to all firms has a customized effect for each firm, targeting its markup level. These effects induce a reallocation of the resources used in production that is welfare-improving.

2.2 Alternative policies

Policy comparison.—I explore now other policy tools that can target firm markups progressively. I focus on the two main types of intervention that can reduce the inefficiencies due to firms' market power, i.e., price controls and output subsidies. Because firms have product market power, the relevant price regulation is a price cap.[10] My main objective is to show that a mandate on the profit-to-cost ratio has an informational advantage compared to the other policies, and an informationally-constrained planner prefers it.

Price controls.—A price cap of the form $p(c,\rho) \leq (1+\rho)c$ enforces uniform markups for the affected firms. It is immediate to notice, therefore, that a mandate on the profit-to-cost ratio replicates the effects of a firm-specific price cap, meaning that implementing this mandate is equivalent to designing and enforcing one targeted price cap for each firm. Such a firm-specific regulation requires observability of the firm marginal cost, and because it holds $p(c,\rho) \leq (1+\rho)c = (1+\rho)[cy(c,\rho)/y(c,\rho)]$, this is equivalent to observability of output (or prices).

REMARK 1. Under Assumption 1, a mandate on the profit-to-cost ratio of a firm can be replicated by a price cap if and only if it is a firm-specific price cap and output is observable.

Proof. See Appendix A. □

Output subsidies.—A firm-specific (additive) subsidy implementing uniform markups across firms can be characterized as follows: $F(y(c,\rho))/(1+\rho) - p(y(c,\rho))y(c,\rho)$, where

[10] While this work focuses on product market power, similar results hold for a context with labor market power when a cap on the profit-to-labor-cost ratio is implemented and compared to a minimum wage.

$F(y(c,\rho)) = \int_0^{y(c,\rho)} p(\xi)d\xi$. Because $p(.)$ is the inverse demand function, the implementation feasibility of these subsidies relies on two observability assumptions. First, the demand structure must be observable. Second, output must be observable.

Therefore, over both categories of alternative policies, a mandate on the profit-to-cost ratio has an informational advantage: achieving a targeted, progressive intervention on firm markups does not require knowledge of the structure of demand or the marginal cost of a firm. In both cases, the observability of output is necessary for implementing firm-specific price regulations or output subsidies.

2.3 Fixed costs of production

As a default, when implementing the policy in practice, profits and costs are computed considering only the variable costs as reported in a firm's financial statement. However, I show that the main results on the effects of the policy are robust to introducing a residual (unobservable) fixed production cost $f \geq 0$, homogeneous across firms.[11] Therefore, firm choices depend on the triple (c, ρ, f). Because it is not possible to disentangle variable and fixed costs, the implementation of a mandate on the profit-to-cost ratio of a firm enforces an optimal production level characterized by

$$\pi(c,\rho,f) - f = \rho[cy(c,\rho,f) + f].$$

Proposition 1 can then be modified as follows:

PROPOSITION 1b. Under assumption 1, when markups are increasing in size, a binding cap on the profit-to-cost ratio implies:

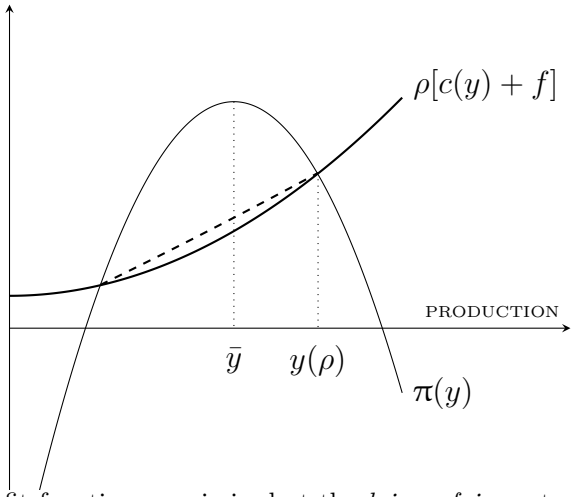
1. An increase in production: $y(c,\rho,0) > y(c,\rho,f) > y(c,\infty,0)$.
2. A progressive reduction in markups: the implied percentage change in markups $\tau(c,\rho,f)$ is increasing in $\mu(y(c,\infty,f))$, with $\mu(y(c,\rho,f)) = [1-\tau(c,\rho,f)]\mu(y(c,\infty,f))$.

Proof. See Appendix A. □

As shown in Figure 1, a cap on the profit-to-cost ratio of a firm induces an incentive to increase production even when the firm features both variable and fixed production costs. In addition, the increase in production is never larger than the one that would be induced if fixed costs were observable, and the regulation could then be applied to variable costs

[11] For example, for publicly-listed firms in the United States included in Compustat, this means that, as a default, the profits and costs relevant for the regulation include the wage bill, intermediaries costs, material costs, and the implied cost of capital but exclude SG&A costs (Sales, General, and Administration), such as R&D, advertising, and executives pay. In Proposition 1b, however, I allow for a residual, unobservable, fixed component of costs f hidden in the reported variable costs. This component is independent of firm size.

FIGURE 1
EFFECTS ON PRODUCTION



NOTE.—The profit function, maximized at the *laissez-faire* output level $\bar{y}$, is given by $\pi(y) = \mathrm{p}(y)y - c(y) - f$, where $c(y)$ is the variable cost function and f is an unobservable fixed cost. The intersection between the profit function and the increasing curve $\rho[c(y) + f]$ determines the output level $y(\rho)$ enforced by capping the profit-to-cost ratio. The dashed curve represents the modified profit function when excess profits are positive under the implementation of the equivalent tax, whose point of maximum determines the output level enforced by the tax. The two policies deliver the same output level.

only. High fixed production costs shield a firm from being subject to such a regulation, lowering its profit-to-cost ratio for the same ρ. This is a conservative result: unobservable fixed costs do not induce additional inefficiencies due to the implementation of the policy; at most, they shield a firm against its effects.

Similarly, the mandate still has a progressive effect on markups, in that a firm featuring higher markup levels absent the policy experiences a larger contraction in the markup after implementing the policy. This result comes from the fact that larger firms also have higher markups. Because the larger the firm, the less effective the shielding power of fixed costs of production, high markup firms are the ones who are also less shielded. In an environment populated by firms heterogeneous in markups, a mandate on the profit-to-cost ratio of a firm implements a customized policy intervention that targets firm margins, reducing their dispersion.[12]

2.4 An equivalent tax

The enforcement of this regulation can be achieved in two ways: directly when compliance is obtained through the threat of a sufficient penalty in case of violation (e.g., criminal prosecution or business shutdown), or indirectly when compliance is achieved by imposing an adequate tax rate on the firm that replicates the same incentives. Let $\hat{c}(y)$ be the

[12] Appendix A shows how while ex-ante markups are increasing in output, ex-post markups are decreasing in output, implying that high-markup firms are the most affected by the policy.

laissez-faire cost function increasing in y, so that profits are $\pi(y) = \mathrm{p}(y)y - \hat{c}(y)$. This equivalent tax can be characterized as an excess-profits tax, as the following Lemma shows:

LEMMA 1. Under Assumption 1, a cap on the profit-to-cost ratio of a firm is implemented by any additive profit tax $T(y,t) = t[\pi(y) - \rho\hat{c}(y)]\mathbb{1}[\pi(y) - \rho\hat{c}(y) > 0]$, with $t \in [1/(1+\rho), 1]$.

Proof. Consider a firm that has any increasing cost function $\hat{c}(y)$ in *laissez-faire*. Suppose by contradiction that, after the introduction of a cap ρ on the profit-to-cost ratio, at the optimum, the firm choice is characterized by $(\tilde{y}, \tilde{c}(.))$, i.e., profits are such that $\mathrm{p}(\tilde{y})\tilde{y} - \tilde{c}(\tilde{y}) = \rho\tilde{c}(\tilde{y})$, with $T(\tilde{y}) = \tilde{c}(\tilde{y}) - \hat{c}(\tilde{y}) > 0$. Because for any y it must always hold $\mathrm{p}(y)y - \hat{c}(y) - T(y) = \rho(\hat{c}(y) + T(y))$, it must be that $T(y) = [1/(1+\rho)][\mathrm{p}(y)y - \hat{c}(y) - \rho\hat{c}(y)]$. As a result, at the optimum, the firm's profits are $\mathrm{p}(\tilde{y})\tilde{y} - \hat{c}(\tilde{y}) - T(\tilde{y}) = [\rho/(1+\rho)]\mathrm{p}(\tilde{y})\tilde{y}$, which is increasing in $\tilde{y}$ under Assumption 1. Therefore, it is not possible that $T(\tilde{y}) > 0$ is optimal. Firms always prefer increasing production rather than increasing costs without producing more. As a special case, therefore, firms never want to pay a tax $T(y,t)$ as characterized by Lemma 1. □

Suppose a firm with variable costs $c(y)$ and a residual unobservable fixed cost f is subject to an excess-profits tax as characterized in Lemma 1. In particular, if $t = 1/(1+\rho)$, net profits are

$$\pi_{\text{tax}}(y) = \mathrm{p}(y)y - c(y) - f - \frac{1}{1+\rho}\left(\mathrm{p}(y)y - c(y) - f - \rho[c(y) + f]\right),$$

whenever $\mathrm{p}(y)y - c(y) - f - \rho[c(y) + f] \geq 0$. Therefore, the following holds:

$$\pi_{\text{tax}}(y) = \begin{cases} \frac{\rho}{1+\rho}\mathrm{p}(y)y, & \text{if } \pi(y) - \rho[c(y) + f] \geq 0 \\ \mathrm{p}(y)y - c(y) - f, & \text{otherwise.} \end{cases}$$

As shown in Figure 1, the firm has the incentive to increase output up to when the gap between $\pi(y)$ and $\rho[c(y) + f]$ closes, i.e., until no taxes need to be paid. The tax changes the firm's profit function such that its new point of maximum coincides with the level of output implied by the cap on the profit-to-cost ratio.

As detailed in the proof, such observations do not only apply to taxes, but they reflect that, under Assumption 1, firms never want to meet the profit-to-cost requirement via cost increases for the same output level; in other words, the firm increases production but does not alter the optimal cost minimization. Examples of this practice include transfers to workers, hiring additional non-productive workers, alterations of the optimal mix of

production factors, or donations. Without this result, a firm could increase transfers, keeping production constant after introducing the policy.

In what follows, I primarily refer to a cap on the profit-to-cost ratio of firms as the policy object of study, but all results apply to the equivalent excess-profits tax as characterized by Lemma 1.

3 Optimal cap in GE oligopolistic competition

So far, I have described the effects of capping the profit-to-cost ratio of a firm in isolation. This section analyzes the impact of the policy in a general equilibrium model of oligopolistic competition. By affecting each firm's pricing choices, the introduction of the policy also affects endogenous factor prices. In addition, reducing firm profits also affects the incentives of firms to enter the market and start producing. I derive a closed-form formula for the optimal cap maximizing the welfare of a representative consumer, given the resource and technological constraints.

In a model with heterogeneous firms that enter an imperfectly competitive market paying a fixed cost of entry, market power induces three allocative inefficiencies. First, it distorts the aggregate supply of production factors. Second, when markups are heterogeneous, it distorts the allocation of factors across firms. Third, it distorts the entry incentives of firms into the market.

A cap on firms' profit-to-cost ratio directly addresses the first two inefficiencies. On the one hand, it incentivizes existing firms to increase production, reducing the aggregate markup. On the other hand, it reduces the dispersion in markups, inducing a desirable reallocation of production factors across firms. In general, however, it leaves entry distorted.

Because this regulation constrains firms' profits, whenever the *laissez-faire* equilibrium features too few firms compared to the social optimum, there exists a trade-off between distorting entry incentives even more while reducing the aggregate markup and the markups dispersion.[13] On the contrary, whenever the *laissez-faire* equilibrium features too many firms compared to the social optimum, this policy addresses all three inefficiencies simultaneously. Oligopolistic competition models belong to this second scenario (e.g., Edmond et al., 2023).[14]

[13] Edmond et al. (2023) suggest that the welfare costs of entry distortions are negligible, while the costs of the aggregate markup and the misallocation of factors of production are significant.

[14] In general, however, when the aggregate markup is high, and factors of production are not supplied inelastically, an economy can also feature aggregate demand externalities. In this context, constraining profits at the firm level increases profits through general equilibrium effects, which increases firms' entry into the market (e.g., for some specifications in Edmond et al., 2023).

3.1 Benchmark model

The decentralized economy is based on Edmond et al. (2023), which allows for normative analysis in a dynamic general equilibrium model with heterogeneous firms and oligopolistic competition.

A representative consumer has preferences over a final consumption good. This good is produced by a perfectly competitive representative firm using inputs from a continuum of sectors. In each sector, imperfectly competitive firms produce differentiated intermediate goods using labor as input. Intermediate firms can be created by paying an irreversible cost of entry. Once this cost has been paid, the new firm receives a one-time productivity draw in a randomly allocated sector. Exit from the market is random, and the economy has no aggregate uncertainty. The representative consumer owns all the firms. The main deviation from Edmond et al. (2023) is that labor is supplied inelastically and is the only factor of production.[15]

Representative consumer.—The representative consumer maximizes

$$\sum_{t=0}^{\infty} \beta^t \log(C_t), \tag{2}$$

subject to

$$C_t = W_t L_t + \Pi_t,$$

where C_t denotes the (numeraire) final consumption good, L_t denotes labor supply, W_t denotes the real wage, $0 < \beta < 1$ denotes the time discount factor, and Π_t denotes aggregate real profits (net of the cost of creating new firms). Labor supply is inelastic; therefore, $L_t = \bar{L} > 0$.

Final-good producer.— The representative firm produces the final good according to the following production technology:

$$Y_t = \left(\int_0^1 y_t(s)^{\frac{\eta-1}{\eta}} ds \right)^{\frac{\eta}{\eta-1}},$$

where Y_t denotes the final-good output, $y_t(s)$ denotes the input from sector $s \in [0, 1]$, and $\eta > 1$ denotes the constant elasticity of substitution (CES) across sectors. In addition, let $P_t = 1$ denote the (normalized) price of the final good, $p_t(s)$ denote the price index for sector s, and $q_t(s) = y_t(s)/Y_t$ denote the relative size of sector s in the economy.

[15] In this section, I consider an environment as in Dixit and Stiglitz (1977) and Dhingra and Morrow (2019), where the focus is on entry distortions and the misallocation of inelastic labor across firms. In Section 4, I discuss the extension to multiple factors of production supplied elastically.

Intermediate-good producers.—In each sector, there are $n_t(s)$ firms, with $n_t(s) \in \mathbb{N}_+$; each firm produces a unique differentiated variety, and it engages in oligopolistic competition *à la Cournot* within the sector. The technology of production of the intermediate good is as follows:

$$y_{it}(s) = z_{it}(s) l_{it}(s),$$

where $y_{it}(s)$ denotes the output of firm i in sector s, $z_{it}(s)$ denotes the productivity of firm i in sector s, and $l_{it}(s)$ denotes labor *employed in production* by firm i in sector s.[16]

Sectoral output $y_t(s)$ is defined by the following production technology:

$$y_t(s) = \left(\sum_{i=1}^{n_t(s)} y_{it}(s)^{\frac{\gamma-1}{\gamma}} ds \right)^{\frac{\gamma}{\gamma-1}},$$

where $\gamma > \eta$ denotes the elasticity of substitution within sectors. Let $q_{it}(s) = y_{it}(s)/y_t(s)$ denote the relative size of firm i in sector s.

Intermediate-good producers maximize

$$\pi_{it}(s) = p_{it}(s) y_{it}(s) - \frac{W_t}{z_{it}(s)} y_{it}(s),$$

where $\pi_{it}(s)$ denotes the profits of firm i in sector s, $p_{it}(s)$ denotes the price of firm i in sector s, and $W_t/z_{it}(s)$ is the marginal cost of firm i in sector s. $p_{it}(s)$ is set considering the final-good producer's inverse demand in a context of oligopolistic competition among firms in sector s. Therefore, at the optimum, a firm's price can be written as a markup $\mu_{it}(s)$ over the marginal cost:

$$p_{it}(s) = \mu_{it}(s) \frac{W_t}{z_{it}(s)}, \; \mu_{it}(s) = \frac{\sigma_{it}(s)}{\sigma_{it}(s) - 1},$$

where $\sigma_{it}(s) = \left([1/\eta] q_{it}(s)^{\frac{\gamma-1}{\gamma}} + [1/\gamma](1 - q_{it}(s)^{\frac{\gamma-1}{\gamma}}) \right)^{-1}$ denotes the demand elasticity to price faced by firm i in sector s.

Most importantly, the ratio of firms' profits to total costs $W_t l_{it}(s)$ can be expressed as:

$$\frac{\pi_{it}(s)}{W_t l_{it}(s)} = \frac{\pi_{it}(s) z_{it}(s)}{W_t y_{it}(s)} = \mu_{it}(s) - 1,$$

[16] Firm-specific productivity is indexed by t even though new firms receive a one-time productivity draw. This is because, over time, the same i identifies different firms with different productivity.

Aggregates.—The following aggregate relationships hold:

$$y_t(s) = z_t(s) l_t(s),$$

where $l_t(s)$ denotes sectoral labor usage, and sectoral productivity is given by $z_t(s) = \left(\sum_{i=1}^{n_t(s)} q_{it}(s)/z_{it}(s)\right)^{-1}$;

$$Y_t = Z_t \tilde{L}_t,$$

where $\tilde{L}_t$ denotes aggregate labor *used in production*, and aggregate productivity is given by $Z_t = \left(\int_0^1 [q_t(s)/z_t(s)]ds\right)^{-1}$.

The misallocation due to heterogeneous markups reduces aggregate productivity, and it is captured by the following relationships:

$$z_t(s) = \left(\sum_{i=1}^{n_t(s)} \left(\frac{\mu_{it}(s)}{\mu_t(s)}\right)^{-\gamma} z_{it}(s)^{\gamma-1}\right)^{\frac{1}{\gamma-1}},$$

$$Z_t = \left(\int_0^1 \left(\frac{\mu_t(s)}{\mathcal{M}_t}\right)^{-\eta} z_t(s)^{\eta-1} ds\right)^{\frac{1}{\eta-1}},$$

where the aggregate markup $\mathcal{M}_t$ and the sectoral markups $\mu_t(s)$ are cost-weighted averages of sectoral and firm-level markups respectively.

Entry and exit.—There is free entry of new firms in the market of intermediate-good producers. Investors can pay a sunk cost κ in units of labor and start up a measure one of firms which, after obtaining a one-time productivity draw $z_{it} \sim G(z)$, will produce a unique new variety of the intermediate good in a randomly allocate sector $s \in [0, 1]$. Let $N_t = \int_0^1 n_t(s)ds$ be the aggregate mass of firms and $M_t = \int_0^1 m_t(s)ds$ be the aggregate mass of entrants. As in Edmond et al. (2023), I assume that entry per sector $m_t(s)$ is IID Poisson with parameter $M_t > 0$ so that each sector has a discrete number of firms. Entrants at time t start producing at time $t + 1$, and, for $j = 0, 1, 2, ...$, they obtain a stream of profits $\pi_{i,t+j}(s)$ for each $t + j$ until they are hit with an IID exit shock, which obtains with probability φ per period. The aggregate mass of firms, therefore, evolves according to $N_{t+1} = (1 - \varphi)N_t + M_t$, and $\mathbb{E}_t n_{t+1}(s) = (1 - \varphi)n_t(s) + \mathbb{E}_t m_t(s)$.

In this environment, net expected profits need to be zero for potential entrants:

$$\kappa W_t = \beta \sum_{j=1}^{\infty} (\beta(1 - \varphi))^{j-1} \frac{C_t}{C_{t+j}} \int_0^1 \bar{\pi}_{t+j}(s)ds, \tag{3}$$

where $\bar{\pi}_{t+j}(s)$ denote expected profits of operating in sector s at time $t+j$.

Equilibrium.—The equilibrium of the decentralized economy is characterized by the following

DEFINITION 1. (Decentralized equilibrium.) Given an initial number of firms $n_0(s) \in \mathbb{N}_+$ per sector, an equilibrium is i) a sequence of firm prices $p_{it}(s)$ and allocations $y_{it}(s)$, $l_{it}(s)$, and ii) aggregate output Y_t, consumption C_t, labor L_t, real wage rate W_t, and mass of entrants M_t such that firms and consumers optimize, the free-entry condition (3) holds with equality, and the goods and the labor market clear at all times t:

i. $Y_t = C_t$;

ii. $L_t = \int \sum_{i=1}^{n_t(s)} l_{it}(s)ds + \kappa M_t$.

Social planner.—The social planner maximizes the welfare of the representative consumer subject to three constraints: i) the resource constraint, ii) the technology of production, and iii) the technology of entry. In particular, the efficient allocation is characterized by the following

DEFINITION 2. (Efficient allocation.) Given an initial number of firms $n_0(s) \in \mathbb{N}_+$ per sector, an efficient allocation is i) a sequence of allocations $y^*_{it}(s)$, $l^*_{it}(s)$, and ii) aggregate output Y^*_t, consumption C^*_t, labor L^*_t, and mass of entrants M^*_t such that:

i. given the aggregate number of firms N^*_t and their distribution $n_t(s)$, the optimal size distributions $q^*_{it}(s)$ and $q^*_t(s)$ are such that $q^*_t(s)$ maximizes Z^*_t subject to $\int_0^1 (q^*_t(s))^{\frac{\eta-1}{\eta}} ds = 1$, and $q^*_{it}(s)$ maximizes $z^*_t(s)$ subject to $\sum_{i=1}^{n_t(s)} (q^*_{i,t}(s))^{\frac{\gamma-1}{\gamma}} = 1$;

ii. given the optimal size distributions $q^*_{it}(s)$ and $q^*_t(s)$, $\{C^*_t\}_{t=0}^{\infty}$ and $\{N^*_{t+1}\}_{t=0}^{\infty}$ maximize (2) subject to the resource constraint $C^*_t \leq Z(N^*_t)(\bar{L}_t - \kappa(N^*_{t+1} - (1-\varphi)N^*_t))$.

3.2 Optimal cap on the profit-to-cost ratio

In what follows, I analyze the effects of a cap on the profit-to-cost ratio of firms. In particular, I derive the closed-form, optimal cap and highlight its informational advantage on alternative policies.

Effect on firm decisions.—Suppose that the following regulation is mandated on firms: the ratio of profits $\pi_{it}(s)$ to costs $W_t l_{it}(s)$ must not exceed a given level $\rho_t \geq 0$. An intermediate-good firm, therefore, maximizes

$$p_{it}(s)y_{it}(s) - \frac{W_t}{z_{it}(s)} y_{it}(s),$$

subject to

$$p_{it}(s) \leq p(y_{it}(s), y_t(s), Y_t),$$

where $p(.)$ characterizes the final-good producer's willingness to pay for firm i's output, given sectoral and aggregate output, and

$$\pi_{it}(s) \leq \rho_t \frac{W_t}{z_{it}(s)} y_{it}(s),$$

which represents the additional constraint implied by the regulation.[17]

If $(1+\rho_t) \leq \mu_{it}(s)$, both constraints bind at the optimum. Indeed, whenever the price $p_{it}(s)$ is strictly lower than the final-good producer's willingness to pay, it is always profitable for a firm to produce more and increase $y_{it}(s)$, as long as the additional units produced are bought, without altering the regulation constraint. In addition, whenever the regulation constraint holds with strict inequality, it is always profitable to cut production $y_{it}(s)$ and increase the price $p_{it}(s)$ such that these changes are compatible with the buyer's willingness to pay, getting closer to the firm optimum under *laissez-faire*.

Therefore, at the optimum, it holds:

$$p_{it}(s) = p(y_{it}(s), y_t(s), Y_t) = \left(\frac{y_{it}(s)}{y_t(s)}\right)^{-\frac{1}{\gamma}} \left(\frac{y_t(s)}{Y_t}\right)^{-\frac{1}{\eta}},$$

which is the same as in *laissez-faire*, and

$$p_{it}(s) = (1+\rho_t)\frac{W_t}{z_{it}(s)},$$

which characterizes the optimal pricing of firm i in sector s after introducing the policy.

Comparing the optimal pricing when the profit-to-cost ratio is capped with the optimal pricing in *laissez-faire*, it is immediate to see that, whenever $\rho_t \leq \mu_{it}(s) - 1$, the new (lower) markup level is given by $(1+\rho_t)$. Therefore, the regulation constrains the markup a firm can set. In this context, the policy increases a firm's incentives to hire labor used to produce its variety of intermediate goods. It does so by introducing a constraint on firms' profits, but, crucially, this constraint also depends on the labor costs. As a result, although the regulation decreases firm profits compared to the *laissez-faire* equilibrium, firms want to ramp up labor (and production), because this allows them to increase the

[17] Note that the final-good producer operates in perfect competition and makes zero profits; therefore, the policy is irrelevant in determining its behavior. Also, note that, in oligopolistic competition with nested-CES demand, the inverse demand satisfies assumption 1. Therefore, the firm does not alter optimal cost minimization.

maximum amount of profits they can achieve. In what follows, I characterize how these effects propagate in a general equilibrium model with heterogeneous firms, oligopolistic competition, and free entry into the market.

Optimal policy.—The following theorem characterizes the optimal cap on the profit-to-cost ratio of firms in an economy featuring a CES technology of production of the consumption good and oligopolistic competition within sectors.

THEOREM 1. There exists a cap $\rho^* > 0$ such that, under $\rho_t = \rho^*$ for all t, the decentralized general equilibrium characterized by Definition 1 is efficient according to Definition 2.

Proof. See Appendix A. □

This result can be decomposed into three parts. First, there exists a cap level ρ_t for all t such that all firms in the economy return to monopolistic-competition pricing, which is $\rho^* = 1/[\gamma - 1]$.[18] This is possible, as shown in Appendix A (Proof of Theorem 1), because monopolistic-competition pricing implies a markup which is a lower bound for the markups in oligopolistic competition, i.e., $\mu^{MC} < \mu_{it}^{OC}(s)$, for all i, s, and t. Therefore, the regulation affects all firms at all times. In this context, the optimal level of the policy eliminates the (negative) consequences of firms' strategic interaction in price formation. In the after-policy optimal pricing, each firm's price choice is independent of other firms' choices.

Second, because firms return to monopolistic-competition prices—which imply constant markups across firms—the cap level ρ^* eliminates misallocation from heterogeneous markups. In that it induces dispersion in the revenue total factor productivity across firms (Hsieh and Klenow, 2009), the heterogeneity of markups is a source of inefficiency in the economy, and this inefficiency is tackled directly by a regulation on the profit-to-cost ratio.

Third, in a context in which the general equilibrium under monopolistic-competition pricing is efficient, the allocation emerging in the decentralized economy after the introduction of the policy is efficient as well, as a result of the fact that a cap ρ^* pushes firms toward monopolistic-competition pricing.[19] This result is, therefore, in the spirit of Dixit and Stiglitz (1977) and Dhingra and Morrow (2019).

Intuitively, in oligopolistic competition, because firms internalize the effect of their choice on sectoral output, they charge too high markups, which allows an excessive number of

[18] In monopolistic competition with CES demand, because firms charge a homogeneous markup over their marginal production cost based on the within-sector elasticity of substitution, optimal firm pricing is given by $p_{it}(s) = \frac{\gamma}{\gamma-1}\frac{W_t}{z_{it}(s)}$.

[19] Note that the efficiency of monopolistic competition requires inelastic labor supply (Dixit and Stiglitz, 1977; Dhingra and Morrow, 2019). Theorem 2 in Section 4 extends the results on the efficiency of the decentralized equilibrium to a context with multiple factors of production supplied elastically.

firms to enter the market compared to the social optimum. As a result, reducing firms' profits up to the monopolistic-competition levels also leads to efficient entry.

Efficiency of monopolistic competition.—Studying the efficiency of monopolistic competition pricing in a context in which the number of firms within a sector is discrete faces some challenges: i) because the law of large numbers does not hold within a sector, establishing a relationship between aggregate variables and expected variables is not immediate; ii) to solve the social planner's problem for the optimal number of firms, differentiability is needed to compute the planner's first order condition.

I derive an efficiency result for monopolistic competition with a discrete number of firms per sector. Appendix A (Proof of Theorem 1) details how these challenges are overcome. In short, i) the law of large numbers still holds if all sectors with the same (discrete) number of firms are grouped. Of these sectors, there exists a continuum, allowing for the application of the law of large numbers to the productivity distribution within this group, which exhibits all possible combinations of productivity draws; ii) as in Edmond et al. (2023), differentiability of aggregate productivity in $n_{t+1}(s)$ for all s and t is retained, even though the number of firms per sector is an integer; however, as a technical remark, I highlight how there must be coherence between how differentiability is retained and the definition of the free-entry condition.

Properties of the optimal cap.—The following corollary highlights some desirable features of the optimal cap on the profit-to-cost ratio of firms.

COROLLARY 1. ρ^* satisfies the following properties: i) it is independent of entry cost κ and exit probability φ; ii) it is invariant to shocks in technology $G(.)$; iii) it is independent of the denomination of prices.

Proof. See Appendix A. □

These robustness properties are not only useful for understanding the effectiveness of the policy, but also for exploring the advantages of this tool compared to other policies.

First, ρ^* does not depend on the structural determinants of market concentration. The interaction of two economic fundamentals ultimately determines market concentration: on the one hand, how costly it is to start up a new business (from a technological point of view); on the other hand, how profitable this business is once it is operating (in other words, how much its products are desirable for consumers, and how costly they are to produce). Market concentration is high when the cost of entry is high for a not-so-profitable product. Market concentration is low when the cost of entry is low for quite a profitable product. The optimal cap is not affected by such structural determinants, but it depends only on the within-sector elasticity of substitution. At the same time, when

these structural features of the economy change, the optimal policy stays the same.

Second, ρ^* does not depend on the productivity distribution within and across sectors. This property highlights that this regulation enforces a targeted intervention without relying on firm-specific characteristics.[20] Indeed, by looking at firms' optimal pricing, it is possible to conclude that the profit-to-cost cap is equivalent to a progressive tax on markups, which homogenizes the markup levels for the firms for which it binds. In addition, ρ^* is also independent of productivity shocks (regardless of whether they affect the whole distribution of firms or just some firms), implying that there is no need to reform it, were the economic environment to change. Often, policy reforms require non-negligible efforts to be implemented—given, for example, the status-quo bias of the political system. In this economy, this policy alleviates such concerns.

Third, an upper bound on the profit-to-cost ratio is not affected by the denomination of prices. Namely, the choice of the numeraire is irrelevant: a mandate expressed in nominal terms has the same behavioral consequences as a mandate expressed in real terms.

3.3 Alternative policies

Policy comparison.—The following result highlights the limitations of enforcing the social optimum using a standard output subsidy or a price control instead of capping the profit-to-cost ratio of firms.

REMARK 2. The social optimum can be enforced by an output subsidy or a price cap only if they are firm-specific and firm output is observable.

The remark highlights two main limitations of standard policies. First, a uniform output subsidy or price control cannot target the misallocation induced by heterogeneous markups. Second, a firm-specific output subsidy or price control requires additional firm-specific information other than revenues and costs, as needed for a cap on the profit-to-cost ratio.

Melitz et al. (2024) and Nocco et al. (2024) show that uniform (multiplicative) output subsidies cannot address the misallocation originating from heterogeneous markups. As a generalization of Edmond et al. (2023), a firm-specific (additive) subsidy implementing uniform markups across firms and enforcing the social optimum can be characterized as follows: $F(q_{it}(s))/(1+\rho^*) - p(q_{it})q_{it}(s)$, where $F(q_{it}(s)) = \int_0^{q_{it}(s)} p(\xi)d\xi$, where $p(.)$ is the inverse demand function, as defined above. Even under the assumption that the demand structure is known, the implementation of this policy requires the observability of prices and output separately.

[20] Note that in such a setting, the observability of the marginal cost of production is equivalent to the observability of productivity or output.

A cap on firms' profit-to-cost ratio implements an intervention targeted to a firm's markup, in that high markup firms are more constrained than low markup firms. It is not possible, therefore, to replicate these effects using a price cap $p_{it}(s) \leq \bar{p}_{it}(s) = \bar{p}_t$, where the cap $\bar{p}_t$ is firm-invariant. To enforce targeted interventions with a price cap, $\bar{p}_{it}(s)$ has to be a function of a firm's marginal cost. Without fixed costs of production, one could then implement $p_{it}(s) \leq (1+\rho^*)[cy_{it}(s)/y_{it}(s)]$, which is, *de facto*, a cap on the price-to-marginal-cost ratio, requiring the observability of output.

4 Optimal policy mix in general equilibrium

In this section, I show how to use a cap on the profit-to-cost ratio to implement the efficient allocation deviating from the assumptions on the market structure or the production technology of the final good from the previous section. In particular, I design an optimal schedule of three *uniform* policies (profit-to-cost cap, sales tax, profit tax) that enforces the social optimum.

The result in Theorem 1 does not crucially rest on labor as the only factor of production or on the favorable environment of oligopolistic competition with CES demand, in which the lower bound of the markup distribution delivers the socially-optimal flow value of varieties, allowing for the simultaneous correction of the factor misallocation and the entry distortion. However, the relaxation of these assumptions forces the social planner to use two additional (uniform) tools, together with a cap on the profit-to-cost ratio: a sales tax τ_s and a profit tax τ_π. First, I present a general framework, nesting the one in Section 3, that relaxes assumptions on the market structure and the production technology. This is a generalized version of Edmond et al. (2023). Second, I characterize the new optimal policy mix. Finally, I present quantitative results on introducing the policy in two main contexts: oligopolistic competition with CES demand and monopolistic competition with Kimball demand.

4.1 Benchmark model

I present the main deviations from the framework in Section 3. The complete description of the decentralized economy is in Appendix A (Proof of Theorem 2.)

Representative consumer.—The first deviation is that I do not impose a specific functional form on consumer preferences. In addition, labor and capital are supplied elastically to firms so that a consumer's utility also internalizes the disutility from working. I also introduce a sales tax and a profit tax that affect a consumer's budget constraint. The representative consumer maximizes

$$\sum_{t=0}^{\infty} \beta^t U(C_t, L_t), \tag{4}$$

subject to

$$(1+\tau_{s,t})(C_t+I_t)=W_tL_t+R_tK_t+\Pi_t-T_t,$$

where C_t denotes the (numeraire) final consumption good, L_t denotes labor supply, $I_t=K_{t+1}-(1-\delta)K_t$ denotes investment, K_t denotes physical capital, δ denotes the depreciation rate, W_t denotes the real wage, R_t denotes the rental rate of capital, $0<\beta<1$ denotes the time discount factor, and Π_t denotes aggregate real profits (net of the cost of creating new firms and net of a profit tax $\tau_{\pi,t}$), $\tau_{s,t}$ is a sales tax on the consumption good and investment, and T_t is a lump sum tax financing $\tau_{s,t}$ and $\tau_{\pi,t}$. Utility $U(.,.)$ is assumed to be strictly increasing and concave in the first argument and strictly increasing and convex in the second argument. Regularity conditions that ensure a well-behaved consumer problem are assumed. [21]

Technology of production.—In the second deviation, I allow for either $n_t(s)\in\mathbb{N}_+$ (oligopolistic competition) or $n_t(s)\in\mathbb{R}_+$ (monopolistic competition). In addition, I do not impose a specific functional form on the within-sector and the between-sector aggregator. In particular, the production technology for the final good is characterized by the between-sector aggregator

$$\int_0^1 \mathcal{A}(s)(q_t(s))ds=1,$$

and the within-sector aggregator

$$\sum_{i=1}^{n_t(s)}\mathcal{B}_i(s)(q_{i,t}(s))=1.$$

where the sum is replaced by an integral if sector s has a continuum of firms of measure $n_t(s)$. Each $\mathcal{A}(s)(.)$ and $\mathcal{B}_i(s)(.)$ is assumed to be increasing and concave in its argument. In addition, the two aggregators are assumed to induce an inverse demand satisfying Assumption 1. The inverse demand is characterized by the following:

$$p_{it}(s)=\frac{\mathcal{A}_q(s)(q_t(s))}{\int_0^1\mathcal{A}_q(s)(q_t(s))q_t(s)ds}\cdot\frac{\mathcal{B}_{q,i}(s)(q_{i,t}(s))}{\sum_1^{n_t(s)}\mathcal{B}_{q,i}(s)(q_{i,t}(s))q_{i,t}(s)}.$$

Differently from Section 3, intermediate producers employ labor, capital, and materials. Therefore, the technology of production for the intermediate goods is as follows:

[21] In particular, utility satisfies Inada conditions, i.e., $\lim_{C\to 0}U_C(C,.)=+\infty$, $\lim_{C\to+\infty}U_C(C,.)=0$, $\lim_{L\to 0}U_L(.,L)=0$, and $\lim_{L\to 6\infty}U_L(.,L)=+\infty$.

$$y_{it}(s) = z_{it}(s)[\phi^{\frac{1}{\theta}} v_{it}(s)^{\frac{\theta-1}{\theta}} + (1-\phi)^{\frac{1}{\theta}} x_{it}(s)^{\frac{\theta-1}{\theta}}]^{\frac{\theta}{\theta-1}},$$

where $z_{it}(s)$ denotes the productivity of firm i in sector s, $x_{it}(s)$ denotes materials used by firm i in sector s, $v_{it}(s)$ denotes value-added by firm i in sector s, and θ is the constant elasticity of substitution between value-added and materials. Value-added is given by:

$$v_{it}(s) = k_{it}(s)^{\alpha} l_{it}(s)^{1-\alpha},$$

where $k_{it}(s)$ is the physical capital employed by firm i in sector s, $l_{it}(s)$ is labor *used in production* employed by firm i in sector s, and α is the constant elasticity of capital to value-added.

As a result, intermediate-good producers' profits are given by

$$\pi_{it}(s) = p_{it}(s) y_{it}(s) - \frac{\Omega_t}{z_{it}(s)} y_{it}(s),$$

where Ω_t is the aggregate index of factor prices, as it results from firms' cost minimization (Appendix A - Proof of Theorem 2), and $\frac{\Omega_t}{z_{it}(s)}$ denotes the marginal production cost.

Technology of entry.—Last, deviating from Section 3, the free-entry condition at time t is characterized as follows:

$$\kappa W_t = \beta \sum_{j=1}^{\infty} (\beta(1-\varphi))^{j-1} \frac{C_t}{C_{t+j}} \int_0^1 (1 - \tau_{\pi,t+j}) \bar{\pi}_{t+j}(s) ds, \tag{5}$$

which is also affected by the profit tax $\{\tau_{\pi,t+j}\}_{j=0}^{\infty}$.

Equilibrium.—The equilibrium of the decentralized economy is characterized by the following

DEFINITION 3. (Decentralized equilibrium.) Given an initial number of firms per sector $n_0(s) \in \mathbb{N}_+$ or $\mathbb{R}_+$ and an aggregate capital stock K_0, an equilibrium is (i) a sequence of firm prices $p_{it}(s)$ and allocations $y_{it}(s)$, $k_{it}(s)$, $l_{it}(s)$,, $x_{it}(s)$ and ii) aggregate output Y_t, consumption C_t, labor L_t, investment I_t, materials X_t, real wage rate W_t, real rental rate R_t and mass of entrants M_t such that firms and consumers optimize, the free-entry condition (5) holds with equality, and the goods, the labor, and capital market clear at all times t:

i. $Y_t = C_t + I_t + X_t$,

ii. $L_t = \int \sum_{i=1}^{n_t(s)} l_{it}(s) ds + \kappa M_t$,

iii. $K_t = \int \sum_{i=1}^{n_t(s)} k_{it}(s)ds$,

iv. $X_t = \int \sum_{i=1}^{n_t(s)} x_{it}(s)ds$,

or the equivalent integral when sectors have a continuum of firms.

Social planner.—The social planner maximizes the welfare of the representative consumer subject to three constraints: i) the resource constraint, ii) the technology of production, and iii) the technology of entry. In particular, the efficient allocation is characterized by the following

DEFINITION 4. (Efficient allocation.) Given an initial number of firms per sector $n_0(s) \in \mathbb{N}_+$ or $\mathbb{R}_+$, an efficient allocation is i) a sequence of allocations $y^*_{it}(s)$, $k^*_{it}(s)$, $l^*_{it}(s)$,, $x^*_{it}(s)$ and ii) aggregate output Y^*_t, consumption C^*_t, labor L^*_t, investment I^*_t, materials X^*_t, and mass of entrants M^*_t such that:

i. given the aggregate number of firms N^*_t and their distribution $n_t(s)$, the optimal size distributions $q^*_{it}(s)$ and $q^*_t(s)$ are such that $q^*_t(s)$ maximizes Z^*_t subject to $\int_0^1 \mathcal{A}(s)(q^*_t(s))ds = 1$, and $q^*_{it}(s)$ maximizes $z^*_t(s)$ subject to $\sum_{i=1}^{n_t(s)} \mathcal{B}_i(s)(q^*_{i,t}(s)) = 1$;

ii. given the optimal size distributions $q^*_{it}(s)$ and $q^*_t(s)$, $\{C^*_t\}_{t=0}^{\infty}$, $\{\tilde{L}^*_t\}_{t=0}^{\infty}$, $\{K^*_{t+1}\}_{t=0}^{\infty}$,, $\{N^*_{t+1}\}_{t=0}^{\infty}$, and $\{X^*_t\}_{t=0}^{\infty}$, maximize (4) subject to the resource constraint $C^*_t + K^*_{t+1} + X^*_t \leq Z(N^*_t)F(K^*_t, \tilde{L}^*_t, X^*_t) + (1-\delta)K^*_t$.

4.2 Optimal policy mix

In the following theorem, I characterize an optimal schedule consisting of three uniform policies that enforces the social optimum.

THEOREM 2. There exist a cap $\rho^*_t > 0$, a profit tax $\tau^*_{\pi,t}$, and sales tax $\tau^*_{s,t}$ for all t such that, under $\{\rho^*_t, \tau^*_{\pi,t}, \tau^*_{s,t}\}_{t=0}^{\infty}$, the decentralized general equilibrium characterized by Definition 3 is efficient according to Definition 4.

Proof. See Appendix A. □

In the first context relevant to the quantitative exercise, oligopolistic competition with CES demand, the optimal policy mix can be characterized as follows:

$$\rho^*_t = \frac{1}{\gamma - 1}, \tag{6}$$

$$\tau^*_{s,t} = -\rho^*_t/[1 + \rho^*_t], \tag{7}$$

$$(1-\tau^*_{\pi,t})\rho^*_t = \frac{1}{\gamma-1}. \tag{8}$$

The first equation describes the optimal cap enforcing monopolistic-competition markups ($\frac{\gamma}{\gamma-1}$). The second equation describes the optimal (negative sales tax) that offsets the residual wedge in the aggregate supply of production factors. The third equation describes the optimal level of the profit tax that delivers optimal entry, and it implies that the optimal profit tax is zero. Because monopolistic-competition pricing delivers optimal entry incentives (e.g., Dixit and Stiglitz, 1977), there is no need for an additional intervention. In the end, the main difference, therefore, with a context in which firms only employ inelastic labor (Section 3) is that the optimal cap on the profit-to-cost ratio needs to be used together with a negative sales tax to offset the aggregate wedge on the supply of factors of production induced by the aggregate markup.

In the second context relevant to the quantitative exercise, I analyze a model of monopolistic competition with Kimball demand (e.g., Klenow and Willis, 2016). In this application, sectors are identical and have a mass of $n_t = N_t$ firms; in addition, the technology of the final good is uniquely characterized by a Kimball aggregator $\int_0^{N_t} \mathcal{A}(q_t(i))ds = 1$, where $\mathcal{A}(.)$ is homogeneous across firms. The optimal policy mix can be characterized as follows:

$$\rho^*_t = \tilde{b} - 1, \tag{9}$$

$$\tau^*_{s,t} = -\rho^*_t/[1+\rho^*_t], \tag{10}$$

$$(1-\tau^*_{\pi,t})\rho^*_t = D^*_t - 1 \tag{11}$$

for any $1 < \tilde{b} \leq \inf\{\mu_{it}(s)\}_{its}$, where D^*_t is the planner's aggregate demand index.[22]

The first equation describes the optimal cap pushing all markups in the economy to the minimum level of the markup distribution before the introduction of the policy. The second equation describes the optimal (negative sales tax) that offsets the residual wedge in the aggregate supply of production factors. The third equation describes the optimal level of the profit tax that delivers optimal entry. Intuitively, because pushing all the markups to the same low level across firms also brings profits to low levels, it is reasonable to expect that a negative profit tax is required to support firms' entry incentives.

In general, to enforce the social optimum, there are three main forces at work. Intuitively,

[22] In this context, the planner's demand index is $D^*_t = \left(\int_0^{N_t} \mathcal{A}_q(q^*_t(i))q^*_t(i)di\right)^{-1}$.

each policy tool addresses one of the three sources of allocative inefficiency of the *laissez-faire* economy: the aggregate wedge in the supply of production factors, the misallocation of production factors across firms, and the distortion in entry incentives.

First, the optimal cap ρ_t^* makes the markups in the economy uniform across firms, eliminating factor misallocation.[23] It also partially reduces the aggregate markup by pushing all firms to produce more. Second, the optimal (negative) sales tax for consumers $\tau_{s,t}^*$ closes the wedge due to the residual aggregate markup, stimulating factor supply. Third, the optimal (negative) profit tax $\tau_{\pi,t}^*$ aligns firms' entry incentives to the socially optimal ones.

In an oligopolistic competition model with CES demand, however, there is no need to rely on the third instrument, namely the profit tax. Because the *laissez-faire* markups are never below the markup level in monopolistic competition, an optimal cap set at $1/(\gamma-1)$ achieves two objectives simultaneously: making the markups homogeneous and enforcing optimal entry incentives. In addition, when labor is the only factor of production, and it is supplied inelastically, there is also no need for the second instrument, namely the sales tax, because monopolistic-competition pricing is efficient (Dixit and Stiglitz, 1977.). Theorem 2, therefore, naturally nests the results and the analysis of Section 3.

In conclusion, it is worth comparing this optimal policy mix with the optimal subsidy derived in Edmond et al. (2023), which also brings the economy to the efficient allocation, addressing all three sources of inefficiency. The comparative advantage of the optimal policy mix characterized in Theorem 2 emerges because it is a non-discriminatory policy under which all firms are subject to the same uniform policy rates. In addition, it relies on the same information set as a profit tax, i.e., revenues $r_{it}(s)$ and costs $c_{it}(s)$: $\frac{\pi_{it}(s)}{c_{it}(s)} = \frac{r_{it}(s)}{c_{it}(s)} - 1 \leq \rho$. On the contrary, the optimal subsidy in Edmond et al. (2023) relies on the knowledge of the demand structure and the observability of prices and quantities separately. However, an optimal output subsidy has the advantage of being unaffected by the presence of unobservable fixed production costs, and it is, therefore, able to address distortions in firm selection, which I do not consider.

In particular, the possibility of relying on three uniform tools to address the three sources of inefficiency (especially the entry distortion) relies on the assumption that entry incentives only depend on the aggregate economic environment, implied by the fact that the equilibrium is determined by one fee-entry condition. The optimal policy mix—in particular, the optimal profit tax—must be differentiated across sectors whenever entry can be directed to specific sectors.

[23] This result relies on assumption 1, which ensures the existence of a lower bound for the markups strictly greater than 1. I show in the extensions (Section 5) how to design an optimal non-discriminatory policy tool that allows even for the relaxation of this assumption.

TABLE 1
OPTIMAL POLICY MIX

	DATA	CASES			
Aggregate markup	$1.1 \sim 1.4$	1.05	1.15	1.25	1.35
Elasticity of substitution within sectors, γ		59.69	12.76	7.16	5.21
Optimal cap, $\boldsymbol{\rho^*}$		**0.02**	**0.09**	**0.16**	**0.23**
Optimal sales tax, $\boldsymbol{\tau_s^*}$		**-0.02**	**-0.08**	**-0.14**	**-0.19**
Welfare (% change)		8.71	14.66	26.76	48.63

NOTES.—The first two rows report calibration targets and estimates from Edmond et al. (2023). Aggregate markups are calibration targets; elasticities of substitution are calibrated parameters. The optimal cap is $\rho^* = 1/[\gamma - 1]$. The optimal sales tax is $\tau_s^* = -\rho^*/[1 + \rho^*]$.

4.3 Quantification: oligopolistic competition with CES demand

Calibration and optimal policy.—I present estimates of the optimal cap on the profit-to-cost ratio and the optimal sales subsidy using calibrated parameters from Edmond et al. (2023).[24] Their estimates rely on data from the US Census of Manufactures (1972-2012), and they assume a multi-factor production of intermediate goods employing labor, capital, and materials.

The optimal policy mix is computed according to equations (6)-(8). As Table 1 shows, the higher the targeted level of the aggregate markup, the less tight the optimal cap on the profit-to-cost ratio of firms. A lower cap is indeed required to constrain smaller markups. At the same time, the higher the targeted level of markups, the higher the optimal sales tax needed to offset the negative wedge in the supply of production factors.

Estimated welfare gains.—The estimated welfare costs of markups imply that the implementation of this policy in an economy with an aggregate markup ranging from 1.05 to 1.35 induces a percentage increase in consumption-equivalent welfare ranging from 8.71 to 48.63.[25] The oligopolistic competition allows firms to charge markups higher than the within-sector elasticity of substitution, so too many firms can survive in the market. In addition, it induces firms to charge heterogeneous markups. Therefore, the optimal cap on the profit-to-cost ratio eliminates all distortions due to imperfect competition: inefficient aggregate markup, misallocation of factors of production, and inefficient entry.

4.4 Quantification: monopolistic competition with Kimball demand

Calibration and optimal policy.—I present estimates of the optimal policy mix using calibrated parameters from Edmond et al. (2023). In this application, sectors are identical and have a mass of $n_t = N_t$ firms; therefore, the technology of the final good is uniquely

[24] For the quantitative exercise $U(C, l) = log(C) - \psi \frac{L^{1+\nu}}{1+\nu}$, where, $\psi > 0$ denotes disutility from labor, $\nu > 0$ denotes the Frisch elasticity of labor supply.

[25] These numbers are computed with a free-entry condition in which the first non-entrant makes zero net expected profits.

TABLE 2
OPTIMAL POLICY MIX

	DATA	CASES			
Aggregate markup	$1.1 \sim 1.4$	1.05	1.15	1.25	1.35
Aggregate demand index, D^*_{ss}		1.06	1.17	1.29	1.41
Optimal cap, $\boldsymbol{\rho^*}$		**0.02**	**0.05**	**0.08**	**0.10**
Optimal sales tax, $\boldsymbol{\tau^*_s}$		**-0.02**	**-0.05**	**-0.07**	**-0.10**
Optimal profit tax, $\boldsymbol{\tau^*_\pi}$		**-1.41**	**-1.85**	**-2.38**	**-3.06**
Welfare (% change)		1.34	8.67	23.63	49.65

NOTES.—The first two rows report calibration targets and estimates from Edmond et al. (2023) for the monopolistic competition model with Kimball demand. Aggregate markups are calibration targets; aggregate demand indexes solve the planner's problem (in the steady state). The optimal cap is the minimum of *laissez-faire* markups. The optimal sale tax is $\tau^*_s = -\rho^*/[1+\rho^*]$. The optimal entry subsidy is such that $(1-\tau^*_\pi)\rho^* = D^*_{\text{ss}} - 1$. Welfare changes are in consumption-equivalent units.

characterized by a Kimball aggregator $\int_0^{N_t} \mathcal{A}(q_t(i))ds = 1$ as in Klenow and Willis (2016). This functional form implies a log-linear demand elasticity $\sigma(q)$:

$$\sigma(q) = \bar{\sigma} q^{-\frac{\epsilon}{\bar{\sigma}}},$$

with $\bar{\sigma} > 1$ and $\epsilon > 0$.

In addition, because the demand elasticity is not always above 1 (i.e., Assumption 1 is not satisfied), I allow firms not to increase production beyond the revenue-maximizing output level, closing the residual gap between profits and capped profits via transfers.

The optimal policy mix is computed according to equations (9)-(11). As Table 2 shows, the smaller the targeted level of the aggregate markup, the tighter the optimal upper bound on the profit-to-cost ratio of firms. A lower cap is indeed required to constrain smaller markups. At the same time, the higher the targeted level of markups, the higher the optimal sales tax needed to offset the aggregate wedge in the supply of production factors. The optimal profit tax is negative and significantly so. Intuitively, as opposed to the oligopolistic competition case, in order to enforce the same homogeneous markup for all firms, the cap significantly compresses firms' profits. As a result, a sizeable negative profit tax is needed to support entry.

Estimated welfare gains.—I also provide estimates of the welfare effects of the implementation of this policy in an economy with an aggregate markup ranging from 1.05 to 1.35, which generates a percentage increase in consumption-equivalent welfare ranging from 1.34 to 49.66. These estimates are the same as those obtained by Edmond et al. (2023) via an optimal, firm-specific output subsidy. Therefore, these results also work as a quantitative counterpart to the theoretical proof in Appendix A of the optimality of the policy as characterized by Theorem 2. In addition, this implies that the violation of assumption 1 (ensuring no alteration of cost minimization) implied by Kimball demand

TABLE 2b
OPTIMAL CAP

	DATA	CASES			
Aggregate markup	$1.1 \sim 1.4$	1.05	1.15	1.25	1.35
Aggregate demand index, D^*_{ss}		1.06	1.17	1.29	1.41
Optimal cap, $\boldsymbol{\rho^*}$		**0.05**	**0.14**	**0.21**	**0.27**
Welfare change (%)		0.60	3.09	7.76	15.56
Uniform subsidy (welfare change)		0.65	5.90	17.36	37.41

NOTES.—The first two rows report calibration targets and estimates from Edmond et al. (2023) for the monopolistic competition model with Kimball demand. Aggregate markups are calibration targets; aggregate demand indexes solve the planner's problem (in the steady state). Welfare changes are in consumption-equivalent units. The last row contains the welfare gains of a uniform subsidy that removes the aggregate markup.

is not quantitatively relevant.

Optimal cap.—One may also wonder what is the welfare-maximizing cap on the profit-to-cost ratio if we restrict a planner's intervention to one tool only, i.e., if output subsidies and profit taxes are zero.

Table 2b reports estimates of the optimal cap for the different specifications and the associated welfare gains. To compare it with a similar, easily implementable policy, I also report the welfare gains of a uniform sales subsidy that removes the aggregate markup, as in Edmond et al. (2023). This is the highest-yielding in terms of welfare gains among the policies they consider. As expected, the cap delivers lower welfare gains (but still sizable) than the uniform sales subsidy, in line with Edmond et al. (2023)'s conclusion that in the monopolistic competition model the aggregate markup distortion contributes more than misallocation to welfare losses.[26]

[26]This is not true in general in oligopolistic competition.

5 Extensions and robustness

This section discusses some extensions of the previous results and their robustness to alternative modeling assumptions. In addition, it discusses potential problems in the implementation of the policy.

5.1 Generalized cap on the profit-to-cost ratio

So far, I have explored the effects of a cap on the profit-to-cost ratio of firms in an environment where firms are price-setters and the regulation is defined on a firm's total costs.

Here I generalize the result on the effects of capping the profit-to-cost ratio of firms without assumptions on the production technology or the market structure. In particular, I also allow for a flexible definition of costs. Such a framework can be used to study the properties of this policy under different market structures (such as price-taking firms or price-setting firms) or different objectives of the planner (such as efficiency or redistribution).

Firms.—Consider a firm that chooses x to maximize profits $\pi(x)$, and $c(x)$ identifies costs of choosing x.

ASSUMPTION 2. (Firm regularity conditions.)

1. $\pi(x)$ is continuous, concave in x and x^* exists such that $\pi(x^*) > \pi(x)$ for every $x \neq x^*$.
2. Unbounded costs: $c(x)$ is continuous, strictly increasing in x, and not bounded above.

Assumption 2.1 ensures that the firm's profit-maximization problem is well-behaved under *laissez-faire*. Assumption 2.2 ensures that the firms' costs are not fixed eventually.[27]

A cap on the profit-to-cost ratio.—I study the effects of a cap $\rho \geq 0$ on the profit-to-cost ratio of a firm. In particular, the following restriction is imposed on the firm profit-maximization: $\pi(x(\rho))/c(x(\rho)) \leq \rho$. If the restriction binds, assumption 2 ensures that at the optimum this relationship holds with equality: concavity of profits implies that whenever $\pi(\rho)/c(x(\rho)) < \rho$, it is always profitable for the firm to push x closer to the unconstrained optimum. Therefore, the optimal choice of the firm is characterized by

$$\pi(x(\rho)) = \rho c(x(\rho)).$$

[27] In other words, the assumptions on the costs ensure that x^* is not independent of costs. For example, under differentiability, if $r(x)$ are the revenues of the firm, it cannot happen that $r'(x^*) = 0$. In addition, when $r(x)$ is strictly increasing in x, no maximum exists if $c(x)$ is eventually fixed.

The following proposition qualifies the impact of the policy on the choices of the firm:

PROPOSITION 2. Under Assumption 2, if $\pi(x^*)/c(x^*) > \rho$, there exists a unique $x(\rho)$ such that $x(\rho) > x^*$.

Proof. Consider a point $x(\rho) > x^*$ such that $\pi(x(\rho)) = \rho c(x(\rho))$. This point exists, and it is unique. It suffices to notice that by concavity of profits, it holds $\pi(x) < \pi(x') < \pi(x^*)$ for every $x > x' > x^*$. Because we have $\pi(x^*) - \rho c(x^*) > 0$, and by monotonicity and unboundedness of costs there exists a $\bar{x} > x^*$ such that $\pi(\bar{x}) - \rho c(\bar{x}) < 0$, we can conclude by continuity of profits and costs that there exists a $x(\rho)$ such that $\pi(x(\rho)) = \rho c(x(\rho))$. In addition, this point is unique for every $x > x$ by concavity of profits and monotonicity of costs.
Note that $x(\rho)$ is optimal. Indeed for any $x > x^*$ such that $\pi(x) < \rho c(x)$, there is always a $x^* < x' < x$ such that $\pi(x') > \pi(x)$ and $\pi(x') \leq \rho c(x')$. We can ignore the points $x < x^*$ because they are never optimal by monotonicity of costs. □

Proposition 2 highlights how, for any definition of costs $c(x)$ that satisfies Assumption 2, a cap on the profit-to-cost ratio increases the incentive of the firm to increase the chosen level of x. Increasing x allows the firm to appropriate more compared to the case in which it decreases its profits leaving $c(x)$ unchanged, e.g., using external donations or transfers to workers.[28]

For example, when $c(x) = c(y)$, where $c(y)$ is the outcome of the firm's cost minimization for the production of output y, this policy makes the firm produce more. Similarly, when $c(x) = wl$, where wl is the firm's wage bill with wage rate w and hired labor l, such policy makes the firm hire more labor.

These results are independent of the firm's pricing as long as Assumption 2 is satisfied. Therefore, they hold both in perfect and imperfect competition. Similarly, they are independent of any specific cost structure (for example, assumptions on the returns to scale or the incidence of fixed costs).

5.2 The optimal profit-to-cost ratio of firms

In this section, I relax Assumption 1, in particular, Assumption 1.2, ensuring that there exists an inferior of the markup distribution greater than 1.

A generalization of a cap on the profit-to-cost ratio of firms (in particular, a policy that combines both an upper bound and a lower bound on the ratio) can implement the social optimum without requiring the existence of an inferior for markups greater than 1 (as

[28]Let $\pi(x) = r(x) - c(x)$. If $r(x)$ is increasing in x, this result holds even when these transfers are included in the definition of costs used in the regulation; in other words, when transfers help the firm meet the mandate not just through lower profits but also higher costs.

implied by assumption 1). In particular, when markups are not bounded below, to bind for all firms in the economy, ρ has to approach 0; firm profits, therefore, go to zero, preventing entry into the market. The generalization also allows the optimal policy mix to employ only two tools: the mandate on the level and a sales tax on consumers, while there is no need for a profit tax.

This version of the policy allows for a discussion of the optimal profit-to-cost ratio of firms in isolation and an answer to the question, "How much should firms earn?"

5.2.1 Optimal non-discriminatory policy

I illustrate the effect of mandating a level $\rho \geq 0$ of the profit-to-cost ratio of firms. Moreover, I characterize its optimal level, which restores the social optimum. The setting is the same as in Section 4, but the within-sector aggregator and the between-sector aggregator are assumed to induce an inverse demand function that satisfies a modified version of Assumption 1:

ASSUMPTION 1b. (Firm regularity conditions.)

1. Revenues $\mathrm{p}(y(c))y(c)$ are continuous, strictly concave in quantity and satisfy Inada conditions, i.e., $\lim_{y\to 0} [\mathrm{p}(y(c))y(c)]' = +\infty$ and $\lim_{y\to+\infty} [\mathrm{p}(y(c))y(c)]' = 0$.
2. The inverse demand elasticity $\epsilon_p(y(c))$ is bounded between 0 and 1.

This modified assumption does not ensure the existence of a minimum for the *laissez-faire* markups larger than 1.

Effect on firm decisions.—I analyze the effects of a regulation mandated on firms. In particular, let $c_{it}(s)$ the marginal cost of firm i in sector s at time t, the ratio of profits $(\pi_{it}(s))$ to costs $(c_{it}(s)y_{it}(s))$ must equate a given level $\rho_t \geq 0$. A (intermediate-good) firm, therefore, maximizes

$$p_{it}(s)y_{it}(s) - c_{it}(s)y_{it}(s),$$

subject to

$$p_{it}(s) \leq p(y_{it}(s), y_t(s), Y_t),$$

where $p(.)$ characterizes the final-good producer's willingness to pay for firm i's output, given sectoral and aggregate output, and

$$\pi_{it}(s) = \rho_t c_{it}(s)y_{it}(s),$$

which represents the additional constraint (at level ρ_t) implied by the regulation.

Note that if $\mu_{it}(s) \neq 1 + \rho_t$, both constraints are binding at the optimum.[29] Indeed, whenever the price $p_{it}(s)$ is strictly lower than the final-good producer's willingness to pay, it is always profitable for a firm to produce more and increase $y_{it}(s)$, as long as the additional units produced are bought.

Therefore, at the optimum, it holds:

$$p_{it}(s) = (1 + \rho_t)c_{it}(s),^{30}$$

which characterizes the optimal pricing of firm i in sector s after introducing the policy.

The pricing equation is, therefore, the same induced by an upper bound on the profit-to-cost ratio of firms. The crucial difference is that a mandate on the level, differently from an upper bound, makes the restriction binding for all firms in the economy rather than just for those with $\mu_{it}(s) > (1+\rho)$. This property is crucial for characterizing the optimal policy that restores the social optimum.

Optimal policy.—The following proposition characterizes the level of the profit-to-cost ratio of firms in an economy without imposing (strong) restrictions on preferences, technology, or the market structure.

THEOREM 2b. There exists a $\rho_t^* > 0$ such that the decentralized general equilibrium under $\rho_t = \rho_t^*$ together with a sales tax $\tau_{s,t} = -\rho^*/[1 + \rho_t^*]$ for all t is efficient.

Proof. See Appendix A. □

For example, in the context of monopolistic competition with Kimball demand (with sectors heterogenous in market concentration $n_t(s)$), the optimal profit-to-cost ratio in the economy is given by $\rho_t^* = \hat{D}_t^* - 1$, with:

$$\hat{D}_t^* - 1 := \frac{\hat{Z}_t^*}{\hat{Z}_{d,t}^*},$$

$$\hat{Z}_{d,t}^* = \left(\int_0^1 (d_t^*(s) - 1) \frac{1}{n_t(s)} q_t^*(s) (z_t^*(s))^{-1} ds \right)^{-1},$$

$$\hat{Z}_t^* = \left(\int_0^1 \frac{1}{n_t(s)} q_t^*(s) (z_t^*(s))^{-1} ds \right)^{-1},$$

[29] Assumption 1 (in particular the Inada condition) ensures the existence of a solution.
[30] Existence of a solution is guaranteed by Inada conditions on firm revenues.

$$d_t^*(s) = \left(\int_0^{n_t(s)} \mathcal{A}_q(q_{it}(s)) q_{it}(s) di \right)^{-1},$$

where $d_t^*(s)$ is the planner's demand index for sector s.[31] Intuitively, $\hat{D}_t^* - 1$ is a corrected measure of the optimal flow value of new firms that a social planner wants to enforce and it ensures that firms have optimal entry incentives.

How much should firms earn?.—The optimal policy derived in Theorem 2b has an attractive, immediate consequence: it allows for a non-ambiguous answer to the question "How much should firms (be allowed to) earn?" I answer this question by characterizing the optimal earnings each firm should have as a function of its costs. This number is indeed constant across firms.

In particular, with Kimball demand, a weighted average of the sectoral demand indexes gives the optimal profit-to-cost ratio of firms in an economy:

$$\rho_t^* = \int_0^1 (d_t^*(s) - 1) \frac{\frac{1}{n_t(s)} q_t^*(s) (z_t^*(s))^{-1}}{\int_0^1 \frac{1}{n_t(s)} q_t^*(s) (z_t^*(s))^{-1} ds} ds.$$

Implementation.—The tax schedule in Section 1 can be adapted to implement a given level ρ of the profit-to-cost ratio of firms.

LEMMA 2. Under assumption 1, a level of the profit-to-cost ratio of a firm is implemented by any additive profit tax $T(t) = t_1[\pi(y) - \rho c(y)]\mathbf{1}(\pi(y) - \rho c(y) \geq 0) + t_2[\rho c(y) - \pi(y)]\mathbf{1}(\pi(y) - \rho c(y) < 0)$, with $t_1 \in [1/(1+\rho), 1]$ and $t_2 \to +\infty$.

To understand how this tax schedule reaches social optimum, it is helpful to compare it with the equivalent optimal policy in Section 4. For such purposes, I will analyze an optimal mix of this additive profit tax and a uniform sales subsidy s_p to producers, equivalent to the consumer negative sales tax. In this way, the effects on both the pricing strategies and the entry incentives are apparent all at once.

First, given the laissez-faire profits $\pi(y)$, the optimal policy mix in the case of oligopolistic competition is such that firm profits $\pi^*(y)$ are as follows:

$$\pi^*(y) = \begin{cases} (1+s_p^*) r(y) - c(y) - \frac{1}{1+\rho^*} [(1+s_p^*) r(y) - c(y) - \rho^* c(y)], & \text{if } \pi(y) - \rho^* c(y) > 0 \\ (1+s_p^*) r(y) - c(y), & \text{otherwise,} \end{cases}$$

where $s_p^* = \rho^*$. The upper bound on the profit-to-cost ratio makes it homogeneous across firms, while the constant sales subsidy closes the gap between prices and marginal costs,

[31] Starred variable refer to the planner's solution.

as usual. This is equivalent to the following:

$$\pi^*(y) = \begin{cases} (1+s_p^*)r(y) - c(y) - \frac{1}{1+\rho}[r(y) - c(y) - \rho c(y)], & \text{if } \pi(y) - \rho c(y) > 0 \\ (1+s_p^*)r(y) - c(y), & \text{otherwise,} \end{cases}$$

where $\rho \to 0$. This schedule works in the context of oligopolistic competition because markups are bounded below. The optimal sales subsidy s_p^* is such that *before* the implementation of the cap on the profit-to-cost ratio, but *after* the implementation of the subsidy, all firms still feature $\pi(y) > \rho c(y)$. In other words, the sales subsidy pushes no firm below marginal-cost pricing.

In the more general context of the current section, however, the introduction of the subsidy, before the implementation of any other policy, pushes some firms (the ones with low *laissez-faire* markups) below marginal-cost pricing, i.e., $\pi(y) < \rho c(y)$. As a result, a cap on the profit-to-cost ratio would be ineffective for these firms, as no threat of taxation is in place.

To push all firms toward the same homogeneous markup, therefore, a penalty for the negative gap between profits and $\rho^* c(y)$ has to be implemented, as follows:

$$\pi^*(y) = \begin{cases} (1+s_p^*)r(y) - c(y) - \frac{1}{1+\rho^*}[(1+s_p^*)r(y) - c(y) - \rho^* c(y)], & \text{if } \pi(y) - \rho^* c(y) > 0 \\ (1+s_p^*)r(y) - c(y) - t_2[-(1+s_p^*)r(y) + c(y) + \rho^* c(y)], & \text{if } \pi(y) - \rho^* c(y) < 0 \\ (1+s_p^*)r(y) - c(y), & \text{otherwise,} \end{cases}$$

where $s_p^* = \rho^*$ and $\lambda \to +\infty$. This extended version naturally nests the one implementing a cap on the profit-to-cost ratio (indeed, it is a double cap). Applying this policy in an oligopolistically competitive market with CES is, then, still optimal: the penalty for negative realizations of the gap just never binds.

5.2.2 Quantification

Calibration and optimal policy.—I present estimates of the optimal level of the profit-to-cost ratio and the optimal (consumer) sales subsidy using calibrated parameters from Edmond et al. (2023).[32] As Table 3 shows, the smaller the targeted level of the aggregate markup, the more tight the optimal level of the profit-to-cost ratio of firms. At the same

[32]The functional form of the Kimball aggregator is from Klenow and Willis (2016). Note that this functional form does not satisfy assumption 1; in particular, Inada conditions are not satisfied. Only firms for which the first-order condition has a zero operate. These firms have productivity $z \geq \frac{\Omega}{D_t} J(0)^{-1}$. This condition is also sufficient to guarantee that the firm's optimal choice exists under the implementation of the optimal policy mix. For any arbitrary ρ, instead, a mandate on the level of the profit-to-cost ratio has a selection effect: only firms for which $z \geq (1+\rho)\frac{\Omega}{D_t} J(0)^{-1}$ operate after the implementation of the policy.

TABLE 3
OPTIMAL POLICY MIX

	DATA	CASES			
Aggregate markup	$1.1 \sim 1.4$	1.05	1.15	1.25	1.35
Aggregate demand index, D^*_{ss}		1.06	1.17	1.29	1.41
Optimal level, $\boldsymbol{\rho^*}$		**0.06**	**0.17**	**0.29**	**0.41**
Optimal sales tax, $\boldsymbol{\tau_s^*}$		**-0.06**	**-0.15**	**-0.23**	**-0.29**
Welfare (% change)		1.34	8.67	23.64	49.66

NOTES.—The first two rows report calibration targets and estimates from Edmond et al. (2023) for the monopolistic competition model with Kimball demand. Aggregate markups are calibration targets; aggregate demand indexes solve the planner's problem (in the steady state). The optimal cap is $\rho^* = D^*_{\text{ss}} - 1$. The optimal sale tax is $\tau_s^* = -\rho^*/[1+\rho^*]$. Welfare changes are in consumption-equivalent units.

time, the higher the targeted level of markups, the higher the optimal sales tax needed to offset the aggregate wedge in the supply of production factors.

Estimated welfare gains.—I also provide quantitative estimates of the welfare gains of eliminating the markups in the context of monopolistic competition with Kimball demand, implying that the implementation of this policy in an economy with an aggregate markup ranging from 1.05 to 1.35 may generate a percentage increase in consumption-equivalent welfare ranging from 1.34 to 49.66. Again, these estimates are the same as those obtained by Edmond et al. (2023) via an optimal, firm-specific output subsidy. Therefore, these results also work as a quantitative counterpart to the theoretical proof in Appendix A of the optimality of the policy as characterized by Theorem 2b.

Comparing these results with the estimates in Section 3, one can notice that, while the optimal policy is such that markups are homogeneous, the optimal profit-to-cost ratios in Table 3 are always higher than the target aggregate markup of the *laissez-faire* decentralized economy. On the contrary, in Table 1, the opposite is true. This is because, in the specification of the monopolistic competition model, the decentralized equilibrium features too few firms with respect to the efficient allocation. Therefore, the equilibrium flow value of varieties (and, therefore, the expected profits of entrants) must increase to sustain more firms entering the market.

5.3 Alternative modelling assumptions

Production technology.— To retain a result on the efficiency of monopolistic competition in *laissez-faire* as in Section 3, CES technology in the production of the final good must be assumed. As discussed in Appendix A (Extension of Proposition 1b), however, the progressivity result relies on the sufficient assumption of constant elasticity of costs to output.

Type of market power.—This work analyzes a context characterized by product market power, in which firms can set output prices. The results, however, also hold in an

environment featuring labor market power, i.e., an environment in which firms can set wages if we implement a profit-to-labor-cost ratio. If labor is the only factor of production, the two policies are equivalent. At the same time, the results also extend to oligopolistic competition *à la Bertrand*.

Sectoral free entry.— This work employs a definition of free entry according to which investors can freely start a measure one of new firms that produce a new differentiated variety of an intermediate good in a randomly allocated sector after receiving a random productivity draw (similarly to Edmond et al., 2023). In other words, entry per sector is not directed. Sectoral net expected profits are, therefore, not zero in general. This assumption is not relevant to the effects of the policy on misallocation but on entry incentives because it allows the existence of a common policy tool that also corrects entry distortions. With sector-specific free-entry conditions, heterogeneous (negative) profit taxes must be used.

Heterogeneous profit shares.—This work does not allow heterogeneous profit shares across firms that come from technology (e.g., heterogeneous DRS). Still, it includes heterogeneous profit shares that come from imperfect competition. One way to microfound the heterogeneity of DRS (and, therefore, to microfound the fact that some firms have a technological reason to make more profits) is the heterogeneity in sectoral entry costs when entry can be directed to specific sectors, as in Atkeson and Burstein (2008). When there are, instead, heterogeneous DRS across firms, the progressivity result can break when high-markup firms also have sufficiently higher returns to scale.

5.4 Policy discussion

5.4.1 Capital owned by the firm

The (extended) model used in the normative analysis features firms renting the capital used in production. In practice, firms own equity, which finances productive capital. A correct measure of profits and costs has to consider this, internalizing the implicit capital cost $(r + \delta)K$ evaluated at the market interest rate r (compensating for depreciation δ). Nimier-David et al. (2024) analyze the case of the mandatory profit-sharing rule in force in France since 1967, in which implicit capital costs are deducted from the accounting profits to share, and provide evidence that optimal capital accumulation is not distorted, as evidence of a correct deduction.

5.4.2 Mergers and acquisitions

While vertical integration is not an effective avoidance strategy, horizontal mergers between a high-markup firm and a low-markup firm may decrease their combined profit-to-cost ratio. For conglomerates of firms operating in different sectors, this policy should

be implemented at a firm's subdivision level, for which firms record different financial statements. Mergers within a sector, instead, are likely to be less problematic. A within-sector merger favors the increase in the markups of all the products the merging firms sell. This happens because of the increase in collusion in the sectoral market and technological synergies that develop after the merging. In this sense, a mandate on the profit-to-cost ratio may also favor welfare-improving, within-sector mergers, in which no increases in the markups charged accompany the productivity gains due to the mergers. As a result, this intervention may also be valuable for competition authorities, not just fiscal authorities, as an additional requirement to be imposed in the context of a merger's conditional approval. Alternatively, this regulation can act as a substitute for competition policy. Once the regulation is in place, firms can merge without requiring additional government interventions.

5.4.3 Fixed costs

The normative analysis of Sections 3 and 4 relies on a class of models featuring fixed entry costs of production but no residual (unobservable) fixed production cost. As shown in Section 2, in an extended model in which firms are also burdened with residual fixed production costs, a cap on the profit-to-cost ratio of firms still has a progressive effect on markups when markups are increasing in firm size.[33]

Implementing the policy, as a default, the measures of profits and costs relevant for the regulation exclude those costs reported as fixed costs in a firm's financial statement, such as R&D, advertising, or executives pay.[34] This is important because, although the after-policy incentive to increase production is robust to any preference or fixed cost structure, it is still true that introducing heterogeneity of demand systems and heterogeneity of fixed costs across differentiated products can break the progressivity result.[35] In addition, when these observable, reported fixed costs are excluded, even if the residual unobservable component of fixed costs hidden in the reported variable costs is independent of size within a sector (by definition, an exogenous fixed cost of production as it is usually defined in macroeconomic models, e.g., in Melitz (2003), has to be orthogonal to size), it may be heterogeneous across sectors. These arguments call for implementing this policy at the

[33]To the best of my knowledge, I am not aware of models—theoretical or quantitative—that can deal at the same time with oligopolistic competition, fixed costs, and entry costs. Edmond et al. (2023) have free entry and oligopolistic competition but drop firm selection. Atkeson and Burstein (2008) have fixed costs and oligopolistic competition but drop entry costs (the free-entry condition is ex-post and defined on fixed costs). Melitz (2003) has fixed and entry costs but does not consider oligopolistic competition.

[34]For publicly-listed firms in the US included in Compustat this implies excluding the costs reported in SG&A (Sales, General, and Administration).

[35]For example, in a context in which bigger firms also charge higher markups before the introduction of the policy and are subject to an unobservable fixed cost, a necessary condition for the heterogeneity of demand systems to break the progressivity is that the introduction of the policy inverts the sales ranking of firms, so that the ex-ante high-sale, high-markup firm is ex-post smaller in sales than the ex-ante low-sale, low-markup firm.

sector level.[36] To conclude, it is worth noticing that, even when the policy is applied to a measure of profits and costs that also include, for example, R&D costs, the policy still might have a progressive effect on markups if the implied profit-to-cost ratio of firms is anyway reliably linked to their markups. De Loecker et al. (2020) suggest that, even though high-markup firms also have higher R&D and advertising as a share of revenues, they also exhibit higher profit rates and markups.

6 Conclusion

Firm market power is a source of allocative inefficiency in the economy that has recently received much attention. In particular, when firms charge heterogeneous markups, production factors are misallocated across firms, reducing aggregate productivity. Because the market cannot self-regulate in this context, correcting these allocative inefficiencies is within the scope of a public regulator.

Firm heterogeneity, however, makes this objective challenging for a social planner. Because of the vast heterogeneity across firms, which also results in heterogeneity in the markups that firms charge over their marginal production costs, optimal policies are generally differentiated. The traditional tools to decrease firms' market power in a targeted way achieve this fine-tuning at the cost of significant complexity and information requirements. In particular, firm-specific price controls (such as product-price caps or minimum wages) and output subsidies cannot progressively decrease firms' markups (so that high-markup firms are more affected than low-markup firms) without firm-specific information that is usually unavailable to a planner.

This paper studies the effects of a regulation mandating a cap on the profit-to-cost ratio of firms as a tool to address these concerns. In Section 2, I first show that introducing this regulation incentivizes the affected firms to increase production, offsetting the output wedge due to more-than-marginal-cost pricing. I then show that under standard macroeconomic assumptions, this intervention reduces firm markups progressively. Most importantly, this result requires the information reported in a firm's financial statement (i.e., revenues and different cost categories) without knowledge of firm-specific unobservable variables such as quantities or inverse demand. Both price controls and output subsidies require knowledge of firm-level quantities to replicate similar results. Output subsidies, in addition, require knowledge of the demand structure.

In Section 3, I show that, in a dynamic general equilibrium model of oligopolistic competition, when the only factor of production is labor supplied inelastically, there is an

[36] An implementation of the policy at the national level may introduce a trade-off between decreasing the markup dispersion within sectors and increasing the markup dispersion across sectors. In addition, when entry is directed to a specific sector, there may be heterogeneous elasticities of substitution across sectors.

optimal cap on the profit-to-cost ratio of firms that implements the efficient allocation in a decentralized economy. In particular, introducing this policy is equivalent to designing and enforcing a different price cap for each firm, proportional to its marginal cost.

The proposed policy tackles two sources of inefficiency: the aggregate wedge in the supply of factors of production (due to the aggregate markup in the economy) and the misallocation of factors of production across firms (due to heterogeneous markups). On the one hand, it pushes firms to charge lower markups; on the other, it pushes firms to charge less dispersed markups. However, this result is obtained at the cost of smaller profits, discouraging firm entry into the market (which may but need not reduce welfare).

In Section 4, I also design an optimal schedule consisting of three uniform policies (a profit-to-cost cap, a sales tax, and a profit tax) that allows the planner to implement the efficient allocation regardless of the underlying market structure (monopolistic or oligopolistic competition) or the demand structure (not just constant elasticity of substitution). Because a cap on the profit-to-cost ratio is equivalent to a well-designed tax on excess profits, this result provides an efficiency-based argument for a comprehensive reform of corporate taxation built around three uniform tax rates (on excess profits, profits, and sales). In particular, the main policy recommendation can be summarized in three parts: i) relying on excess-profits taxes (i.e., on the gap between profits and a corrected measure of costs) to correct factor misallocation; ii) relying on a negative VAT tax to correct the residual wedge in the supply of production factors; iii) relying on a zero (or even negative) profit tax to support firm entry.

Finally, this paper does not explore the consequences of introducing this policy on R&D investments by firms, maintaining, however, that there is bound to be ambiguity on this relationship depending on the specific assumptions on the R&D technology. However, when the regulation is not used in isolation but as part of an optimal policy mix, the measures taken to support firms' entry incentives, for example, cutting the profit tax, also sustain innovation incentives against the adverse effects of capping profits. Nevertheless, the interaction between a firm's profit-to-cost ratio and its innovation incentives is a promising starting point for future research on the effects of the proposed policy. Similarly, this paper does not explore the effects of the policy on redistribution and inequality. A cap on the profit-to-cost ratio on its own achieves non-negligible welfare gains at the expense of profits, as opposed to output subsidies that require the economical and political feasibility of subsidizing (especially large) firms.

References

1. Atkeson, Andrew, and Ariel Burstein. “Pricing-to-Market, Trade Costs, and International Relative Prices.” The American Economic Review 98, no. 5 (2008): 1998–2031.
2. 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 135, no. 2 (2020): 645–709.
3. Baqaee, David R., and Emmanuel Farhi. "Productivity and Misallocation in General Equilibrium". The Quarterly Journal of Economics 135, no. 1 (2020): 105–163.
4. Baron, David P., and Roger B. Myerson. "Regulating a Monopolist with Unknown Costs". Econometrica 50, no. 4 (1982): 911-930.
5. Bilbiie, Florin O., Fabio Ghironi, and Marc J. Melitz. "Monopoly Power and Endogenous Product Variety: Distortions and Remedies." American Economic Journal: Macroeconomics 11, no. 4 (2019): 140–74.
6. Boar, Corina, and Virgiliu Midrigan. "Markups and Inequality." Forthcoming Review of Economic Studies, forthcoming Review of Economic Studies.
7. De Loecker, Jan, Jan Eeckhout, and Gabriel Unger. "The Rise of Market Power and the Macroeconomic Implications." The Quarterly Journal of Economics 135, no. 2 (2020): 561–644.
8. Dixit, Avinash K., and Joseph E. Stiglitz. “Monopolistic Competition and Optimum Product Diversity.” The American Economic Review 67, no. 3 (1977): 297–308.
9. Dhingra, Swati, and John Morrow. "Monopolistic Competition and Optimum Product Diversity under Firm Heterogeneity." Journal of Political Economy 127, no. 1 (2019): 196-232.
10. Eeckhout, Jan, Chunyang Fu, Wenjian Li, and Xi Weng. "Optimal Taxation and Market Power."
11. Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu. "How Costly Are Markups?." Journal of Political Economy 131, no. 7 (2023): 1619-1675.
12. Hsieh, Chang-Tai, and Peter J. Klenow. “Misallocation and Manufacturing TFP in China and India.” The Quarterly Journal of Economics 124, no. 4 (2009): 1403–48.
13. Klenow, Peter J., and Jonathan L. Willis. "Real Rigidities and Nominal Price Changes." Economica, 83 no. 331 (2016): pages 443-472.

14. Mankiw, N. G., and M. D. Whinston. “Free Entry and Social Inefficiency.” The RAND Journal of Economics 17, no. 1 (1986): 48–58.

15. Melitz, Marc J. “The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity.” Econometrica 71, no. 6 (2003): 1695–1725.

16. Melitz, Marc J., Gianmarco I.P. Ottaviano, Mikhail Oshmakashvili, and Davide Suverato. "Markup Distortions and Optimal Non-discriminatory Industrial Policy".

17. Nocco, Antonella, Gianmarco I.P. Ottaviano, Matteo Salto, and Atsushi Takadoro. "Leaving the Global Playing Field through Non-discriminatory Corporate Taxes and Susbidies."

18. Restuccia, Diego, and Richard Rogerson. "Policy Distortions and Aggregate Productivity with Heterogeneous Plants." Review of Economic Dynamics, Elsevier for the Society for Economic Dynamics, 11, no. 4 (2008): 707-720.

19. Zhelobodko, Evgeny, Sergey Kokovin, Mathieu Parenti, and Jacques-François Thisse. “Monopolistic Competition: Beyond the Constant Elasticity of Substitution.” Econometrica 80, no. 6 (2012): 2765–84.

Appendix

A Omitted proofs

Proof of Proposition 1.

1.1 Note that we assume $\mathrm{p}'(y) < 0$. Note also that $\mu(c,\rho) = 1+\rho < \mu(c,\infty)$ because we assume the mandate is binding. Therefore, $\mathrm{p}(y(c,\rho)) < \mathrm{p}(y(c,\infty))$, which implies $y(c,\rho) > y(c,\infty)$.

1.2 It is enough to notice that the markup after the implementation of a cap on the profit-to-cost ratio is given by: $p(c,\rho)/c = \mathrm{p}(y(c,\rho))y(c,\rho)/cy(c,\rho) = 1+\rho$. Therefore it holds $(1-\tau(c,\rho)) = \mu(y(c,\rho))/\mu(y(c)) = (1+\rho)/\mu(y(c))$, which implies that $\tau(c,\rho)$ is increasing in $\mu(y(c))$.

Proof of Remark 1.

If. Suppose we can design a firm-specific price cap, and output is observable. We can then implement for each firm c $p(c,\rho) \leq (1+\rho)c = (1+\rho)(cy(c,\rho)/y(c,\rho))$ for any $\rho \geq 0$. Then such restriction is equivalent to $\pi(c,\rho)/cy(c,\rho) \leq \rho$.

Only if. Suppose we can design a price cap that replicates the restrictions imposed by a mandate on the profit-to-cost ratio of a firm. Such a price cap is $p(c,\rho) \leq \bar{p} = (1+\rho)c$, which is firm-specific because the threshold $\bar{p}$ is a function of the firm-specific marginal cost, and output must be observable because it is equivalent to the ratio of two observable variables $cy(c,\rho)/c$.

Proof of Proposition 1b.

Note $\pi(c,\rho,f)/(cy(c,\rho,f)+f) = \rho$ implies:

$$\frac{py(c,\rho,f)}{cy(c,\rho,f)+f} = 1+\rho$$

$$\frac{cy(c,\rho,f)+f}{py(c,\rho,f)} = \frac{1}{1+\rho}$$

$$\frac{1}{\mu(y(c,\rho,f))} + \frac{f}{p(c,\rho,f)y(c,\rho,f)} = \frac{1}{1+\rho}$$

Because $[\pi(c)-f]/[cy(c)+f] > \rho$, it must be $\mu(y(c)) > \mu(y(c,\rho,f))$. In addition, $\mu(y(c,\rho,f)) > 1+\rho = \mu(y(c,\rho))$. Since markups are increasing in output, it holds $y(c,\rho) > y(c,\rho,f) > y(c)$. In addition, demand elasticity is always larger than 1, this

implies that revenues are increasing in output. Ex-post markups are, therefore, decreasing in output. Because ex-ante markups are increasing in output and because ex-post markups are always lower than ex-ante markups, the implied tax on markups is increasing in ex-ante markups.

Extension of Proposition 1b.

I prove all the results on the effects of the introduction of a cap on the profit-to-cost ratio (existence and uniqueness of a solution; binding constraints; increase in production; and progressive reduction on markups) in a context in which the elasticity of variable costs to output is constant (still allowing for the presence of fixed costs of production.) Variable costs are given by the twice continuously differentiable and strictly increasing function $c(y)$ for $y > 0$, and the constant elasticity of variable costs to output is $\frac{c'(y)y}{c(y)} = \epsilon_c$. The fixed cost is given by $f \geq 0$ and is constant across firms. To describe firm heterogeneity, I still index the productivity of a firm by c, meaning that for two firms (c_a, c_b) with $(c_a < c_b)$, it holds $c'_a(y) < c'_b(y)$ for all y, with $c_a(0) = c_b(0) = 0$.

When firms do not have a unitary cost of production, the Inada conditions must be replaced to ensure existence and uniqueness of a solution of the firm problem, so that Assumption 1 becomes:

ASSUMPTION 1 - MODIFIED. (Firm regularity conditions.)

1. Profits $\pi(y(c)) = \mathrm{p}(y(c))y(c) - c(y(c)) - f$ are continuous, strictly concave in quantity and satisfy $\lim_{y \to 0} \pi'(y) > 0$ and $\lim_{y \to +\infty} \pi'(y) < 0$.

2. The inverse demand elasticity $\epsilon_p(y(c))$ is bounded between $m > 0$ and $1 - m < 1$.

Existence and Uniqueness.—First, given a binding $\rho > 0$, there exists a $y(c, \rho, f)$ such that $\pi(y(c, \rho, f)) = \rho[c(y(c, \rho, f)) + f]$. At the unconstrained optimum, it holds $\pi(y(c, \infty)) > \rho[c(y(c, \infty)) + f]$. Note that by assumption 1.1 (modified), because profits are strictly concave and have a negative derivative at the limit, it holds that $\pi(y)$ is decreasing in y for any $y > y(c, \infty)$. Therefore, because $c(y)$ is strictly increasing in y, there must exist a $\bar{y} > y(c, \infty)$ such that $\pi(\bar{y}) < \rho[c(\bar{y}) + f]$. By continuity of $\pi(y) - \rho[c(y) + f]$, therefore, there exists a $\bar{y} > y(c, \rho, f) > y(c, \infty)$ such that $\pi(y(c, \rho, f) = \rho[c(y(c, \rho, f)) + f]$. Because $\pi(y)$ is strictly decreasing in y and $c(y)$ is strictly increasing in y, such point is unique for $y \geq y(c, \infty)$. I disregard the range $y < y(c, \infty)$ because an intersection of $\pi(y)$ and $\rho[c(y) + f]$ in this lower range delivers lower profits and is always dominated by the profits implied by the intersection in the upper range.

Binding constraints.—I define a profit function $\pi(p, y) = py - c(y) - f$ that characterizes the profits of the firm for arbitrary p and y. The firm has to satisfy two constraints: $p \leq \mathrm{p}(y)$ and $\pi(p, y) \leq \rho[c(y) + f]$. Suppose by contradiction that at least one of the

two constraints is not binding, i.e., a firm chooses $\tilde{p}$ and $\tilde{y}$ such that either $\tilde{p} < \mathrm{p}(\tilde{y})$ or $\pi(\tilde{p}, \tilde{y}) < \rho[c(\tilde{y}) + f]$. Note that it holds $\pi(p, y) \leq \pi(y)$ for any p and y because $\mathrm{p}(y) \geq p$ by definition and $\pi(p, y)$ is increasing in p. Since it must always hold that $\pi(p, y) \leq \rho[c(y) + f]$, for any $y(c, \infty) < y < y(c, \rho, f)$ it holds $\pi(p, y) \leq \rho[c(y) + f] < \rho[c(y(c, \rho, f)) + f] = \pi(y(c, \rho, f))$; therefore this levels of output will not be chosen. In addition, for any $y > y(c, \rho, f)$, $\pi(p, y) \leq \pi(y) < \pi(y(c, \rho, f))$, because $\pi(y)$ is decreasing in y. As a result, $(\tilde{p}, \tilde{y})$ cannot be optimal unless $\tilde{y} = y(c, \rho, f)$. Given $y = y(c, \rho, f)$, profit maximization implies $\tilde{p} = \mathrm{p}(y(c, \rho, f))$ because $\pi(y(c, \rho, f)) \geq \pi(p, y(c, \rho, f))$ for any p and this is feasible by construction. Therefore, at the optimum $(\tilde{y}, \tilde{p})$ are such that both constraints are binding.

Production increase.—As shown proving existence, because the intersection in the lower range $y < y(c)$ is never optimal (because costs are strictly increasing and $\pi(y(\rho, c, f)) = \rho[c(y(c, \rho, f)) + f])$, $y(c, \rho, f)$ is chosen, and $y(c, \rho, f) > y(c, \infty)$.

Progressive reduction in markups.—Note that $\pi(y(c, \rho, f))/[c(y(c, \rho, f) + f] = \rho$ implies:

$$\frac{\mathrm{p}(y(c, \rho, f))y(c, \rho, f)}{c(y(c, \rho, f)) + f} = 1 + \rho$$

$$\frac{c(y(c, \rho, f)) + f}{p(y(c, \rho, f))y(c, \rho, f)} = \frac{1}{1 + \rho}$$

$$\frac{c'(y(c, \rho, f))}{\mathrm{p}(y(c, \rho, f))} \frac{c(y(c, \rho, f))}{c'(y(c, \rho, f))y(c, \rho, f)} + \frac{f}{p(c, \rho, f)y(c, \rho, f)} = \frac{1}{1 + \rho}$$

$$\frac{1}{\mu(y(c, \rho, f))} \frac{1}{\epsilon_c} + \frac{f}{p(c, \rho, f)y(c, \rho, f)} = \frac{1}{1 + \rho}$$

Note that revenues are increasing in output, which implies that more productive firms have lower markups after the introduction of the policy. Indeed, take two firms c_a and c_b, with $c_a < c_b$. If for firm b it holds:

$$\mathrm{p}(y(c_b, \rho, f))y(c_b, \rho, f) = (1 + \rho)[c_b(y(c_b, \rho, f)) + f],$$

for firm a it holds:

$$\mathrm{p}(y(c_b, \rho, f))y(c_b, \rho, f) > (1 + \rho)[c_a(y(c_b, \rho, f)) + f].$$

Therefore, because profits are strictly decreasing and the costs are strictly increasing in

quantities, we have that $y(c_a, \rho, f) > y(c_b, \rho, f)$.

Finally, progressivity requires a higher tax rate on markups for firm c_a:

$$(1 - \tau_a) = \frac{\mu(y(c_a, \rho, f))}{\mu(y(c_a, \infty))} < \frac{\mu(y(c_b, \rho, f))}{\mu(y(c_b, \infty))} = (1 - \tau_b)$$

which is equivalent to

$$\frac{\mu(y(c_b, \infty))}{\mu(y(c_a, \infty))} < \frac{\mu(y(c_b, \rho, f))}{\mu(y(c_a, \rho, f))},$$

ensured by the fact that the more productive firm has higher ex-ante markups, but lower ex-post markups.

Proof of Theorem 1.

Structure of the proof.

We need to show that the decentralized equilibrium under the policy level $\rho^* = \frac{1}{\gamma - 1}$ is (constrained) efficient. First, we show that ρ^* implies monopolistic-competition pricing. Then, we show that the decentralized equilibrium under monopolistic-competition pricing with a finite number of firms per sector is (constrained) efficient.

ρ^* implies monopolistic-competition pricing.

Note that in equilibrium in the oligopolistic model, when the policy is not in place, it holds

$$\mu_{it}^{\mathrm{OC}}(s) = \frac{\sigma_{it}(s)}{\sigma_{it}(s) - 1} = \frac{1}{\frac{\gamma-1}{\gamma} - (\frac{\gamma-1}{\gamma} - \frac{\eta-1}{\eta}) q_{it}(s)^{\frac{\gamma-1}{\gamma}}} > \frac{\gamma}{\gamma - 1} = \mu_{it}^{\mathrm{MC}}(s)$$

because $q_{it}(s)^{\frac{\gamma-1}{\gamma}} > 0$, where $\mu_{it}^{\mathrm{OC}}(s)$ is markup for firm i, in sector s, at time t under oligopolistic competition, and $\mu_{it}^{\mathrm{MC}}(s)$ is markup for firm i, in sector s, at time t under monopolistic competition. In addition, we know that, whenever $\rho < \mu_{it}(s)$, optimal pricing of firms is given by $p_{it}(s) = (1 + \rho)\frac{W_t}{z_{it}(s)}$. Therefore, under $\rho^* = \frac{1}{\gamma - 1}$, we have $p_{it}(s) = \frac{\gamma}{\gamma - 1}\frac{W_t}{z_{it}(s)} = \mu_{it}^{\mathrm{MC}}(s)\frac{W_t}{z_{it}(s)}$. We can then conclude that ρ^* enforces monopolistic-competition pricing, and, therefore, the decentralized equilibrium of an economy in which the Maximum Ratio is in place at this level is equivalent to the decentralized equilibrium of an economy where a finite number of firms per sector do not interact strategically in profit maximization and choose optimal prices as they were ignoring the impact of their choices on sectoral variables.

At this point, we need to prove that the equilibrium conditions of the decentralized equilibrium under monopolistic competition are the same as the optimality conditions in

the social planner problem. This proof builds on Edmond et al. (2023).

Monopolistic competition is efficient: preliminary results.

Throughout the proof, decentralized equilibrium variables are not marked by a star (without $*$), and social planner's variables are marked by a star (with $*$). Given a sector s with $n_t(s)$ firms and productivity levels $\{z_{it}(s)\}_{i=1}^{n_t(s)}$ We write $\underline{z}_{t,-i}(s)$ for the vector of productivity levels of the $n_t(s)-1$ firms excluding i. Also, we write $\mu = \frac{\gamma}{\gamma-1}$. Remember that $Z_t = \left(\int_0^1 q_t(s)\frac{1}{z_t(s)}\right)^{-1}$ and $z_t(s) = \left(\sum_{i=1}^{n_t(s)} q_{it}(s)\frac{1}{z_{it}(s)}\right)^{-1}$. Because under homogeneous markups for a given distribution $n_t(s)$ we have $z_t(s) = z_t^*(s) = \left(\sum_1^{n_t(s)} z_{it}(s)^{\gamma-1}\right)^{\frac{1}{\gamma-1}}$ and $Z_t = Z_t^* = \left(\int_0^1 z_t(s)^{\eta-1}\right)^{\frac{1}{\eta-1}}$, we know that it holds:

$$q_{it}(s)^{-\frac{1}{\gamma}} = \frac{p_{it}(s)}{p_t} = \frac{\mu\frac{W_t}{z_{it}(s)}}{\mu\frac{W_t}{z_t(s)}} = \frac{z_t(s)}{z_{it}(s)} = q(z_{it}(s), \underline{z}_{-i}(s), n_t(s))$$

$$q_{it}^*(s)^{-\frac{1}{\gamma}} = \frac{z_t(s)}{z_{it}(s)} = q(z_{it}(s), \underline{z}_{-i}(s), n_t(s))$$

$$q_t(s)^{-\frac{1}{\eta}} = \frac{p_t(s)}{P_t} = \frac{\mu\frac{W_t}{z_t(s)}}{\mu\frac{W_t}{Z_t}} = \frac{Z_t}{z_t(s)} = Q(\underline{z}_t(s), n_t(s), \{\underline{z}_t(s), n_t(s)\}_s)$$

$$q_t(s)^{-\frac{1}{\eta}} = \frac{Z_t}{z_t(s)} = Q(\underline{z}_t(s), n_t(s), \{\underline{z}_t(s), n_t(s)\}_s)$$

where X_t summarizes aggregates that are the same for all firms in all sectors, and noticing that functions $q(.)$ and $Q(.)$ are the same for the decentralized equilibrium and the social planner's problem, meaning that knowing sector sizes and productivity distribution you compute relative sizes in the same way. We also define conveniently $\tilde{q}_{it}(s) = q_{it}(s)q_t(s) = \tilde{q}(z_{it}(s), \underline{z}_{t,-i}(s), n_t, X_t)$.

These equivalences allow us to write, for given N_t and $\{n_t(s)\}_s$:

$$Z_t = Z_t^* = \left(\int_0^1 \sum_{i=1}^{n_t(s)} \tilde{q}(z_{it}(s), \underline{z}_{t,-i}(s), n_t(s), X_t)\frac{1}{z_{it}(s)}ds\right)^{-1}$$

Now, we notice that since $n_t(s)$ is IID distributed according to a pdf Pr(.) over $n_t(s) \in \{0, 1, 2, ...\}$ with parameter N_t[37], there is a measure $\Pr(n_t; N_t)$ of sectors characterized by exactly $n_t(s)$ firms. Therefore, in a sector with $n_t(s)$ firms, there has been $n_t(s)\Pr(n_t(s); N_t)$

[37]E.g., if entrants per sector $m_t(s)$ are IID distributed as a Poisson pdf with parameter M_t, $n_t(s)$ is a sum of IID Poisson random variables, and it is therefore distributed as a Poisson with parameter $N_t = \sum_{s=0}^{t-1}(1-\varphi)^{t-1-s}M_t + (1-\varphi)^t N_0$

productivity draws, or better, there have been $\Pr(n_t(s); N_t)$ draws of productivity vectors $\underline{z}_t$ of dimension $n_t(s)$. As a result, we can establish a relationship between aggregates and expectations applying the law of large numbers within a sector with $n_t(s)$ firms, as follows[38]:

$$
\begin{aligned}
Z_t = Z_t^* = & && (12)\\
&= \left(\int_0^1 \sum_{i=1}^{n_t(s)} \tilde{q}(z_{it}(s), \underline{z}_{t,-i}(s), n_t, X_t) \frac{1}{z_{it}(s)} ds \right)^{-1} && (13)\\
&= \left(\sum_{n_t=0}^{\infty} \int_{\{s:n_t(s)=n_t\}} \sum_{i=1}^{n_t(s)} \tilde{q}(z_{it}(s), \underline{z}_{t,-i}(s), n_t, X_t) \frac{1}{z_{it}(s)} ds \right)^{-1} && (14)\\
&= \left(\sum_{n_t=0}^{\infty} \int_{\{s:n_t(s)=n_t\}} \sum_{i=1}^{n_t} \tilde{q}(z_{it}(s), \underline{z}_{t,-i}(s), n_t, X_t) \frac{1}{z_{it}(s)} ds \right)^{-1} && (15)\\
&= \left(\sum_{n_t=0}^{\infty} \sum_{i=1}^{n_t} \int_{\{s:n_t(s)=n_t\}} \tilde{q}(z_{it}(s), \underline{z}_{t,-i}(s), n_t, X_t) \frac{1}{z_{it}(s)} ds \right)^{-1} && (16)\\
&= \left(\sum_{n_t=0}^{\infty} \sum_{i=1}^{n_t} \int ... \int \Pr(n_t; N_t) \tilde{q}(z_t, \underline{z}_{t,-z}, n_t, X_t) \frac{1}{z} dG_{n_t} \right)^{-1} && (17)\\
&= \left(\sum_{n_t=0}^{\infty} n_t \int ... \int \Pr(n_t; N_t) \tilde{q}(z_t, \underline{z}_{t,-z}, n_t, X_t) \frac{1}{z} dG_{n_t} \right)^{-1} && (18)\\
&= \left(\sum_{n_t=0}^{\infty} \Pr(n_t; N_t) \int ... \int n_t \tilde{q}(z_t, \underline{z}_{t,-i}, n_t, X_t) \frac{1}{z} dG_{n_t} \right)^{-1}, && (19)
\end{aligned}
$$

or, equivalently:

$$Z_t = Z_t^* = \left(\sum_{n_t=0}^{\infty} \Pr(n_t; N_t) \tilde{z}^{-1}(n_t) \right)^{-1}$$

where $\tilde{z}^{-1}(n_t) = \int ... \int n_t \tilde{q}(z_t, \underline{z}_{t,-i}, n_t, X_t) \frac{1}{z} dG_{n_t}$.

Lastly, we conclude that, for the same N_t, $Z_t = Z_t^* = Z(N_t)$.

Standard free-entry

Monopolistic competition is efficient: free entry.

Assuming, $M_t > 0$ for all t, the free-entry condition is defined as follows:

[38] Note that moving from the third to the fourth step (40-41) we are summing over sectors keeping fixed the same i. Therefore, we are summing over $\Pr(n_t; N_t)$ independent draws of the productivity vector.

$$\kappa W_t = \beta \sum_{j=1}^{\infty} (\beta(1-\varphi))^{j-1} \frac{C_t}{C_{t+j}} \int_0^1 \bar{\pi}_{t+j}(s) ds$$

where $\bar{\pi}_{t+j}$, the expected profits of operating at time $t+j$ in sector s

$$\bar{\pi}_{t+j}(s) = \int ... \int \pi_{t+j}(z_{i,t+j}(s), \underline{z}_{t+j}(s), n_{t+j}(s)+1, X_{t+j}) dG_{n_{t+j}(s)+1}(z_{i,t+j}(s), \underline{z}_{t+j}(s))$$

where $(z_{i,t+j}(s), \underline{z}_{t+j(s)}, n_{t+j}(s)+1)$ identifies a potential entrant firm i with productivity $z_{i,t+j}(s)$, operating in sector s, with other $n_{t+j}(s)$ firms characterized by productivity levels $\underline{z}_{t+j}(s)$.

We can, therefore, write

$$\int_0^1 \bar{\pi}_{t+j}(s) ds = \sum_{n_{t+j}=0}^{\infty} \Pr(n_{t+j}; N_{t+j}) \int ... \int \pi_{t+j}(z_{t+j}, \underline{z}_{t+j}, n_{t+j}+1, X_{t+j}) dG_{n_{t+j}+1}(z_{t+j}, \underline{z}_{t+j}),$$

We can, therefore, establish the following relationship:

$$\begin{aligned}
\int_0^1 \bar{\pi}_{t+j}(s) ds = & \qquad (20)\\
= & \sum_{n_{t+j}=0}^{\infty} \Pr(n_{t+j}; N_{t+j}) \int ... \int \pi_{t+j}(z_{t+j}, \underline{z}_{t+j}, n_{t+j}+1, X_{t+j}) dG_{n_{t+j}} \qquad (21)\\
= & \sum_{n_{t+j}=0}^{\infty} \Pr(n_{t+j}; N_{t+j}) \int ... \int (\mu-1) \frac{W_{t+j}}{z} y(z_{t+j}, \underline{z}_{t+j}, n_{t+j}+1, X_{t+j}) dG_{n_{t+j}} \qquad (22)\\
= & (\mu-1) W_{t+j} \sum_{n_{t+j}=0}^{\infty} \Pr(n_{t+j}; N_{t+j}) \int ... \int \frac{1}{z} y(z_{t+j}, \underline{z}_{t+ji}, n_{t+j}+1, X_{t+j}) dG_{n_{t+j}} \qquad (23)\\
= & (\mu-1) W_{t+j} Y_{t+j} \sum_{n_{t+j}=0}^{\infty} \Pr(n_{t+j}; N_{t+j}) \int ... \int \frac{1}{z} \tilde{q}(z_{t+j}, \underline{z}_{t+j}, n_{t+j}+1, X_{t+j}) dG_{n_{t+j}} \qquad (24)\\
= & (\mu-1) W_{t+j} Y_{t+j} \sum_{n_{t+j}=0}^{\infty} \Pr(n_{t+j}; N_{t+j}) \frac{1}{(n_{t+j}+1)} \tilde{z}(n_{t+j}+1)^{-1} \qquad (25)\\
= & (\mu-1) W_{t+j} Y_{t+j} (\hat{Z}^{+}_{t+j})^{-1} \qquad (26)
\end{aligned}$$

Note that indeed, $\hat{Z}^{+}_{t+j}(N_t) \neq Z_{t+j}(N_t)$ because of the weighting factors $\frac{1}{(n_{t+j}+1)}$ and because there is one additional firm in each sector. The free-entry condition is then given

by:

$$\kappa W_t = \beta \sum_{j=1}^{\infty} (\beta(1-\varphi))^{j-1} \frac{C_t}{C_{t+j}} (\mu - 1) W_{t+j} Y_{t+j} (\hat{Z}^+_{t+j})^{-1}$$

Note that, because $W_{t+j} = \frac{Z_{t+j}}{\mu}$ for all j, we finally have:

$$\kappa Z_t = \beta \sum_{j=1}^{\infty} (\beta(1-\varphi))^{j-1} \frac{C_t}{C_{t+j}} (\mu - 1) \frac{Z_{t+j}}{\hat{Z}^+_{t+j}} Y_{t+j}$$

Monopolistic competition is efficient: social planner's choice of aggregate number of firms.

The planner's choice of the aggregate number of firms $\{N_{t+j}\}_{j=1}^{\infty}$ is given by:

$$\kappa W^*_t = \beta \sum_{j=1}^{\infty} (\beta(1-\varphi))^{j-1} \frac{C^*_t}{C^*_{t+j}} \frac{dZ^*_{t+j}}{dN^*_{t+j}} \frac{1}{Z^*_{t+j}} Y^*_{t+j}$$

and the following holds from the static allocation problem:

$$q^*_{it}(s)^{-\frac{1}{\gamma}} = \frac{z^*_t(s)}{z_{it}(s)}$$

$$q^*_t(s)^{-\frac{1}{\eta}} = \frac{Z^*_t}{z_t(s)}$$

As in Edmond et al. (2023), $\frac{dZ^*_{t+j}}{dN^*_{t+j}} \frac{1}{Z^*_{t+j}}$ is given by an application of the envelope theorem, which requires ignoring the integer constrain when differentiating with respect to $n_{t+j}(s)$. In particular, an increase of measure 1 in the aggregate number of firms N_{t+j} induces a flow $\epsilon(s)$ of firms within each sector s, with $\mathbb{E}(\epsilon(s)) = 1$. This additional firms in sector s, therefore, induce a gain in aggregate productivity equal to $\frac{dZ^*_{t+j}}{dn_{t+j}(s)} \epsilon(s)$. The expected gain in aggregate productivity induced by an increase of measure 1 in the aggregate number of firms is given by the (expected) total derivative:

$$\frac{dZ^*_{t+j}}{dN^*_{t+j}} \frac{1}{Z^*_{t+j}} = \mathbb{E}_\epsilon \left(\int_0^1 \frac{dZ^*_{t+j}}{dn_{t+j}(s)} \epsilon(s) \frac{1}{Z^*_{t+j}} ds \right) = \int_0^1 \frac{dZ^*_{t+j}}{dn_{t+j}(s)} \mathbb{E}_\epsilon(\epsilon(s)) \frac{1}{Z^*_{t+j}} ds = \int_0^1 \frac{dZ^*_{t+j}}{dn_{t+j}(s)} \frac{1}{Z^*_{t+j}} ds$$

Edmond et al. (2023) show that, applying the envelope theorem (and ignoring the integer constraint):

$$\frac{dZ^*_{t+j}}{dn_{t+j}(s)} \frac{1}{Z^*_{t+j}} = (\mu - 1) \frac{1}{n_{t+j}(s)} q^*_{t+j}(s) \frac{Z^*_{t+j}}{z^*_{t+j}(s)}$$

For such derivation to be consistent with our definition of the free-entry condition, the integer constraint must be relaxed such that, in the derivative, sector-level variables also contain the additional firm the sector gets on average:

$$\frac{dZ^*_{t+j}}{dn_{t+j}(s)}\frac{1}{Z^*_{t+j}} = (\mu-1)\frac{1}{(n_{t+j}(s)+1)}q^{+,*}_{t+j}(s)\frac{Z^*_{t+j}}{z^{+,*}_{t+j}(s)}$$

Therefore:

$$\frac{dZ^*_{t+j}}{dN^*_{t+j}}\frac{1}{Z^*_{t+j}} = \tag{27}$$

$$= \int_0^1 \frac{dZ^*_{t+j}}{dn_{t+j}(s)}\frac{1}{Z^*_{t+j}}ds \tag{28}$$

$$= Z^*_{t+j}(\mu-1)\int_0^1 \frac{1}{(n_{t+j}(s)+1)}q^{+,*}_{t+j}(s)(z^{+,*}_{t+j}(s))^{-1}ds \tag{29}$$

$$= Z^*_{t+j}(\mu-1)\sum_{n_{t+j}}\Pr(n_{t+j};N^*_{t+j})\frac{1}{(n_{t+j}+1)}\int\dots\int(n_{t+j}+1)\tilde{q}^*(z_{t+j},\underline{z}_{t+j},n_{t+j}+1,X_{t+j})\frac{1}{z}dG_{n_{t+j}+1} \tag{30}$$

$$= Z^*_{t+j}(\mu-1)\sum_{n_{t+j}}\Pr(n_{t+j};N^*_{t+j})\frac{1}{(n_{t+j}+1)}\tilde{z}^*_{t+j}(n_{t+j}+1) \tag{31}$$

$$= (\mu-1)\frac{Z^*_{t+j}}{\hat{Z}^{+,*}_{t+j}} \tag{32}$$

The social planner's choice can then be expressed by:

$$\kappa W^*_t = \beta\sum_{j=1}^{\infty}(\beta(1-\varphi))^{j-1}\frac{C^*_t}{C^*_{t+j}}(\mu-1)\frac{Z^*_{t+j}}{\hat{Z}^{+,*}_{t+j}}Y^*_{t+j}$$

Because $W^*_{t+j} = Z^*_{t+j}$ for all j, we finally have[39]:

$$\kappa Z^*_t = \beta\sum_{j=1}^{\infty}(\beta(1-\varphi))^{j-1}\frac{C^*_t}{C^*_{t+j}}(\mu-1)\frac{Z^*_{t+j}}{\hat{Z}^{+,*}_{t+j}}Y^*_{t+j}$$

We have shown that the free-entry condition is equivalent to the social planner' optimal condition for the aggregate number of firms, when evaluated at the same aggregate variables.

The definition of the free-entry condition and the mode in which the integer constraint is ignored in computing the derivative of aggregate productivity to sectoral concentration need to be consistent. In the subsection "Adjusted free-entry," I discuss a free-entry

[39]Note that this result also holds with elastic labor supply.

condition consistent with evaluating the derivative at initial values.

Monopolistic competition is efficient: equilibrium conditions.

We now have to compare the equilibrium conditions of the decentralized problem to the equilibrium conditions of the social planner's problem. In the decentralized equilibrium, equilibrium conditions are as follows:

$$\kappa Z_t = \beta \sum_{j=1}^{\infty} (\beta(1-\varphi))^{j-1} \frac{C_t}{C_{t+j}} (\mu - 1) \frac{Z_{t+j}}{\hat{Z}^+_{t+j}} Y_{t+j}$$

$$C_t = Y_t$$

$$Y_t = Z_t \tilde{L}_t$$

$$\tilde{L}_t = L_t - \kappa(N_{t+1} - (1-\varphi)N_t)$$

$$L_t = \bar{L}$$

In the social planner's problem, the equilibrium conditions are:

$$\kappa Z^*_t = \beta \sum_{j=1}^{\infty} (\beta(1-\varphi))^{j-1} \frac{C^*_t}{C^*_{t+j}} (\mu - 1) \frac{Z^*_{t+j}}{\hat{Z}^{+,*}_{t+j}} Y_{t+j}$$

$$C^*_t = Y^*_t$$

$$Y^*_t = Z^*_t \tilde{L}_t$$

$$\tilde{L}^*_t = L^*_t - \kappa(N^*_{t+1} - (1-\varphi)N^*_t)$$

$$L^*_t = \bar{L}$$

Because, as previously shown, for the same N_t, $Z_t = Z(N_t) = Z^*_t$ (and, similarly, $\hat{Z}^+_t = \hat{Z}^+(N_t) = \hat{Z}^{+,*}_t$), this implies that the decentralized equilibrium is efficient.

Note that, when labor supply is elastic, the last equilibrium condition is $\psi C_t L_t^{\nu} = \frac{Z_t}{\mu}$ in the decentralized equilibrium and $\psi C_t^* L_t^{*\nu} = Z_t^*$ in the social planner's equilibrium. Aggregate labor is, therefore, inefficient in the decentralized equilibrium. To restore efficiency, combining ρ^* and a uniform wage subsidy τ to workers paid by lump sum transfers is enough. In particular, $(1+\tau) = \mu = (1+\rho^*)$, such that the last equilibrium condition of the decentralized problem becomes $\psi C_t L_t^{\nu} = (1+\tau) W_t = \frac{(1+\tau) Z_t}{\mu} = Z_t$.

Adjusted free-entry

Monopolistic competition is efficient: free entry.

Assuming, $M_t > 0$ for all t, the free-entry condition is defined as follows:

$$\kappa W_t = \beta \sum_{j=1}^{\infty} (\beta(1-\varphi))^{j-1} \frac{C_t}{C_{t+j}} \int_0^1 \bar{\pi}_{t+j}(s) ds$$

where $\bar{\pi}_{t+j}$, the expected profits of operating at time $t+j$ in sector s

$$\bar{\pi}_{t+j}(s) = \int ... \int \pi_{t+j}(z_{i,t+j}(s), \underline{z}_{t+j,-i}(s), n_{t+j}(s), X_{t+j}) dG_{n_{t+j}(s)}$$

and we can, therefore, write

$$\int_0^1 \bar{\pi}_{t+j}(s) ds = \sum_{n_{t+j}=0}^{\infty} \Pr(n_{t+j}; N_{t+j}) \int ... \int \pi_{t+j}(z_{i,t+j}, \underline{z}_{t+j,-i}, n_{t+j}, X_{t+j}) dG_{n_{t+j}},$$

meaning that, at the margin, new entrants at time $t+j$ that bring the aggregate number of firms to N_{t+j} have an ex-ante probability of $\Pr(n_{t+j}; N_{t+j})$ to end up in a sector with a total number of firms (including themselves) equal to n_{t+j}.

Assuming that i) the last entrants at the margin have zero expected profits and that ii) investors correctly anticipate equilibrium future market concentration, we have the following relationship:

$$\int_0^1 \bar{\pi}_{t+j}(s)ds = \tag{33}$$

$$= \sum_{n_{t+j}=0}^{\infty} \Pr(n_{t+j}; N_{t+j}) \int ... \int \pi_{t+j}(z_{i,t+j}, \underline{z}_{t+j,-i}, n_{t+j}, X_{t+j}) dG_{n_{t+j}} \tag{34}$$

$$= \sum_{n_{t+j}=0}^{\infty} \Pr(n_{t+j}; N_{t+j}) \int ... \int (\mu - 1)\frac{W_{t+j}}{z} y(z_{i,t+j}, \underline{z}_{t+j,-i}, n_{t+j}, X_{t+j}) dG_{n_{t+j}} \tag{35}$$

$$= (\mu - 1)W_{t+j} \sum_{n_{t+j}=0}^{\infty} \Pr(n_{t+j}; N_{t+j}) \int ... \int \frac{1}{z} y(z_{i,t+j}, \underline{z}_{t+j,-i}, n_{t+j}, X_{t+j}) dG_{n_{t+j}} \tag{36}$$

$$= (\mu - 1)W_{t+j}Y_{t+j} \sum_{n_{t+j}=0}^{\infty} \Pr(n_{t+j}; N_{t+j}) \int ... \int \frac{1}{z} \tilde{q}(z_{i,t+j}, \underline{z}_{t+j,-i}, n_{t+j}, X_{t+j}) dG_{n_{t+j}} \tag{37}$$

$$= (\mu - 1)W_{t+j}Y_{t+j} \sum_{n_{t+j}=0}^{\infty} \Pr(n_{t+j}; N_{t+j}) \frac{1}{n_{t+j}} \tilde{z}(n_{t+j})^{-1} \tag{38}$$

$$= (\mu - 1)W_{t+j}Y_{t+j}\hat{Z}_{t+j}^{-1} \tag{39}$$

Note that indeed, $\hat{Z}_{t+j} \neq Z_{t+j}$ because of the correction factor $\frac{1}{n_{t+j}}$. The free-entry condition is then given by:

$$\kappa W_t = \beta \sum_{j=1}^{\infty} (\beta(1-\varphi))^{j-1} \frac{C_t}{C_{t+j}} (\mu - 1) W_{t+j} Y_{t+j} \hat{Z}_{t+j}^{-1}$$

Note that, because $W_{t+j} = \frac{Z_{t+j}}{\mu}$ for all j, we finally have:

$$\kappa Z_t = \beta \sum_{j=1}^{\infty} (\beta(1-\varphi))^{j-1} \frac{C_t}{C_{t+j}} (\mu - 1) \frac{Z_{t+j}}{\hat{Z}_{t+j}} Y_{t+j}$$

Monopolistic competition is efficient: social planner's choice of aggregate number of firms.

The planner's choice of the aggregate number of firms $\{N_{t+j}\}_{j=1}^{\infty}$ is given by:

$$\kappa W_t^* = \beta \sum_{j=1}^{\infty} (\beta(1-\varphi))^{j-1} \frac{C_t^*}{C_{t+j}^*} \frac{dZ_{t+j}^*}{dN_{t+j}^*} \frac{1}{Z_{t+j}^*} Y_{t+j}^*$$

and the following holds from the static allocation problem:

$$q_{it}^*(s)^{-\frac{1}{\gamma}} = \frac{z_t^*(s)}{z_{it}(s)}$$

$$q_t^*(s)^{-\frac{1}{\eta}} = \frac{Z_t^*}{z_t(s)}$$

As in Edmond et al. (2023), $\frac{dZ_{t+j}^*}{dN_{t+j}^*}\frac{1}{Z_{t+j}^*}$ is given by an application of the envelope theorem, which requires ignoring the integer constrain when differentiating with respect to $n_{t+j}(s)$. In particular, an increase of measure 1 in the aggregate number of firms N_{t+j} induces a flow $\epsilon(s)$ of firms within each sector s, with $\mathbb{E}(\epsilon(s)) = 1$. This additional firms in sector s, therefore, induce a gain in aggregate productivity equal to $\frac{dZ_{t+j}^*}{dn_{t+j}(s)}\epsilon(s)$. The expected gain in aggregate productivity induced by an increase of measure 1 in the aggregate number of firms is given by the (expected) total derivative:

$$\frac{dZ_{t+j}^*}{dN_{t+j}^*}\frac{1}{Z_{t+j}^*} = \mathbb{E}_\epsilon\left(\int_0^1 \frac{dZ_{t+j}^*}{dn_{t+j}(s)}\epsilon(s)\frac{1}{Z_{t+j}^*}ds\right) = \int_0^1 \frac{dZ_{t+j}^*}{dn_{t+j}(s)}\mathbb{E}_\epsilon(\epsilon(s))\frac{1}{Z_{t+j}^*}ds = \int_0^1 \frac{dZ_{t+j}^*}{dn_{t+j}(s)}\frac{1}{Z_{t+j}^*}ds$$

Not that, in assuming differentiability, we are also assuming that the evaluation of a change $e(s)$ in the number of firms of sector s happens at distribution $\{n_{t+j}\}_s$ so that you do not internalize the change in the number of firms per sector when evaluating $\frac{dZ_{t+j}^*}{dn_{t+j}(s)}$. This is the equivalent of evaluating the marginal benefit of an additional measure 1 of aggregate firms in the free-entry condition at $\int_0^1 \bar{\pi}_{t+j}(s)ds$.

Edmond et al. (2023) show that, applying the envelope theorem (and ignoring the integer constraint[40]):

$$\frac{dZ_{t+j}^*}{dn_{t+j}(s)}\frac{1}{Z_{t+j}^*} = (\mu - 1)\frac{1}{n_{t+j}(s)}q_{t+j}^*(s)\frac{Z_{t+j}^*}{z_{t+j}^*(s)}$$

Therefore:

[40]Note that such derivation is actually consistent only with a definition of free-entry where expected net profits are zero for the last entrants. If the free-entry condition is defined as in Edmond et al. (2023), where expected net profits are zero for the first non-entrant, the derivative must be such that sectoral variables are evaluated also taking into account the additional firm entering the market. In general, the relaxation of the integer constraint requires a stance on how handling differentiability with discrete increases.

$$\frac{dZ^*_{t+j}}{dN^*_{t+j}}\frac{1}{Z^*_{t+j}} = \tag{40}$$

$$= \int_0^1 \frac{dZ^*_{t+j}}{dn_{t+j}(s)}\frac{1}{Z^*_{t+j}}ds \tag{41}$$

$$= Z^*_{t+j}(\mu-1)\int_0^1 \frac{1}{n_{t+j}(s)}q^*_{t+j}(s)(z^*_{t+j}(s))^{-1}ds \tag{42}$$

$$= Z^*_{t+j}(\mu-1)\sum_{n_{t+j}} \Pr(n_{t+j};N^*_{t+j})\frac{1}{n_{t+j}}\int ... \int n_{t+j}\tilde{q}(z_{t+j},\underline{z}_{t+j},n_{t+j},X_{t+j})\frac{1}{z}dG_{n_{t+j}} \tag{43}$$

$$= Z^*_{t+j}(\mu-1)\sum_{n_{t+j}} \Pr(n_{t+j};N^*_{t+j})\frac{1}{n_{t+j}}\tilde{z}^*_{t+j}(n_{t+j}) \tag{44}$$

$$= (\mu-1)\frac{Z^*_{t+j}}{\hat{Z}^*_{t+j}} \tag{45}$$

The social planner's choice can then be expressed by:

$$\kappa W^*_t = \beta\sum_{j=1}^{\infty}(\beta(1-\varphi))^{j-1}\frac{C^*_t}{C^*_{t+j}}(\mu-1)\frac{Z^*_{t+j}}{\hat{Z}^*_{t+j}}Y^*_{t+j}$$

Because $W^*_{t+j} = Z^*_{t+j}$ for all j, we finally have[41]:

$$\kappa Z^*_t = \beta\sum_{j=1}^{\infty}(\beta(1-\varphi))^{j-1}\frac{C^*_t}{C^*_{t+j}}(\mu-1)\frac{Z^*_{t+j}}{\hat{Z}^*_{t+j}}Y^*_{t+j}$$

We have shown that the free-entry condition is equivalent to the social planner' optimal condition for the aggregate number of varieties.

The definition of the free-entry condition and the mode in which the integer constraint is ignored in computing the derivative of aggregate productivity to sectoral concentration need to be consistent. If one were to use Edmond et al. (2023)'s specification of the free-entry condition, in which the first non-entrants have zero net expected profits—as opposed to the last (realized-in-equilibrium) entrants—it would be enough to adapt the way in which the integer constraint is ignored: the derivative must be evaluated such that sectoral variables include the potential entrant.[42]

[41] Note that this result also holds with elastic labor supply.

[42] Indeed, both monopolistic-competition models and oligopolistic-competition models have tackled the problem of the integer constraint in different ways. Monopolistic-competition models that exhibit a continuum of firms are not subject to it, both in the decentralized and social planner equilibrium (e.g., in studying the efficiency of Melitz, 2003). Interestingly, Dixit and Stiglitz (1977) consider a finite number of firms and ignore the integer constraint in deriving both the decentralized equilibrium and the social planner's choices. Decentralized equilibria of oligopolistic models ignore the integer constraint when the free-entry condition is assumed to hold with equality at the sector level. At the same time, they are not

Monopolistic competition is efficient: equilibrium conditions.

We now have to compare the equilibrium conditions of the decentralized problem to the equilibrium conditions of the social planner's problem. In the decentralized equilibrium, equilibrium conditions are as follows:

$$\kappa Z_t = \beta \sum_{j=1}^{\infty} (\beta(1-\varphi))^{j-1} \frac{C_t}{C_{t+j}} (\mu - 1) \frac{Z_{t+j}}{\hat{Z}_{t+j}} Y_{t+j}$$

$$C_t = Y_t$$

$$Y_t = Z_t \tilde{L}_t$$

$$\tilde{L}_t = L_t - \kappa(N_{t+1} - (1-\varphi)N_t)$$

$$L_t = \bar{L}$$

In the social planner's problem, the equilibrium conditions are:

$$\kappa Z_t^* = \beta \sum_{j=1}^{\infty} (\beta(1-\varphi))^{j-1} \frac{C_t^*}{C_{t+j}^*} (\mu - 1) \frac{Z_{t+j}^*}{\hat{Z}_{t+j}^*} Y_{t+j}$$

$$C_t^* = Y_t^*$$

$$Y_t^* = Z_t^* \tilde{L}_t$$

$$\tilde{L}_t^* = L_t^* - \kappa(N_{t+1}^* - (1-\varphi)N_t^*)$$

$$L_t^* = \bar{L}$$

Because, as previously shown, for the same N_t, $Z_t = Z(N_t) = Z_t^*$ (and, similarly, $\hat{Z}_t =$

subject to the integer constraint when the free-entry condition (at the sector level) is in the form of a complementary slackness (as in Atkeson and Burstein, 2008). Social planner's problems in oligopolistic models ignore the integer constraint when the social planner is assumed to differentiate with respect to an integer variable and when the optimality condition is assumed to hold with equality at the sector level.

$\hat{Z}(N_t) = \hat{Z}_t^*$), this implies that the decentralized equilibrium is efficient.

Note that, when labor supply is elastic, the last equilibrium condition is $\psi C_t L_t^{\nu} = \frac{Z_t}{\mu}$ in the decentralized equilibrium and $\psi C_t^* L_t^{*\nu} = Z_t^*$ in the social planner's equilibrium. Aggregate labor is, therefore, inefficient in the decentralized equilibrium. To restore efficiency, combining ρ^* and a uniform wage subsidy τ to workers paid by lump sum transfers is enough. In particular, $(1+\tau) = \mu = (1+\rho^*)$, such that the last equilibrium condition of the decentralized problem becomes $\psi C_t L_t^{\nu} = (1+\tau) W_t = \frac{(1+\tau) Z_t}{\mu} = Z_t$.

Proof of Corollary 1

The proof is immediate considering that $\rho^* = \frac{1}{\gamma - 1}$, which is independent of κ and $G(.)$. In addition, it is also immediate to see that $\pi_{it}(s) \leq \rho W_t l_{it}(s) = \frac{\pi_{it}(s)}{K_t} \leq \frac{\rho W_t l_{it}(s)}{K_t}$, for any strictly positive K_t.

Proof of Theorem 2

Decentralized economy

The model is based on Edmond et al. (2023), but relaxes several assumptions. In particular, it does not commit to specific functional forms for the preferences of the representative consumer, or the aggregators of final-good production within and between sectors.

A representative consumer has preferences over a final consumption good and the supply of labor. The final good is produced by a perfectly competitive representative firm using inputs from a continuum of sectors, indexed by $s \in [0,1]$. In each sector, a number (or a mass) $n_t(s)$ of imperfectly competitive firms produce differentiated intermediate goods using labor, capital, and materials as inputs. Intermediate firms can be created by paying an irreversible cost of entry. Once such cost has been paid, a new firm receives a one-time productivity draw in a randomly allocated sector. Exit from the market is random, and the economy has no aggregate uncertainty. The representative consumer owns all the firms.

Representative consumer.—The representative consumer maximizes

$$\sum_{t=0}^{\infty} \beta^t U(C_t, L_t),$$

subject to

$$(1+\tau_{s,t})(C_t + I_t) = W_t L_t + R_t K_t + \Pi_t - T_t,$$

where C_t denotes the (numeraire) final consumption good, L_t denotes labor supply,

$I_t = K_{t+1} - (1-\delta)K_t$ denotes investment, K_t denotes physical capital, δ denotes the depreciation rate, W_t denotes the real wage, R_t denotes the rental rate of capital, $0 < \beta < 1$ denotes the time discount factor, and Π_t denotes aggregate real profits (net of the cost of creating new firms and net of a profit tax $\tau_{\pi,t}$), $\tau_{s,t}$ is a sales tax on the consumption good, and T_t is a lump sum tax financing $\tau_{s,t}$ and $\tau_{\pi,t}$. Utility $U(.,.)$ is assumed to be strictly increasing and concave in the first argument and strictly increasing and convex in the second argument. Regularity conditions that ensure a well-behaved consumer problem are assumed. [43].

The optimal labor supply choice satisfies:

$$\frac{-U_L(C_t, L_t)(1+\tau_{s,t})}{U_C(C_t, L_t)} = W_t.$$

Optimal investment choice satisfies:

$$\beta \frac{U_C(C_{t+1}, L_{t+1})}{U_C(C_t, L_t)} \left(\frac{R_{t+1}}{(1+\tau_{s,t+1})} + 1 - \delta\right) = 1.$$

Final-good producer.— The representative firm produces the final good Y_t according to the following production technology:

$$\int_0^1 \mathcal{A}(s)(q_t(s))ds = 1,$$

where $q_t(s) = \frac{y_t(s)}{Y_t}$ denotes the relative size of sector s, $y_t(s)$ denotes the input from sector $s \in [0,1]$. The sector-specific function $\mathcal{A}(s)(.)$ is assumed to be increasing and concave. $P_t = 1$ denotes the (normalized) price of the final good. In addition, $p_t(s)$ denotes the price index for sector s, so that:

$$1 = \int_0^1 p_t(s) q_t(s) ds$$

The final good is used for consumption, investment, or materials in production X_t:

$$Y_t = C_t + I_t + X_t$$

Within-sector aggregator.—Sectoral output $y_t(s)$ is defined implicitly by the following sector-specific aggregator:

[43]For the quantitative analysis, consumer utility is assumed to be separable in consumption and labor supply, with $U(C, L) = \log(C) - \psi \frac{L^{1+v}}{1+v}$, where ψ denotes the cost of effort and v the Frisch elasticity of labor supply.

$$\sum_{i=1}^{n_t(s)} \mathcal{B}_i(s)(q_{i,t}(s)) = 1,$$

where $q_{it}(s) = y_{it}(s)/y_t(s)$ denotes the relative size of firm i in sector s, $y_{it}(s)$ denotes the output of firm i in sector s, and the firm-specific function $\mathcal{B}_i(s)(.)$ is assumed to be increasing and concave. The sum is replaced with an integral if there is a mass of firms within sectors. The between-sector and within-sector aggregators, linking firm output to aggregate output, are also assumed to induce an inverse demand satisfying assumption 1. In addition, $p_t(s)$ is such that:

$$p_t(s) = \sum_{1}^{n_t(s)} p_{it}(s) q_{it}(s) ds.$$

The within-sector inverse demand of the final-good producers is given by:

$$\frac{p_{it}(s)}{p_t(s)} = \frac{\mathcal{B}_{q,i}(s)(q_{i,t}(s))}{\sum_1^{n_t(s)} \mathcal{B}_{q,i}(s)(q_{i,t}(s)) q_{i,t}(s)},$$

while the between-sector inverse demand of the final-good producers is given by:

$$p_t(s) = \frac{\mathcal{A}_q(s)(q_t(s))}{\int_0^1 \mathcal{A}_q(s)(q_t(s)) q_t(s) ds}.$$

Ultimately, the inverse demand is equal to:

$$p_{it}(s) = \frac{\mathcal{A}_q(s)(q_t(s))}{\int_0^1 \mathcal{A}_q(s)(q_t(s)) q_t(s) ds} \cdot \frac{\mathcal{B}_{q,i}(s)(q_{i,t}(s))}{\sum_1^{n_t(s)} \mathcal{B}_{q,i}(s)(q_{i,t}(s)) q_{i,t}(s)},$$

Intermediate-good producers.—In each sector, there are $n_t(s)$ firms, with $n_t(s) \in \mathbb{N}$ or $n_t(s) \in \mathbb{R}_+$; each firm produces a unique differentiated variety, and it engages in either monopolistic competition or oligopolistic competition within the sector. The technology of production of the intermediate good is as follows:

$$y_{it}(s) = z_{it}(s)[\phi^{\frac{1}{\theta}} v_{it}(s)^{\frac{\theta-1}{\theta}} + (1-\phi)^{\frac{1}{\theta}} x_{it}(s)^{\frac{\theta-1}{\theta}}]^{\frac{\theta}{\theta-1}},$$

where $z_{it}(s)$ denotes the productivity of firm i in sector s[44], $x_{it}(s)$denotes materials used by firm i in sector s, $v_{it}(s)$ denotes value-added by firm i in sector s, and θ is the constant elasticity of substitution between value-added and materials. Value-added is given by:

[44]Firm-specific productivity is indexed by t even though new firms receive a one-time productivity draw. This is because, over time, the same i identifies different firms and s.

$$v_{it}(s) = k_{it}(s)^{\alpha} l_{it}(s)^{1-\alpha},$$

where $k_{it}(s)$ is the physical capital employed by firm i in sector s, $l_{it}(s)$ is labor *used in production* employed by firm i in sector s, and α is the constant elasticity of capital to value-added.

Input demands.—Inputs demands are standard as in Edmond et al. (2023), but they have to internalize the fact that, with the sales tax on the final good, the price for materials at time t is $(1 + \tau_{s,t})$. The input price index is therefore:

$$\Omega_t = \left\{ \phi \left[\left(\frac{R_t}{\alpha} \right)^{\alpha} \left(\frac{W_t}{1-\alpha} \right)^{1-\alpha} \right]^{1-\theta} + (1-\phi)(1+\tau_{s,t})^{1-\theta} \right\}^{\frac{1}{1-\theta}},$$

so that the firm marginal cost can be expressed as $\frac{\Omega_t}{z_{it}(s)}$.

Profit maximization.—Intermediate-good producers maximize

$$\pi_{it}(s) = p_{it}(s) y_{it}(s) - \frac{\Omega_t}{z_{it}(s)} y_{it}(s),$$

where $\pi_{it}(s)$ denotes the profits of firm i in sector s, $p_{it}(s)$ denotes the price of firm i in sector s, and $\Omega_t / z_{it}(s)$ is the marginal cost of firm i in sector s. $p_{it}(s)$ is set considering the final-good producer's inverse demand in a context of monopolistic or oligopolistic competition among firms in sector s. Therefore, at the optimum, a firm price can be written as a markup $\mu_{it}(s)$ over the marginal cost:

$$p_{it}(s) = \mu_{it}(s) \frac{\Omega_t}{z_{it}(s)}, \; \mu_{it}(s) = \frac{\sigma_{it}(s)}{\sigma_{it}(s) - 1},$$

where $\sigma_{it}(s)$ denotes the demand elasticity to price faced by firm i in sector s, satisfying assumption 1.

Most important, the ratio between firms' profits and total costs $\frac{\Omega_t}{z_{it}(s)} y_{it}(s)$ can be expressed as:

$$\frac{\pi_{it}(s) z_{it}(s)}{\Omega_t y_{it}(s)} = \mu_{it}(s) - 1,$$

Aggregates.—Given firm gross output per unit of firm TFP $\frac{y_{it}(s)}{z_{it}(s)}$:

$$F(k_{it}(s), l_{it}(s), x_{it}(s)) = [\phi^{\frac{1}{\theta}} v_{it}(s)^{\frac{\theta-1}{\theta}} + (1-\phi)^{\frac{1}{\theta}} x_{it}(s)^{\frac{\theta-1}{\theta}}]^{\frac{\theta}{\theta-1}}$$

the sector-level and economy-level aggregates

$$F(k_t(s), l_t(s), x_t(s)) = \sum_{1}^{n_t(s)} F(k_{it}(s), l_{it}(s), x_{it}(s))$$

$$F(K, \tilde{L}, X_t) = \int_0^1 F(k_t(s), l_t(s), x_t(s))ds$$

only depend on sector-level and aggregate-level factors of production respectively, with $h_t(s) = \sum_1^{n_t(s)} h_{it}(s)$ and $H = \int_0^1 h_t(s)$ for $h = k, l, x$. Labor employed *in production* is denoted by $\tilde{L}$.

This implies the following formulas for sector-level and aggregate productivity :

$$y_t(s) = z_t(s)F(k_t(s), l_t(s), x_t(s)),$$
$$Y_t = Z_t F(K, \tilde{L}, X_t),$$

with sectoral productivity given by

$$z_t(s) = \left(\sum_{i=1}^{n_t(s)} \frac{q_{it}(s)}{z_{it}(s)} \right)^{-1};$$

and aggregate productivity given by

$$Z_t = \left(\int_0^1 \frac{q_t(s)}{z_t(s)} ds \right)^{-1}.$$

As in Edmond et al. (2023), sector-level and aggregate markups are sales-weighted harmonic averages:

$$\mu_t(s) = \left(\sum_{i=1}^{n_t(s)} \frac{1}{\mu_{it}(s)} \frac{p_{it}(s)y_{it}(s)}{p_t(s)y_t(s)} \right)^{-1},$$

where $\mu_t(s)$ represents sectoral markup;

$$\mathcal{M}_t = \left(\int_0^1 \frac{1}{\mu_t(s)} \frac{p_t(s)y_t(s)}{Y_t} ds \right)^{-1},$$

where $\mathcal{M}_t$ represents aggregate markup.

In addition, the following holds:

$$p_t(s) = \mu_t(s)\frac{\Omega_t}{z_t(s)};$$

$$1 = \mathcal{M}_t\frac{\Omega_t}{Z_t};$$

The aggregate counterpart of the firm-specific factor allocations are :

$$Z_t F_L = \mathcal{M}_t W_t$$

$$Z_t F_K = \mathcal{M}_t R_t$$

$$Z_t F_X = \mathcal{M}(1 + \tau_{s,t})$$

Entry and exit.—There is free entry of new firms in the market of intermediate-good producers. Investors can pay a sunk cost κ in units of labor and start up a measure one of firms which, after obtaining a one-time productivity draw $z_{it} \sim G(z)$, will produce a unique new variety of the intermediate good in a randomly allocate sector $s \in [0, 1]$. Let $N_t = \int_0^1 n_t(s)ds$ be the aggregate mass of firms and $M_t = \int_0^1 m_t(s)ds$ be the aggregate mass of entrants. As in Edmond et al. (2023), I assume that entry per sector $m_t(s)$ is IID Poisson with parameter $M_t > 0$ so that each sector has a discrete number of firms. Entrants at time t start producing at time $t + 1$, and, for $j = 0, 1, 2, ...,$ they obtain a stream of profits $\pi_{i,t+j}(s)$ for each $t + j$ until they are hit with an IID exit shock, which obtains with probability φ per period. The aggregate mass of firms, therefore, evolves according to $N_{t+1} = (1 - \varphi)N_t + M_t$, and $\mathbb{E}_t n_{t+1}(s) = (1 - \varphi)n_t(s) + \mathbb{E}_t m_t(s)$.

In this environment, zero net expected profits need to be zero for potential entrants:

$$\kappa W_t = \beta \sum_{j=1}^{\infty} (\beta(1 - \varphi))^{j-1} \frac{C_t}{C_{t+j}} \int_0^1 \bar{\pi}_{t+j}(s)ds,$$

where $\bar{\pi}_{t+j}(s)$ denote expected profits of operating in sector s at time $t + j$, equivalent to

$$\bar{\pi}_{t+j}(s) = \int ... \int \pi_{t+j}(z_{i,t+j}(s), \underline{z}_{t+j}(s), n_{t+j}(s) + 1)dG_{n_{t+j}(s)},$$

where $(z_{i,t+j}(s), \underline{z}_{t+j(s)}, n_{t+j}(s)+1)$ identifies a potential entrant firm i with productivity $z_{i,t+j}(s)$, operating in sector s, with other $n_{t+j}(s)$ firms characterized by productivity levels $\underline{z}_{t+j}(s)$.

If there is a mass of firms $n_t(s) \in \mathbb{R}_+$ in each sector, expected profits are given by:

$$\bar{\pi}_{t+j}(s) = \int \pi_{t+j}(z_{i,t+j}, s) dG.$$

Government budget constraint.—The lump-sum tax T_t balances the government budget:

$$T_t = -\tau_{s,t}(C_t + I_t + X_t) - \tau_{\pi,t}(Y_t - W_t\tilde{L}_t - R_tK_t - (1+\tau_{s,t})X_t).$$

Note that aggregate profits (net of entry costs and profit tax) are:

$$\Pi_t = (1-\tau_{\pi,t})(Y_t - W_t\tilde{L}_t - R_tK_t - (1+\tau_{s,t})X_t) - W_t(L_t - \tilde{L}_t).$$

Social planner

The social planner maximizes the welfare of the representative consumer conditional on three constraints: i) resource constraint, ii) technology of production, and iii) technology of entry. The problem is split into two parts: the static problem, in which the planner chooses aggregate productivity by choosing the relative size of firms, taking as given the distribution of firms per sector; the dynamic problem, in which the planner chooses aggregate variables, given the relationship between the aggregate number of firms and aggregate productivity induced by the static problem, denoted by $Z_t^* = Z^*(N_t)$.

Dynamic problem.—The planner maximizes:

$$\sum_{t=0}^{\infty} \beta^t U(C_t^*, \tilde{L}_t^* + \kappa(N_{t+1}^* - (1-\varphi)N_t^*)),$$

subject to

$$C_t^* + K_{t+1}^* + X_t^* \leq Z(N_t^*)F(K_t^*, \tilde{L}_t^*, X_t^*) + (1-\delta)K_t^*.$$

The optimality conditions are therefore:

$$-\frac{U_{L,t}^*}{U_{C,t}^*} = Z_t^*(N_t^*)F_{L,t}^*$$

$$1 = \beta \frac{U^*_{C,t+1}}{U^*_{C,t}} (Z^*_{t+1}(N^*_{t+1}) F^*_{K,t+1} + 1 - \delta)$$

$$Z^*_t(N^*_t) F^*_{Xt} = 1$$

$$\kappa W^*_t = \beta \sum_{j=1}^{\infty} (\beta(1-\varphi))^{j-1} \frac{U^*_{C,t+j}}{U^*_{C,t}} \frac{dZ^*_{t+j}(N^*_{t+j})}{dN^*_{t+j}} \frac{Y^*_{t+j}}{Z^*_{t+j}}$$

with $W^*_t = -\frac{U^*_{C,t}}{U^*_{L,t}}$.

Static problem.—The social planner chooses $\{q^*_{it}(s)\}_{i+1}^{n_t(s)}$ for all s, and $\{q^*_t(s)\}_{s\in[0,1]}$ to maximize:

$$\left(\int_0^1 q^*_t(s) \frac{1}{z_t(s)} \right)^{-1}$$

subject to

$$\int_0^1 \mathcal{A}(s)(q^*_t(s))ds = 1,$$

and

$$\left(\sum_1^{n_t(s)} q^*_{it}(s) \frac{1}{z_{it}(s)} \right)^{-1}$$

subject to

$$\sum_{i=1}^{n_t(s)} \mathcal{B}_i(s)(q^*_{i,t}(s)) = 1.$$

The optimal allocation of relative output across firms can be then characterized by:

$$\frac{z_t(s)}{z_{it}(s)} = \frac{\mathcal{B}_{q,i}(s)(q_{i,t}(s))}{\sum_1^{n_t(s)} \mathcal{B}_{q,i}(s)(q_{i,t}(s)) q_{i,t}(s)},$$

$$\frac{Z_t}{z_t(s)} = \frac{\mathcal{A}_q(s)(q_t(s))}{\int_0^1 \mathcal{A}_q(s)(q_t(s)) q_t(s) ds}.$$

Cap on the profit-to-cost ratio

Under assumption 1, after the introduction of firms' optimal pricing can be characterized by:

$$p_{it}(s) = (1+\rho)\frac{\Omega_t}{z_{it}(s)},$$

implying

$$\pi_{it}(s) = \rho\frac{\Omega_t}{z_{it}(s)}y_{it}(s).$$

Aggregate productivity

I derive some useful results on aggregate productivity. In particular, I want to show that, if in the decentralized equilibrium and the social planner solution the distribution of firms' relative sizes is the same, the relationship between aggregate productivity and the aggregate number of firms is the same, i.e., for the same N_t, $Z_t = Z^d(N_t) = Z(N_t) = Z^*(N_t) = Z_t^*$. Throughout the proof, I will focus on the case where sectors have a discrete number of firms and are heterogeneous in concentration because the aggregation results are more challenging when applying the law of large numbers.

Throughout the proof, decentralized equilibrium variables are not marked by a star (without $*$), and social planner's variables are marked by a star (with $*$). Given a sector s with $n_t(s)$ firms and productivity levels $\{z_{it}(s)\}_{i=1}^{n_t(s)}$ We write $\underline{z}_{t,-i}(s)$ for the vector of productivity levels of the $n_t(s)-1$ firms excluding i. Remember that $Z_t = \left(\int_0^1 q_t(s)\frac{1}{z_t(s)}\right)^{-1}$ and $z_t(s) = \left(\sum_{i=1}^{n_t(s)} q_{it}(s)\frac{1}{z_{it}(s)}\right)^{-1}$. The optimal relative sizes of firms and sectors in the decentralized equilibrium and in the social planner solution are defined implicitly by the following:

$$\frac{\mathcal{B}_{q,i}(s)(q_{i,t}(s))}{\sum_1^{n_t(s)}\mathcal{B}_{q,i}(s)(q_{i,t}(s))q_{i,t}(s)} = \frac{p_{it}(s)}{p_t} = \frac{\mu\frac{\Omega_t}{z_{it}(s)}}{\mu\frac{\Omega_t}{z_t(s)}} = \frac{z_t(s)}{z_{it}(s)} \implies q(z_{it}(s), \underline{z}_{-i}(s), n_t(s))$$

$$\frac{\mathcal{B}_{q,i}(s)(q_{i,t}(s))}{\sum_1^{n_t(s)}\mathcal{B}_{q,i}(s)(q_{i,t}(s))q_{i,t}(s)} = \frac{z_t(s)}{z_{it}(s)} \implies q(z_{it}(s), \underline{z}_{-i}(s), n_t(s))$$

$$\frac{\mathcal{A}_q(s)(q_t(s))}{\int_0^1\mathcal{A}_q(s)(q_t(s))q_t(s)ds} = \frac{p_t(s)}{P_t} = \frac{\mu\frac{\Omega_t}{z_t(s)}}{\mu\frac{\Omega_t}{Z_t}} = \frac{Z_t}{z_t(s)} \implies Q(\underline{z}_t(s), n_t(s), \{\underline{z}_t(s), n_t(s)\}_s)$$

$$\frac{\mathcal{A}_q(s)(q_t(s))}{\int_0^1\mathcal{A}_q(s)(q_t(s))q_t(s)ds} = \frac{Z_t}{z_t(s)} \implies Q(\underline{z}_t(s), n_t(s), \{\underline{z}_t(s), n_t(s)\}_s)$$

We can note that the functions $q(.)$ and $Q(.)$ are the same for the decentralized equilibrium and the social planner's problem, meaning that knowing the sector sizes and the productivity distribution, you can compute relative sizes in the same way. We also define conveniently $\tilde{q}_{it}(s) = q_{it}(s)q_t(s) = \tilde{q}(z_{it}(s), \underline{z}_{t,-i}(s), n_t(s), X_t)$, where X_t summarizes aggregates that are the same for all firms in all sectors.

These equivalences allow us to write, for given N_t and $\{n_t(s)\}_s$:

$$Z_t = Z_t^* = \left(\int_0^1 \sum_{i=1}^{n_t(s)} \tilde{q}(z_{it}(s), \underline{z}_{t,-i}(s), n_t(s), X_t) \frac{1}{z_{it}(s)} ds \right)^{-1}$$

Now, we notice that since $n_t(s)$ is IID distributed according to a pdf $\Pr(.)$ over $n_t(s) \in \{0, 1, 2, ...\}$ with parameter N_t[45], there is a measure $\Pr(n_t; N_t)$ of sectors characterized by exactly $n_t(s)$ firms. Therefore, in a sector with $n_t(s)$ firms, there has been $n_t(s) \Pr(n_t(s); N_t)$ productivity draws, or better, there have been $\Pr(n_t(s); N_t)$ draws of productivity vectors $\underline{z}_t$ of dimension $n_t(s)$. As a result, we can establish a relationship between aggregates and expectations applying the law of large numbers within a sector with $n_t(s)$ firms, as follows[46]:

[45] E.g., if entrants per sector $m_t(s)$ are IID distributed as a Poisson pdf with parameter M_t, $n_t(s)$ is a sum of IID Poisson random variables, and it is therefore distributed as a Poisson with parameter $N_t = \sum_{s=0}^{t-1} (1-\varphi)^{t-1-s} M_t + (1-\varphi)^t N_0$

[46] Note that moving from the third to the fourth step (40-41) we are summing over sectors keeping fixed the same i. Therefore, we are summing over $\Pr(n_t; N_t)$ independent draws of the productivity vector.

$$
\begin{aligned}
Z_t = Z_t^* = & && (46)\\
&= \left(\int_0^1 \sum_{i=1}^{n_t(s)} \tilde{q}(z_{it}(s), \underline{z}_{t,-i}(s), n_t, X_t) \frac{1}{z_{it}(s)} ds \right)^{-1} && (47)\\
&= \left(\sum_{n_t=0}^{\infty} \int_{\{s:n_t(s)=n_t\}} \sum_{i=1}^{n_t(s)} \tilde{q}(z_{it}(s), \underline{z}_{t,-i}(s), n_t, X_t) \frac{1}{z_{it}(s)} ds \right)^{-1} && (48)\\
&= \left(\sum_{n_t=0}^{\infty} \int_{\{s:n_t(s)=n_t\}} \sum_{i=1}^{n_t} \tilde{q}(z_{it}(s), \underline{z}_{t,-i}(s), n_t, X_t) \frac{1}{z_{it}(s)} ds \right)^{-1} && (49)\\
&= \left(\sum_{n_t=0}^{\infty} \sum_{i=1}^{n_t} \int_{\{s:n_t(s)=n_t\}} \tilde{q}(z_{it}(s), \underline{z}_{t,-i}(s), n_t, X_t) \frac{1}{z_{it}(s)} ds \right)^{-1} && (50)\\
&= \left(\sum_{n_t=0}^{\infty} \sum_{i=1}^{n_t} \int ... \int \Pr(n_t; N_t) \tilde{q}(z_t, \underline{z}_{t,-z}, n_t, X_t) \frac{1}{z} dG_{n_t}(\underline{z}_t) \right)^{-1} && (51)\\
&= \left(\sum_{n_t=0}^{\infty} n_t \int ... \int \Pr(n_t; N_t) \tilde{q}(z_t, \underline{z}_{t,-z}, n_t, X_t) \frac{1}{z} dG_{n_t}(\underline{z}_t) \right)^{-1} && (52)\\
&= \left(\sum_{n_t=0}^{\infty} \Pr(n_t; N_t) \int ... \int n_t \tilde{q}(z_t, \underline{z}_{t,-z}, n_t, X_t) \frac{1}{z} dG_{n_t}(\underline{z}_t) \right)^{-1}, && (53)
\end{aligned}
$$

with the law of large numbers applied between line 5 and 6 (the productivity distribution across sectors is IID), or, equivalently:

$$
Z_t = Z_t^* = \left(\sum_{n_t=0}^{\infty} \Pr(n_t; N_t) \tilde{z}^{-1}(n_t) \right)^{-1}
$$

where $\tilde{z}^{-1}(n_t) = \int ... \int n_t \tilde{q}(z_t, \underline{z}_{t,-i}, n_t, X_t) \frac{1}{z} dG_{n_t}(\underline{z}_t)$.

Lastly, we conclude that, for the same N_t, $Z_t = Z_t^* = Z(N_t)$.

Free-entry condition

In this part of the proof, we express the free-entry condition in terms of aggregates.

Assuming, $M_t > 0$ for all t, the free-entry condition is defined as follows:

$$
\kappa W_t = \beta \sum_{j=1}^{\infty} (\beta(1-\varphi))^{j-1} \frac{C_t}{C_{t+j}} (1 - \tau_{\pi,t}) \int_0^1 \bar{\pi}_{t+j}(s) ds
$$

where $\bar{\pi}_{t+j}$, the expected profits of operating at time $t+j$ in sector s

$$\bar{\pi}_{t+j}(s) = \int ... \int \pi_{t+j}(z_{i,t+j}(s), \underline{z}_{t+j}(s), n_{t+j}(s)+1, X_{t+j}) dG_{n_{t+j}(s)+1}(z_{i,t+j}(s), \underline{z}_{t+j}(s))$$

and we can, therefore, write

$$\int_0^1 \bar{\pi}_{t+j}(s)ds = \sum_{n_{t+j}=0}^{\infty} \Pr(n_{t+j}; N_{t+j}) \int ... \int \pi_{t+j}(z_{t+j}, \underline{z}_{t+j}, n_{t+j}+1, X_{t+j}) dG_{n_{t+j}+1}(z_{t+j}, \underline{z}_{t+j}),$$

We can, therefore, establish the following relationship:

$$\begin{aligned}
\int_0^1 \bar{\pi}_{t+j}(s)ds = & \qquad (54)\\
= & \sum_{n_{t+j}=0}^{\infty} \Pr(n_{t+j}; N_{t+j}) \int ... \int \pi_{t+j}(z_{t+j}, \underline{z}_{t+j}, n_{t+j}+1, X_{t+j}) dG_{n_{t+j}} \qquad (55)\\
= & \sum_{n_{t+j}=0}^{\infty} \Pr(n_{t+j}; N_{t+j}) \int ... \int \rho \frac{\Omega_{t+j}}{z} y(z_{t+j}, \underline{z}_{t+j}, n_{t+j}+1, X_{t+j}) dG_{n_{t+j}} \qquad (56)\\
= & \rho\Omega_{t+j} \sum_{n_{t+j}=0}^{\infty} \Pr(n_{t+j}; N_{t+j}) \int ... \int \frac{1}{z} y(z_{t+j}, \underline{z}_{t+ji}, n_{t+j}+1, X_{t+j}) dG_{n_{t+j}} \qquad (57)\\
= & \rho\Omega_{t+j} Y_{t+j} \sum_{n_{t+j}=0}^{\infty} \Pr(n_{t+j}; N_{t+j}) \int ... \int \frac{1}{z} \tilde{q}(z_{t+j}, \underline{z}_{t+j}, n_{t+j}+1, X_{t+j}) dG_{n_{t+j}} \qquad (58)\\
= & \rho\Omega_{t+j} Y_{t+j} \sum_{n_{t+j}=0}^{\infty} \Pr(n_{t+j}; N_{t+j}) \frac{1}{(n_{t+j}+1)} \tilde{z}(n_{t+j}+1)^{-1} \qquad (59)\\
= & \rho\Omega_{t+j} Y_{t+j} (\hat{Z}^{+}_{t+j})^{-1} \qquad (60)
\end{aligned}$$

Note that indeed, $\hat{Z}^{+}_{t+j}(N_t) \neq Z_{t+j}(N_t)$ because of the weighting factors $\frac{1}{(n_{t+j}+1)}$ and because there is one additional firm in each sector. The free-entry condition is then given by:

$$\kappa W_t = \beta \sum_{j=1}^{\infty} (\beta(1-\varphi))^{j-1} \frac{C_t}{C_{t+j}} \rho(1-\tau_{\pi,t}) \Omega_{t+j} Y_{t+j} (\hat{Z}^{+}_{t+j})^{-1}$$

Characterization of optimal policy mix

Let $\underline{\mu} = \inf \{\mu_{its}(s)\}$ for all i, t, s, whose existence is ensured by assumption 1. Let $\tilde{b}$ be any number such that $1 < \tilde{b} \leq \underline{\mu}$.

Then, consider the following policy mix:

$$\rho_t^* = \tilde{b} - 1,$$

$$\tau_{s,t} = -\rho_t^*/[1+\rho_t^*],$$

$$(1-\tau_{\pi,t}^*)\rho_t^* = \frac{dZ_t^*}{dN_t^*}\frac{\frac{\hat{Z}_t^{+,*}}{Z_t^*}}{Z_t^*}$$

The free entry condition under such a policy mix is :

$$\kappa W_t = \beta \sum_{j=1}^{\infty} (\beta(1-\varphi))^{j-1} \frac{C_t}{C_{t+j}} (1-\tau_{\pi,t+j}^*)\rho_{t+j}^* \frac{\Omega_{t+j}}{\hat{Z}_{t+j}} Y_{t+j}.$$

Note that given $1 = \mathcal{M}_t \frac{\Omega_t}{Z_t}$ and $Z_t F_L = \mathcal{M}_t W_t$,the free-entry condition is

$$\kappa Z_t F_L = \beta \sum_{j=1}^{\infty} (\beta(1-\varphi))^{j-1} \frac{C_t}{C_{t+j}} \frac{dZ_{t+j}^*}{dN_{t+j}^*} \frac{\frac{\hat{Z}_{t+j}^{+,*}}{Z_{t+j}^*}}{Z_{t+j}^*} \frac{Z_{t+j}}{\hat{Z}_{t+j}^+} Y_{t+j}.$$

On the other hand, the planner's choice of the aggregate number of firms, given $W_t^* = Z_t^* F_L^*$, is given by:

$$\kappa Z_t^* F_L^* = \beta \sum_{j=1}^{\infty} (\beta(1-\varphi))^{j-1} \frac{C_t^*}{C_{t+j}^*} \frac{dZ_{t+j}^*}{dN_{t+j}^*} \frac{1}{Z_{t+j}^*} Y_{t+j}^*.$$

The two conditions are the same when evaluated at the social planner solution.

Equilibrium conditions

Social planner solution.—The social planner solution is identified by the following system of equations:

$$-\kappa \frac{U_{C,t}^*}{U_{L,t}^*} = \beta \sum_{j=1}^{\infty} (\beta(1-\varphi))^{j-1} \frac{U_{C,t+j}^*}{U_{C,t}^*} \frac{dZ(N_{t+j}^*)}{dN_{t+j}^*} \frac{Y_{t+j}^*}{Z(N_{t+j}^*)}$$

$$Y_t^* = C_t^* + I_t^* + X_t^*$$

$$-\frac{U_{L,t}^*}{U_{C,t}^*} = Z(N_t^*) F_{L,t}^*$$

$$1 = \beta \frac{U^*_{C,t+1}}{U^*_{C,t}} (Z(N^*_{t+1}) F^*_{K,t+1} + 1 - \delta)$$

$$Y^*_t = Z(N^*_t) F(K^*_t, \tilde{L}^*_t, X^*_t)$$

$$\tilde{L}^*_t = L^*_t - \kappa(N^*_{t+1} - (1-\varphi)N^*_t)$$

$$Z_t(N^*_t) F^*_{Xt} = 1$$

$$I^*_t = K^*_{t+1} - (1-\delta)K^*_t$$

Decentralized economy.—Under the optimal policy mix, the equilibrium conditions are as follows:

$$\kappa W_t = \beta \sum_{j=1}^{\infty} (\beta(1-\varphi))^{j-1} \frac{C_t}{C_{t+j}} \frac{dZ(N^*_{t+j})}{dN^*_{t+j}} \frac{\frac{\hat{Z}^+(N^*_{t+j})}{Z^*_{t+j}}}{Z(N^*_{t+j})} \frac{\Omega_{t+j}}{\hat{Z}^+(N_{t+j})} Y_{t+j}.$$

$$\frac{1}{(1+\rho^*_t)}(C_t + I_t) = W_t L_t + R_t K_t + \Pi_t - T_t,$$

$$T_t = \frac{\rho^*_t}{(1+\rho^*_t)}(C_t + I_t + X_t) - \tau_{\pi,t}(Y_t - W_t \tilde{L}_t - R_t K_t - \frac{1}{(1+\rho^*_t)} X_t)$$

$$\Pi_t = (1-\tau_{\pi,t})(Y_t - W_t \tilde{L}_t - R_t K_t - \frac{1}{(1+\rho^*_t)} X_t) - W_t(L_t - \tilde{L}_t)$$

$$Y_t = C_t + I_t + X_t$$

$$1 = (1+\rho^*_t)\frac{\Omega_t}{Z(N_t)}$$

$$\frac{1}{(1+\rho^*_t)} Z(N_t) F_{L,t} = W_t$$

$$\frac{1}{(1+\rho_t^*)} Z(N_t) F_{K,t} = R_t$$

$$\frac{1}{(1+\rho^*)} Z(N_t) F_{X,t} = \frac{1}{(1+\rho_t^*)}$$

$$\frac{-U_L(C_t, L_t)}{U_C(C_t, L_t)} \frac{1}{(1+\rho_t^*)} = W_t.$$

$$\beta \frac{U_C(C_{t+1}, L_{t+1})}{U_C(C_t, L_t)} ((1+\rho_t^*) R_{t+1} + 1 - \delta) = 1.$$

$$\tilde{L}_t = L_t - \kappa(N_{t+1} - (1-\varphi) N_t)$$

$$Y_t = Z(N_t) F(K_t, \tilde{L}_t, X_t)$$

The social planner solution solves the system of equations of the decentralized economy. Therefore, the optimal policy mix enforces the efficient allocation.

Adjusted free-entry condition

I also report the derivations under the free-entry condition compatible with evaluating the derivative of aggregate productivity with respect to the aggregate number of firms at the equilibrium levels (without internalizing the additional firm that enters the market on average.)

Assuming, $M_t > 0$ for all t, the free-entry condition is defined as follows:

$$\kappa W_t = \beta \sum_{j=1}^{\infty} (\beta(1-\varphi))^{j-1} \frac{C_t}{C_{t+j}} \int_0^1 \bar{\pi}_{t+j}(s) ds$$

where $\bar{\pi}_{t+j}$, the expected profits of operating at time $t+j$ in sector s

$$\bar{\pi}_{t+j}(s) = \int ... \int (1 - \tau_{\pi,t+j}) \pi_{t+j}(z_{i,t+j}(s), \underline{z}_{t+j,-i}(s), n_{t+j}(s), X_{t+j}) dG_{n_{t+j}(s)}$$

and we can, therefore, write

$$\int_0^1 \bar{\pi}_{t+j}(s) ds = \sum_{n_{t+j}=0}^{\infty} \Pr(n_{t+j}; N_{t+j}) \int ... \int \pi_{t+j}(z_{i,t+j}, \underline{z}_{t+j,-i}, n_{t+j}, X_{t+j}) dG_{n_{t+j}},$$

meaning that, at the margin, new entrants at time $t+j$ that bring the aggregate number of firms to N_{t+j} have an ex-ante probability of $\Pr(n_{t+j}; N_{t+j})$ to end up in a sector with a total number of firms (including themselves) equal to n_{t+j}.

Assuming that i) the last entrants at the margin have zero expected profits and that ii) investors correctly anticipate equilibrium future market concentration, we have the following relationship:

$$\int_0^1 \bar{\pi}_{t+j}(s)ds = \tag{61}$$

$$= \sum_{n_{t+j}=0}^{\infty} \Pr(n_{t+j}; N_{t+j}) \int ... \int \pi_{t+j}(z_{i,t+j}, \underline{z}_{t+j,-i}, n_{t+j}, X_{t+j}) dG_{n_{t+j}} \tag{62}$$

$$= \sum_{n_{t+j}=0}^{\infty} \Pr(n_{t+j}; N_{t+j}) \int ... \int \rho \frac{\Omega_{t+j}}{z} y(z_{i,t+j}, \underline{z}_{t+j,-i}, n_{t+j}, X_{t+j}) dG_{n_{t+j}} \tag{63}$$

$$= \rho \Omega_{t+j} \sum_{n_{t+j}=0}^{\infty} \Pr(n_{t+j}; N_{t+j}) \int ... \int \frac{1}{z} y(z_{i,t+j}, \underline{z}_{t+j,-i}, n_{t+j}, X_{t+j}) dG_{n_{t+j}} \tag{64}$$

$$= \rho \Omega_{t+j} Y_{t+j} \sum_{n_{t+j}=0}^{\infty} \Pr(n_{t+j}; N_{t+j}) \int ... \int \frac{1}{z} \tilde{q}(z_{i,t+j}, \underline{z}_{t+j,-i}, n_{t+j}, X_{t+j}) dG_{n_{t+j}} \tag{65}$$

$$= \rho \Omega_{t+j} Y_{t+j} \sum_{n_{t+j}=0}^{\infty} \Pr(n_{t+j}; N_{t+j}) \frac{1}{n_{t+j}} \tilde{z}(n_{t+j})^{-1} \tag{66}$$

$$= \rho \Omega_{t+j} Y_{t+j} \hat{Z}_{t+j}^{-1} \tag{67}$$

Note that indeed, $\hat{Z}_{t+j} \neq Z_{t+j}$ because of the correction factor $\frac{1}{n_{t+j}}$. The free-entry condition is then given by:

$$\kappa W_t = \beta \sum_{j=1}^{\infty} (\beta(1-\varphi))^{j-1} \frac{C_t}{C_{t+j}} (1 - \tau_{\pi,t+j}) \Omega_{t+j} Y_{t+j} \hat{Z}_{t+j}^{-1}.$$

Proof of Theorem 2b

I only characterize the effects of the optimal policy on the free-entry condition in the decentralized equilibrium and compare it to the socially optimal condition for entry. I focus on the case of monopolistic competition with Kimball demand relevant for the quantitative exercise. The rest follows from the proof of Theorem 2 with

]

$$\rho^* = \frac{dZ_t^*}{dN_t^*} \frac{\frac{\hat{Z}_t^{+,*}}{Z_t^*}}{Z_t^*}$$

The free entry condition under the policy $\rho^*_{t+j} = D^*_{t+j} - 1$ is :

$$\kappa W_t = \beta \sum_{j=1}^{\infty} (\beta(1-\varphi))^{j-1} \frac{C_t}{C_{t+j}} (D^*_{t+j} - 1) \frac{\Omega_{t+j}}{\hat{Z}_{t+j}} Y_{t+j}.$$

In addition, from Edmond et al. (2023)'s derivations we have that the planner's choice of the aggregate number of firms $\{N^*_{t+j}\}_{j=1}^{\infty}$ is given by:

$$\kappa W^*_t = \beta \sum_{j=1}^{\infty} (\beta(1-\varphi))^{j-1} \frac{C^*_t}{C^*_{t+j}} \frac{dZ^*_{t+j}}{dN^*_{t+j}} \frac{1}{Z^*_{t+j}} Y^*_{t+j}$$

with

$$\frac{dZ^*_{t+j}}{dN^*_{t+j}} \frac{1}{Z^*_{t+j}} = \int_0^1 \frac{dZ^*_{t+j}}{dn_{t+j}(s)} \frac{1}{Z^*_{t+j}} ds = \int_0^1 (d_t(s) - 1) \frac{1}{n_{t+j}(s)} q^*_{t+j}(s) \frac{Z^*_{t+j}}{z^*_{t+j}(s)} ds$$

from the definition of $D^*_{t+j} - 1$ it follows:

$$\frac{dZ^*_{t+j}}{dN^*_{t+j}} \frac{1}{Z^*_{t+j}} = (D^*_{t+j} - 1) \frac{Z_{t+j}}{\hat{Z}_{t+j}},$$

Because when sectors are homogeneous it follows that $Z^*_t = \hat{Z}^*_t$, this concludes the proof.

B Recent policy discussion on business taxation

In addition to proposals to raise standard corporate taxes ("Kamala Harris backs plan to raise US corporate tax rate to 28%", Financial Times, August 2024), proposals or adoptions of mandatory profit-sharing rules or excess-profits taxes have been widespread in the last years, primarily but not exclusively as an emergency reaction to energy price hikes. In November 2023, France adopted a structural mandatory profit-sharing scheme for firms making profits higher than 1% of revenues for three consecutive years. The European Union Council Regulation 2022/1854 of 6 October 2022 mandated Member States to adopt a windfall tax levied at a rate of at least 33% over the profits of companies in the crude petroleum, natural gas, coal and refinery sector. Italy introduced a one-off 40% tax on banks' profits resulting from high interest rates ("Italian banks hit with surprise windfall tax", BBC News, August 2023). Spain also raised 3 billion euros with a windfall tax on excess profits of banks and is currently considering additional measures ("Spanish Banks Face Extra Payments After Windfall-Tax Review", Bloomberg, October 2024). Proposals and adoptions were not limited to the energy and banking sectors: "more

than 30 windfall taxes, several of which now cover multiple sectors, have been introduced or proposed across Europe since the start of 2022", including in the UK ("Europe's thriving businesses face mounting windfall tax hit", Financial Times, August 2023). In addition, François et al. (2022) propose an excess-profits tax on companies' capitalization ("A Modern Excess Profit Tax"). Similar proposals have been raised in the United States ("U.S. Senate finance chair to propose tax on excess oil profits", Reuters, June 2022), and Canada ("NDP's proposed tax on 'excess' profits could rake in $\$8B$: budget watchdog", CBC, April 2021).