CIRJE-F-1092

Functional Forms for Tractable Economic Models and the Cost Structure of International Trade

Michal Fabinger
The University of Tokyo

E. Glen Weyl
Microsoft Research New York City

August 2018

Functional Forms for Tractable Economic Models and the Cost Structure of International Trade*

Michal Fabinger† E. Glen Weyl‡

August 2018

Abstract

We present functional forms allowing a broader range of analytic solutions to common economic equilibrium problems. These can increase the realism of pen-and-paper solutions or speed large-scale numerical solutions as computational subroutines. We use the latter approach to build a tractable heterogeneous firm model of international trade accommodating economies of scale in export and diseconomies of scale in production, providing a natural, unified solution to several puzzles concerning trade costs. We briefly highlight applications in a range of other fields. Our method of generating analytic solutions is a discrete approximation to a logarithmically modified Laplace transform of equilibrium conditions.

*This paper replaces its previous working versions "Pass-Through and Demand Forms" / "A Tractable Approach to Pass-Through Patterns" / "The Average-Marginal Relationship and Tractable Equilibrium Forms". We are grateful to many colleagues and seminar participants for helpful comments. We appreciate the research assistance of Konstantin Egorov, Eric Guan, Franklin Liu, Eva Lyubich, Yali Miao, Daichi Ueda, Ryo Takahashi, Huan Wang, and Xichao Wang. This research was funded by the Kauffman Foundation, the Becker Friedman Institute for Research in Economics, the Japan Science and Technology Agency and the Japan Society for the Promotion of Science to which we are grateful. We are particularly indebted to Jeremy Bulow for detailed discussion and for inspiring this work and to James Heckman for advice on relevant theorems in duration analysis and nonparametric estimation. All errors are our own.

†Graduate School of Economics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan and CERGE-EI, Prague, the Czech Republic: fabinger@e.u-tokyo.ac.jp.

‡Microsoft Research New York City, 641 Avenue of the Americas, New York, 10011 USA and Department of Economics, Princeton University: glenweyl@microsoft.com.

1 Introduction

Analytic solutions have played a major role in many fields of economics. They are useful both as closed-form, pencil-and-paper solutions to applied theory models, and as components (subroutines) of larger models, making them more computationally tractable.[1] In this paper, we substantially expand the class of known analytic solutions to a broad class of standard economic models. We then illustrate the application of these ideas to a computationally-intensive model of international trade that helps resolve a long-standing puzzle about trade costs by allowing more realistic functional forms of such costs.[2]

We observe that most frequently used functional forms that lead to analytic solutions in economics, namely linear and constant-elasticity functions, share a convenient property: They preserve functional forms under transformations that we refer to as "average-marginal transformations". That is, the functional form of the average value of the function is the same as that of its derivative. Formally, we say that a functional form class is preserved by average-marginal transformations if for any function $F(q)$ the class also contains any linear combination of $F(q)$ and $qF'(q)$.[3] We then find all functions that have such property.

Within this class, we identify functional forms that have a given level of "algebraic tractability", a property we define. These are linear combinations of power functions satisfying certain conditions. When used to represent demand and cost curves they lead to economic optimization conditions that may be transformed to polynomial equations of a degree smaller than 5. These, in turn, may be solved explicitly by the method of radicals. This substantially generalizes the simple analytic solutions that economists are familiar with in the case of constant marginal cost and either linear or constant-elasticity demand. Even beyond degree 5, precise solutions to such polynomial equations are available at minimal cost in standard mathematical software.

We show that elements of functional form classes preserved by average-marginal transformations also have advantageous properties when applied to aggregation over heterogeneous firms in monopolistically competitive models: They lead to closed-form aggregation integrals under very flexible assumptions. This means that a problem with a continuum of heterogeneous firms may be reduced to a set of explicit equations at the macroeconomic level.

In our method, the existence of closed-form solutions to optimization conditions sometimes

[1] This later type of use is particularly important in the closely allied computationally-intensive field of Bayesian statistics where closed-form tractable priors are typically used to approximate otherwise computationally intractable probability models.

[2] In this main, computationally intensive application we find that analytic characterization of the solutions of sub-problems in larger-scale models is particularly useful in conjunction with analytic-differentiation software, graphics processing units (GPUs), and related optimization algorithms. GPU computing has witnessed dramatic developments over the last few years, which our work benefitted from.

[3] A simple economic interpretation would be to identify $F(q)$ with the price $P(q)$ of a good sold by a monopolist, i.e. with the *average* revenue the firm receives per unit sold, in which case $F(q) + qF'(q) = P(q) + qP'(q)$ is the *marginal* revenue. The name of the transformation is chosen to be consistent with this and similar examples. The Bulow-Pfleiderer demand class (Bulow and Pfleiderer, 1983) discussed later is also invariant under average-marginal transformations.

requires parameter restrictions involving parameters both from the supply side and the demand side. These restrictions may or may not be approximately satisfied in a particular market. If they are not satisfied, one may be tempted to conclude that our method is not applicable. Most likely this is the reason why economists have not found (or have not attempted to find) the kind of solutions we discuss in our paper.

We explain, however, that the range of applicability of our method is larger than it may seem at first sight as this issue does not pose a large problem. Even if the parameter restrictions are not satisfied, analytic solutions at other parameter values may be used to construct an interpolation that covers parameter values of interest. In this way one can extend the usefulness of our analytic method. Another way is to realize that a given demand or cost function may be *approximated* by functions that satisfy our restrictions, in which case the restrictions are satisfied by choice.[4]

While our approach is useful in many fields of economics, as we illustrate, the main application we focus on in this paper belongs to the field of international trade. Analytic tractability has been important for international trade to the extent that almost all models assumed constant marginal costs of both production and logistics/shipping. Under such assumptions trade models are much more straightforward to solve. Yet, as we discuss in detail, these assumptions contrast with models of cost used by the logistics managers that economists are presumably attempting to describe. As we show, our functional forms preserve analytical tractability while allowing the realism of matching such models.

Our primary application in this paper shows how such more realistic models of cost can help resolve the trade cost puzzle in a model of world trade flows with heterogeneous firms.[5] Standard models of international trade attribute the observed rapid falloff of trade flows with distance to trade costs that increase dramatically with distance. But we have no reason to believe that such dramatic distance dependence of trade costs exists in the real world. Container shipping charges depend on distance only modestly, and in any case, represent only a tiny fraction of the value of the transported goods. A similar statement holds for the so-called iceberg trade costs, i.e., the damage of goods during transportation: We know goods typically do not get damaged during transport, and if they do, the damage probability is unlikely to strongly increase with the distance a shipping container traveled over the ocean. While trade costs may be sizeable, they are much more likely to be associated with shipment preparation and coordination or with loading and unloading, rather than with the distance traveled over the ocean. For this reason, the rapid falloff of trade with distance represents a puzzle from the point of view of standard models of international trade.[6]

Our model resolves this puzzle in a very natural way. Firms find it costly to produce larger

[4]Yet way of extending the usefulness of the solutions is to use Taylor series expansions around them, which may be useful for certain types of models.

[5]Even though we do need considerable computational power to fit our model to the data, without the tractability of our functional forms the computations would be significantly harder and we would not have attempted to obtain a calibration of world trade flows that we discuss below.

[6]This puzzle in various forms has been discussed by many authors; see Disdier and Head (2008) for an overview and Head and Mayer (2013) for an in-depth discussion of the problem.

quantities due to increasing marginal costs of production. At the same time, they find it beneficial to concentrate their exports to a few destinations due to economies of scale in shipping. With this cost structure, even a small cost advantage of a particular destination will be enough to make the firm export there instead of other destinations. If trade costs are slightly smaller for closer destinations, this cost advantage will lead to substantially larger trade flows at smaller distances and substantially smaller trade flows at larger distances.

The model also resolves a puzzle related to firm entry into export markets. Although it is not as widely discussed as the trade cost puzzle, it is a clear empirical regularity that models with constant marginal costs cannot address in a natural way. In the data, one can often see two similar firms, say, from China, one exporting to, say, Portugal and not to Greece and the other exporting to Greece and not to Portugal. To reconcile such patterns with the assumption of constant marginal cost of production, standard international trade models introduce stochastic cost shocks specific to each firm-destination pair. These cost shocks have to be dramatically large. For the second firm they need to offset the entire profit the first firm makes from its exports to Portugal. In the absence of any real-world phenomenon that could lead to cost shocks of this kind, this represents a puzzle.[7]

Our model explains this puzzle in a straightforward way. With increasing marginal costs of production and economies of scale in shipping, firms need to solve a combinatorially difficult problem of choosing export destinations.[8] Different approximate solutions of this choice problem can lead to different sets of export destinations, even if the maximized profits are almost the same. One approximately optimal set of export destinations may include Portugal, while another one may include Greece.

We solve the model using an iterative method involving an outer loop and an inner loop. The outer loop requires an evaluation of firms' profit functions for many discrete choices of export destinations. Our functional forms bring a tractability advantage that makes these evaluations fast. In the inner loop, we solve for a general equilibrium of the world economy keeping the discrete choices fixed. There using our functional forms is helpful because it allows for an analytic calculation of derivatives that are needed for accelerated gradient descent algorithms.

Separately, we develop many other applications of the proposed functional forms. For the model of outsourcing decisions in a sequential supply chain constructed by Antràs and Chor (2013), we reformulate the theory to simplify the analysis and use this new formulation to apply our functional forms. This allows us to show that for more realistic demand functions, outsourcing occurs at both the early (viz. raw materials) and late (viz. final commercial sales) stages of production, while intermediate stages are performed in-house, corresponding to common observation of outsourcing patterns. For a model of labor bargaining by Stole and Zwiebel (1996a,b), we tractably generalize the closed-form solutions that have been found for linear or constant-elasticity demand and show that for realistic demand patterns the employment effects of bargaining have interesting and intuitive

[7]We discuss alternative mechanisms in Section 4.

[8]In economics there are many combinatorially difficult problems, and we expect our methods to be useful there.

cyclical patterns. We also discuss applications to imperfectly competitive supply chains, two-sided platforms, selection markets, auctions, and, extensively, monopolistic competition.

Finally, we show that our method may be thought of as a discrete approximation to a logarithmically modified Laplace transform. It may also be thought of as a sieve method of non-parametric econometrics. In addition, the transformed variables reveal economic properties of demand functions that would appear accidental otherwise.

The next section provides a quick illustration of our functional forms with a focus on modeling demand under income inequality. Section 3 presents our main theoretical results. Section 4 focuses on modeling world trade. Section 5 discusses other applications. Section 6 develops the theory connecting our tractable functions to a logarithmically modified Laplace transform. The paper also includes an appendix and supplementary material.

2 Example: Replacing Constant-Elasticity Demand

2.1 Constant-elasticity demand and its flexible replacement

The most canonical and widely-used demand form in economic analysis is the constant-elasticity specification, corresponding to inverse demand $P(q) = aq^{-b}$. It is frequently used because of its analytic tractability. Historically, it appeared in the economic literature because in discrete-choice settings it is plausible that product's valuations follow the income distribution and the income distribution was believed to be approximately Pareto, i.e., power-law.[9] Modern data of income, however, led to different conclusions on the shape of the income distribution.[10]

In this section we discuss another demand form that is also highly analytically tractable but has more flexibility. This flexibility allows us to get a much better match to the income distribution.[11]

[9] In cases where each individual can consume at most one unit of an indivisible product, the inverse demand function equals the reversed quantile function of the distribution of valuations (willingness to pay), up to constant rescaling. The reversed income quantile function here refers to the function that maps a given quantile q measured starting at the top of the income distribution to the corresponding valuation level. Note also that, of course, we do not wish to say that the most important property of constant-elasticity demand lies in the context in which it first appeared. We are merely using this example as an illustration of our approach to demand functions.

[10] The origin of constant-elasticity demand historically appears to be the argument by Say (1819) that willingness to pay for a typical discrete-choice product is likely to be proportional to income, and thus that the distribution of the willingness to pay should have the same shape as the income distribution. (Say's assumption is likely to be approximately correct for example for products that save a fixed amount of time to the owner, independently of their wealth.) Since early probate measurements of top incomes exhibited power laws (i.e., Pareto distributions) (Garnier, 1796; Say, 1828), by extrapolation Dupuit (1844) and Mill (1848) suggested that demand would have a constant elasticity. This observation appears to be the origin of the modern focus on constant elasticity demand form (Ekelund and Hébert, 1999; Lloyd, 2001). However evidence on broader income distributions that became available in the 20th century as the tax base expanded (Piketty, 2014) shows that, beyond the top incomes that were visible in 19th century data, the income distribution is roughly lognormal through the mid-range and thus has a probability density function that is bell-shaped, rather than power-law. Distributions that accurately match income distributions throughout their full range (Reed and Jorgensen, 2004; Toda, 2012, 2017) have a similar bell shape but incorporate the Pareto tail measured in the 19th century data.

[11] Similarly, this flexibility could allow us to get a better match to a distribution of valuations in cases where it differs from the exact income distribution.

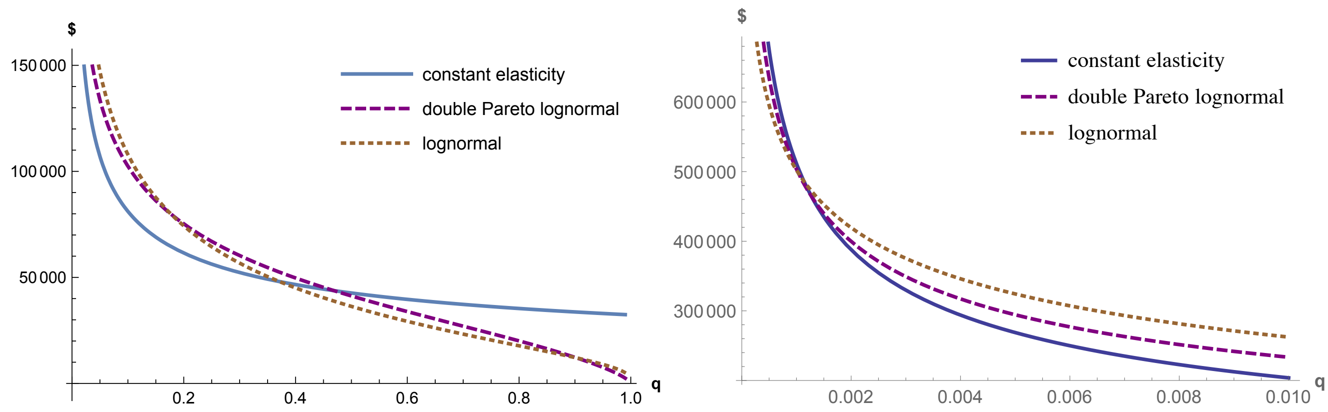


Figure 1: Comparing the fit of the best-fit lognormal and the best-fit constant-elasticity form to a double-Pareto lognormal estimation of the 2012 US income distribution, represented as a demand function (reversed quantile function). Dollars at any reversed quantile represent the income of the corresponding individual. On the left is the fit for the full income distribution, while the right shows the upper tail. We used a standard calibration of a double Pareto lognormal proposed by Reed (2003) and used the generalized method of moments to find the constant-elasticity demand function that best fits this throughout the full range of the income distribution. In the upper tail the constant elasticity approximation is a bit better of a fit than the lognormal. However, in the rest of the income distribution its fit is terrible, while the lognormal fits quite well (although in economic models it is harder to work with).

As an illustration, we show that our proposed demand form leads to substantially different policy implications than the constant-elasticity form in the socially important case of bias of innovative technical progress.

Consider the task of representing the empirical income distribution using a corresponding demand function. Constant-elasticity demand fails this purpose, as illustrated in Figure 1. This is because the income distribution is not Pareto but approximately double Pareto lognormal (Reed and Jorgensen, 2004; Toda, 2012, 2017). Working with the double Pareto lognormal distribution (or with the more loosely fitting lognormal distribution) in economic models would be quite difficult.[12] To overcome this difficulty, we propose a functional form that allows for the same basic shape, but leads to calculations almost as easy as for the constant-elasticity form:

$$P(q) = m - ma_- \left(\frac{q}{q_0}\right)^{-b} - ma_+ \left(\frac{q}{q_0}\right)^{b}, \qquad a_- \equiv 1 - a_+. \tag{1}$$

A set of parameter values that matches the income distribution (for the US in 2012) very well is $a_- = -1/2, a_+ = 5/2, b = 2/5$. We obtained these values by a generalized-method-of-moments fit and rounding the results. The match is illustrated in Figure 2.

[12]See Footnote 14.

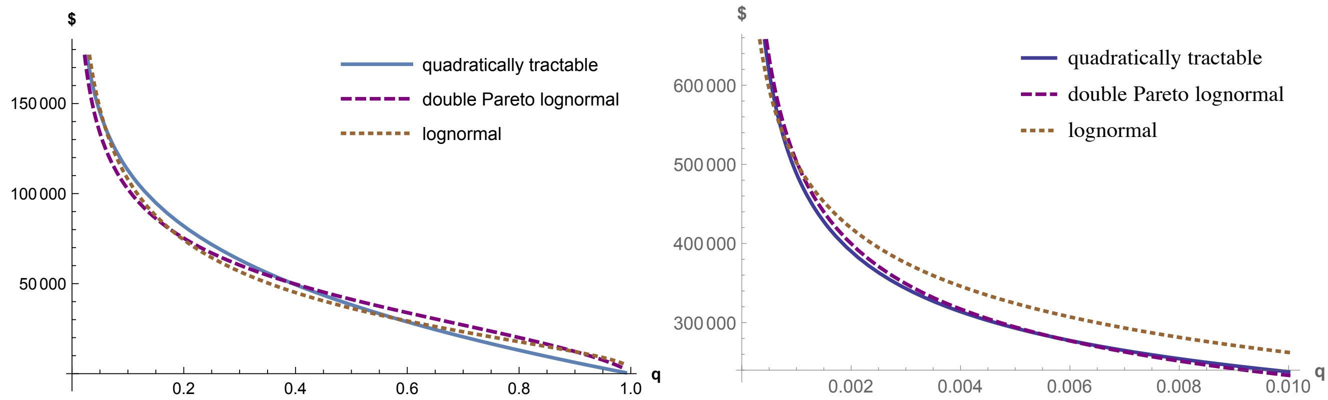


Figure 2: Comparing the fit of the best-fit lognormal and the best-fit quadratically solvable form to a double-Pareto lognormal estimation of the US Income distribution, represented as a demand function (reversed quantile function). Dollars at any reversed quantile represent the income of the corresponding individual. On the left is the fit for the full income distribution, while the right shows the upper tail.

2.2 Bias in technological progress

As a simple, illustrative application, we discuss the case of bias in technological progress described in Kremer and Snyder (2015, 2017). First, we do that for the case of constant-elasticity demand and explain why it is highly tractable. Then we turn to our proposed demand form in Equation 1 and show that it preserves a high degree of tractability that constant-elasticity demand has.

When the private sector decides which products to develop, it chooses profit-making products, not necessarily those products that create the greatest social value. This bias in technological progress depends on the discrepancy between private and social gains. Kremer and Snyder (2015, 2017) consider the fraction of the social gains from creating a new product that may be appropriated by a monopolist[13], referred to as the appropriability ratio, and show that the maximal fraction of potential surplus that may be lost due to imperfect appropriability is equal to one minus this appropriability ratio. They compare different demand functions since they lead to different bias in research and development, but always assume no costs. Here we assume a fixed demand function and consider biases at different levels of marginal production cost. We walk quite didactically through the process of solving the model in order to illustrate the source of the tractability of the constant-elasticity form and why it carries over to our proposed generalized form but not to the lognormal distribution form. We then follow Kremer and Snyder (2017)'s argument that a sensible demand function is one matching the world income distribution and use this as motivation for using our form to study the impact of cost on the appropriability ratio, which is very different under our form than under constant elasticity.

Consider a monopolist with a constant marginal cost c and constant-elasticity (inverse) demand $P(q) = aq^{-b}$. Her marginal revenue is $P(q) + P'(q)q$. Under the constant elasticity form, $P'(q)q =$

[13]Budish, Roin and Williams (2015) studied this problem recently in a different context.

$-abq^{-b}$, which has the same form as $P(q)$, just a different multiplicative constant out front. For this reason, the marginal revenue has the same form as well: $\mathrm{MR}(q) = a(1-b)q^{-b}$. The monopolist optimally equates it to the marginal cost, so the optimal quantity may be determined by solving the linear equation $a(1-b)x = c$ with $x \equiv q^{-b}$, yielding $q = (a(1-b)/c)^{1/b}$. From this it follows by substitution that the firm's absolute markup is $\overline{\mathrm{PS}} \equiv \mathrm{PS}/q = cb/(1-b)$, where PS stands for the producer surplus. Furthermore, the average consumer surplus also has the same form as $P(q)$, differing only by a multiplicative constant:

$$\overline{\mathrm{CS}} \equiv \frac{\mathrm{CS}}{q} = \frac{\int_0^q P(\tilde{q})\, d\tilde{q} - P(q)q}{q} = \frac{\frac{a}{1-b}q^{1-b} - aq^{1-b}}{q} = \frac{ab}{1-b}q^{-b}.$$

Evaluated at the optimal quantity, the average consumer surplus is $\overline{\mathrm{CS}} = cb(1-b)^{-2}$. The appropriability ratio, i.e., the ratio of producer surplus and the total surplus, may be evaluated as $\overline{\mathrm{PS}}/\left(\overline{\mathrm{PS}} + \overline{\mathrm{CS}}\right) = (1-b)/(2-b)$, which is a constant independent of cost. Thus all products have precisely the same appropriability ratio, and cost is irrelevant to the bias of investments in research and development.

If we tried to investigate this problem in a tractable way for more general demand functions that have been used in the economic literature, we could use the Bulow-Pfleiderer demand introduced in the next section, which includes both constant-elasticity and linear demand as special cases. However, we would again find that the cost c has no impact on the bias of technical progress.

If instead, we tried to investigate the implications of demand curves corresponding to other distributions of product valuations, such as the lognormal distribution or the double-Pareto lognormal distribution (which fits the income distribution), we would quickly find that such investigation cannot be carried out analytically.[14]

Here we show that working with the demand form in Equation 1 is much easier and elegant and leads to substantive economic results. Its marginal form $P'(q)q$ has the same functional form as $P(q)$ itself:

$$P'(q)q = mba_{-}\left(\frac{q}{q_0}\right)^{-b} - mba_{+}\left(\frac{q}{q_0}\right)^{b}.$$

If we introduce the notation

$$a_{n,-} \equiv (1-b)^n a_{-}, \qquad a_{n,+} \equiv (1+b)^n a_{+}, \qquad x \equiv \left(\frac{q}{q_0}\right)^{b},$$

[14]For a lognormal distribution with mean μ_ℓ and standard deviation σ_ℓ of the exponent, the inverse demand is $P(q) = \exp\left(\sigma_\ell \Phi^{-1}(1-q) + \mu_\ell\right)$, where Φ is the standard normal cumulative distribution function and where we normalized maximum demand to 1. There is no analytic solution to the monopolist's optimization condition $\mathrm{MR} = c$ because the following expression is too complicated: $P'(q)q = -\sqrt{2\pi}\sigma_\ell q \exp\left(\sigma_\ell \Phi^{-1}(1-q) + \mu_\ell + \frac{1}{2}\left[\Phi^{-1}(1-q)\right]^2\right) = -P(q)\sqrt{2\pi}\sigma_\ell q \exp\left(\frac{1}{2}\left[\Phi^{-1}(1-q)\right]^2\right)$. The more realistic double-Pareto lognormal distribution leads to even more complicated expressions. Clearly, if demand functions of this kind were used inside larger models, the absence of analytic tractability could quickly become a significant obstacle.

the monopolist's first-order condition is just the quadratic equation

$$-\ a_{1,-} + \left(1 - \frac{c}{m}\right) x - \ a_{1,+} x^2 = 0.$$

This leads to the closed-form solution

$$q = q_0 x^{1/b}, \quad x = \frac{1}{2a_{1,+}} \left(1 - \frac{c}{m} + \sqrt{\left(1 - \frac{c}{m}\right)^2 - 4a_{1,-}a_{1,+}}\right).$$

The per-unit consumer and producer surplus again take the same functional form:

$$\overline{\text{CS}} = -mb\ \tilde{a}_- \left(\frac{q}{q_0}\right)^{-b} + mb\ \tilde{a}_+ \left(\frac{q}{q_0}\right)^{b}, \quad \overline{\text{PS}} = m - c - ma_- \left(\frac{q}{q_0}\right)^{-b} - ma_+ \left(\frac{q}{q_0}\right)^{b}.$$

The appropriability ratio is then

$$\frac{\overline{\text{PS}}}{\overline{\text{PS}} + \overline{\text{CS}}} = \frac{m - c - m\ a_- \left(q / q_0\right)^{-b} - m\ a_+ \left(q / q_0\right)^{b}}{m - c + ma_{-1,-} \left(q / q_0\right)^{-b} + ma_{-1,+} \left(q / q_0\right)^{b}} = \frac{(1 - b^2)\left(-a_- + a_+ \left(q / q_0\right)^{2b}\right)}{-\left(2 + b - b^2\right) a_- + \left(2 - b - b^2\right) a_+ \left(q / q_0\right)^{2b}},$$

where the last equality was obtained by substituting for the marginal cost from the first-order condition. Substituting the parameter values we specified right after Equation 1 gives

$$\frac{\overline{\text{PS}}}{\overline{\text{PS}} + \overline{\text{CS}}} = \frac{21 + 105 \left(q / q_0\right)^{4/5}}{56 + 180 \left(q / q_0\right)^{4/5}}.$$

This equals $21/56 \approx 37.5\%$ for $q = 0$ (when the product serves a tiny fraction of the population) and monotonically increases in q to $53/118 \approx 53.4\%$ for $q = q_0$ (when most of the population is served). This suggests a bias towards cheap, mass-market products and away from expensive products that mostly cater to the rich; of course, all this analysis is based, like Kremer and Snyder's, on aggregate surplus and might well reverse if distributional concerns were incorporated.

While we focused here on biases from the appropriability ratio, it can be shown (in closed-form) that many other aspects of standard intellectual property policy differ substantially under our form from the results under the constant pass-through class. For example, under our form the ratio of consumer surplus to monopoly deadweight loss is much greater (usually by several times) than under the Bulow-Pfleiderer class so that patents are more desirable and optimal patent protection is greater than under the standard forms. Similarly allowing pharmaceutical producers to price discriminate often increases deadweight loss under the standard forms (Aguirre et al., 2010), while it is always beneficial under our form. Thus the standard forms are substantively misleading on a number of issues and the added complexity of using our form is minimal.

3 Central Results

In the previous section we focused on a particular functional form derived from our theory, a particular calibration target (the US income distribution) and a particular application. However, our approach applies much more broadly. We characterize all functional forms that have the useful property of the form above: namely that, in the language of demand curves, linear combinations of marginal revenue and inverse demand take the same form as inverse demand itself. Within these we then identify functional forms that lead to closed-form solutions utilizing power functions and the method of radicals.

3.1 Form preservation under the average-marginal transformation

Let us denote by $F(q)$ the average of an economic variable that depends on q, where a baseline interpretation of q is a quantity of a good. The marginal variable is then $(qF(q))' = F(q) + qF'(q)$. We now formally define what it means for these two variables to have the same functional form, as we alluded to in the previous section.

Definition 1. (Form Preservation) *We say that a functional form class $\mathcal{C}$ is* form-preserving *under average-marginal transformations if for any function $F(q) \in \mathcal{C}$, the class also contains any linear combination of $F(q)$ and $qF'(q)$. In other words, $F \in \mathcal{C} \Rightarrow \forall (a,b) \in \mathbb{R}^2 : \; aF + bqF' \in \mathcal{C}$. In economic terms, we interpret $F(q)$ as the average of the variable $qF(q)$, such as revenue or cost, and $F(q) + qF'(q)$ as its marginal counterpart. This definition thus states that any linear combination of the average and marginal variables belong to the defined class of functions.*[15]

Obviously, if $\mathcal{C}$ is taken to be a sufficiently large (e.g. infinite-dimensional) class of functions it may be form-preserving in a fairly mechanical way. For example, if it is the set of all analytic functions with the domain $(0, \overline{q})$ for some $\overline{q}$ then we know that $aF(q) + bqF'(q)$ is also analytic and has at least as large a domain. This observation is not very useful for the purposes of tractability because the set of all analytic functions with this domain contains many that, as we discussed in the previous section, are not tractable using standard analytic and computational methods.

Thus we will naturally wish to consider smaller classes. It is, therefore, useful to identify the most general set of finite-dimensional functional form classes that are form-preserving under the average-marginal transformations $F \to aF + bqF'$. Before stating the characterization theorem, let us briefly clarify what we mean by the dimensionality of a functional form class. For example, a functional form class $a_1 e^{-a_2 q}$, where a_1 and a_2 are continuously varying real numbers is two-dimensional, while $a_1 e^{-a_2 q} q^{-a_3}$ with continuously varying real a_1, a_2, and a_3 is three-dimensional.[16]

[15] Note that any form-preserving class is also form-preserving under multiple applications of operators of this type.

[16] While this intuitive description is sufficient for practical purposes, more formally we say that an *m-dimensional functional form class* is a subset of a space of functions (of a scalar, continuous variable) that is homeomorphic to an m-dimensional manifold, possibly with a boundary. Such manifold, with or without a boundary, is often referred to as the *moduli space*.

Theorem 1. (Characterization of Form-Preserving Functions) *Any real finite-dimensional functional form class with domain* $(0, \infty)$ *(or an open subinterval of it) that is form-preserving under average-marginal transformations must be a set of linear combinations of*

$$(\log q)^{a_{jk}}\, q^{-t_j}, \quad a_{jk} = 0, 1, ..., n_j^{(1)}, \quad j = 1, 2, ..., N_1,$$

$$(\log q)^{b_{jk}} \cos\left(\tilde{t}_j \log q\right) q^{-\hat{t}_j}, \quad b_{jk} = 0, 1, ..., n_j^{(2)}, \quad j = 1, 2, ..., N_2,$$

$$(\log q)^{c_{jk}} \sin\left(\tilde{t}_j \log q\right) q^{-\hat{t}_j}, \quad c_{jk} = 0, 1, ..., n_j^{(2)}, \quad j = 1, 2, ..., N_2,$$

where $\{t_j\}_{j=1}^{N_1}$*,* $\{\tilde{t}_j\}_{j=1}^{N_2}$*, and* $\{\hat{t}_j\}_{j=1}^{N_2}$ *are fixed sets of real numbers and* $N_1, N_2 \in \mathbb{N}$ *. If we exclude functions oscillating as* $q \to 0_+$*, only the functions in the first row are allowed. In that case the most general form is the set of linear combinations of*

$$q^{-t_j}, \quad q^{-t_j} \log q, \quad q^{-t_j} (\log q)^2, \quad ... \quad , q^{-t_j} (\log q)^{n_j}, \quad j = 1, 2, ..., N_1.$$

The proof is provided in Appendix A.

3.2 Tractability

We now provide a specific formal definition of "tractability" that allows us to characterize the class of form-preserving functional forms that have various levels of such tractability. While the term tractability is constantly invoked in economics papers to justify various "simplifying" assumptions, it is almost never defined formally.[17]

A potential reason for this is that there is no standard, clear definition within applied mathematics of the notion of tractability of the solution of mathematical equations. The theory of polynomial equations establishes that generic polynomial equations of degree at most four have solutions in terms of "the method of radicals" (roots of different orders) and that generic polynomial equations of higher degree have no such solutions, according to the Abel–Ruffini theorem. But this theory does not imply that one could not extend the list of "closed-form" functions by adding some other functions (other than roots) to provide solutions to higher order polynomials. In practice, polynomial equations of any reasonably low order (say less than a hundred) can be solved extremely rapidly by standard mathematical software (Kubler et al., 2014).[18]

For this reason, we use a specific definition of tractability, which we call *algebraic tractability*, that is very simplistic: an equation is algebraically tractable at some level k if it can be solved

[17] Of course, in other contexts the word "tractability" may have other meanings that are also useful. We specify below what we mean by "tractability" in this paper.

[18] Of course, the notion of "tractability" and "closed-form solutions" is subjective to some extent. Equations whose solutions may be expressed in terms of functions that are familiar enough are often said to have closed-form solutions. That does not imply, however, that such notion is meaningless. Familiar functions are easier to work with for researchers thanks to existing intuition, as well as thanks to their implementation in symbolic or numerical software. In this paper we made definite choices to resolve the terminological ambiguity.

using power functions and a solution to a polynomial equation of degree no greater than k. While this definition eliminates many other functions with known solutions, it does a good job capturing existing forms that are widely considered tractable while allowing an extension to richer forms in a pragmatic manner given the ease with which polynomial equations can be solved both analytically and computationally (Kubler and Schmedders, 2010).

An important feature of the (non-oscillating) class of functional forms in Theorem 1 is that if we include terms with powers of logarithms we must also include all terms with powers of logarithms below this. That is, if the class includes linear combinations of $q\left(\log q\right)^2$ and $q^{-1/2}\left(\log q\right)^2$ it must also include linear combinations $q \log q$, $q^{-1/2}\log q$, q and $q^{-1/2}$. With a small number of (explicitly enumerable) exceptions, classes of functional forms like this can rarely be solved in closed-form because of the combination of power and logarithmic terms.[19]

On the other hand, the even-simpler class of sums of power functions nests all frequently-used tractable forms in the economic literature, namely constant-elasticity demand combined with constant marginal cost, linear demand combined with linear marginal cost as in Farrell and Shapiro (1990), and the "constant pass-through" demand of Bulow and Pfleiderer (1983) (henceforth BP) with constant marginal cost.[20,21] As a result, we focus on functional form classes composed of linear combinations of power functions q^{-t_j}.[22]

The BP demand corresponds to $P\left(q\right) = p_0 + p_t q^{-t}$ for some real constants t, p_0 and p_t, not necessarily all positive. In a monopolist's first-order condition, constant marginal cost enters in the same way as $-p_0$. In this sense, constant marginal cost is compatible with this demand side specification. (Similarly, linear marginal cost would be fully compatible with the demand side in the special case of $t = -1$, i.e., linear demand.)

Using the BP demand form with constant marginal cost leads both to tractability and to an important substantive implication: the constancy of the pass-through rate of the constant marginal cost to price. However, it is clearly possible to preserve the former property without the latter.

[19] The most notable exception is the case when only a single power of q is used which can be divided out of the equation to yield a polynomial in $\log q$. While this class is of some interest, we do not focus on it here because it has the unappealing property that if one wishes to include a constant term (which is often desirable as we discuss below) one is limited to a small number of powers of logarithms and all other parameters are set. There are other specific exceptions and exploring the use of these is an interesting direction for future research, but none offers the flexibility afforded by power functions that we focus on below. This is likely why they have formed the basis of so much prior work. We thus see the logarithm-based forms instead as limits of the power forms that are worth including but not focusing on.

[20] In this section, for expositional purposes, we discuss tractability from the point of view of monopoly problems. But it is worth noting that tractability considerations would be exactly the same for Cournot oligopoly and very similar in the many applications we discuss in this paper.

[21] The BP demand, defined below, gives constant pass-through rate of specific taxes to monopolist's prices only in the case of a constant marginal cost. For this reason, we prefer to use the term Bulow-Pfleiderer (BP) demand, instead of the frequently used term "constant pass-through demand".

[22] There are a few cases not nested in the forms of Theorem 1 for which the firm's first-order condition may be solved. Hyperbolic demand curves used by Simonovska (2015) are one of them. Cases where the solution is in terms of the Lambert W function are the exponential utility function of Behrens and Murata (2007, 2012) and single-product versions of the Almost Ideal Demand System and of translog demand.

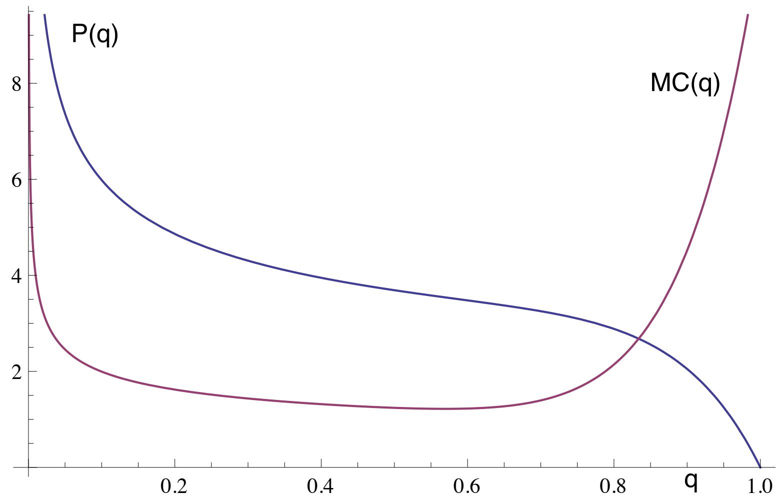


Figure 3: Example of a bell-shaped-distribution-generated demand and U-shaped cost curve contributing to equilibrium conditions that can be solved linearly: $P(q) = 3\left(q^{-0.3} - q^{10}\right)$ and $MC(q) = q^{-0.3} + 10q^{10}$.

For example, consider inverse demand and average cost of the form $P(q) = p_s q^{-s} + p_t q^{-t}$ and $AC(q) = ac_s q^{-s} + ac_t q^{-t}$.[23] Then the monopolist's first-order condition gives

$$(p_s - ac_s)(1-s)\, q^{-s} + (p_t - ac_t)(1-t) q^{-t} = 0 \implies q = \left(-\frac{(p_s - ac_s)(1-s)}{(p_t - ac_t)(1-t)}\right)^{\frac{1}{s-t}}.$$

This more general form thus still leads to a closed-form solution but offers substantially more flexibility. For example, it can accommodate simultaneously U-shaped cost curves and demand curves generated by a bell-shaped valuation distribution (in the sense of discrete choice). Figure 3 provides an example. A disadvantage of this form, however, is that it does not include a constant term. A constant term would have been useful for studying the pass-through rate and similar comparative statics. Another disadvantage is the absence of an explicit expression for the direct demand $Q(p) = P^{-1}(p)$.

It is thus useful to look beyond systems that lead to a linear equation (after a substitution using a power function). Quadratic, cubic and quartic equations also yield closed-form solutions by the method of radicals. Furthermore, polynomials of higher, but still small, order can be solved extremely quickly by most mathematical software without resorting to numerical search. For this reason, we define tractability in terms of the degree of polynomial solution a form admits.

Definition 2. (Tractability) *We say that an economic problem involving a scalar q is* algebraically tractable at level k *if a definite power of q is the solution of a polynomial equation of order k. For short we often refer to this simply as "tractability" and use adverbial forms for low k (e.g. linearly or quadratically tractable). By classical results of the theory of polynomial equations, only for $k \leq 4$*

[23] Mrázová and Neary (2014) studied the properties such bi-power form applied to inverse demand functions in combination with constant marginal cost. Their goal was not to obtain closed-form solutions.

Form	Tractability properties	Flexibility	Special cases	Historical notes
$F(q) = f_0 + f_{-1}q$	Linearly tractable Linearly invertible	Linear MC	Constant MC	Farrell and Shapiro (1990)
$F(q) = f_0 + f_t q^{-t}$	Linearly tractable Linearly invertible	Any constant pass-through	Linear Constant elasticity Exponential	BP constant pass-through demand
$F(q) = f_t q^{-t} + f_s q^{-s}$	Linearly tractable	Demand generated by bell-shaped distribution U-shaped cost	BP	Mrázová and Neary (2014) bi-power demand
$F(q) = f_t q^{-t} + f_0 + f_{-t} q^t$	Quadratically tractable Quadratically invertible	Income distribution U-shaped cost	BP	This paper
$F(q) = f_0 + f_t q^{-t} + f_{2t} q^{-2t}$	Quadratically tractable Quadratically invertible	Demand generated by bell-shaped distribution U-shaped cost	BP	Fabinger and Weyl (2012) APT demand

Table 1: Various classes of linearly or quadratically tractable, form-preserving equilibrium systems discussed in this or previous papers.

can such an equation be explicitly solved by the method of radicals and thus we refer to economic problems that are algebraically tractable at level $k \leq 4$ as analytically tractable*.*

We now characterize the set of functional forms from the power class that are tractable at level k for any positive integer k. A very naive conjecture based on the above discussion is that this is simply the set of forms that can be written as the sum of $k+1$ powers. To see why this is wrong, consider the equation

$$q + 1 + q^{-1/2} = 0.$$

This does not admit a quadratic solution, but can be solved cubically by defining $x \equiv q^{-1/2}$, transforming the equation into

$$x^{-2} + 1 + x = 0 \iff x^3 + x^2 + 1 = 0.$$

While the quadratic solution fails here, the cubic succeeds, because the gap between the power of the first and second term $(1 - 0 = 1)$ is not equal to that between the second and third term $(0 - (-1/2) = 1/2)$; instead it is twice the second gap, implying that there is a "missing" term $q^{1/2}$ in the equation. On the other hand, the equation

$$q^{1/2} + 1 + q^{-1/2} = 0$$

is quadratically tractable because the gap between the first and second powers equals that between the second and third. More broadly the number of such *evenly-spaced* powers sufficient to represent the class determines its level of tractability.

Theorem 2. (Closed-Form Solutions) *A functional form class $\mathcal{C}$ composed of all linear combinations of a finite set of powers of q is algebraically tractable at level k for generic linear coefficients if and only if the powers included are $\{a + bi\}_{i \in J}$ for some fixed real numbers a and b and some fixed set of integers $J \subseteq \{0, \ldots, j\}$ for a fixed integer $j \leq k$. More informally, a class of sum of power laws is tractable at level k if it consists of at most $k+1$ evenly-spaced powers of q.*

One example of applying this theorem was given in the previous section: our tractable form involves 3 evenly spaced power laws and thus is quadratically tractable. Table 1 summarizes a rich set of other possibilities covered by this theorem. The demand side of some of these has appeared in previous literature as we cite in the paper, though only in the case of Farrell and Shapiro (1990) are we aware of authors harnessing the accompanying cost-side flexibility.

3.3 Aggregation over heterogeneous firms

Models of international trade involving firm heterogeneity frequently use the framework of Melitz (2003) or Melitz and Ottaviano (2008), which assume respectively constant elasticity and linear demand. While these forms clearly play a role in the tractability of those models, the models are not always explicitly solvable even under these forms. Instead, the key properties these allow is that the firms' optimization problems may be solved explicitly and that aggregation integrals over heterogeneous firms may be expressed in closed form, assuming Pareto-distributed firm productivity.

We present a theorem that shows that substantial generalizations of these models can still lead to closed-form aggregation. We defer a full model set-up to Supplementary Material I.7.5, but it may be thought of simply as the Melitz (2003) model with relaxed functional form assumptions on the shape of demand, supply, and firm productivity distributions.

Theorem 3. (Aggregation) *Suppose that the utility structure implies an inverse demand curve* $P(q)$ *and that firms have marginal cost functions* $MC(q) = aMC_1(q) + MC_0(q)$*, where* a *is an idiosyncratic parameter influencing the firm's productivity, distributed according to a cumulative distribution function* $G(a)$*. Assume that* P*,* MC_0*,* MC_1*, and* G *are linear combinations of powers of their arguments, with the second order condition for the firm's profit maximization satisfied. Furthermore, suppose that the powers are such that* MC_1 *and the difference between marginal revenue and* MC_0 *are both of the form* $q^\beta N(q^\alpha)$ *with common* α*, but possibly differing* β *and polynomials* N*. Then the aggregation integrals for the firms' revenue, cost, and profit may be performed explicitly. The resulting expressions may contain special functions, namely the standard hypergeometric function, the standard Appell function, or more generally Lauricella functions, and in the case of high-order polynomials (higher-order tractable specifications), increasingly high-degree polynomial root functions.*

While this result is closely related to our general theory and our other applications (in particular, because this aggregation is possible when the relevant variables have our proposed forms), there are also a few differences worth noting. First, aggregation is still possible when the heterogeneous component $MC_1(q)$ of marginal cost is "shifted" (in the exponent space) by a uniform multiplicative power factor relative to $MR(q) - MC_0(q)$. This corresponds to the "possibly differing β" in the statement of the theorem. Second, our results here are about aggregation, not solution, and the resulting functions are not, therefore, solutions to polynomial equations but rather various functions that may be exotic to some economists, but are widely used in mathematics and related applied fields. Finally, as the complexity of the forms rises, it is the complexity of these functions that rises.

Closed-form aggregation is useful for at least three reasons. First of all, in the simplest cases the resulting aggregation integrals are just polynomial functions, which means that at the aggregate level the economic equations are relatively simple. Second, when the aggregation integrals lead to commonly used special functions, these are likely to be implemented in numerical software of the researcher's choice. The researcher gets a fast and numerically reliable implementation of these functions and their derivatives without spending time on approximation methods. Third, it is possible to take advantage of the properties of these functions that have been studied in the mathematical literature.

3.4 Interpolation between solutions

We have discussed how to obtain closed-form solutions in economic modeling. We used linear combinations of power function and imposed conditions on their exponents. It is natural to ask what happens if these conditions are not satisfied. Suppose we have a computationally intensive model whose numerical solutions rely on closed-form solutions to its sub-problems. If we relax our assumptions on the exponents we just mentioned, the sub-problems will not be solvable in closed form and obtaining numerical solutions to the full problem may require an excessive amount of time. Here we would like to point out that in this case we have another way to proceed: we can solve the full problem at special loci where the conditions on exponents are satisfied, and then interpolate between the resulting solutions.

Let us illustrate this approach with a toy example that is not computationally intensive. Consider a monopolistic firm with marginal revenue $\mathrm{MR}(q) = \mathrm{MR}_0 q^{-b_R}$, $b_R > 0$ and marginal cost $\mathrm{MC}(q) = \mathrm{MC}_0 + \mathrm{MC}_1 q^{-b_C}$, $\mathrm{MC}_1 > 0$. After the substitution $x = q^{-b_R}$, the firm's first-order condition becomes $\mathrm{MC}_0 + \mathrm{MC}_1 x^b = \mathrm{MR}_0 x$, with $b \equiv b_C/b_R$. This equation admits closed-form solutions by the method of radicals for $b \in \{-3, -2, -1, -\frac{1}{2}, -\frac{1}{3}, 0, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1, 2, 3, 4\}$. The second-order condition is satisfied only for $b < 1$, so we restrict our attention to the first 9 values. For illustration, consider the simple goal of finding numerical values of q for b between these points. Instead of solving the first-order conditions by usual numerical methods, we may interpolate between the closed-form solutions, say, using cubic splines. Figure 4 shows the result of such interpolation, as well as the true solutions to the first-order condition, for $\mathrm{MC}_0 = \mathrm{MC}_1 = \mathrm{MR}_0 = 1$. The agreement is extremely good, with average absolute value of relative deviations equal to 0.00013 and maximum absolute value of relative deviations equal to 0.00056.[24]

If variables of interest in large scale, computationally intensive problems are similarly well behaved, then clearly the interpolation method could save remarkable amounts of computation time and research budgets. There are other ways to extend the usefulness of the closed-form solutions to other parameter values. For example, it may be possible to perform a Taylor expansion around a given closed-form solution. Such approach may also be combined with the interpolation method.

[24] The corresponding values for Mathematica's Hermite polynomial interpolation were 0.00069 and 0.0021.

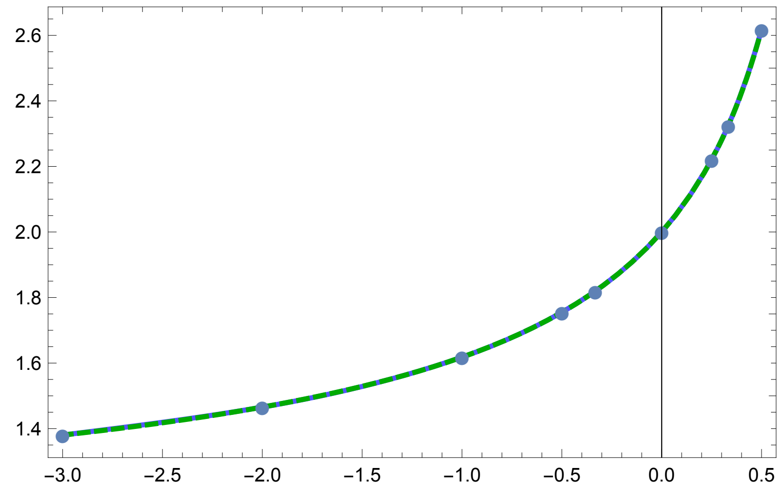


Figure 4: Comparison of an interpolation between analytic solutions and the correct values. The blue dots represent analytic solutions. The blue solid line corresponds to an interpolation using cubic splines. The green dashed line represents correct values.

More broadly, one can view our approach to economic modeling as resembling pragmatic approaches in Bayesian Statistics. In that literature, it is usually impossible to compute the posterior probability distribution associated with most prior distributions given the observed, often large, data set. It is therefore common to approximate the prior by one selected from a class of prior distributions which are known to update to another prior within that class in closed form as this minimizes the computational requirements of updating. In a similar manner, our tractable equilibrium forms may approximate arbitrary cost and demand curves, while allowing solutions in closed-form which allow nesting inside of computationally intensive models.

In the next two sections we explore concrete applications of our approach to closed-form solutions in economics. We will return to more theoretical matters in Section 6.

4 World Trade

4.1 Overview

In this section we present a large-scale empirical application of our analytic approach to flexible functional forms in economics: a model of world trade with a realistic cost structure for heterogeneous firms.

International trade researchers almost always postulated constant marginal costs. Firms were assumed to have a constant marginal cost of production. They were also assumed to face constant marginal costs of trade, either in the "iceberg" form (i.e., damage of goods as they are transported) or in a per-unit form. Both of these assumptions are unrealistic. When we depart from them, we find an interpretation of world trade flows that is dramatically different from the conventional view. The model's parameters and predictions take realistic values, which resolves empirical puzzles in the international trade literature.

Our model describes a world with multiple countries, with a general setup analogous to Melitz (2003). Our two important modifications are as follows. First, in addition to the usual iceberg cost, we allow for a specific cost of trade that varies non-linearly with the traded quantity. Second, production is subject to increasing marginal cost, designed to capture the difficulty of scaling the firm, e.g. due to internal agency problems. After discussing computational considerations and, separately, the two economic cost effects, we return to the setup of the main model in Subsection 4.6.

4.2 Computational considerations

In applied fields of economics, such as the study of international trade, researchers can quickly reach the limits of what is computationally feasible because of the number of economic agents and the high dimensionality of their choice sets (or state spaces). In our case, we study trade flows between many countries involving heterogeneous firms, each of which is facing a combinatorially difficult decision problem. With powerful hardware, software, and efficient algorithms, we were able to get a model fit for given parameter values in about a month and at a non-trivial cost. We were utilizing our analytic solutions to sub-problems, without which the computation would be substantially longer and more costly.

Our functional forms help us in two ways: First, to evaluate firms' sales decisions conditional on the level of their marginal cost of production and export entry decisions, we just need to evaluate closed-form solutions. This is crucial for being able to quickly evaluate a large number of alternative sales patterns a firm may consider, and to find some of the best ones. Second, conditional of all firms' export entry decisions, we can solve for the resulting general equilibrium of world trade by accelerated gradient descent algorithms, the Adam algorithm in our case. For this algorithm to be useful, we need to be able to calculate gradients of candidate solutions' loss functions (error functions) analytically. Because of the scale of the problem, we do not perform the gradient calculation by hand. Instead, we rely on automatic analytic differentiation software, namely the neural-network optimization package of PyTorch, which allows us to run all computations in a highly parallel fashion on graphics processing units (GPUs).[25]

[25] PyTorch is an open source software framework developed by Facebook primarily for deep learning in artificial neural networks. Its first version was released in 2017.

4.3 Firm-level economies of scale in shipping: a generalized Economic Order Quantity model

Most models of international trade assume that the costs of trade are of the "iceberg" type: a fraction of all goods transported is assumed to be destroyed in transit. It seems implausible that most of true trade costs would scale with trade volume and value in this manner.[26] A certain fraction of international trade papers, e.g. Melitz and Ottaviano (2008), allow for constant marginal per unit costs of trade.[27] However, the adoption of standardized shipping containers has made such constant marginal per-unit costs of transportation extremely low relative to the trade costs necessary to explain the rates of global trade flows.

We work with the assumption that most important trade costs come from coordination (shipment preparation) costs and inventory costs, which is why the logistics literature focuses on them. These costs depend on the frequency of shipping. If a firm ships too infrequently, it will face large costs associated with idle inventory. If it ships too frequently, shipment preparation costs will add up to a large number. Knowing this trade-off, the firm will choose an optimal frequency of shipping that balances these two effects. The resulting effective cost of trade then exhibits economies of scale: a firm wishing to ship only a small quantity on average per unit of time will find shipping to be costly per unit of quantity.

To gain empirical insight into the scale economies of international trade, we estimate a model of optimal shipping frequency using monthly international shipment data. Our approach generalizes the classic Economic Order Quantity model of Ford W. Harris, which is widely taught in operations management courses in business schools and applied by logistics planning managers in corporations.[28] Consider a firm that produces a single good in one country and wishes to ship to a different country quantity q per year, on average. The firm faces a tradeoff between inventory costs and coordination costs associated with frequent shipping. The average annual inventory cost C_i is linearly proportional to q and to the time T a typical unit of the good needs to remain in

[26] Tariffs would depend on value in the same way as iceberg trade costs, although the details of their impact would be different, as goods are not destroyed and governments collect tariff revenue. That said, most trade costs modeled as iceberg trade costs in the literature are not supposed to represent tariffs and we will not focus on tariffs in this paper, although, of course, they may be incorporated in our model.

[27] Per-unit costs of trade seem more realistic than costs of trade proportional to the goods' value, as documented by, e.g., Hummels and Skiba (2004). Note that this reference did not allow for non-linearity of trade costs.

[28] Despite not appearing in the international trade literature, the Economic Order Quantity (EOQ) model (Harris, 1913) is perhaps the most classical model of trade costs in the operations research literature and is regularly taught to business students as a method of optimizing their inventory decisions; see e.g. Cárdenas-Barrón et al. (2014) for highlights of its importance. Judging from the absence of citations, the academic international trade community is largely unaware of Harris' publication. When fixed costs per shipment are included in international trade models, they are incorporated in theoretical models with different structures. Those models are similar in spirit, but do not strictly speaking contain the EOQ model or its generalizations (Kropf and Sauré, 2014; Hornok and Koren, 2015a). These papers provide very useful insights into shipment frequency issues, and so does the purely empirical paper Hornok and Koren (2015b). Note also that economies of scale in shipping were studied by Anderson et al. (2014) and Forslid and Okubo (2016), but those approaches were not based on shipping frequency and in the former case involved external (i.e., not within-firm) economies of scale.

storage. If the size of each shipment is q_s, then T, in turn, is linearly proportional to q_s/q, implying $C_i = \kappa_i q_s$, for some constant κ_i. The coordination cost C_s of each shipment is proportional to its size: $C_s = \kappa_t q_s^\gamma$, $\gamma \in [0,1)$. (In addition, we could assume an additional term proportional to q_s, but this would not affect the optimal choice of q_s for given q.) The resulting average annual coordination cost is $C_t = C_s q/q_s = \kappa_t q q_s^{\gamma-1}$. Minimizing the sum of the inventory cost and the coordination cost leads to the optimal choice $q_s = \left(q(1-\alpha)\kappa_t\kappa_i^{-1}\right)^{\frac{1}{2-\gamma}}$, the minimized value $(2-\gamma)(1-\gamma)^{-\frac{1-\gamma}{2-\gamma}}\kappa_i^{\frac{1-\gamma}{2-\gamma}}\kappa_t^{\frac{1}{2-\gamma}}q^{\frac{1}{2-\gamma}}$, and the optimal frequency of shipping $f_s = q/q_s$ equal to

$$f_s = (1-\alpha)^{-\frac{1}{2-\gamma}}\kappa_i^{\frac{1}{2-\gamma}}\kappa_t^{-\frac{1}{2-\gamma}}q^{\frac{1-\gamma}{2-\gamma}}.$$

This result implies that we can infer the coordination cost exponent γ by examining the relationship between the average annual quantity shipped and the frequency of shipping. If we regress the logarithm of shipping frequency f_s on the logarithm of average annual quantity q, the resulting slope coefficient should equal $\beta \equiv (1-\gamma)/(2-\gamma)$. The model predicts that this coefficient always lies between 0 and 1/2, since $\gamma \in [0,1)$.

Our simple model of shipping frequency choice nests two important extreme cases. The original Economic Order Quantity model, in which the cost per shipment is fixed, corresponds to $\gamma = 0$ and $\beta = 1/2$, implying effective cost of trade (here inventory and coordination) proportional to $\sqrt{q}$. The other extreme case has $\gamma \to 1$ and $\beta \to 0$ and corresponds to effective cost of trade linearly proportional to q, i.e., constant marginal cost of trade, as assumed in almost all of the international trade literature.

To estimate β and to test the prediction that $\beta \in (0,1/2]$, we used a dataset on monthly shipments from China to Japan during years 2000-2006. We focus on firms in one narrowly-defined product category.[29] Our point estimate of β (averaged across industries) is 0.39 with a 95% confidence interval of [0.36,0.42].[30] We can thus clearly reject the null hypothesis that $\gamma = 1$ and $\beta = 0$, which would correspond to trade costs being linearly proportional to quantity shipped, as assumed in the vast majority of the international trade literature. We can also reject the original EOQ model, which would correspond to $\gamma = 0$ and $\beta = 1/2$. We see, however, that the original EOQ model is closer to reality than the linear proportionality assumption.

In our main trade model, we round the resulting value for β from 0.39 to 0.4. This estimate implies that increasing quantity by 10% reduces (the variable component of) the marginal cost of trade by 4%. We refer to the effective cost of trade as "cost of shipping", remembering that it

[29] We selected firms by requiring that they specialize in one product category (one 8-digit HS code). The exporting firm had to be active for more than two years to be included in our estimation sample. We selected industries that included at least 10 firms meeting these criteria, in order to work with industries that allow for a precise estimate of β. We were also careful to take into account potential effects of seasonality, which could affect our estimates. We constructed a measure of seasonal variations of exports for individual industries. Our estimates of β did not differ almost at all between industries with larger and smaller seasonality. We discuss more details in Appendix B.

[30] The confidence interval corresponds to a simple statistical model in which β for different industries is drawn from a normal distribution.

arises from per-shipment coordination costs and from inventory costs with optimally chosen shipping frequency. In the rest of this section, we use the notation ν_{LT} for what was $1-\beta = 1/(2-\gamma)$ here, and set $\nu_{\text{LT}} = 0.6$.

4.4 Export quantity determination

For clarity of exposition, let us now consider the problem of export quantity determination for a firm that faces trade costs found in the previous subsection. Similar ingredients will appear also in our main model described in Subsection 4.6.

A firm considers exporting to one foreign country. If it delivers quantity q_f there, it will receive revenue $R(q_f)$, for which we choose the form $R(q_f) = \nu_R^{-1}\kappa_R q_f^{\nu_R}$, where $\nu_R = 1 - 1/\sigma$ and $\sigma = 5$. The elasticity of demand $\sigma = 5$ is consistent with the typical range in the trade literature. The firm faces an iceberg trade cost factor τ, meaning that it needs to send τq_f in order for q_f to arrive. The shipping requires $L_T(q_f)$ units of labor, which translates into a cost $wL_T(q_f)$. We choose $L_T(q) = \nu_{\text{LT}}^{-1}\kappa_{\text{LT}} q^{\nu_{\text{LT}}}$, with $\nu_{\text{LT}} = 3/5$, in agreement with the previous subsection. In this illustrative example, we assume constant marginal cost MC.

The second derivative of the profit function $R(q_f) - \tau \text{MC} q_f + wL_T(q_f)$ is $\frac{2}{5}w\kappa_{\text{LT}} q_f^{-7/5} - \frac{1}{5}\kappa_R q_f^{-6/5}$, so the profit function is convex for $q_f \in \left(0, 32w^5\kappa_{\text{LT}}^5\kappa_R^{-5}\right)$ and concave for $q_f \in \left(32w^5\kappa_{\text{LT}}^5\kappa_R^{-5}, \infty\right)$. To identify the maximum, we just need to find potential local maxima in the second region and to check whether they are larger than zero. This is because at $q_f = 0$ the profit is zero, and as $q_f \to \infty$ it goes to $-\infty$.

The firm's first-order condition is

$$R'(q_f) - \tau \text{MC} - wL_T'(q_f) = 0 \quad \Longrightarrow \quad -\frac{w\kappa_{\text{LT}}}{\tau} q_f^{-\frac{2}{5}} + \frac{\kappa_R}{\tau} q_f^{-\frac{1}{5}} - \text{MC} = 0.$$

We recognize that the function of q_f on the left-hand side is one of our proposed tractable functional forms. We can, therefore, solve the first-order condition in closed-form, in this case using the quadratic formula. If $\text{MC} > \kappa_R^2/(4w\tau\kappa_{\text{LT}})$ there is no real solution and the firm will choose not to export. If $\text{MC} \leq \kappa_R^2/(4w\tau\kappa_{\text{LT}})$, the solution that lies in the $\left[32w^5\kappa_{\text{LT}}^5\kappa_R^{-5}, \infty\right)$ region equals

$$q_f = \left(\frac{\kappa_R + \sqrt{\kappa_R^2 - 4\tau \text{MC}\kappa_{\text{LT}} w}}{2\tau \text{MC}}\right)^5.$$

Plugging this position of the local maximum into the profit function gives

$$\frac{1}{192\text{MC}^4\tau^4}\left(\kappa_R + \sqrt{\kappa_R^2 - 4\text{MC}w\tau\kappa_{\text{LT}}}\right)^3\left(-16\text{MC}w\tau\kappa_{\text{LT}} + 3\kappa_R\left(\kappa_R + \sqrt{\kappa_R^2 - 4\text{MC}w\tau\kappa_{\text{LT}}}\right)\right)$$

The first two factors are positive, and the last one is positive if and only if $\text{MC} < 15\kappa_R^2/(64w\tau\kappa_{\text{LT}}) \approx 0.234\kappa_R^2/(w\tau\kappa_{\text{LT}})$. If this condition is satisfied, the firm will export the quantity satisfying the first-order condition, otherwise, it will export zero.[31] Thus for any level of marginal cost, the quantity chosen by the firm may be written compactly as

$$q_f = \left(\frac{\kappa_R + \sqrt{\kappa_R^2 - 4\tau \mathrm{MC} \kappa_{\mathrm{LT}} w}}{2\tau \mathrm{MC}} \right)^5 1_{\mathrm{MC} < \frac{15\kappa_R^2}{64 w \tau \kappa_{\mathrm{LT}}}},$$

where the second factor represents an indicator function. We see that the functional form allowed for a very simple and straightforward analysis. We also see that exporting may not be profitable even if there is no fixed cost of exporting. This implies that such model with constant elasticity of demand can generate an export cutoff without fixed costs of exporting.

In our main model described in Subsection 4.6, which no longer assumes that the marginal cost of production is constant, we still benefit from the closed-form characterization of the solution to the first-order condition in terms of the level of marginal production cost. This is both for the evaluation of the solution and for taking derivatives of the solution, as needed by gradient descent algorithms. Of course, the degree of the benefit grows in proportion to the number of potential export destinations.

4.5 Increasing marginal cost of production

Economies of scale, modeled using fixed costs of production, are present in most models of firms in the international trade literature. By contrast, diseconomies of scale almost never appear in that literature. Yet there are many reasons to believe that increasing marginal costs of production are similarly important in shaping firms' behavior. This is presumably why introductory economics classes frequently illustrate increasing marginal cost schedules. Beyond short-to-medium term capacity constraints and adjustment costs usually discussed in such courses, even in the longer term if a company decides to scale up its production an order of magnitude, it needs to introduce an additional layer of management hierarchy, which brings with it non-trivial agency problems. In a large organization incentives are diluted, and maintaining motivation, discipline, and output quality becomes harder.[32] Of course, managers of firms are intuitively aware of these problems, at least to some extent, and take them to account when shaping the structure of the firm.

[31] If exporting leads to zero profit just like not exporting, we specify that the firm chooses not to export.

[32] The restaurant industry is an obvious example: few people would associate chain restaurants with outstanding culinary experience. Another fairly obvious example is the automobile industry: there are many automakers in the world, each having a relatively small market share, very stable over time, even though cars produced by different automakers are highly substitutable from costumers' perspective. With constant marginal costs of production this would require a remarkably small dispersion of marginal costs across firms, which is especially hard to rationalize given the large observed fluctuations of currency exchange rates. Also, the increasing nature of marginal costs of production was one of the reasons why socialist economies were unsuccessful: state-controlled monopolies avoid duplication of effort in product design and other fixed costs of production, yet they suffer from severe agency problems that private sector competition can mitigate. Although here we emphasize increasing marginal costs of production in the long term, they are also interesting at short time scales; see Almunia, Antràs, Lopez-Rodriguez and Morales (2018) and references therein.

Issues of this kind are the subject of interest to vast literature within organizational economics, which includes Williamson (1967), Calvo and Wellisz (1978), and Tirole (1986).[33]

Estimating how much marginal costs increase with production volume is non-trivial since both economies and diseconomies of scale play a role in firm behavior. Our model provides a unique opportunity to obtain such estimates by matching firm-level multi-destination export data with world trade model solutions.

4.6 Model setup

Apart from the cost structure of the firms, our model is closely analogous to Melitz (2003), which many readers are familiar with. For this reason, we keep the description of the modeling setup succinct.

The world consists of N_c countries, indexed by k. In each country, different varieties ω of a differentiated good are produced by monopolistically competitive heterogeneous single-product firms using a single factor of production, for simplicity referred to as labor.

Consider a firm located in country k and identified by an index i. In order to produce a quantity q_i, the firm needs to pay a variable cost of $\frac{1}{1+\alpha}\kappa_{C,i}w_k q_i^{1+\alpha}$, where w_k is the competitive wage the firm's country k and $\kappa_{C,i}$ is a positive constant that depends on the firm. Importantly, the constant α determines how quickly marginal costs increase when any firm decides to scale up production; it is the elasticity of the marginal cost of production with respect to quantity. In addition to the variable cost, there is a fixed cost f_o of operation and a fixed cost f_x of exporting to a destination country k_d, expressed in units of domestic labor.[34]

Entry into the industry is unrestricted, but involves a sunk cost of entry f_e, again in units of domestic labor. Only after the entry cost has been paid does the firm learn its variable cost parameter $\kappa_{C,i}$, drawn from a distribution with cumulative distribution function $\tilde{G}(\kappa_C)$. When the value of $\kappa_{C,i}$ is revealed, the firm decides whether or not to exit the industry, and if it does not exit, whether to export to any of the other countries. In addition to endogenous exit, with a probability of δ_e per period the firm is exogenously forced to exit (starting from the end of the first period).

Trade costs have two components. The first corresponds to standard iceberg trade costs: in order of one unit of the good to arrive in the destination country k_d, τ_{k,k_d} units need to be shipped.[35] The second component requires using an amount of labor given by $L_{T,k,k_d}(q) = \nu_{\mathrm{LT}}^{-1}\kappa_{\mathrm{LT},k,k_d}q^{\nu_{\mathrm{LT}}}$, where we set $\nu_{\mathrm{LT}} = \frac{3}{5}$ to be consistent with the empirical value, as in Subsection 4.4.[36]

Consumers in each country have a CES utility function $U = \left(\int q_\omega^{1-\frac{1}{\sigma}} d\omega\right)^{\frac{\sigma}{\sigma-1}}$ that depends on the

[33] Oliver E. Williamson's Nobel lecture (Williamson (2009)) provides an excellent, compact discussion of the many things that may go wrong in a large organization. For a related discussion, see Tirole (1988).

[34] In general, we can allow for country-dependence of these costs: $f_{o,k}$ and f_{x,k,k_d}. We chose to make them country-independent for simplicity, not for tractability or computational feasibility reasons.

[35] Including also per-unit trade costs would not affect the computational feasibility of the model.

[36] The cost $L_{T,k,k_d}(q)$ is associated with coordination/shipment preparation tasks and with inventory costs. Its form is motivated by the empirical results of Subsection 4.3.

quantity q_ω of each variety ω consumed. As in Subsection 4.4, we set the elasticity of substitution σ equal to 5, which is consistent with the typical range in the existing empirical literature of about 4 to 8. This exact choice is motivated by analytic tractability. Each country k has an endowment of labor $L_{E,k}$, which is supplied at a country-specific competitive wage rate w_k mentioned above.

The revenue a firm can earn by selling a quantity q in a given market is $R_{k_d}(q) = \frac{\kappa_{R,k_d}}{\nu_R} q^{\nu_R}$, where $\nu_R = 1 - \frac{1}{\sigma}$. The factor κ_{R,k_d} is endogenously determined and depends on the price index and the consumption expenditures in the destination country.

The firm may choose to exit the industry (to save on the fixed cost f_o) or to operate and sell its product in a number of countries, earning a non-negative profit π per period of operation. In expectation, an entrant needs to break even: $\delta_e f_e = \mathrm{E}\pi$, which determines the equilibrium measure of firms in each country.[37] Similarly, labor markets in each country k need to clear, which means that the total labor demanded by firms at wage w_k needs to equal the labor endowment $L_{E,k}$. If we impose balanced budget conditions, consumers' expenditures equal their wage earnings, as firms earn zero ex-ante profits.[38]

4.7 The exporting firm's problem

Let us discuss the nature of the exporting firm's problem. Increasing marginal costs will limit the scale of the firm's production. Since trade is subject to decreasing marginal costs, the firm will concentrate its exports into a limited number of countries. The overall production level q_i of firm i as well as export market entry decisions are endogenous. For now let us consider the relation between of export quantities and q_i, conditional on having paid fixed costs of exporting to a number of countries.

The first-order condition for choosing the quantity q_{f,i,k_d} that should reach a foreign market k_d equates the marginal revenue and the comprehensive marginal cost that depends on the overall production level q_i:

$$R'_{k_d}\left(q_{f,i,k_d}\right) = \tau_{k,k_d}\mathrm{MC}_i\left(q_i\right) + w_k L'_{T,k,k_d}\left(q_{f,i,k_d}\right) \Rightarrow \frac{\kappa_{R,k_d}}{\tau_{k,k_d}} q_{f,i,k_d}^{-\frac{1}{5}} = \mathrm{MC}_i\left(q_i\right) + \frac{w_k \kappa_{\mathrm{LT},k,k_d}}{\tau_{k,k_d}} q_{f,i,k_d}^{-\frac{2}{5}},$$

in analogy with Subsection 4.4. The solution for q_{f,i,k_d} given q_i is:

$$q_{f,i,k_d} = \frac{1}{\left(2\tau_{k,k_d}\mathrm{MC}_i\left(q_i\right)\right)^5}\left(\kappa_{R,k_d} + \sqrt{\kappa_{R,k_d}^2 - 4 w_k \kappa_{\mathrm{LT},k,k_d}\tau_{k,k_d}\mathrm{MC}_i\left(q_i\right)}\right)^5$$

If the marginal cost of production $\mathrm{MC}_i\left(q_i\right)$ exceeds $\kappa_{R,k_d}^2/(4 w_k \kappa_{\mathrm{LT},k,k_d}\tau_{k,k_d})$, the first-order condition cannot be satisfied. For domestic sales we assume $\kappa_{\mathrm{LT},k,k} = 0$ and $\tau_{k,k} = 1$, so the optimal quantity sold domestically is simply $q_{i,k} = \left(\kappa_{R,k}/\mathrm{MC}_i\left(q_i\right)\right)^5$.

The total quantity q_i produced should equal the total of quantity sold domestically and sent

[37] The model has no explicit discounting of future utility, but δ_e plays a role similar to a discount rate.

[38] In our empirical setting we allow for budget imbalances that reflects similar imbalances in the data.

abroad: $q_i = q_{i,k} + \sum_{k_d \neq k} \tau_{k,k_d} q_{i,k,k_d}$, with q_{f,i,k_d} given by the formula above. This represents one equation for one unknown: q_i. Each root of this equation represents a candidate optimum for the firm.[39] The profit-maximizing choice(s) of destinations may then be found by evaluating total profits at these candidate optima. For a small number of countries this is simple, but for large N_c the problem becomes combinatorially difficult.[40] For this reason, when we solve the model for a large number of countries, we use approximate algorithms instead of an exhaustive search.[41]

4.8 Solution strategy

We solve the model using an iterative algorithm that has an outer loop and an inner loop.[42] In the outer loop firms decide whether or not they pay fixed costs of operation and fixed costs of exporting and commit to their decision. In the inner loop, we then solve for the general equilibrium of the world economy given these fixed-cost decisions.

Finding this general equilibrium without the tractable functional forms is computationally difficult since a multi-level nested iteration is very time-consuming. However, thanks to the analytic nature of our model, we were able to obtain the general equilibrium much faster using accelerated gradient descent in a space parametrized by quantities q, wages w, measures of firms M, price-index related variables κ_R, and country-level expenditures E. We used the Adam optimizer of Kingma and Ba (2014), as implemented in PyTorch, a neural network optimization software for GPU computing.[43] The gradients are computed analytically by automatic differentiation (autograd, in this

[39]Mathematically, the firm's choice of destinations in order to maximize profit is a submodular function maximization. This is because serving an additional set of markets A is less attractive if the initial set of markets S_l is larger: $\pi(S_2 \cup A) - \pi(S_2) \leq \pi(S_1 \cup A) - \pi(S_1)$ for $S_1 \subseteq S_2$ and $A \cap S_2 = \emptyset$. Here $\pi(S)$ denotes the optimal profit a firm can earn if it serves a set of markets S. If instead our problem was supermodular function maximization, it would be algorithmically easy. International trade papers such as Antràs et al. (2017) take advantage of supermodular function maximization being straightforward.

[40]For an in-depth discussion of combinatorial discrete choice problems in economics, see Eckert and Arkolakis (2017). The method that Eckert and Arkolakis propose would be useful for us if the number of countries we consider were substantially smaller. This is because the method reduces the exponent of an exponentially difficult problem, but does not change its exponential nature; submodular function maximization is NP-hard in general.

[41]We should clarify that even conditional on having made export fixed-cost payments, the number of candidate optima is still combinatorially large. This is because for some destinations it may be impossible to satisfy the FOC and in those cases we allow the firm to export zero amount there. To avoid this difficulty, when we consider candidate optima, we restrict attention to those that satisfy a particular ordering condition, without loss of generality. We rank export destinations by the level of (constant) marginal cost that would make them a profitable destination, in descending order. Then we require that if a firm exports a positive amount to a given destination, it also exports to all preceding destinations. Imposing this condition is without loss of generality because if a firm decides to export zero amount to a destination, it should not have paid the associated fixed cost of exporting in the first place.

[42]Due to its combinatorial nature, the exact version of our model is computationally extremely difficult. It may seem natural to try to obtain approximate solutions by first fixing aggregate variables in the model, solving for firm decisions given these aggregates, and then updating the aggregate variables based on the firms' behavior. We attempted to do that, but could not get results within a reasonable amount of time and budget. This is because for any values of aggregates, we needed to solve separate discrete choice problems by many firms, which requires a lot of time. For this reason, we used a different nesting of loops: we moved all discrete choice decisions into an outer loop of an iterative algorithm, and given these discrete choices, we solved for all continuous variables in an inner loop.

[43]We tried several accelerated and non-accelerated gradient descent algorithms. Adam performed the best.

case) and backpropagation.[44] , [45]

Given a solution to the general equilibrium problem, we then let firms reconsider their fixed cost payments. For numerical stability, we do not update at once the fixed cost payment decisions of all firms. Instead, for each productivity level in a country we introduce $N_v = 10$ versions (copies) of firms, which can differ by their fixed cost commitments. Updating fixed-cost commitments then proceeds in cohorts. In one iteration of the outer loop, version 1 firms will be able to reconsider the fixed cost payment. In the second iteration, version 2 firms will do so, etc. Keeping different versions of firms comes at a computational cost, of course, but we found this necessary.

Finding the best fixed cost decision is a combinatorially difficult problem. Given that there are $N_c - 1$ potential export destinations, this leads to 2^{N_c-1} possibilities for the exports. With $N_c = 100$, this is more than 10^{29}. To obtain an approximate optimum, we use Algorithm 2 of Buchbinder et al. (2015), which is stochastic in nature. We consider 9 (random) runs of that algorithm, and if the best of them is better than the firm's previous fixed cost decision, we update it. After the update, we again solve for a new general equilibrium involving continuous variables.

4.9 Fitting the model

We work with $N_c = 100$ countries. This choice is motivated by data availability and parameter fit considerations: For a substantially larger number of countries, the trade data would be too noisy and unreliable. For a substantially smaller number, it would be impossible to read off the elasticity of the marginal production cost from the firms' export pattern using our method.

The labor endowment in the model is interpreted as an efficiency-adjusted number of units of a single production factor, which in practice would include labor, capital, and the related productivity. This effective labor endowment and the trade cost prefactors $\kappa_{\mathrm{LT},k,k_d}$ are recovered by fitting the model to data on country GDP and world trade flows for the year 2006, as described below.

To match the typical empirical firm size distribution, which we take as Pareto distribution with Pareto index $\mu_R = 1.05$, we choose the firm size distribution to be another Pareto distribution with Pareto index $\mu_R(\sigma - 1)/(1 + \sigma\alpha)$.[46] The productivity distribution is the same for every country in

[44] Our model, as detailed in the next subsection, had more than 20,000 variables and described 2 million potential trade flows. Newton's method would not be feasible here, given that the Hessian of the loss function would have 400 million entries, although light-weight second order methods, such as L-BFGS, could potentially be useful. They would again benefit from the analytic nature of our model. For an overview of optimization algorithms, see the excellent book by Goodfellow, Bengio and Courville (2016).

[45] Given firms' sunk cost decisions, we need to solve for the general equilibrium of the world economy, i.e., we need to solve for wages, price indexes, and the measure of firms of each type in each country, as well as for production levels of each firm. What makes our calculation fast is the fact that we have explicit formulas for quantities sent to individual destinations conditional on the firm's marginal cost, and that these formulas and their derivatives may be evaluated extremely fast.

[46] The value of 1.05 for the Pareto index of the firm size distribution has empirical support in Aoyama et al. (2010), at least for the advanced economies studied there. In our open-economy model, there is no simple closed-form expression for the firm size (revenue) as a function of firm productivity. For this reason we use a formula that would hold exactly for closed economies, as well as in the absence of trade costs for the world. Simple algebra shows that the required value of the productivity Pareto index is $\mu_R(\sigma - 1)/(1 + \sigma\alpha)$ and we use this value. The calibration results

N_c	100	N_p	20	N_v	10	α	multiple values
σ	5	ν_R	0.8	ν_{LT}	0.6	μ_R	1.05
$\delta_e f_e$	0.05	f_o	0.1	f_x	10^{-5}	τ	1.05

Table 2: Calibration parameter values

the model; any real-world overall firm productivity differences across countries are represented by adjustments to the countries' effective labor endowments.

For computational purposes we discretize the productivity distribution to $N_p = 20$ discrete values, each representing the same probability mass.[47]

In addition, we need to specify (the flow value of) the costs of entry, fixed costs of production, export market entry costs, as well as iceberg trade costs. We make these choices as simple as possible, independent of the country or country pair. Their values are given in Table 2. The flow value of the cost of entry is set to one half of the fixed cost of operation. The fixed cost of exporting is set to be negligible. The iceberg trade cost $\tau - 1$ is non-zero but small enough to be consistent with prices firms in practice pay for insurance or as tariffs. In general, the parameters are chosen to reflect a long-term interpretation of the model, with timescales of many years.[48]

We use importer-reported data on international trade flows for the year 2006 from the UN Comtrade database. We select 100 countries/economies with the largest GDP, as reported by the IMF in its World Economic Outlook database, subject to trade and GDP data availability. We adjust the countries' GDP for tradability using the United Nations' gross value added database; see Appendix C.

4.10 Elasticity of the marginal cost of production

We solve for the model fit for different values of the parameter α, the quantity-elasticity of the marginal cost of production. Then we compare the resulting pattern of firm trade with that of Chinese firm-level export data for 2006 in order to find what value of α leads to a good agreement.

We obtained fits to the data on world trade flows and adjusted GDP levels for values of α ranging from 0.15 to 0.3; see Figure 5. In each case, we computed power-law best-fit curves that describe the dependence of the median size of Chinese firms that export to a destination as a function of the popularity rank of that export destination.[49] The popularity is computed as the fraction of Chinese

in a good match for the firm size distribution of Chinese exporters (both for all firms and for single-HSID firms, as these have the same empirical shape). Given this encouraging result, we have not explored other specifications for the productivity distribution.

[47]Initially, we tried $N_p = 10$, but such crude discretization led to numerical errors that were too large. Also note that even though for simplicity we sometimes refer to the probability masses as "firms", they really represent collections of firms in monopolistic competition, not a few discrete firms in an oligopoly model.

[48]We do not attempt to model high-frequency phenomena in international trade (except that shipping frequency considerations provide a micro-foundation for our trade costs). For studying month-to-month or year-to-years changes, it would not be appropriate to assume that the sunk fixed cost of exporting is fairly negligible.

[49]More precisely, we use the generalized method of moments to fit functions of the form $c_0\,(\text{rank})^{-c_1} + c_2$.

firms in the data that choose to export to the given destination. We also computed such best-fit curve for the data. The results are intuitive: For smaller α, the difference between the median (log) size of firms exporting to unpopular destinations and to popular destinations is larger because in this case the most productive firms will dominate world trade, and if a less productive firm decides to export at all, it will choose a few of the popular destinations.

The data corresponds roughly to $\alpha \approx 0.225$ if we consider all 99 export destinations when computing the best-fit curves, or to $\alpha \approx 0.25$, if we consider the first third of them by popularity rank. The first estimate has the advantage of taking into account a large range of export destinations. We include the second estimate because the top third of the destinations account for a vast majority of Chinese export and because the model's precision is lower for very small countries. The values $\alpha \approx 0.225$ or $\alpha \approx 0.25$ would imply that if a firm decides to scale up production by an order of magnitude, its marginal cost increases by about 68% or 78%, respectively. These values seem very realistic, given that such a dramatic expansion of the firm would require an additional layer of management hierarchy with related principal-agent problems. Note that these inefficiencies would be partially offset by savings on the fixed cost of production.

4.11 The gravity equation of trade and the dependence of trade costs on distance

The model fit results have important implications for the gravity equation of trade and for the trade cost puzzle (discussed in detail by Disdier and Head (2008) and Head and Mayer (2013)).[50] The gravity equation of trade implied by the data[51]

$$\log x_{ij} \approx -0.77 \log d_{ij} + 1.12 \log y_i + 1.10 \log y_j + \text{const.}$$

matches well the gravity equation implied by the fitted model[52]

$$\log x_{ij} \approx -0.71 \log d_{ij} + 1.08 \log y_i + 1.02 \log y_j + \text{const.}$$

[50] See Head and Mayer (2014) for a recent overview of the literature on the gravity equation of trade. Our purpose here is to highlight the consequences of our model's mechanisms, so we focus on the baseline gravity equation that describes the dependence of trade flows on distance and effective GDPs of countries. A comprehensive, in-depth investigation of our model that parallels detailed studies in the gravity-equation literature will be reported separately. It is, of course, worth investigating gravity equations with added controls, such as common language. Similarly, it is good to account for the "multilateral resistance" phenomenon (i.e.more isolated countries being more eager to trade with a given partner) already when designing the regression/estimation equations to study. Our model provides very different structural equations than other models, so the matter of multilateral resistance is quite involved. In addition, it is good to explicitly consider trade flow zeros in constructing the regression/estimation equations, although that makes little difference here as almost all trade flows are non-zero in our sample of 100 economies.

[51] This is for 30 largest economies. For 100 economies the results would be more noisy.

[52] Here we used $\alpha = 0.225$.

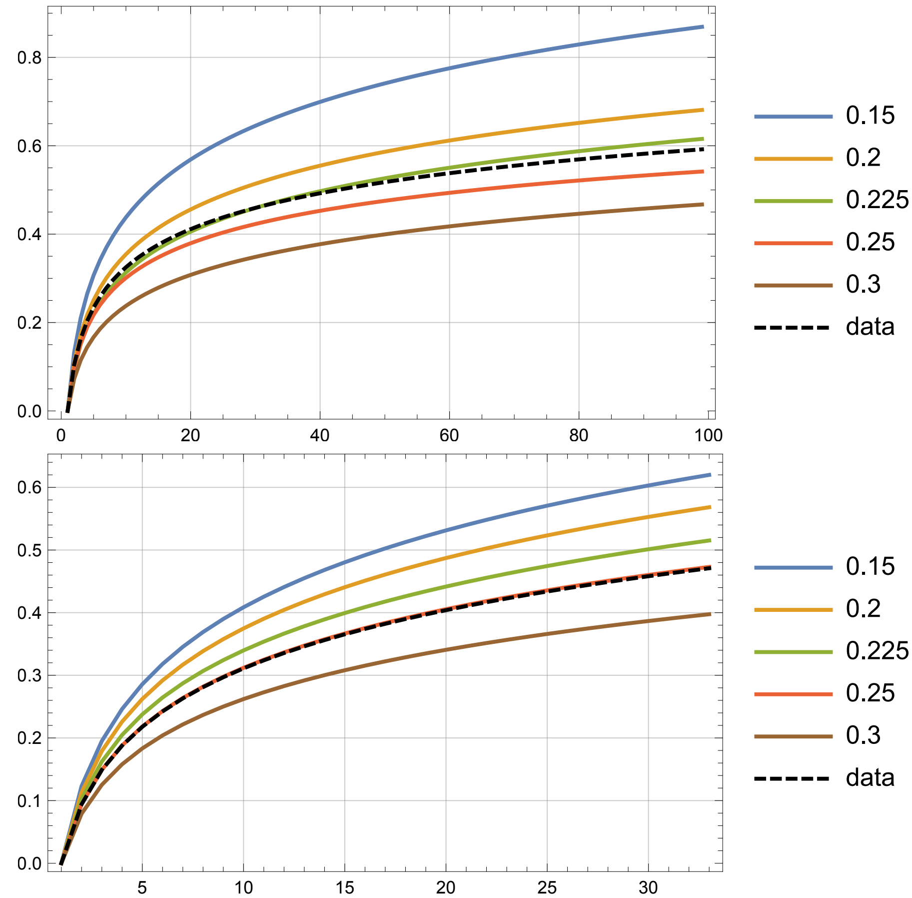


Figure 5: Median revenue of Chinese exporting firm (base-10 log scale) by destination country for different values of α and for the observed values. For the first export destination (United States), the median log revenue is normalized to 0 to make visual comparisons easier. The top figure corresponds to all 99 export destinations in the model, while the bottom figure corresponds to the top third by export popularity.

Of course, this is not surprising given that the world trade flows were a target of our model fit; if the fit was perfect, the two equations would coincide. What is interesting, though, is that the trade cost prefactors (i.e. the factors $\kappa_{\mathrm{LT},k,k_d}$ in $L_{T,k,k_d}(q) = \nu_{\mathrm{LT}}^{-1}\kappa_{\mathrm{LT},k,k_d}q^{\nu_{\mathrm{LT}}}$) depend on distance only very weakly:

$$\log \kappa_{\mathrm{LT}} \approx 0.048 \log d_{ij} + 0.032 \log y_i + 0.048 \log y_j + \text{const.}$$

We see that trade flows decrease rapidly with distance despite only a very mild increase of trade cost prefactors κ_{LT} with distance. Although this may look surprising at first sight, there is clear intuition for this phenomenon: Due to increasing marginal costs of production, firms effectively have only a limited output to sell and due to economies of scale in shipping, they need to concentrate their exports to only a few countries. They choose close countries because the trade costs are slightly lower, which leads to strong effects for the decrease of trade with distance.[53]

[53]We briefly discuss related literature and mechanisms in Appendix C.2.

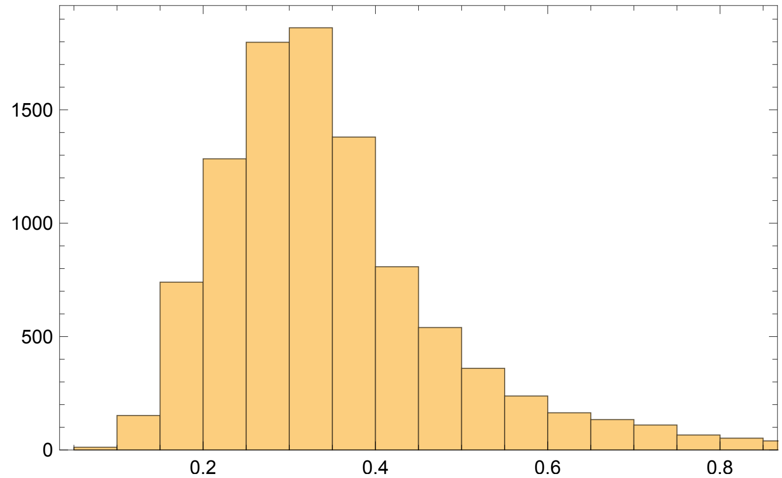


Figure 6: Histogram of the (symmetrized) trade cost prefactors.

From the histogram in Figure 6 we see that the dispersion of κ_{LT} is very small, which is only possible with a very small dependence on distance. This small dispersion is very much consistent with sea shipping over large distances being only mildly more expensive than over short distances. This provides a very natural resolution to the trade cost puzzle in the international trade literature.

4.12 Choice of export destinations

In the Introduction, we briefly discussed an empirical pattern of firm entry into export markets that would seem puzzling in standard models of international trade. Our model naturally implies such pattern. Figure 7 illustrates export market entry choices in the fitted model for pairs of identical firms, i.e. firms from the same country and with the same productivity.[54] These would be impossible in a corresponding model with constant marginal cost unless we introduced unrealistically large firm-destination specific cost shocks (or other similar shocks).[55] It is straightforward to see why this is the case. With constant marginal costs, the decision of whether or not to enter a particular export destination is independent of such decisions for other destinations, as long as the firm does not shut down. If there were no firm-destination specific shocks, then two identical firms with the same constant marginal costs would reach the same conclusions about the profitability of each export destination. In order to make one of the firms give up on a particular destination, we would have to introduce a firm-destination specific shock that would offset all the profit the firm was about to make from selling at that destination.

Our model naturally delivers the export destination choice pattern that would seem puzzling

[54]The choice of countries for the figure is not completely arbitrary. China was chosen for the figure because it is a large country and we see patterns of this kind in its firm-level data. We chose the Czech Republic since it is a small country with many neighbors and we know of patterns of this kind based on a series of interviews with Czech exporters featured in *Hospodarske noviny*, a newspaper.

[55]Although for identical firms this would be impossible if we introduce differences between the firms, there are other phenomena that may play a role. We briefly discuss them in Appendix C.3.

otherwise. With increasing marginal costs of production, a destination that is profitable for one firm may not be profitable for another identical firm, if that firm already serves other locations. Of course, if there were no economies of scale in shipping (and no significant export entry fixed costs), firms would dilute their exports over more destinations and would not face a combinatorial discrete choice problem. In that case, two identical firms would serve the same destinations, unless, again, there were firm-destination specific shocks. For this reason, both increasing marginal costs of production and economies of scale in shipping are crucial for our model's ability to resolve the export destination choice puzzle.

A mechanism of this kind also has the potential to explain why personal relationships can play a large role in international trade. Just like shorter distance, knowing someone trustworthy to cooperate with at a potential export destination can provide a mild profit advantage for exporting there instead of other destinations. This modest advantage may then have a large effect on trade flows, given increasing marginal costs of production and economies of scale in shipping.[56]

The subject of the export destination choice pattern is, of course, very rich and calls for an in-depth empirical and theoretical investigation, which will be provided in a separate, monothematic paper.

4.13 Implications for modeling international trade

A vast majority of models of international trade (and spatial economics) assume constant marginal cost of production within firm, even though empirical evidence for such constancy is lacking and even though organizational economics is telling us that scaling up a firm is highly nontrivial, given all the internal agency problems. An obvious reason for making the assumption of constant marginal cost is that it decouples firms' behavior in different export destinations and makes the models easy to solve. Without such decoupling we need to solve combinatorial discrete choice problems, which are hard in the case of submodular function maximization (corresponding to increasing marginal cost).

We have seen that working directly with increasing marginal costs leads to a dramatically different perspective on quantitative and qualitative aspects of international trade. It is computationally challenging, but the results are worth it. Some of the puzzles are no longer puzzling, as trade costs behave the way we would expect.

In the future, working with trade models that impose constant marginal production costs will not be as appealing to us as before. It suddenly has a flavor of the proverbial searching for keys under a streetlight. Once we accept the idea of working with models with increasing marginal costs, there are many questions to address. It would be good to re-think many topics in international trade, such as the impact of various policies, interventions or technological changes on the equilibria and

[56] Similarly, the mechanism may help explain why in the long run trade liberalization can have dramatic effects on trade flows, as for example in the case of the 2001 US-Vietnam trade liberalization; see McCaig and Pavcnik (2018).

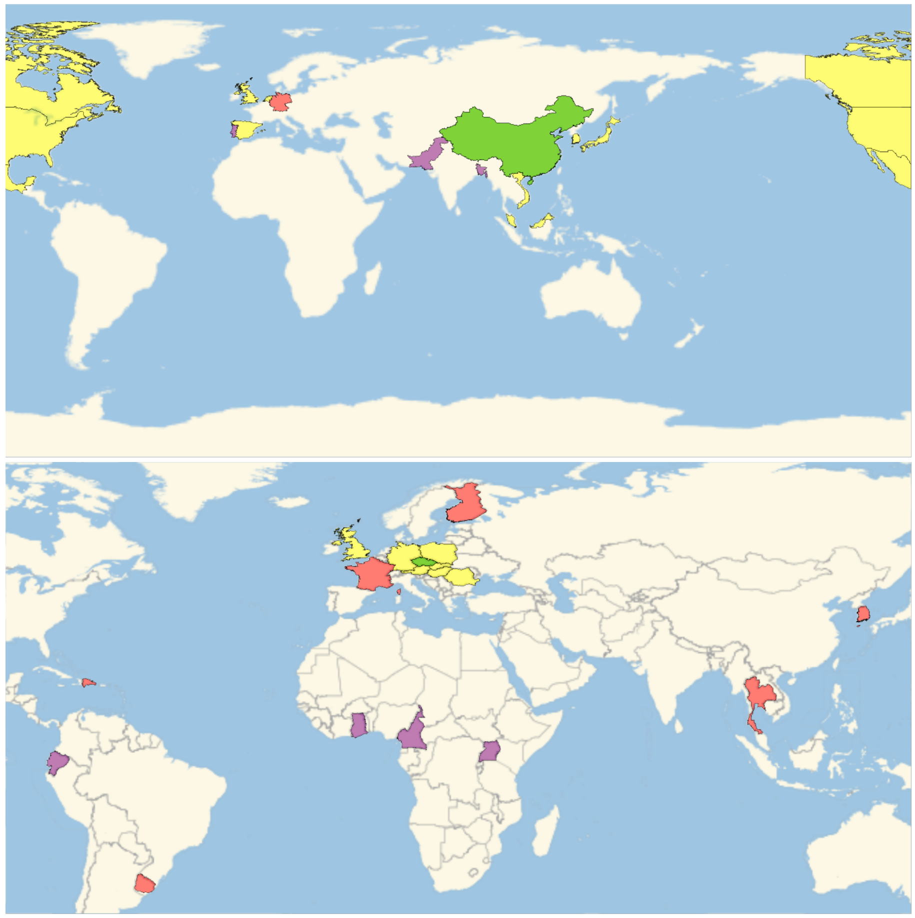


Figure 7: The top map highlights export destinations of two Chinese firms in the fitted model. Specifically, for two identical firms the map shows destinations to which both firms export (yellow), destinations to which only firm 1 exports (red), destinations to which only firm 2 exports (purple), and the country of origin (green). The bottom map shows similar information for two firms from the Czech Republic.

on the welfare of different agents in the economy.[57] The model we worked with is very parsimonious, but including multiple locations per country, multiple sectors, supply chains, and/or foreign direct investment would be desirable. Some of these ingredients would bring their own combinatorial discrete choice problems. The models will certainly be even more computationally intensive than the one we worked with. Improvements in algorithms and hardware, hopefully, will make solving the models feasible.

[57]For example, for welfare consequences we can no longer use simple, elegant formulas such as those derived by Arkolakis, Costinot and Rodríguez-Clare (2012).

5 Breadth of Application

5.1 Overview of applications

In this section we provide a brief overview of numerous other applications.

5.1.1 Supply chains with hold-up (Antràs and Chor, 2013)

Antràs and Chor (2013) develop a model of continuum sequential supply chains where a main firm organizing its production needs to decide whether to outsource or insource (i.e. perform in-house) each stage of the production process. Production requires relationship-specific investment, which leads to a hold-up problem in the spirit of Grossman and Hart (1986). Outsourcing a production stage has the advantage of giving high-powered incentives to the producers, while insourcing has the advantage of mitigating the hold-up problem.

The paper works with constant-elasticity demand and concludes that there can be only one production stage at which the main firm switches production mode; depending on the parameter values, either all of the upstream or all of the downstream (but not both) is outsourced and the rest is insourced. This, of course, clashes with the fact that for many manufacturing supply chains both the upstream (say, elementary components) and the downstream (say, retail) are outsourced, while the core of the production process is insourced.

In Appendix D.1 and in Supplementary Material I.1, we introduce a transformation of economic variables that makes the mathematics of the model dramatically simpler, in particular connecting the analysis to the classical monopsony problem, whose cost-side aspects are analogous to the demand-side aspects of a monopoly problem.[58] This allows us to observe by insights analogous to ours above that constant-elasticity demand may be replaced by our tractable functional forms without almost any loss of analytic power. We find that for a realistic functional form of this kind (where the product has a "saturation point" in terms of quality), the model implies that both upstream and downstream parts of the supply chain are optimally outsourced, while the middle (core) of the supply chain is optimally insourced, as our intuition suggests in many real-world cases.

5.1.2 Labor bargaining without commitment (Stole and Zwiebel)

Stole-Zwiebel bargaining, as introduced in Stole and Zwiebel (1996a,b), has become one of the standard ways of modeling labor bargaining in relation to unemployment. In their model, if an employee leaves the firm after an unsuccessful wage bargaining, the remaining employees may renegotiate their wage, and they will choose to do so since the firm's bargaining position is weakened. For this reason, the firm will choose to employ more workers than it would if labor markets are competitive; employing an additional worker lowers negotiated wages for the others.

[58]Using our transformed variables would have saved at least 10 pages of the original paper Antràs and Chor (2013). But of course, relative to these authors we have the benefit of hindsight.

While this model appears to differ from previous examples we considered, as it is not a straightforward monopsony model, we show that behavior under the Stole-Zwiebel model corresponds to a "partial" application of the marginal-average transformation ("partial monopolization") and thus remains tractable under our forms. Thus it is common to use standard, form-preserving tractable forms to analyze this model, especially constant-elasticity. The downside of the assumption is that interesting effects are suppressed: the overemployment ratio (ratio of actual employment and employment under competitive labor markets) is a constant independent of economic conditions.

We introduce richer functional forms that preserve the tractability of the model. We find that for a plausible parameterization, changes in the overemployment ratio can account for a non-trivial fraction of employment changes over the business cycle. These results are discussed in Supplementary Material I.2.

5.1.3 Imperfectly competitive supply chains

Imperfectly competitive supply chains, as described in Salinger (1988), are a very natural and popular way of modeling multi-stage production. We find that models of this kind may be solved in closed form not only for linear or constant-elasticity demand but also for our proposed, much more flexible functional forms. Intuitively, behavior at each level of the supply chain is derived by applying the marginal-average transformation to behavior at the preceding level, as each step of the supply chain forms the demand for the level above it. We discuss this application in Appendix D.2 and provide the details in Supplementary Material I.3.

5.1.4 Two-sided platforms à la Rochet and Tirole (2003)

Rochet and Tirole (2003) developed a model of two-sided platforms that allows for understanding pricing decisions for the two sides of the market and their surplus consequences. The model used linear demand. We find that our more flexible functional forms preserve the tractability of the model. This can lead to very different conclusions, as discussed in Supplementary Material I.4.

5.1.5 Auction Theory

Symmetric independent private values first-price auctions. First price auctions with symmetric independent private values may be solved explicitly for uniform or Pareto value distributions. We find that the tractable functional forms we propose still lead to closed-form solutions, and at the same time they allow for more realistic (i.e. bell-shaped) value distributions. We discuss these results in Supplementary Material I.5.1.

Auctions v. posted prices (Einav, Farronato, Levin and Sundaresan, 2018). Einav et al. (2018) develop a model in which online sellers choose either auctions or posted prices. They use a uniform distribution in their model. We find that our proposed functional forms also lead to tractable models but allow a richer set of possibilities for the sellers' optimal behavior that better match the data. We explain the details in Supplementary Material I.5.2.

5.1.6 Selection markets

In selection markets (markets with adverse or advantageous selection) as in Mahoney and Weyl (2017)'s generalization of Einav et al. (2010) and Einav and Finkelstein (2011), the equilibrium conditions are such that again our proposed tractable functional forms lead to closed-form solutions. This allows for modeling possibilities that provide a better match to the empirical evidence, as explained in Supplementary Material I.6.

5.1.7 Monopolistic competition

Tractable functional forms are very useful in the case of monopolistic competition beyond what we discussed in the previous section. Supplementary Material I.7 contains an extensive discussion of other possible modeling choices that generalize, say, the Melitz model or the Krugman model. These calculations may be used as a basis for new research projects on international trade.

6 General Approximation and the Laplace-Log Transform

In most of the examples in the previous sections, we have focused on average-marginal form-preserving classes of relatively low dimensions that are tractable at low orders. While these are useful in many applications and reasonably flexible, they have limits in their ability to fit arbitrary equilibrium systems. In this section we show that this limitation arises from the desired tractability of these forms, rather than any underlying rigidity of our average-marginal form-preserving classes. Under weak conditions we formulate here, arbitrary (univariate) equilibrium forms can be approximated arbitrarily well by members of form-preserving classes. The limit of this approximation is the inverse Laplace-log transform of the equilibrium condition. Highly tractable forms may thus be seen as ones with "simple" inverse Laplace-log transforms. We show how the special, policy-relevant features of many common demand forms can be characterized in terms of their transforms. Proofs of the theorems in this, more abstract, section appear in Appendix A. A number of these proofs are straightforward adaptations of theorems in the existing literature. We include those theorems here for completeness and for the reader's convenience.

In the next subsection we provide definitions of the Laplace-log transform, utilizing existing mathematical literature. Identifying the most important connections between what is useful in economics and the mathematical literature is non-trivial. While a reasonable number of economists are familiar with Laplace transform based on the Riemann-Stieltjes integral, a theory based on that integral would exclude, say, the exponential demand function, which is a popular modeling choice in the economics literature. For a more complete theory we need to utilize the distribution theory by Laurent Schwartz, which has not been used in economics.

6.1 The Laplace-log transform and arbitrary approximation

Under quite general conditions, univariate equilibrium conditions may be expressed as linear combinations of average-marginal form-preserving functions. To make this statement precise, we focus on the demand side here and write an inverse demand curve of interest as $P(q) = U'(q)$, where $U(q)$ is a function primitive to $P(q)$. We assume that $P(q)$ is non-increasing, which implies that such primitive function exists. Depending on the model of choice, $U(q)$ may or may not be proportional to the utility of an agent, but to keep the terminology simple, here we refer to $U(q)$ as the utility.[59] Even though we explicitly discuss the demand side here, the mathematical theorems below apply to the cost side as well, with a straightforward reinterpretation.

We observe that virtually all shapes of demand functions that are useful in economics may be associated with a utility function of the form[60]

$$U(q) = \int_{-\infty}^{0} u(t)\, q^{-t} dt, \tag{2}$$

for an appropriate $u(t)$, where we work on some arbitrarily chosen finite interval $[0, \bar{q}]$. This integral may be interpreted as a Laplace transform in terms of the variable $s \equiv \log q$, and for this reason, we refer to $u(t)$ as the inverse Laplace-log transform of $U(q)$.[61,62] At the same time, the integral may

[59] $U(q)$ would literally be a term in the utility function $U(q) + \tilde{q}\tilde{P}$ in a model with two goods q and $\tilde{q}$, where $\tilde{q}$ is treated as a numéraire good with price $\tilde{P}$ normalized to 1. In this case the marginal utility of q equals its price $P(q)$.

[60] The Laplace-log representation (2) of a given utility function $U(q)$ exists under various conditions. Theorem 18b in Section VII.18 of Widder (1941) states general necessary and sufficient conditions on $U(q)$ for the existence of $u_I(t)$ such that (3) is satisfied; *almost all* utility functions we may encounter in economic applications *do satisfy these conditions*. Sections VII.12-17 of Widder (1941) provide conditions that guarantee that $u_I(t)$ exists and has certain properties, such as being of bounded variation, nondecreasing, or belonging to the functional space L^p. Additional conditions may be found in Chapter 2 of the book by Arendt et al. (2011), which contains recent developments in the theory. In situations when utility unbounded below is desired, e.g. for constant demand elasticity smaller than 1, we can depart from (2) and instead use the bilateral specification $U(q) = \int_{-\infty}^{\infty} u(t)\, q^{-t} dt$. However this generalization requires the use of more technically involved bilateral Laplace transforms and thus we do not discuss it in greater detail here, though analogous results are available on request.

[61] Our use of t for exponents throughout the text and our use of $s \equiv \log(q)$ here match the standard notation in the literature on Laplace transforms.

[62] After an extensive literature search of hundreds of articles and talking to numerous economists, including highly accomplished econometricians, we concluded that this is almost certainly the first time (inverse) Laplace transform in log quantity is used in the economic literature. Note, however, that a different transform, namely (inverse) Laplace transform in quantity, as opposed to log quantity, has been used in economics. These transforms have different properties and should not be confused. Note also that the way we use Laplace transform is different from, say, engineering fields in the sense that, because of the additional logarithm, functions of main interest for us in economics typically would not be of interest in engineering, and vice versa. For this reason, books containing detailed tables of Laplace transform were not of help to us. Except for trivial cases, we needed to derive the transforms listed in Supplementary Material E by ourselves.

be thought of as expressing $U(q)$ as a linear combination of form-preserving functions of Theorem 1.

Technical Clarification (Integral Definition).[63] Here we define the integral (2) to be the Riemann-Stieltjes integral

$$U(q) = \int_{-\infty}^{0} q^{-t} du_I(t) \tag{3}$$

for some function $u_I(t)$, not necessarily nonnegative, such that the integral converges. If this function is differentiable, its derivative $u_I'(t)$ is the $u(t)$ that appears on the right-hand side of (2). If $u_I(t)$ is only piecewise differentiable, then $u(t)$ is not an ordinary function but involves Dirac delta functions (i.e. point masses) at the points of discontinuity of $u_I(t)$.

The corresponding inverse demand curve is $P(q) = U'(q) = -\int_{-\infty}^{0} t\, u(t)\, q^{-t-1} dt$, or

$$P(q) = \int_{-\infty}^{1} p(t)\, q^{-t} dt, \tag{4}$$

where we defined $p(t) \equiv (1-t)\, u(t-1)$. We see that $P(q)$ is a linear combination of form-preserving functions of Theorem 1. The following theorem summarizes convenient properties of this approach to demand curves: uniqueness, inclusion of linear combinations of power functions, approximations to arbitrary functions, and analyticity.

Theorem 4. (Laplace-log Transform with Riemann-Stieltjes Integrals)
(A) *For each function $U(q)$ that may be represented in the form (2) in the sense of (3), there exists just one normalized*[64] *function $u_I(t)$ such that (3) holds.* **(B)** *Any polynomial utility function may be written in the form (2).* **(C)** *All functions of the form (2) are analytic. In particular, their derivatives of any order exist.* **(D)** *An arbitrary utility function $\tilde{U}(q)$ continuous on an interval $[0, \bar{q}]$ may be approximated with an arbitrary precision by utility functions of the form (2), in the sense of uniform convergence*[65] *on $[0, \bar{q}]$.*

Note that part D of this theorem is a simple consequence of the Weierstrass approximation theorem.[66] The reader may ask why we do not instead work simply with polynomials in q and use

[63] Note that in certain parts of the paper we need a more general definition of the integral (2) than the definition (3). In those cases, e.g. in the proof of Theorem 1, we use the Schwartz distribution theory instead of the Riemann-Stieltjes integral theory.

[64] Normalization here means that $u_I(0+) = 0$ and $u_I(t) = (u_I(t+) + u_I(t-))/2$. See Section I.6 of Widder (1941).

[65] By *uniform convergence* we mean that for any continuous $\tilde{U}(q)$ there exists a sequence $\{U_j(q), j \in \mathbb{N}\}$ of functions of the form (2) such that for any $\epsilon > 0$, all elements of the sequence after some position n_ϵ satisfy $\sup_{q\in[0,\bar{q}]} |\tilde{U}(q) - U_j(q)| < \epsilon$.

[66] There is also a related, more powerful theorem, the Müntz-Szász theorem. Barnett and Jonas (1983) use it to propose to write direct demand as Müntz-Szász polynomials of prices. Here we write inverse demand as polynomials of powers of quantities (times possibly another power of quantity), but the same logic would apply here: we could use Müntz-Szász polynomials.

them as approximations. Even though this would be possible in principle, it would not be practical. This is because in economics we often need flexibility in the $q \to 0_+$ limit behavior of the inverse demand function. With any (finite-order) polynomial, we would always get finite $\lim_{q \to 0_+} P(q)$, i.e., a choke price; to allow for $\lim_{q \to 0_+} P(q) = \infty$, we could not stay within a finite-order approximation.

Theorem 1 allowed for functions other than linear combinations of power functions, such as $q^{-\alpha}(\log q)^n$ or $\log q$, that are also useful in economics.[67] Although according to part D of the last theorem, such functions may be approximated by functions of the Riemann-Stieltjes interpretation (3) of (2), it is convenient to be able to write them *exactly* in the form (2) by using a more powerful definition of the integral. This is achieved by the following counterpart of Theorem 4, which goes beyond the theory of the Riemann-Stieltjes integral and instead discusses Laplace transform of generalized functions based on the distribution theory by Laurent Schwartz. In the following, $\bar{s}$ is a real number smaller than $\log \bar{q}$.

Theorem 5. (Laplace-log Transform with Schwartz Integrals) *A function $U(q)$ such that the related function $U_{[s]}(s) \equiv U(e^s)$ considered in the half-complex-plane domain $\mathbb{C}_{\bar{s}}^- \equiv \{s | Re\ s < \bar{s}\}$ is analytic (i.e. holomorphic) and bounded by a polynomial function may be expressed in the form (2) with u representing a distribution, i.e. a generalized function, or more precisely an element of $\mathcal{D}'$ as defined by Zemanian (1965).*[68] *This distribution is unique. Conversely, for any Laplace-transformable distribution u, the integral (2) viewed as a function of $s \equiv \log q$ in the domain $\mathbb{C}_{\bar{s}}^-$ is analytic and bounded by a polynomial of s.*

Definition 3. (Laplace Versions of Economic Variables) *For a variable $V(q)$ that may be expressed as an integral of the form $V(q) = \int_a^b v(t)\, q^{-t} dt$, we use the adjective* Laplace *to refer to $v(t)$. For example, $u(t)$ of (2) would be referred to as* Laplace utility*, and $p(t)$ of (4) as* Laplace inverse demand *or* Laplace price*.*

Here we present a theorem describing the relationship of the integral and its discrete approximation. Its proof is constructed using the Euler-Maclaurin formula related to the trapezoidal rule for numerical integration. Following the same logic, it is possible to derive and prove other approximation theorems by adapting numerous theorems on numerical integration that exist in the applied mathematics literature.

Theorem 6. (Discrete Approximation) *The Laplace-log transform of a function $f(t)$ may be expressed as*

$$\int_{-\infty}^{t_{\max}} q^{-t} f(t)\, dt = \Delta t \sum_{t \in T} q^{-t} f(t) - \frac{1}{2} q^{-t_{\max}} \Delta t f(t_{\max}) - \frac{1}{2} q^{-t_{\min}} \Delta t f(t_{\min}) + R,$$

[67] For example, $P(q) = a - b \log q$ corresponds to exponential demand, studied by many authors, including Aguirre et al. (2010). Similarly, inverse demand functions $P(q) = a - b(\log q)^n$ have interesting implications for market failure in sequential supply chains such as Cournot's multiple-marginalization problem.

[68] Here "bounded by a polynomial" refers to the absolute value of $U_{[s]}(s)$ being no greater than the absolute value of some polynomial of s in the domain $\mathbb{C}_{\bar{s}}^-$.

where $T \equiv \{t_{\min}, t_{\min} + \Delta t, ..., t_{\max}\}$ is an evenly spaced grid with at least two points, m is an integer such that f is $(2m+1)$-times continuously differentiable on $[t_{\min}, t_{\max}]$ and where the remainder R is described below.

The remainder in the theorem consists of three parts: $R \equiv R_1 + R_2 + R_3$. The first part R_1 is simply the difference of $\int_{-\infty}^{t_{\max}} q^{-t} f(t)\,dt$ and $\int_{t_{\min}}^{t_{\max}} q^{-t} f(t)\,dt$, and can be made very small, since $\int_{-\infty}^{t_{\min}} q^{-t} f(t)\,dt = q^{-t_{\min}} \int_{-\infty}^{0} q^{-t} f(t + t_{\min})\,dt$, which is exponentially suppressed for $t_{\min}$ chosen sufficiently negative and for a well-behaved $f(t)$. The second part R_2 may be expressed using derivatives of $h(t) \equiv f(t) q^{-t}$ at $t_{\min}$ and $t_{\max}$:

$$R_2 = \sum_{k=1}^{m} \frac{B_{2k}}{(2k)!} \left(\Delta \mathrm{t}^{2k} h^{(2k-1)}(t_{\min}) - \Delta \mathrm{t}^{2k} h^{(2k-1)}(t_{\max}) \right),$$

where B_{2k} represent Bernoulli numbers. These terms are suppressed by powers of $\Delta \mathrm{t}$ as well as by the factorial in the denominator.[69] The last part R_3 may be expressed and bounded using integrals of high derivatives of $h(t)$:

$$R_3 = -\tfrac{\Delta \mathrm{t}^{2m+1}}{(1+2m)!} \int_{t_{\min}}^{t_{\max}} P_{1+2m}(t) h^{(1+2m)}(t)\,dt \quad , \quad |R_3| \leq \tfrac{2\zeta(2m+1)\Delta \mathrm{t}^{2m+1}}{(2\pi)^{2m+1}} \int_{t_{\min}}^{t_{\max}} |h^{(2m+1)}(t)|\,dt \ ,$$

where ζ is the Riemann zeta function and P_{1+2m} are periodic Bernoulli functions.

Note that this theorem provides a prescription for the weights of the power terms that approximate the integral and gives a bound for the associated error. Of course, by leaving the weights flexible and fitting them using a generalized method of moments, it is possible to get a better approximation with a smaller error. It is also possible to use alternative prescribed weights that correspond to other numerical integration methods. The fact that very different weight choices can all give good approximations is related to the fact that the problem of finding optimal weights is a case of so-called ill-posed problems, for which regularization is typically used in the applied mathematics and econometrics literature.[70]

6.2 Complete monotonicity and pass-through behavior

Continuous representations of inverse demand functions introduced in the previous subsection provide more conceptual clarity than discrete approximations, which have their idiosyncrasies depending on precisely how many terms are included. These representations in terms of inverse Laplace-log transform can provide useful intuition. For example, if a researcher wishes to find a good discrete approximation to a particular inverse demand function, the researcher may compute the exact inverse Laplace-log transform (or consult Supplementary Material E) to see where the Laplace inverse

[69] Moreover, it is possible to rescale q by a constant factor to keep $\log q$ small in absolute value for the range of quantities of interest.

[70] As mentioned above, the validity of such approximations may be proved along the lines of the proof given here.

demand function $p(t)$ is positive or negative. Choosing a few evenly spaced mass points with a similar positivity/negativity pattern is then likely to lead to a tractable approximation to the original inverse demand function that has similar qualitative properties.

Inverse Laplace-log transform representations of inverse demand functions are useful also for another reason: Many demand curves have economic properties (determining many policy implications) that are easily understood in terms of the inverse Laplace-log transform. To develop the related theory, we start with a standard definition of completely monotone functions and then discuss relations between complete monotonicity, the form of Laplace inverse demand, and economic consequences for the pass-through rate.[71] We classify many commonly used demand functions using this property, given that, as we discussed in the previous section, many policy questions turn on properties of the pass-through rate tied down by complete monotonicity.

Definition 4. (Completely Monotone Function) *A function $f(x)$ is completely monotone iff for all $n \in \mathbb{N}$ its nth derivative exists and satisfies $(-1)^n f^{(n)}(x) \geq 0$.*

It turns out that many commonly used demand functions are such that the consumer surplus is completely monotone as a function of negative log quantity. For this reason, we make the following definition.

Definition 5. (Complete Monotonicity of the Demand Specification)[72]

We say that a demand function (or a utility function) satisfies the complete monotonicity criterion iff the associated consumer surplus is a completely monotone function of $-s$, i.e. for all $n \in \mathbb{N}$, $CS_{[s]}^{(n)}(s) \geq 0$, or equivalently[73] *$U_{[s]}^{(n)}(s) - U_{[s]}^{(n+1)}(s) \geq 0$. Strict complete monotonicity criterion then refers to these inequalities being strict.*

Theorem 7. (Nonnegativity of Laplace Consumer Surplus) *A (single-product) utility function is bounded below and satisfies the complete monotonicity criterion iff the Laplace consumer surplus $cs(t)$ is nonnegative and supported on $(-\infty, 0)$, i.e. $CS(q) = \int_{-\infty}^{0} cs(t)\, q^{-t} dt$ for some $cs(t) \geq 0$.*

Theorem 8. (Monotonicity of the Pass-Through Rate) *The complete monotonicity criterion for demand functions implies the pass-through rate decreasing with quantity in the case of constant-*

[71]Brockett and Golden (1987) also discuss relations between complete monotonicity and a type of Laplace transform. The Laplace transform used there is in terms of quantity q, whereas in our discussion, it is in terms of the logarithm of quantity. These two transforms are distinct and should not be confused. Similarly, the mathematical notion of complete monotonicity has very different economic manifestations in Brockett and Golden (1987) and in our work.

[72]In principle, it is possible to empirically test whether an empirical demand curve satisfies the complete monotonicity criterion. The relevant empirical test has been developed by Heckman et al. (1990). It would just have to be translated from the duration analysis context to our demand theory context.

[73]The fact that these definitions are equivalent may be seen as follows: With the marginal utility of the outside good normalized to one and $U(0)$ is set to zero, we have $CS(q) = -qP(q) + \int_0^q P(q_1)\, dq_1 = -qU'(q) + \int_0^q U'(q_1)\, dq_1 = U(q) - qU'(q)$. This translates into $CS_{[s]}(s) = U_{[s]}(s) - U'_{[s]}(s)$, where we use the subscript $[s]$ to emphasize that the variable is to be treated as a function of s. The equivalence for any $n \in \mathbb{N}$ then follows by differentiation.

marginal-cost monopoly. The only exception is BP demand, for which the pass-through rate is constant.

Theorem 9. (Complete Monotonicity of Demand Specification) *The following demand functions satisfy the complete monotonicity criterion:*[74]
Pareto/constant elasticity ($\epsilon > 1$), BP ($\epsilon > 1$), logistic distribution, log-logistic distribution ($\gamma > 1$), Gumbel distribution ($\alpha > 1$), Weibull distribution ($\alpha > 1$), Fréchet distribution ($\alpha > 1$), gamma distribution ($\alpha > 1$), Laplace distribution[75]*, Singh-Maddala distribution ($a > 1$), Tukey lambda distribution ($\lambda < 1$), Wakeby distribution ($\beta > 1$), generalized Pareto distribution ($\gamma < 1$), Cauchy distribution.*

Corollary. (Monotonicity of the Pass-Through Rate) *The last two theorems imply that the demand functions listed in Theorem 9 lead to constant-marginal-cost pass-through rate decreasing in quantity, with the exception of Pareto/constant elasticity as well as the more general BP demand, which are known to lead to constant pass-through.*

Theorem 10. (Absence of Complete Monotonicity of Demand Specification) *The following demand functions do **not** satisfy the complete monotonicity criterion: normal distribution, lognormal distribution, constant superelasticity (Klenow and Willis), Almost Ideal Demand System (either with finite or infinite surplus), log-logistic distribution ($\gamma < 1$), Fréchet distribution ($\alpha < 1$), Weibull distribution ($\alpha < 1$), Gumbel distribution ($\alpha < 1$), Pareto/constant elasticity ($\varepsilon > 1$), gamma distribution ($\alpha < 1$), Singh-Maddala distribution ($a < 1$), Tukey lambda distribution ($\lambda > 1$), Wakeby distribution ($\beta < 1$), generalized Pareto distribution ($\gamma > 1$).*

In our Supplementary Material J we provide a more complete taxonomy of pass-through properties of some of the demand forms mentioned here. Interestingly, the normal distribution has economic properties close to those of forms that satisfy the complete monotonicity criterion, since the non-complete monotonicity manifests itself only for very high-order derivatives.[76] The lognormal distribution is not quite so well-behaved, but the more realistic income model (the double Pareto lognormal) behaves similarly for calibrated parameter values.

[74]The parameter names are chosen as in Mathematica.

[75]Each half of the distribution separately, or the full distribution smoothed by arccosh to ensure the existence of the derivatives.

[76]In particular we found that the normal distribution of consumer values has properties very close to those satisfying the complete monotonicity criterion: constant-marginal-cost pass-through is increasing in price (as we show below), and low-order derivatives of $CS(s)$ with respect to $-s$ are positive. We concluded that the complete monotonicity criterion is not satisfied based on examining the sign on the tenth derivative of $CS(s)$. The absence of complete monotonicity is consistent with our expression to the corresponding Laplace inverse demand, which does not seem to satisfy $t\,cs(t) \geq 0$. In most economic applications, the difference from completely monotone problems is inconsequential because it manifests itself only in very high derivatives of $CS(s)$.

7 Conclusion

We have shown that the set of analytic solutions to many common economic problems is substantially richer than typically assumed. They let economists work with flexible, realistic models, instead of imposing restrictive, unrealistic assumptions in order to get analytic solutions of traditional kinds. Our approach to getting analytic solutions is also useful when applied to sub-problems of larger economic models. In those cases it leads to the ability to solve those models numerically in a much more efficient way, as in our international trade application.

The international trade model provides a perspective on the gravity equation of trade that is completely different from the rest of the literature. The model resolves economic puzzles related to the cost of trade since its parameters take realistic values and at the same time the model matches well firm-level and country-level trade patterns.

Of course, there are many other applications of our method, some of which we briefly discussed here, some of which we will report in separate papers, and some of which, hopefully, the reader will develop on his/her own.

References

Adachi, Takanori and Takeshi Ebina, "Cost Pass-Through and Inverse Demand Curvature in Vertical Relationships with Upstream and Downstream Competition," *Economics Letters*, 2014, *124* (3), 465–468.

_ **and** _ , "Double Marginalization and Cost Pass-Through: Weyl–Fabinger and Cowan Meet Spengler and Bresnahan–Reiss," *Economics Letters*, 2014, *122* (2), 170–175.

Aguirre, Iñaki, Simon George Cowan, and John Vickers, "Monopoly Price Discrimination and Demand Curvature," *American Economic Review*, 2010, *100* (4), 1601–1615.

Akerlof, George, "The Market for "Lemons": Quality Uncertainty and the Market Mechanism," *Quarterly Journal of Economics*, 1970, *84* (3), 488–500.

Almunia, Miguel, Pol Antràs, David Lopez-Rodriguez, and Eduardo Morales, "Venting Out: Exports during a Domestic Slump," 2018.

Anderson, James, Mykyta Vesselovsky, and Yoto Yotov, "Gravity with Scale Economies," Technical Report, LeBow College of Business, Drexel University 2014.

Antràs, Pol, "Firms, Contracts, and Trade Structure," *Quarterly Journal of Economics*, 2003, *118* (4), 1375–1418.

_ **and Davin Chor**, "Organizing the Global Value Chain," *Econometrica*, 2013, *81* (6), 2127–2204.

_ **and Elhanan Helpman**, "Global Sourcing," *Journal of Political Economy*, 2004, *112* (3), 552–580.

— and —, "Contractual Frictions and Global Sourcing," in Elhanan Helpman, Dalia Marin, and Thierry Verdier, eds., *The Organization of Firms in a Global Economy*, Cambridge, MA: Harvard University Press, 2008.

—, **Teresa C. Fort, and Felix Tintelnot**, "The Margins of Global Sourcing: Theory and Evidence from US Firms," *American Economic Review*, 2017, *107* (9), 2514–64.

Aoyama, Hideaki, Yoshi Fujiwara, Yuichi Ikeda, Hiroshi Iyetomi, and Wataru Souma, *Econophysics and Companies: Statistical Life and Death in Complex Business Networks*, Cambridge University Press, 2010.

Apostol, Tom M, "An Elementary View of Euler's Summation Formula," *The American Mathematical Monthly*, 1999, *106* (5), 409–418.

Arendt, Wolfgang, Charles JK Batty, Matthias Hieber, and Frank Neubrander, *Vector-valued Laplace Transforms and Cauchy problems*, Basel: Birkäuser, 2011.

Arkolakis, Costas, "Market Penetration Costs and the New Consumers Margin in International Trade," *Journal of Political Economy*, 2010, *118* (6), 1151–1199.

—, **Arnaud Costinot, and Andrés Rodríguez-Clare**, "New Trade Models, Same Old Gains?," *American Economic Review*, 2012, *102* (1), 94–130.

Atkinson, Anthony B., Thomas Piketty, and Emmanuel Saez, "Top Incomes in the Long Run of History," *Journal of Economic Literature*, 2011, *49* (1), 3–71.

Bagnoli, Mark and Ted Bergstrom, "Log-Concave Probability and its Applications," *Economic Theory*, 2005, *26* (2), 445–469.

Barnett, William A. and Andrew B. Jonas, "The Müntz-Szatz Demand System: An Application of a Globally Well Behaved Series Expansion," *Economics Letters*, 1983, *11* (4), 337–342.

Behrens, Kristian and Yasusada Murata, "General Equilibrium Models of Monopolistic Competition: A New Approach," *Journal of Economic Theory*, 2007, *136* (1), 776–787.

— and —, "Trade, Competition, and Efficiency," *Journal of International Economics*, 2012, *87* (1), 1–17.

Bellman, Richard Ernest, Robert E Kalaba, and Jo Ann Lockett, *Numerical inversion of the Laplace transform*, North-Holland, 1966.

Berry, Stephen, James Levinsohn, and Ariel Pakes, "Automobile Prices in Market Equilibrium," *Econometrica*, 1995, *63* (4), 841–890.

Brockett, Patrick L. and Linda L. Golden, "A Class of Utility Functions Containing All the Common Utility Functions," *Management Science*, 1987, *33* (8), 955–964.

Buchbinder, Niv, Moran Feldman, Joseph Seffi, and Roy Schwartz, "A Tight Linear Time (1/2)-Approximation for Unconstrained Submodular Maximization," *SIAM Journal on Computing*, 2015, *44* (5), 1384–1402.

Budish, Eric, Benjamin Roin, and Heidi Williams, "Do Firms Underinvest in Long-Term Research? Evidence from Cancer Clinical Trials," *American Economic Review*, 2015, *105* (7), 2044–2085.

Bulow, Jeremy I. and Paul Pfleiderer, "A Note on the Effect of Cost Changes on Prices," *Journal of Political Economy*, 1983, *91* (1), 182–185.

Calvo, Guillermo A and Stanislaw Wellisz, "Supervision, Loss of Control, and the Optimum Size of the Firm," *Journal of Political Economy*, 1978, *86* (5), 943–952.

Caplin, Andrew and Barry Nalebuff, "Aggregation and Imperfect Competition: On the Existence of Equilibrium," *Econometrica*, 1991, *59* (1), 25–59.

_ **and** _ , "Aggregation and Social Choice: A Mean Voter Theorem," *Econometrica*, 1991, *59* (1), 1–23.

Cárdenas-Barrón, Leopoldo Eduardo, Kun-Jen Chung, and Gerardo Treviño-Garza, "Celebrating a Century of the Economic Order Quantity Model in Honor of Ford Whitman Harris," *International Journal of Production Economics*, 2014, *155*, 1–7.

Cassola, Nuno, Ali Hortaçsu, and Jakub Kastl, "The 2007 Subprime Market Crisis Through the Lens of European Central Bank Auctions for Short-Term Funds," *Econometrica*, 2013, *81* (4), 1309–1345.

Cournot, Antoine A., *Recherches sur les Principes Mathematiques de la Theorie des Richess*, Paris, 1838.

Crawford, Greg, Robin Lee, Michael Whinston, and Ali Yurukoglu, "The Welfare Effects of Vertical Integration in Multichannel Television Markets," *Econometrica*, 2018, *86* (3), 891–954.

Cutler, David M. and Sarah J. Reber, "Paying for Health Insurance: the Trade-Off between Competition and Adverse Selection," *Quarterly Journal of Economics*, 1998, *113* (2), 433–466.

Diamond, Peter and Emmanuel Saez, "The Case for a Progressive Tax: From Basic Research to Policy Recommendations," *Journal of Economic Perspectives*, 2011, *25* (4), 165–190.

Disdier, Anne-Célia and Keith Head, "The Puzzling Persistence of the Distance Effect on Bilateral Trade," *The Review of Economics and Statistics*, 2008, *90* (1), 37–48.

Dupuit, Arsène Jules Étienne Juvénal, *De la Mesure de L'utilité des Travaux Publics*, Paris, 1844.

Eaton, Jonathan, Samuel Kortum, and Francis Kramarz, "An Anatomy of International Trade: Evidence from French Firms," *Econometrica*, 2011, *79* (5), 1453–1498.

Eckert, Fabian and Costas Arkolakis, "Combinatorial Discrete Choice," in "2017 Meeting Papers" Society for Economic Dynamics 2017.

Egonmwan, Amos Otasowie, "The Numerical Inversion of the Laplace Transform," Master's thesis, Johannes Kepler University, Linz 2012.

Einav, Liran, Amy Finkelstein, and Mark R. Cullen, "Estimating Welfare in Insurance Markets Using Variation in Prices," *Quarterly Journal of Economics*, 2010, *125* (3), 877–921.

— and — , "Selection in Insurance Markets: Theory and Empirics in Pictures," *Journal of Economic Perspectives*, 2011, *25* (1), 115–138.

— **, Chiara Farronato, Jonathan Levin, and Neel Sundaresan**, "Auctions versus Posted Prices in Online Markets," *Journal of Political Economy*, 2018, *126* (1), 178–215.

— **, Theresa Kuchler, Jonathan Levin, and Neel Sundaresan**, "Assessing Sale Strategies in Online Markets Using Matched Listings," *American Economic Journal: Microeconomics*, 2015, *7* (2), 215–47.

Ekelund, Robert B. and Robert F. Hébert, *Secret Origins of Modern Microeconomics*, Chicago: University of Chicago Press, 1999.

Entrepreneur Magazine, *Start Your Own Import/Export Business*, Entrepreneur Press, 2003.

Fabinger, Michal and E. Glen Weyl, "Pass-Through and Demand Forms," 2012. https://goo.gl/FPvslo.

Farrell, Joseph and Carl Shapiro, "Horizontal Mergers: An Equilibrium Analysis," *American Economic Review*, 1990, *80* (1), 107–126.

Feller, Willliam, *An Introduction to Probability Theory and its Applications*, Vol. 2, John Wiley & Sons, 2008.

Forslid, Rikard and Toshihiro Okubo, "Big is Beautiful when Exporting," *Review of International Economics*, 2016.

Gabaix, Xavier, David Laibson, Deyuan Li, Hongyi Li, Sidney Resnick, and Caspar G. de Vries, "The Impact of Competition on Prices with Numerous Firms," *Journal of Economic Theory*, 2016, *165*, 1–24.

Garnier, Germaine, *Abrégé Élémentaire des Principes de L'Économie Politique*, Paris: H. Agasse, 1796.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville, *Deep Learning*, Vol. 1, MIT press Cambridge, 2016.

Grossman, Sanford J. and Oliver D. Hart, "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration," *Journal of Political Economy*, 1986, *9* (4), 691–719.

Haile, Philip A. and Elie Tamer, "Inference with an Incomplete Model of English Auctions," *Econometrica*, 2003, *111* (1), 1–51.

Hajargasht, Gholamreza and William E. Griffiths, "Pareto-Lognormal Distributions: Inequality, Poverty, and Estimation from Grouped Income Data," *Economic Modeling*, 2013, *33*, 593–604.

Harris, F. W., "How Many Parts to Make at Once," *Magazine of Management*, 1913, *10* (2), 135–136.

Hart, Oliver and John Moore, "Property Rights and the Nature of the Firm," *Journal of Political Economy*, 1990, *98* (6), 1119–1158.

Head, Keith and Thierry Mayer, "What Separates Us? Sources of Resistance to Globalization," *Canadian Journal of Economics/Revue canadienne d'économique*, 2013, *46* (4), 1196–1231.

_ **and** _ , "Gravity Equations: Workhorse, Toolkit, and Cookbook," in "Handbook of International Economics," Vol. 4, Elsevier, 2014, pp. 131–195.

Heckman, James J., Richard Robb, and James R. Walker, "Testing the Mixture of Exponentials Hypothesis and Estimating the Mixing Distribution by the Method of Moments," *Journal of the American Statistical Association*, 1990, *85* (410), 582–589.

Helpman, Elhanan, Marc Melitz, and Yona Rubinstein, "Estimating Trade Flows: Trading Partners and Trading Volumes," *The Quarterly Journal of Economics*, 2008, *123* (2), 441–487.

_ **, Oleg Itskhoki, and Stephen Redding**, "Inequality and Unemployment in a Global Economy," *Econometrica*, 2010, *78* (4), 1239–1283.

Holmström, Bengt and John Roberts, "The Boundaries of the Firm Revisited," *Journal of Economic Perspectives*, 1998, *12* (4), 73–94.

Hornok, Cecília and Miklós Koren, "Administrative Barriers to Trade," *Journal of International Economics*, 2015, *96*, S110–S122.

_ **and** _ , "Per-Shipment Costs and the Lumpiness of International Trade," *Review of Economics and Statistics*, 2015, *97* (2), 525–530.

Hummels, David and Alexandre Skiba, "Shipping the Good Apples Out? An Empirical Confirmation of the Alchian-Allen Conjecture," *Journal of Political Economy*, 2004, *112* (6), 1384–1402.

Kingma, Diederik P. and Jimmy Ba, "Adam: A Method for Stochastic Optimization," 2014. http://arxiv.org/abs/1412.6980.

Klenow, Peter and Jon Willis, "Real Rigidities and Nominal Price Changes," 2006. http://klenow.com/RealRigidities.pdf.

Kremer, Michael and Christopher M. Snyder, "Preventives Versus Treatments," *Quarterly Journal of Economics*, 2015, *130* (3), 1167–1239.

_ **and** _ , "Worst-Case Bounds on R&D and Pricing Distortions: Theory and Disturbing Conclusions if Consumer Values Follow the World Income Distribution," 2017. http://www.dartmouth.edu/%7Ecsnyder/general19.pdf.

Kropf, Andreas and Philip Sauré, "Fixed Costs per Shipment," *Journal of International Economics*, 2014, *92* (1), 166–184.

Krugman, Paul, "Scale Economies, Product Differentiation, and the Pattern of Trade," *American Economic Review*, 1980, *70* (5), 950–959.

Kubler, Felix and Karl Schmedders, "Competitive Equilibria in Semi-Algebraic Economies," *Journal of Economic Theory*, 2010, *145* (1), 301–330.

— , **Philipp Renner, and Karl Schmedders**, "Computing All Solution to Polynomial Equatoins in Economics," in Karl Schmedders and Kenneth L. Judd, eds., *Handbook of Computational Economics*, Vol. 3, Amsterdam, Holland: Elsevier, 2014.

Lehmer, Derrick H., "On the Maxima and Minima of Bernoulli Polynomials," *The American Mathematical Monthly*, 1940, *47* (8), 533–538.

Lloyd, P. J., "The Origins of the von Thünen-Mill-Pareto-Wicksell-Cobb-Douglas Function," *History of Political Economy*, 2001, *33* (1), 1–19.

Mahoney, Neale and E. Glen Weyl, "Imperfect Competition in Selection Markets," *Review of Economics and Statistics*, 2017, *99* (4), 637–651.

McCaig, Brian and Nina Pavcnik, "Export Markets and Labor Allocation in a Low-Income Country," *American Economic Review*, 2018, *108* (7), 1899–1941.

McFadden, Daniel, "Conditional Logit Analysis of Qualitative Choice Behavior," in Paul Zarembka, ed., *Fronteirs in Econometrics*, New York: Academic Press, 1974, pp. 105–142.

Melitz, Marc J., "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity," *Econometrica*, 2003, *71* (6), 1695–1725.

— **and Gianmarco I. P. Ottaviano**, "Market Size, Trade and Productivity," *Review of Economic Studies*, 2008, *75* (1), 295–316.

Mill, John Stuart, *Principles of Political Economy with some of their Applications to Social Philosophy*, London: Longmans, Green and Co., 1848.

Miller, Kenneth S and Stefan G Samko, "Completely Monotonic Functions," *Integral Transforms and Special Functions*, 2001, *12* (4), 389–402.

Mrázová, Monika and J. Peter Neary, "Not So Demanding: Preference Structure, Firm Behavior, and Welfare," 2014. https://www.economics.ox.ac.uk/materials/papers/13172/paper691.pdf.

Myerson, Roger B., "Optimal Auction Design," *Mathematics of Operations Research*, 1981, *6* (1), 58–73.

Natalini, Pierpaolo and Biagio Palumbo, "Inequalities for the Incomplete Gamma Function," *Mathematical Inequalities and Applications*, 2000, *3* (1), 69–77.

Piketty, Thomas, *Capital in the Twenty-First Century*, Cambridge, MA: Harvard University Press, 2014.

Quint, Daniel, "Imperfect Competition with Complements and Substitutes," *Journal of Economic Theory*, 2014, *152*, 266–290.

Reed, William J., "The Pareto Law of Incomes – an Explanation and an Extension," *Physica A*, 2003, *319*, 469–486.

— **and Murray Jorgensen**, "The Double Pareto-Lognormal Distribution – A New Parametric Model for Size Distributions," *Communicatoins in Statistics – Theory and Methods*, 2004, *33* (8), 1733–1753.

Rochet, Jean-Charles and Jean Tirole, "Platform Competition in Two-Sided Markets," *Journal of the European Economic Association*, 2003, *1* (4), 990–1029.

Saez, Emmanuel, "Striking it Richer: The Evolution of Top Incomes in the United States," 2013. https://eml.berkeley.edu/%7Esaez/saez-UStopincomes-2012.pdf.

Salinger, Michael A., "Vertical Mergers and Market Foreclosure," *Quarterly Journal of Economics*, 1988, *5* (1), 345–356.

Say, Jean-Baptiste, *Traité D'Économie Politique ou Simple Exposition de la Maniè̀e dont se Forment, se Distribuent et se Consomment les Richesses*, Paris: Guillaumin., 1819.

— , *Cours Complet D'Économie Politique Practique*, Paris: Rapilly, 1828.

Scheuer, Florian and Kent Smetters, "Could a Website Really Have Doomed Health Exchanges? Multiple Equilibria, Initial Conditions and the Construction of the Fine," 2014. http://ssrn.com/abstract=2386402.

Schilling, René L, Renming Song, and Zoran Vondracek, *Bernstein Functions: Theory and Applications*, Berlin: Walter de Gruyter, 2010.

Simonovska, Ina, "Income Differences and Prices of Tradables: Insights from an Online Retailer," *The Review of Economic Studies*, 2015, *82* (4), 1612–1656.

Spengler, Joseph J., "Vertical Integration and Antitrust Policy," *Journal of Political Economy*, 1950, *50* (4), 347–352.

Stole, Lars A. and Jeffrey Zwiebel, "Intra-firm Bargaining under Non-Binding Contracts," *Review of Economic Studies*, 1996, *63* (3), 375–410.

— **and** — , "Organizational Design and Technology Choice under Intrafirm Bargaining," *American Economic Review*, 1996, *86* (1), 195–222.

Tikhonov, Andrey, "Solution of Incorrectly Formulated Problems and the Regularization Method," *Soviet Math. Dokl.*, 1963, *5*, 1035–1038.

Tirole, Jean, "Hierarchies and Bureaucracies: On the Role of Collusion in Organizations," *JL Econ. & Org.*, 1986, *2*, 181.

— , *The Theory of Industrial Organization*, MIT Press, 1988.

Toda, Alexis Akira, "The Double Power Law in Income Distribution: Explanations and Evidence," *Journal of Economic Behavior & Organization*, 2012, *84* (1), 364–381.

— , "A Note on the Size Distribution of Consumption: More Double Pareto than Lognormal," *Macroeconomic Dynamics*, 2017, *21* (6), 1508–1518.

Weyl, E. Glen, "Double Marginalization in Two-Sided Markets," 2008. http://papers.ssrn.com/sol3/papers.cfm?abstract%5Fid=1324412.

— , "Monopoly, Ramsey and Lindahl in the Rochet and Tirole (2003)," *Economics Letters*, 2009, *103* (2), 99–100.

_ **and Jean Tirole**, "Market Power Screens Willingness-to-Pay," *Quarterly Journal of Economics*, 2012, *127* (4), 1971–2003.

_ **and Michal Fabinger**, "Pass-Through as an Economic Tool: Principles of Incidence under Imperfect Competition," *Journal of Political Economy*, 2013, *121* (3), 528–583.

Widder, David Vernon, *The Laplace Transform*, Princeton: Princeton University Press, 1941.

Williamson, Oliver E., "Hierarchical Control and Optimum Firm Size," *Journal of Political Economy*, 1967, *75* (2), 123–138.

_ , "Transaction Cost Economics: The Natural Progression," 2009. https://goo.gl/JJ3cvf.

Zemanian, Armen H, *Distribution Theory and Transform Analysis: an Introduction to Generalized Functions, with Applications*, New York: McGraw Hill, 1965.

Appendix

A Proofs of Theorems

Proof of Theorem 1 (Characterization of Form-Preserving Functions). Here we present a constructive proof of the theorem that exactly traces the steps we first used to derive the theorem's statement. It is instructive for readers familiar with Fourier transform or Laplace transform because it highlights the properties of functions we emphasize in this paper and shows how using the transforms, calculations may be conveniently performed just in a couple of lines. Other readers may prefer reading Supplementary Material H, where we discuss how the theorem may be proven without functional transforms.

Here we derive the structure of m-dimensional functional form classes $\mathcal{C}$ that are invariant under average-marginal transformations. We take as the domain of the functions an open interval I of positive real numbers, which may include all positive real numbers.[77] For convenience we express the (infinitely differentiable) functions $F(q)$ on I in terms of functions $G(s)$ defined in the corresponding logarithmically transformed domain, with the identification $s \equiv \log q$, $F(q) \equiv G(\log q)$. Consider a function $F(q) \in \mathcal{C}$ and its counterpart $G(s)$. In terms of G, the average-marginal form-preservation requires that the counterpart of $aG + bG'$ belong to the class $\mathcal{C}$, if the counterpart of G does so. For technical reasons, we will work with $G(s)$ multiplied by the characteristic function $1_S(s)$ of an arbitrarily chosen finite non-empty interval $S \equiv (s_1, s_2) \in I$, i.e. with $G_S(s) \equiv G(s)\, 1_S(s)$.[78] We denote by $\hat{G}_S(\omega)$ the Fourier transform of $G_S(s)$, which in turn may be expressed as the inverse Fourier transform $G_S(s) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} \hat{G}_S(\omega)\, e^{-i\omega s} d\omega$.[79]

By iterating the defining property of average-marginal form-preservation, we know that the class $\mathcal{C}$ contains also counterparts of the derivatives $G^{(n)}(s)$. We will consider the first m of them, in

[77] If functions of negative numbers were of interest, we could simply switch to working in terms of $(-q)$ instead of q and derive analogous results.

[78] If we worked with infinite intervals, the convergence of the integrals below would not be always guaranteed.

[79] The Fourier transform used in the proof is equivalent to the Laplace transform with imaginary s. Both transforms may be thought of as parts of the holomorphic Fourier-Laplace transform.

addition to $G(s)$. For $n = 1, 2, ..., m$, we denote by $G_S^{(n)}(s)$ the truncation of $G^{(n)}(s)$ to the interval S, i.e. $G_S^{(n)}(s) \equiv G^{(n)}(s)\, 1_{s\in S}$. Inside the interval S,

$$G_S^{(n)}(s) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} (-i\omega)^n \hat{G}_S(\omega)\, e^{-i\omega s} d\omega, \quad \text{for } s \in S,\ n \in \{0, 1, 2, ..., m\}. \tag{5}$$

The $m+1$ functions $G_S(s), G_S^{(1)}(s), G_S^{(2)}(s), ..., G_S^{(m)}(s)$ span a vector space with dimensionality $m+1$ or less. Dimensionality equal to $m+1$ would contradict the assumption of having an m-dimensional functional form class, which implies that the set of functions $G_S(s), G_S^{(1)}(s), G_S^{(2)}(s)$, ..., $G_S^{(m)}(s)$ must be linearly dependent on the interval S. As a result, there must exist a polynomial $T_0(.)$ (with real coefficients), such that

$$\int_{-\infty}^{\infty} T_0(-i\omega)\, \hat{G}_S(\omega)\, e^{-i\omega s} d\omega \tag{6}$$

is zero for any $s \in S$. This expression vanishes not only for $s \in S \equiv (s_1, s_2)$, but also for $s \in (-\infty, s_1)$ and $s \in (s_2, \infty)$. This is because the right-hand-side of (5) when extended to arbitrary $s \in \mathbb{R}$ represents the nth derivative of $G_S(s)$ in the sense of the Schwartz distribution theory, and given that $G_S(s)$ vanishes for $s \in (-\infty, s_1)$ and $s \in (s_2, \infty)$, so must its nth derivative. Given that the expression (6) is a generalized function[80] of s that gives zero when integrated against any test function[81] supported on $(-\infty, s_1 - \epsilon] \cup [s_1 + \epsilon, s_2 - \epsilon] \cup [s_2 + \epsilon, \infty)$ for any $\epsilon > 0$, we may write it as a linear combination of Dirac delta functions and a finite number of their derivatives located at s_1 and s_2. By computing its Fourier transform we find that $T_0(-i\omega)\, \hat{G}_S(\omega)$ must be of the form $T_1(\omega)\, e^{is_1\omega} + T_2(\omega)\, e^{is_2\omega}$ with some polynomials $T_1(\omega)$ and $T_2(\omega)$, with complex coefficients in general. Consequently, $\hat{G}_S(\omega)$ may be written as

$$\hat{G}_S(\omega) = \frac{T_1(\omega)}{T_0(-i\omega)} e^{is_1\omega} + \frac{T_2(\omega)}{T_0(-i\omega)} e^{is_2\omega}.$$

The polynomial $T_0(-i\omega)$ may have a common factor with $T_1(\omega)$ or $T_2(\omega)$ or both. If we cancel these common factors, we may rewrite the expression as

$$\hat{G}_S(\omega) = \frac{T_3(\omega)}{T_5(\omega)} e^{is_1\omega} + \frac{T_4(\omega)}{T_6(\omega)} e^{is_2\omega} \tag{7}$$

for some polynomials T_3, T_4, T_5, and T_6, such that T_3 has no common divisors with T_5 and similarly for T_4 with T_6. Let us compute the inverse Fourier transform of the last expression for $\hat{G}_S(\omega)$ using the residue theorem. To perform the integration, we consider each of the two terms in (7) separately and specialize to $s \in S$. We close the integration contour by semicircles at infinity of the complex plane, correctly chosen so that their contribution to the integral vanishes. The integral value is then equal to the sum of the pole (residue) contributions, which give exponentials of s multiplied by polynomials of s. We see that for $s \in S$, $G_S(s) = \sum_{j=1}^{N} D_j(s)\, e^{-ist_j}$, for some integer N, complex numbers t_j and polynomials $D_j(s)$. Since the interval S was chosen arbitrarily, not just $G_S(s)$, but also $G(s)$ itself must take this form. In the last expression the constants may be complex. Without loss of generality, we can assume that the first N_1 numbers t_j are real and the remaining ones have an imaginary part. By combining individual terms into real contributions so

[80] By a generalized function we mean an element of the space $\mathcal{S}'(\mathbb{R})$ of distributions.

[81] A test function here refers to an element of the space $\mathcal{S}(\mathbb{R})$ of space of rapidly decreasing functions.

that $G(s)$ is real, we get

$$G(s) = \sum_{j=1}^{N_1} A_j(s) e^{-st_j} + \sum_{j=1}^{N_2} \left(B_j(s) \cos \tilde{t}_j s + C_j(s) \sin \tilde{t}_j s \right) e^{-\hat{t}_j s},$$

where $A_j(s)$, $B_j(s)$, and $C_j(s)$ are polynomials, and $N_1 + 2N_2 = N$. This form of $G(s)$ translates into the following form of $F(q)$:

$$F(q) = \sum_{j=1}^{N_1} A_j(\log q) q^{-t_j} + \sum_{j=1}^{N_2} \left(B_j(\log q) \cos\left(\tilde{t}_j \log q\right) + C_j(\log q) \sin\left(\tilde{t}_j \log q\right) \right) q^{-\hat{t}_j}. \quad (8)$$

If we wish to exclude the possibility of oscillations, e.g. in economic applications where we allow the functional form to be valid arbitrarily close to $q = 0$, we can set the polynomials B_j and C_j to zero and consider only functions of the form $F(q) = \sum_{k=1}^{N_1} A_j(\log q) q^{-t_j}$. An example of functional forms of this kind is $aq^{-t} + bq^{-u} + cq^{-u} \log q + dq^{-u}(\log q)^2$. The reader can easily verify that this is a four-dimensional functional form class invariant under average-marginal transformations. In general, it is now straightforward to check that the result (8) implies the statement of the theorem. □

Proof of Theorem 2 (Closed-Form Solutions). The proof is straightforward. By assumption, there exists some definite power b such that $x \equiv q^b$ satisfies an algebraic equation of order k: $P_k(x) = 0$, where $P_k(x)$ is a polynomial of order at most k. For this to be true, all elements of the functional form class must factorize as $q^a P_k(q^b)$ for some definite a. When expanded, the powers of q in individual terms lie on the grid $a, a+b, ..., a+bk$. □

Proof of Theorem 3 (Aggregation). The firm's revenue $qP(q)$, cost $\int \mathrm{MC}(q)\, dq$, and profit are all linear combinations of powers of q. For this reason, it suffices to show that it is possible to perform explicitly aggregation integrals $\mathcal{I}$ for powers of q (the quantity optimally chosen by a firm with productivity parameter a): $\mathcal{I} \equiv \int q(a)^{\gamma_1}\, dG(a)$. Changing the integration variable to q gives: $\mathcal{I} = \int q^{\gamma_1} G'(a(q)) a'(q)\, dq$. The firm's first-order condition equates the marginal revenue $R'(q) = P(q) + qP'(q)$ to the marginal cost $\mathrm{MC}_0(q) + a\mathrm{MC}_1(q)$ and implies

$$a = \frac{R'(q) - \mathrm{MC}_0(q)}{\mathrm{MC}_1(q)} \Rightarrow a'(q) = \frac{R''(q) - \mathrm{MC}_0'(q)}{\mathrm{MC}_1(q)} - \frac{R'(q) - \mathrm{MC}_0(q)}{\mathrm{MC}_1(q)^2} \mathrm{MC}_1'(q).$$

Substituting these expressions into the integral gives

$$\mathcal{I} = \int q^{\gamma_1} \left(\frac{R''(q) - \mathrm{MC}_0'(q)}{\mathrm{MC}_1(q)} - \frac{R'(q) - \mathrm{MC}_0(q)}{\mathrm{MC}_1(q)^2} \mathrm{MC}_1'(q) \right) G' \left(\frac{R'(q) - \mathrm{MC}_0(q)}{\mathrm{MC}_1(q)} \right) dq.$$

Since $G'(a)$ is a mixture of powers of a, and $(R'(q) - \mathrm{MC}_0(q))\, \mathrm{MC}_1'(q)$ and $R''(q) - \mathrm{MC}_0'(q)$ are mixtures of powers of q, the integral on the right-hand side may be written as a linear combination of integrals of the type

$$\int q^{\gamma_5} \mathrm{MC}_1(q)^{\gamma_7} \left(-\mathrm{MC}_0(q) + R'(q) \right)^{\gamma_6} dq,$$

where γ_7 equals $-\gamma_6 - 1$ or $-\gamma_6 - 2$. Given our assumptions, up to a known multiplicative constant this integral equals $\int q^{\gamma_8} N_1(q^\alpha)^{\gamma_9} N_2(q^\alpha)^{\gamma_{10}}\, dq$. If we change the integration variable to $x \equiv q^\alpha$, the problem reduces to computing the integral $\int x^{\gamma_{11}} N_1(x)^{\gamma_{12}} N_2(x)^{\gamma_{13}}\, dx$. To complete the proof, it suffices to examine the structure of this intergral for different structures of the polynomials.

Depending on the structure of the polynomials, the following six non-exclusive cases may arise:

(1) If the polynomials N_1 and N_2 are trivial, the integral reduces to a power function of q, without any special functions.

(2) If either N_1 or N_2 is trivial and the other polynomial is linear, the integral leads to the standard hypergeometric function, denoted ${}_2F_1$, since up to an additive constant

$$\int x^{\gamma_{11}} \left(1+\gamma_{14}x\right)^{\gamma_{13}} dx = \frac{x^{1+\gamma_{11}}}{1+\gamma_{11}} {}_2F_1\left(1+\gamma_{11}, -\gamma_{13}; 2+\gamma_{11}; -x\gamma_{14}\right)$$

(3) If both N_1 and N_2 are linear, the integral leads to the standard Appell function, denoted F_1, since up to an additive constant

$$\int x^{\gamma_{11}} \left(1+\gamma_{18}x\right)^{\gamma_{12}} \left(1+\gamma_{19}x\right)^{\gamma_{13}} dx = \frac{x^{1+\gamma_{11}}}{1+\gamma_{11}} F_1\left(1+\gamma_{11}; -\gamma_{12}, -\gamma_{13}; 2+\gamma_{11}; -x\gamma_{18}, -x\gamma_{19}\right)$$

(4) If either N_1 and N_2 is trivial and the other polynomial is quadratic, the integral again leads to the standard Appell function, denoted F_1:

$$\int x^{\gamma_{11}} \left(1+\gamma_{14}x+\gamma_{15}x^2\right)^{\gamma_{13}} dx =$$

$$\frac{\gamma_{15}^{\gamma_{13}} x^{1+\gamma_{11}}}{1+\gamma_{11}} \left(\frac{1+x\gamma_{14}+x^2\gamma_{15}}{\gamma_{15}+x\gamma_{14}\gamma_{15}+x^2\gamma_{15}^2}\right)^{\gamma_{13}} F_1\left(1+\gamma_{11}; -\gamma_{13}, -\gamma_{13}; 2+\gamma_{11}; \gamma_{16}x, \gamma_{17}x\right)$$

where $\gamma_{16} = -2\gamma_{15}\left(\gamma_{14}+\sqrt{\gamma_{14}^2-4\gamma_{15}}\right)^{-1}$, and $\gamma_{17} = 2\gamma_{15}\left(-\gamma_{14}+\sqrt{\gamma_{14}^2-4\gamma_{15}}\right)^{-1}$.

(5) If N_1 and N_2 are both of order less than five, we can factorize them into products of linear polynomials with the factorization performed in closed form by the method of radicals. The resulting integral may be expressed using Lauricella functions. In particular, by the fundamental theorem of algebra, N_1 and N_2 may be written as products of linear functions. This means that up to a multiplicative constant, $x^{\gamma_{11}} \left(1+\gamma_{18}x\right)^{\gamma_{12}} \left(1+\gamma_{19}x\right)^{\gamma_{13}}$ equals $x^{b-1}\left(1-u_1x\right)^{-b_1}\ldots\left(1-u_nx\right)^{-b_n}$, where u_i represent the reciprocals of the roots of the polynomials. These roots, as well the constants b, b_1, ..., b_n may be found explicitly using the standard formulas for solutions to quadratic, cubic, or quartic equations. Up to an additive constant, the corresponding integral equals

$$\int x^{b-1}\left(1-u_1x\right)^{-b_1}\ldots\left(1-u_nx\right)^{-b_n} dx = \frac{x^b}{b} {F_D}^{(n)}\left(b, b_1, \ldots, b_n, b+1; u_1x, \ldots, u_nx\right)$$

This is because in general the Lauricella function ${F_D}^{(n)}$ is defined as

$${F_D}^{(n)}\left(b, b_1, \ldots, b_n, c; x_1, \ldots, x_n\right) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int_0^1 y^{b-1}(1-y)^{c-b-1}\left(1-x_1y\right)^{-b_1}\ldots\left(1-x_ny\right)^{-b_n} dy$$

with Γ denoting the standard gamma function, and in the special case of $c = b+1$ this definition becomes

$${F_D}^{(n)}\left(b, b_1, \ldots, b_n, b+1; x_1, \ldots, x_n\right) = b\int_0^1 y^{b-1}\left(1-x_1y\right)^{-b_1}\ldots\left(1-x_ny\right)^{-b_n} dy$$

Substituting $y \to x_0/x, x_1 \to u_1 x$ and $x_n \to u_n x$ then leads to the desired result for the integral:

$$\int_0^x x_0^{b-1} \left(1-u_1 x_0\right)^{-b_1} \ldots \left(1-u_n x_0\right)^{-b_n} dx_0 = \frac{x^b}{b} F_D{}^{(n)} \left(b, b_1, \ldots, b_n, b+1; u_1 x, \ldots, u_n x\right)$$

(6) Finally, if either N_1 or N_2 is of order five or higher, the factorization involves root functions, since the method of radicals can no longer be used. However, the resulting integral may still be expressed using Lauricella functions as described above.

We conclude that the structure of the resulting expressions for the integral agrees with the statement of Theorem 3. □

Proof of Theorem 4 (Laplace-log Transform with Riemann-Stieltjes Integrals). (A) This follows from Theorem I.6.3 of of Widder (1941). **(B)** If we choose $u_I(t)$ appearing in Equation 3 from the paper to be piecewise constant with a finite number N of points of discontinuity $\{t_j, j = 1, 2, ..., N\}$, the integral becomes $U(q) = \sum_{j=1}^{N} a_j q^{-t_j}$, where a_j is the (signed) magnitude of the discontinuity at point t_j, i.e. the magnitude of the mass that $u(t)$ has at point t_j. If we choose t_j to be nonpositive integers, $U(q)$ will be a polynomial of q. By appropriate choices of N and a_j, any polynomial of q may be expressed in this way. **(C)** Given that polynomials are included in Equation 2 from the paper, the theorem follows from the Weierstrass approximation theorem, which states that polynomials are dense in the space of continuous functions on a compact interval. For a constructive proof of the theorem due to Bernstein, see e.g. Section VII.2 of Feller (2008). **(D)** This follows from Theorem I.5a of Widder (1941). □

Proof of Theorem 5 (Laplace-log Transform with Schwartz Integrals). The three sentences of the theorem are implied by the following statements in Zemanian (1965): (1) Theorem 8.4-1 and Corollary 8.4-1a, (2) Theorem 8.3-1a, (3) Theorem 8.3-2 and the text following Corollary 8.4-1a. □

Proof of Theorem 6 (Discrete Approximation). This theorem follows straightforwardly from Theorem 4 of Apostol (1999). That theorem provides in its Equation 25 a convenient form of the Euler-Maclaurin formula, which may be written, after a small change of notation, as:

$$\begin{gathered}\sum_{k=1}^{n_T} F(k) = \int_1^{n_T} F(x)\,dx + \mathcal{C}(F) + E_F\left(n_T\right), \\ \mathcal{C}(F) = \tfrac{1}{2}F(1) - \sum_{r=1}^{m} \tfrac{B_{2r}}{(2r)!} F^{(2r-1)}(1) + \tfrac{1}{(2m+1)!}\int_1^{\infty} P_{2m+1}(x) F^{(2m-1)}(x)\,dx, \\ E_F\left(n_T\right) = \tfrac{1}{2}F\left(n_T\right) - \sum_{r=1}^{m} \tfrac{B_{2r}}{(2r)!} F^{(2r-1)}\left(n_T\right) + \tfrac{1}{(2m+1)!}\int_{n_T}^{\infty} P_{2m+1}(x) F^{(2m-1)}(x)\,dx.\end{gathered}$$

We can use this form of the Euler-Maclaurin formula to prove the discrete approximation theorem. The relationship we want to prove is

$$\sum_{t\in T} q^{-t} f(t) = \tfrac{1}{\Delta \mathrm{t}} \int q^{-t} f(t)\,dt + \tfrac{1}{2} q^{-t_{\min}} f\left(t_{\min}\right) + \tfrac{1}{2} q^{-t_{\max}} f\left(t_{\max}\right) - \tfrac{R_1+R_2+R_3}{\Delta \mathrm{t}},$$

where $T \equiv \{t_{\min}, t_{\min}+\Delta\mathrm{t}, ..., t_{\max}\}$ and n_T is the number of points in the grid T. Equivalently,

$$\sum_{t\in T} q^{-t} f(t) = \tfrac{1}{\Delta \mathrm{t}} \int_{t_{\min}}^{t_{\max}} q^{-t} f(t)\,dt + \tfrac{1}{2} q^{-t_{\min}} f\left(t_{\min}\right) + \tfrac{1}{2} q^{-t_{\max}} f\left(t_{\max}\right) - \tfrac{R_2+R_3}{\Delta \mathrm{t}}.$$

If we use the notation

$$F(k) \equiv q^{-t_{\min}-k\Delta \mathrm{t}} f\left(t_{\min} + (k-1)\Delta \mathrm{t}\right)$$

we can rewrite the individual terms in the desired formula as

$$\begin{gathered}\sum_{t\in T} q^{-t} f(t) = \sum_{k=1}^{n_T} F(k), \\ \tfrac{1}{2} q^{-t_{\min}} f\left(t_{\min}\right) + \tfrac{1}{2} q^{-t_{\max}} f\left(t_{\max}\right) = \tfrac{F(1)}{2} + \tfrac{F(n_T)}{2}, \\ \tfrac{R_2}{\Delta \mathrm{t}} = \sum_{r=1}^{m} \tfrac{B_{2r}}{(2r)!}\left(F^{(-1+2r)}(1) - F^{(-1+2r)}\left(n_T\right)\right), \\ \tfrac{R_3}{\Delta \mathrm{t}} = -\tfrac{\Delta \mathrm{t}^{2m}}{(1+2m)!}\int_{t_{\min}}^{t_{\max}} P_{1+2m}(t) h^{(1+2m)}(t)\,dt = -\tfrac{1}{(2m+1)!}\int_1^n P_{1+2m}(x) F^{(1+2m)}(x)\,dx.\end{gathered}$$

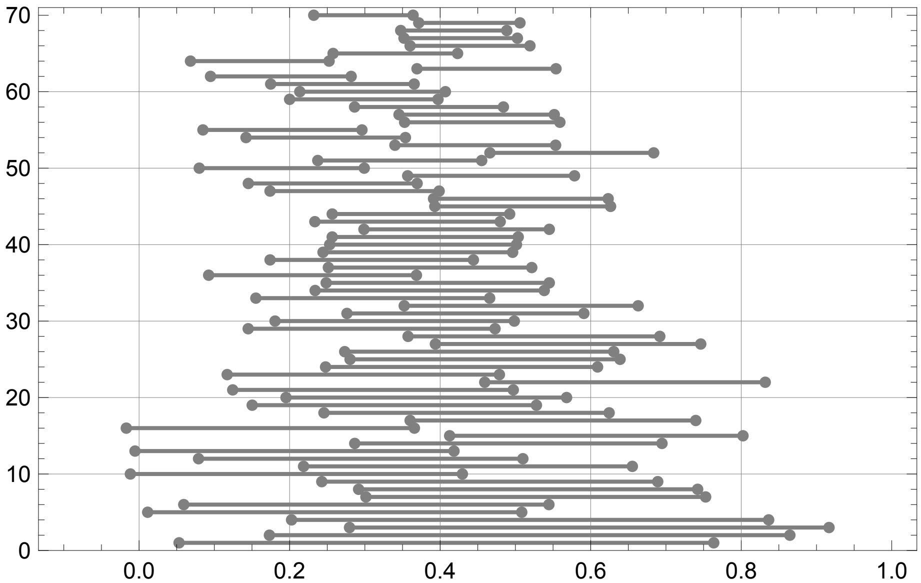


Figure 8: Confidence intervals for the cost exponent γ for individual industries at the 95% level. For visualization purposes, the industries are ordered by the standard deviation of γ and stacked vertically.

By comparing these expressions with those of Theorem 4 of Apostol (1999), we see that the main statement of Theorem 6 is valid. The bound on R_3 then simply follows from the formula $|P_{2m+1}(x)| \leq 2(2m+1)!(2\pi)^{-2m-1}$; see p. 538 of Lehmer (1940). $\square$

Proof of Theorem 7 (Nonnegativity of Laplace Consumer Surplus). This theorem follows from Bernstein's theorem on completely monotone functions, formulated e.g. as Theorem IV.12a of Widder (1941) or Theorem 1.4 of Schilling et al. (2010). $\square$

Proof of Theorem 8 (Monotonicity of the Pass-Through Rate). Constant marginal cost monopoly pass-through rate may be expressed as $\rho = CS'_{[s]}(s)/CS''_{[s]}(s)$, which is straightforward to verify from the basic definitions. For a completely monotone problem, Laplace consumer surplus $cs\,(t)$ is nonnegative. For this reason, the inverse of ρ may be expressed as a weighted average of t with nonnegative weight $w\,(t,s) \equiv t\,cs\,(t)\,e^{-st}/\int_{-\infty}^{0} t\,cs\,(t)\,e^{-st}dt$ as follows

$$\frac{1}{\rho} = \frac{CS''_{[s]}(s)}{CS'_{[s]}(s)} = -\frac{\int_{-\infty}^{0} t^2 cs\,(t)\,e^{-st}dt}{\int_{-\infty}^{0} t\;cs\,(t)\,e^{-st}dt} = -\int_{-\infty}^{0} t\,w\,(t,s)\,dt.$$

In response to an increase in s, the weight gets shifted towards more negative t,[82] and $1/\rho$ decreases. We conclude that ρ is decreasing in q. Only if $t\,cs\,(t)$ is supported at one point will there be no shift in weight and ρ remains constant. That case corresponds to BP demand. $\square$

Proof of Theorem 9 (Complete Monotonicity of Demand Specification). The complete monotonicity properties follow by straightforwardly recognizing that in these cases $t\,p\,(t)$ is nonnegative and supported on $(-\infty, 1)$, with the corresponding Laplace inverse demand functions $p\,(t)$ listed in our Supplementary Material E, which also contains additional discussion. Note that for most of the inverse demand functions listed in the theorem, it is also possible to prove complete monotonicity using Theorems 1–6 of Miller and Samko (2001). $\square$

[82]In the same mathematical sense as in the definition of first order stochastic dominance.

$N_{f,\min}$	N_I	β	σ_β
5	192	0.39	0.20
10	70	0.39	0.12
15	45	0.39	0.10
20	23	0.41	0.10
25	14	0.39	0.10
30	11	0.42	0.07
35	9	0.42	0.08

Table 3: Sensitivity to the cutoff $N_{f,\min}$ of the number of firms per industry. The cutoff influences the number of industries N_I that satisfy the sample selection criteria and the resulting mean β and the corresponding standard deviation σ_β.

Proof of Theorem 10 (Absence of Complete Monotonicity of Demand Specification) . The statement of the theorem follows by inspection of the Laplace inverse demand functions, as in the previous proof. Additional discussion may be found in Supplementary Material E. □

B Details of the Generalized EOQ Model Estimation

Here we provide additional details of the estimation of the cost parameter $\beta = (1-\gamma)/(2-\gamma)$. As mentioned in the main text, we selected industries that included at least 10 firms satisfying our criteria. The corresponding confidence intervals corresponding to individual industries are plotted in Figure 8.

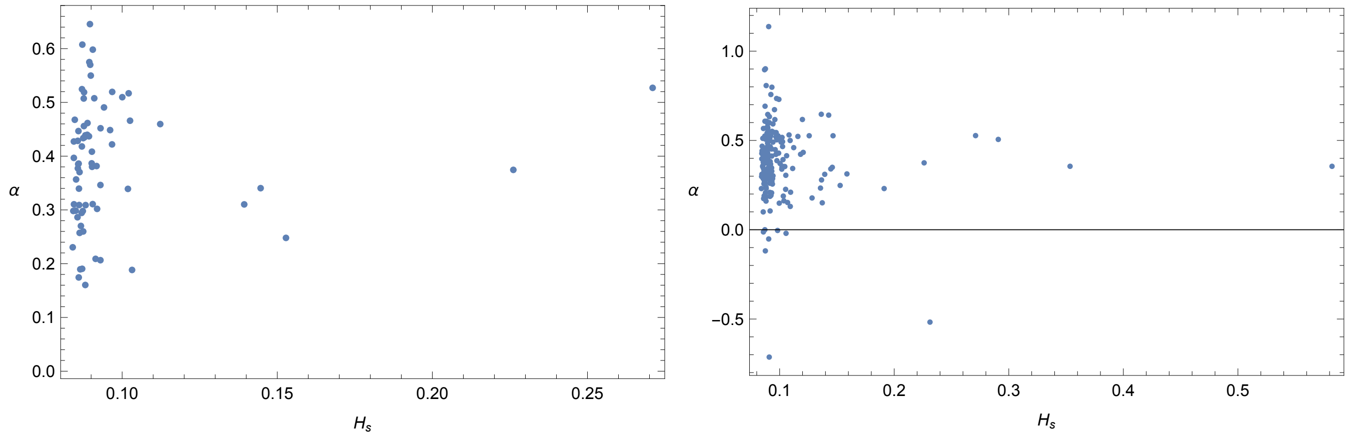
In principle, the value of average estimated β could be sensitive to the cutoff on the number of firms per industry. Table 3 summarizes the dependence of the resulting average β on the choice of the cutoff. It turns out that the average β remains roughly the same even for large changes of the cutoff on the number of firms.

The estimated value of γ could be, in principle, also influenced by seasonality patterns. To investigate this issue, we construct a measure of seasonality of individual industries. In particular, we calculate a Herfindahl-like seasonality index based on the shares of trade in individual months of the year, defined as $H_s = \sum_{i=1}^{12} v_i^2$, where v_i is the average share of month i in the average annual trade value. A high value of the index means that trade flows are very unevenly distributed across months. Then we regress γ on this measure. We find that the 95% confidence interval of the slope coefficient is [-0.69,1.21] and the corresponding p-value is 0.58. For robustness, we change the cutoff to 5 firms, getting the confidence interval [-0.91,0.30] and the p-value of 0.32. In both cases we do not reject the hypothesis that the slope coefficient is zero. The data is plotted in Figure 9.

C World Trade

C.1 Details of data construction

Here we provide details of the data construction for Section 4. The economies used to fit our model are, in descending order by 2006 GDP, United States, Japan, Germany, China, United Kingdom, France, Italy, Canada, Spain, Brazil, Russia, South Korea, Mexico, India, Australia, Netherlands, Turkey, Switzerland, Sweden, Belgium, Saudi Arabia, Norway, Poland, Austria, Denmark,



(a) (b)

Figure 9: The relationship of the cost exponent α for specific industries and the industry seasonality index H_s. Figure (a) corresponds to the sample used for the main estimation, which is based on industries with at least 10 firms satisfying the sample selection criteria. We do not observe any systematic pattern relating α and H_s. Figure (b) corresponds to a cutoff set to 5 firms as a robustness check. Also, in this case the values of α do not seem to be influenced by H_s.

Greece, South Africa, Iran, Argentina, Ireland, Nigeria, United Arab Emirates, Thailand, Finland, Portugal, Hong Kong, Venezuela, Malaysia, Colombia, Czech Republic, Chile, Israel, Singapore, Pakistan, Romania, Algeria, Hungary, New Zealand, Kuwait, Peru, Kazakhstan, Bangladesh, Morocco, Vietnam, Qatar, Slovakia, Croatia, Ecuador, Luxembourg, Slovenia, Dominican Republic, Oman, Belarus, Tunisia, Bulgaria, Syria, Sri Lanka, Serbia/Serbia and Montenegro, Lithuania, Guatemala, Kenya, Costa Rica, Lebanon, Latvia, Azerbaijan, Cyprus, Ghana, Uruguay, Yemen, Tanzania, El Salvador, Bahrain, Trinidad and Tobago, Panama, Cameroon, Ivory Coast, Iceland, Estonia, Ethiopia, Jordan, Macau, Zambia, Bosnia and Herzegovina, Bolivia, Jamaica, Uganda, Honduras, Paraguay, Gabon, and Senegal. These countries were selected based on data availability. We computed the tradable share (percentage) of GDP by selecting tradable sectors from United Nations gross value added database. We fit the GDP in the model to the tradable portion of GDP, computed as GDP reported by IMF World Economic Outlook database multiplied by the tradable share of GDP.[83] This means that, for example, education revenue and education expenditures are not counted towards the model's GDP and expenditures, which is appropriate for a model designed to capture manufacturing and similar industries. Multi-sector extensions including services are, of course, possible. Also, note that we exclude re-imports and re-exports from the trade flows data.

[83] In the model, all imports are consumed domestically, which implies that exports cannot be larger than tradable GDP. However, such situation might arise for small, highly open economies. To avoid this discrepancy, when the calculated portion of tradable GDP that goes to domestic consumption is smaller than five percent of the tradable GDP in the data, we increment it so that it reaches that level. This is done by correspondingly increasing both the (adjusted) tradable GDP and the (adjusted) consumption in the economy. This criterion was satisfied for just one economy, Hong Kong. Of course, a more realistic way of modeling this situation is to include multi-stage production and/or multi-stage transportation in the model. This will require some additional research work, but it is a clear direction to pursue.

C.2 Related literature

Here we briefly discuss connections of the results of Subsection 4.11 to related issues in the literature, as mentioned in Footnote 53. Helpman, Melitz and Rubinstein (2008) studied the role of the extensive margin of trade for the estimation of the distance-dependence of trade costs based on world trade flows. The authors found the distance effect to be 27 to 30 percent smaller than in benchmark estimates based on the gravity equation of trade without extensive margin effects. Although this is an important correction, it is not enough to resolve the trade cost puzzle. We get much stronger effects because of the increasing marginal cost of production. Moreover, unlike that paper we do not need unrealistically high export market entry costs that would be inconsistent with the everyday experience that even sole entrepreneurs with very limited capital (for example, 25,000 USD) are able to start an import/export business, a fact that is explained in many resources, such as Entrepreneur Magazine (2003) .

Separately, Arkolakis (2010) builds an elegant model of international trade where fixed costs of exporting are indeed negligible (and marginal costs of production are constant). Even though the demand is CES, some firms choose not to export to a particular destination because before serving a customer, they need to pay a sizeable per-customer advertising cost, which can make serving that customer unprofitable. An argument against this mechanism is that it would not work if targeted advertising was possible. Empirical evidence in the industrial organization literature shows that the main portion of observed aggregate demand elasticity comes from heterogeneity in the consumers' valuation of products, not from elasticity of demand by a given individual; an individual's demand is quite inelastic in the data. If firms could reach high-valuation customers and advertize directly to them, they would export to that destination. Especially in recent years targeted advertising via the Internet is quite easy and widespread, so it is hard to justify the modeling assumption that it is impossible. For this reason, it is better to think of the insightful paper Arkolakis (2010) in a more abstract way: as an investigation of situations where effective demand departs from CES. In principle, we could remove economies of scale in shipping from our model and instead modify the demand. In this case, again, we could combine this with our assumption of increasing marginal costs of production, and using our proposed tractable functional forms for demand we could proceed with computations in the same way. But of course, we already have empirical evidence on the economies of scale in shipping, and we know that logistics costs as a proportion of world GDP are very large. Note that the influential study of export decisions Eaton, Kortum and Kramarz (2011) also uses the Arkolakis (2010) mechanism in theoretical modeling.

C.3 Firm export patterns

Here we mention other possible mechanisms potentially leading to patterns similar to those in Figure 7 of Subsection 4.12, as referenced in Footnote 55. If the countries significantly differ and we break the symmetry between the firms (in terms of how their products enter utility functions), it is possible to explain patterns resembling those in Figure 7. For example, windows imported by Finland are likely to be very different from windows imported by Portugal. If a firm specializes in only one kind of windows, it is natural for them to export to only one of these destinations. Another possible phenomenon that could lead to similar patterns in the data would be distribution centers in export destinations. For example, a firm may serve both Spain and Portugal from one distribution center based in Spain. In that case international trade flow data would not record such sales in Portugal as exports to Portugal, but instead as exports to Spain and then exports from Spain to Portugal. Yet another possibility is the case of very large firms. If these firms were so large that monopolistic competition description of the market was inappropriate and we needed to model it as an oligopoly, there could be an alternative explanation for choosing different export destinations. In this case strategic effects of market entry could potentially play a role. A firm may not choose to serve Greece because Greece is already served by its rival and the market the is not profitable enough for two firms to enter. The puzzle would still remain for smaller firms that cannot influence the entire industry. More generally, these three explanations may be valid in some cases but are not powerful enough to explain the majority of the empirical regularity in the data, especially in the case of smaller firms that directly export goods that are not geographically specialized. Case studies of individual exporters also make it clear that the export pattern is typically not explained by those three explanations. A detailed investigation of these issues will be reported separately.

D Applications

D.1 Supply chains with hold-up (Antràs and Chor, 2013)

We consider a generalization of the supply chain model of Antràs and Chor (2013, henceforth AC). Instead of the variables introduced in the original paper, we use a different set of variables that makes the mathematics and intuition substantially simpler.[84]

A firm produces a final good by sequentially using a continuum of customized inputs each provided by a different supplier indexed by $j \in [0,1]$, with small j representing initial stages of production (upstream) and large j representing final stages (downstream). If production proceeds smoothly, the *effective* quality-adjusted quantity q of the final good is the integral of the effective quality-adjusted quantity contributed by intermediate input j, which we denote $q_s(j)$: $q = \int_0^1 q_s(j)\, dj$. This effective quantity represents both the quantity of the good and its quality level. But we will refer to it simply as "quality", since this will make the discussion sound more natural. If production is "disrupted" by the failure of some supplier $\bar{j} \in [0,1)$ to cooperate, then only the quality accumulated to that point in the chain is available, with all further quality-enhancement impossible: $q = \int_0^{\bar{j}} q_s(j)\, dj$. The firm faces an inverse demand function $P(q)$, which does not necessarily have to be decreasing because, for example, consumers may have little willingness-to-pay for an improperly finished product. If there is no disruption in production, $q = q(1)$.

Following the property rights theory of the firm (Grossman and Hart, 1986; Hart and Moore, 1990; Antràs, 2003), input production requires relationship-specific investments. The marginal revenue from additional quality brought by supplier j, $MR\,(q(j))\, q_s(j)$ is therefore split between the firm and supplier j, where $MR = P + P'q$.[85] In particular, the supplier receives a fraction $1 - \beta(j)$ (its bargaining power).

The cost of producing quality $q_s(j)$ is homogeneous across suppliers and equal to $C(q_s(j))$, which is assumed strictly convex.[86] Thus the first-order condition of supplier j equates the share of

[84] The relationship between our variables introduced in the next paragraph (in a notation compatible with the rest of this paper) and the variables in Antràs and Chor (2013) is as follows. Let us use the symbol $\tilde{q}$ to refer to a quantity measure denoted q in AC, which is *distinct* from what we call effective quality-adjusted quantity q. In order to recover AC's original model as a special case, we identify their output $\tilde{q}$ with $q^{1/\alpha}$, where $\alpha \in (0,1)$ is a constant defined there. For the present discussion we do not need q to be linearly proportional to the number of units produced. It is just some measure of the output, which may or may not be quite abstract. A similar statement applies to the customized intermediate input. Our measure $q_s\,(j)$ of a particular input is related to AC's measure $x(j)$ by $q_s\,(j) = \theta^\alpha (x\,(j))^\alpha$, where θ is a positive productivity parameter defined in their original paper.

[85] See AC's Subsection 3.1 for a discussion of why only marginal revenue, and not the full-downstream revenue, is the pie that is bargained over and an alternative micro-foundation of this model.

[86] The AC model corresponds to $C(q_s) = (q_s)^{1/\alpha} c/\theta$, where c and θ are positive constants defined in their paper. In our notation, the suppliers' cost is convex but their contributions towards the final output are linear. In the original

marginal revenue she bargains for with her marginal cost:

$$MC\left(q_s(j)\right) \equiv C'\left(q_s(j)\right) = \left[1 - \beta\left(j\right)\right] MR\left(q\left(j\right)\right). \tag{9}$$

The cost to the firm of obtaining a contribution $q_s(j)$ from supplier j is, therefore, the surplus it must leave in order to induce $q_s(j)$ to be produced, $q_s MC\left(q_s(j)\right)$.

The firm chooses $\beta\left(j\right)$ through the nature of the contracting relationship optimally for each supplier to maximize its profits. Following AC and Antràs and Helpman (2004, 2008), we mostly focus on the *relaxed* problem where $\beta\left(j\right)$ may be adjusted freely and continuously. This provides most of the intuition for what happens when the firm is constrained to choose between two discrete levels of β corresponding to outsourcing (low β) and insourcing (high β) and may be more realistic given the complexity of real-world contracting (Holmström and Roberts, 1998). Note that by convexity, $MC' > 0$, while each q_s makes a linearly separable contribution to q. Thus for any fixed q the firm wants to achieve, it does so most cheaply by setting all $q_s = q$ by Jensen's Inequality. Thus Equation 9 becomes, at any optimum $q^\star$,

$$\beta^*\left(j\right) = 1 - \frac{MC\left(q^\star\right)}{MR\left(jq^\star\right)}. \tag{10}$$

From this we immediately see that $\beta^\star$ is co-monotone with MR: in regions where marginal revenue is increasing, $\beta^\star$ will be rising and conversely when marginal revenue is decreasing. The marginal revenue associated with constant elasticity demand is in a constant ratio to inverse demand. This implies AC's principal result that when revenue elasticity is less than unity the firm will tend to outsource upstream and when revenue elasticity is less than unity the firm will tend to outsource downstream. However, it seems natural to think that $P(q)$ would initially rise, as consumers are willing to pay very little for a product that is nowhere near completion, and would eventually fall as the product is completed according to the standard logic of downward-sloping demand. We now solve in an equally-simple form a model allowing this richer logic.

Equation 10 implies that the surplus left to each supplier is $q_s MC(q)$ and thus total cost is $qMC(q)$. The problem reduces to choosing q to maximize revenue $qP(q)$ less cost $qMC(q)$, giving first-order condition

$$MR(q) = MC\left(q\right) + q\, MC'\left(q\right). \tag{11}$$

This differs from the familiar neoclassical first-order condition $MR\left(q\right) = MC\left(q\right)$ only by the presence of the (positive) term $q\, MC'\left(q\right)$. Note that $MC + qMC'$ bears the same relationship to MC that MC bears to AC; this equation therefore similarly inherits the tractability properties of the standard monopoly problem. The reason is that the hold-up makes multi-part tariff pricing impossible, creating a linear-price monopsony structure by forcing the firm to pay suppliers the marginal cost of the last unit of quality for all units produced.

Let us now consider $P(q) = p_{-t}q^t + p_{-u}q^u$ and $MC(q) = mc_{-t}q^t + mc_{-u}q^u$. This includes AC's specification as the special case when $p_{-t} = 0$ and $mc_{-u} = 0$ so that each has constant elasticity.[87] However, let us focus instead on the case when $t, u, mc_{-u}, p_{-t} > 0 = mc_{-t} > p_{-u}$ and $u > t$ so that

paper the suppliers' cost is linear, but their contributions towards the final output have diminishing effects. These are two alternative interpretations of the same economic situation from the point of view of two different systems of notation. As mentioned before, in our interpretation, the product of a supplier is q_s, whereas in the original paper the supplier's product is x, related to q_s by $q_s\left(j\right) = \theta^\alpha (x\left(j\right))^\alpha$.

[87] In particular, in their notation, AC have $t = \frac{1}{\alpha}$, $u = 1 + \frac{\rho}{\alpha}$, $mc_{-t} = {}^{c}/_{\alpha\theta}$ and $p_{-u} = A^{1-\rho}$, where θ and ρ is are positive constants defined in AC, not to be confused with the pass-through rate denoted by ρ or the conduct parameter denoted by θ in other parts of this paper.

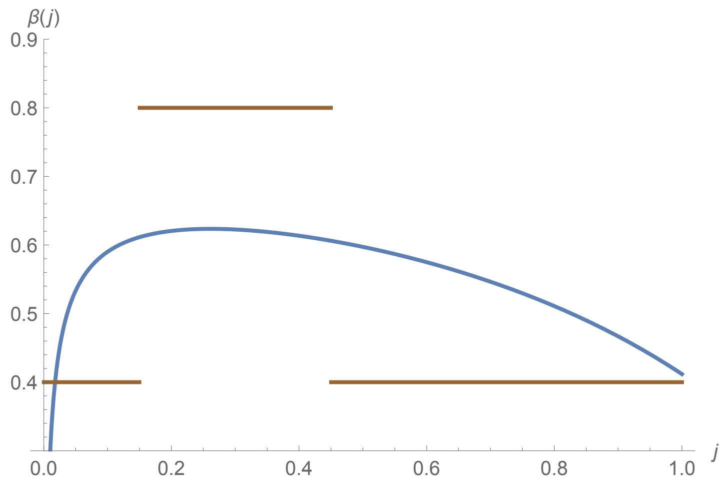


Figure 10: Optimal relaxed and restricted $\beta^\star$ in the AC model when $t = 0.35, u = 0.7, \frac{p_{-u}}{mc_{-u}} = -4$.

the first term of the inverse demand dominates at small quantities while the second dominates at large quantities. The expression resulting for $\beta^* (j)$ is:

$$\beta^* (j) = 1 - \frac{1}{(1+u)\left[\left(1 - \frac{p_{-u}}{mc_{-u}}\right) j^t + \frac{p_{-u}}{mc_{-u}} j^u\right]}. \tag{12}$$

Note that because $mc_{-u} > 0 > p_{-u}$, the first denominator term is positive and the second denominator term is negative. This implies that at small j (where j^t dominates), $\beta^\star$ increases in j, while at large j, it decreases in j. In the AC complements case when $p_{-u} = 0$, or even if p_{-u} is sufficiently small, this large j behavior is never manifested and all outsourcing (low $\beta^\star$) occurs at early stages. Also note that only the ratio of coefficients $\frac{p_{-u}}{mc_{-u}}$ matters for the sourcing pattern; p_{-t} is irrelevant, as the joint level of p_{-u} and mc_{-u}.

However, for many parameters an inverted U-shape emerges. For example, Figure 10 shows the case when $t = 0.35, u = 0.7, p_{-t} = 1.8, \frac{p_{-u}}{mc_{-u}} = -4$. The curve corresponds to the shape of the relaxed solution. Depending on precisely which values of β we take insourcing and outsourcing to correspond to, this can lead to insourcing in the middle of the production and outsourcing at either end. In Supplementary Material I.1 we study in detail the constrained problem using largely closed-form methods for the case when outsourcing gives $\beta_O = 0.8$ and insourcing gives $\beta_I = 0.4$. This is illustrated by the lines in Figure 10, which show the constrained optimum. This gives the same qualitative answer as the relaxed problem, as expected.

D.2 Imperfectly competitive supply chains

The models that founded the field of industrial organization were Cournot (1838)'s of symmetric oligopoly and complementary monopoly. Equilibrium in these models is characterized by

$$P + \theta P' q = MC.$$

Under Cournot competition, $\theta = 1/n$, where n is the number of competing firms and MC is interpreted as the common marginal cost of all producers. Under Cournot complements (which does not require symmetry) $\theta = m$, where m is the number of complementary producers and MC is interpreted as the aggregated marginal cost of all producers.[88] Note that $P + \theta P'q$ is just a linear combination of P and $P'q$ and thus has the same form as either of these components in a form-preserving class of functional forms. Thus either problem yields exactly the same characterization of tractability as the monopoly problem.

In the last half-century a variant on Cournot (1838)'s complementary monopoly problem proposed by Spengler (1950) has been more commonly used. In this model one firm sells an input to another who in turn sells to a consumer. The difference from Cournot's model is principally in the timing; namely the "upstream" firm is assumed to set her price prior to the downstream firm. In this case the upstream firm effectively sets part of the downstream firm's marginal cost. Her first-order condition is

$$P + P'q = MC + \hat{P},$$

where $\hat{P}$ is the sales price set by the upstream firm. Thus the effective inverse demand faced by the upstream firm is $\hat{P}(q) \equiv P(q) + P'(q)q - MC(q)$. The upstream firm then solves a monopoly problem with this inverse demand. This yields an upstream marginal revenue curve bearing the same relationship to $\hat{P}$ that MR bears to P. Because the form-preserving feature may be applied an arbitrary number of times, however, this transformation does not change our characterization of tractability. Thus a form-preserving class has the same tractability characterization in Spengler's model as in the standard Cournot model.

We can go further and allow for many layers of production and arbitrary imperfect competition (or complements) at each later as in Salinger (1988). The same characterization of tractability continues to apply. In Supplementary Material I.3 we provide an explicit expression for the coefficients in the polynomial equation for any tractable form. Adachi and Ebina (2014a,b) argue that flexible functional forms are particularly important in such models because many important and policy-relevant properties are imposed by standard tractable forms. For example, the markup of the upstream firm in Spengler's model is identical to that of the two firms if they merged under the BP demand class, but the upstream firm will typically charge a lower markup than an integrated firm under reasonable conditions (bell-shaped-distribution-generated demand and U-shaped cost curves).

[88] With constant marginal cost, and in some other special cases, the asymmetric Cournot competition model may also be solved if both demand is specified in an appropriate form. To maintain the generality of our analysis we do not discuss this solvable, asymmetric special case.

Supplementary Material

E Laplace Inverse Demand Functions

The following table contains Laplace inverse demand functions corresponding to inverse demand functions used in the literature. Although for most Laplace inverse demand functions we include only a few terms, closed-form expressions for all terms exist. Here p_a refers to a mass-point of magnitude p_a at location a. In the alternative notation on the lower lines, $\delta(x-a)$ refers to a mass-point of magnitude 1 at location a, i.e. to a Dirac delta function centered at a. We use standard notation for special functions: Γ stands for the gamma function and W for the Lambert W function.

Constant elasticity / Pareto: $q(P) = \left(\frac{P}{\beta}\right)^{-\epsilon}$ $P(q) = \beta q^{-1/\epsilon}$

$p(t):$ $p_{\frac{1}{\epsilon}} = \beta$

$p(t):$ $\beta\delta\left(t - \frac{1}{\epsilon}\right)$

Constant pass-through / BP: $q(P) = \left(\frac{P-\mu}{\beta}\right)^{-\epsilon}$ $P(q) = \mu + \beta q^{-1/\epsilon}$

$p(t):$ $p_0 = \mu,\ \ p_{\frac{1}{\epsilon}} = \beta$

$p(t):$ $\beta\delta\left(t - \frac{1}{\epsilon}\right) + \mu\delta(t)$

Gumbel distribution: $q(P) = \exp\left(-\exp\left(\frac{P-\alpha}{\beta}\right)\right)$ $P(q) = \alpha + \beta\log(-\log(q))$

$p(t):$ $p_0 = \mu,\ \ p(t) = -\frac{\beta}{t}$ for $t < 0$

$p(t):$ $\alpha\delta(t) - \frac{\beta 1_{t<0}}{t}$

Weibull distribution: $q(P) = e^{-\left(\frac{P}{\beta}\right)^{\alpha}}$ $P(q) = \beta(-\log(q))^{\frac{1}{\alpha}}$

$p(t):$ $\frac{(-1)^{\frac{1}{\alpha}}\beta t^{-\frac{1}{\alpha}-1}}{\Gamma\left(-\frac{1}{\alpha}\right)}$ for $t < 0$

$p(t):$ $\frac{(-1)^{\frac{1}{\alpha}}\beta 1_{t<0} t^{-\frac{1}{\alpha}-1}}{\Gamma\left(-\frac{1}{\alpha}\right)}$

Fréchet distribution: $q(P) = 1 - e^{-\left(\frac{P-\mu}{\beta}\right)^{-\alpha}}$ $P(q) = \mu + \beta(-\log(1-q))^{-1/\alpha}$

$p(t):$ $p_0 = \mu,\ \ p_{\frac{1}{\alpha}} = \beta,\ \ p_{\frac{1}{\alpha}-1} = -\frac{\beta}{2\alpha},\ \ p_{\frac{1}{\alpha}-2} = \frac{\beta}{8\alpha^2} - \frac{5\beta}{24\alpha},\ \ \ldots$

$p(t):$ $\left(\frac{\beta}{8\alpha^2} - \frac{5\beta}{24\alpha}\right)\delta\left(t - \frac{1}{\alpha} + 2\right) + \beta\delta\left(t - \frac{1}{\alpha}\right) - \frac{\beta\delta\left(t - \frac{1}{\alpha} + 1\right)}{2\alpha} + \mu\delta(t) + \ldots$

Logistic distribution: $q(P) = \left(\exp\left(\frac{P-\mu}{\beta}\right) + 1\right)^{-1}$ $P(q) = \mu - \beta\log\left(\frac{1}{1-q} - 1\right)$

$p(t):$ $p_0 = \mu,\ \ p_0^{(1)} = -\beta,\ \ p_{-1} = -\beta,\ \ p_{-2} = -\frac{\beta}{2},\ \ p_{-3} = -\frac{\beta}{3},\ \ p_{-4} = -\frac{\beta}{4},\ \ \ldots$

$p(t):$ $-\beta\sum_{j=1}^{\infty}\frac{\delta(j+t)}{j} + \mu\delta(t) - \beta\delta'(t)$

Log-logistic distribution: $q(P) = \left(\left(\frac{P}{\sigma}\right)^{\gamma} + 1\right)^{-1}$ $P(q) = \sigma\left(\frac{q}{1-q}\right)^{-1/\gamma}$

$p(t):$ $p_{\frac{1}{\gamma}} = \sigma,\ \ p_{\frac{1}{\gamma}-1} = -\frac{\sigma}{\gamma},\ \ p_{\frac{1}{\gamma}-2} = \frac{\sigma}{2\gamma^2} - \frac{\sigma}{2\gamma},\ \ p_{\frac{1}{\gamma}-3} = -\frac{\sigma}{6\gamma^3} + \frac{\sigma}{2\gamma^2} - \frac{\sigma}{3\gamma},\ \ \ldots$

$p(t):$ $\left(\frac{\sigma}{2\gamma^2} - \frac{\sigma}{2\gamma}\right)\delta\left(t - \frac{1}{\gamma} + 2\right) + \sigma\delta\left(t - \frac{1}{\gamma}\right) - \frac{\sigma\delta\left(t - \frac{1}{\gamma} + 1\right)}{\gamma} + \ldots$

Laplace distribution ($q < \frac{1}{2}$): $q(P) = \frac{1}{2}\exp\left(\frac{\mu-P}{\beta}\right)$ $P(q) = \mu - \beta\log(2q)$

$p(t):$ $p_0 = \mu - \beta\log(2),\ \ p_0^{(1)} = -\beta$

$p(t):$ $\delta(t)(\mu - \beta\log(2)) - \beta\delta'(t)$

Laplace distribution ($q > \frac{1}{2}$): $q(P) = 1 - \frac{1}{2}\exp\left(\frac{P-\mu}{\beta}\right)$ $P(q) = \mu + \beta\log(2(1-q))$

$p(t):$ $p_0 = \beta\log(2) + \mu,\ \ p_{-1} = -\beta,\ \ p_{-2} = -\frac{\beta}{2},\ \ p_{-3} = -\frac{\beta}{3},\ \ p_{-4} = -\frac{\beta}{4},\ \ \ldots$

$p(t):$ $\delta(t)(\beta\log(2) + \mu) - \beta\sum_{j=1}^{\infty}\frac{\delta(j+t)}{j}$

Normal distribution: $q(P) = \operatorname{erfc}\left(\frac{P-\mu}{\sqrt{2}\sigma}\right)$ $P(q) = \mu - \sqrt{2}\sigma\operatorname{erfc}^{-1}(2-q)$

$p(t):$ $p_0^{(1)} = -\sqrt{\frac{\pi}{2}}\sigma,\ \ p_0^{(2)} = -\frac{1}{2}\sqrt{\frac{\pi}{2}}\sigma,\ \ p_0^{(3)} = \frac{1}{24}\left(-\sqrt{2}\pi^{3/2} - 2\sqrt{2\pi}\right)\sigma,\ \ \ldots$

$p(t):$ $-\sqrt{\frac{\pi}{2}}\sigma\delta'(t) - \frac{1}{2}\sqrt{\frac{\pi}{2}}\sigma\delta''(t) + \frac{1}{24}\left(-\sqrt{2}\pi^{3/2} - 2\sqrt{2\pi}\right)\sigma\delta^{(3)}(t) + \ldots$

Lognormal distribution: $q(P) = \operatorname{erfc}\left(\frac{\log(P)-\mu}{\sqrt{2}\sigma}\right)$ $P(q) = \exp\left(\mu - \sqrt{2}\sigma\operatorname{erfc}^{-1}(2-q)\right)$

$p(t):$ $p_0^{(1)} = \sqrt{\frac{\pi}{2}}\left(-e^{\mu}\right)\sigma\delta'(t),\ \ p_0^{(2)} = \frac{1}{4}\pi e^{\mu}\sigma^2 - \frac{1}{2}\sqrt{\frac{\pi}{2}}e^{\mu}\sigma,\ \ \ldots$

$p(t):$ $\left(\frac{1}{4}\pi e^{\mu}\sigma^2 - \frac{1}{2}\sqrt{\frac{\pi}{2}}e^{\mu}\sigma\right)\delta''(t) - \sqrt{\frac{\pi}{2}}e^{\mu}\sigma\delta'(t) + \ldots$

Almost Ideal Demand System: $q(P) = \frac{\alpha+\beta\log(P)}{P} \quad P(q) = -\frac{\beta W\left(-\frac{q e^{-\frac{\alpha}{\beta}}}{\beta}\right)}{q}$

$p(t): \quad p_0 = e^{-\frac{\alpha}{\beta}}, \quad p_{-1} = \frac{e^{-\frac{2\alpha}{\beta}}}{\beta}, \quad p_{-2} = \frac{3e^{-\frac{3\alpha}{\beta}}}{2\beta^2}, \quad p_{-3} = \frac{8e^{-\frac{4\alpha}{\beta}}}{3\beta^3}, \quad p_{-4} = \frac{125e^{-\frac{5\alpha}{\beta}}}{24\beta^4}, \quad \ldots$

$p(t): \quad \frac{125e^{-\frac{5\alpha}{\beta}}\delta(t+4)}{24\beta^4} + \frac{8e^{-\frac{4\alpha}{\beta}}\delta(t+3)}{3\beta^3} + \frac{3e^{-\frac{3\alpha}{\beta}}\delta(t+2)}{2\beta^2} + e^{-\frac{\alpha}{\beta}}\delta(t) + \frac{e^{-\frac{2\alpha}{\beta}}\delta(t+1)}{\beta} + \ldots$

Constant superelasticity: $q(P) = \left(\epsilon\log\left(\frac{\theta-1}{\theta P}\right)+1\right)^{\frac{\theta}{\epsilon}} \quad P(q) = \frac{(\theta-1)e^{\frac{1}{\epsilon}-\frac{q^{\epsilon/\theta}}{\epsilon}}}{\theta}$

$p(t): \quad p_0 = e^{\frac{1}{\epsilon}} - \frac{e^{\frac{1}{\epsilon}}}{\theta}, \quad p_{-\frac{\epsilon}{\theta}} = \frac{e^{\frac{1}{\epsilon}}}{\theta\epsilon} - \frac{e^{\frac{1}{\epsilon}}}{\epsilon}, \quad p_{-\frac{2\epsilon}{\theta}} = \frac{e^{\frac{1}{\epsilon}}}{2\epsilon^2} - \frac{e^{\frac{1}{\epsilon}}}{2\theta\epsilon^2}, \quad p_{-\frac{3\epsilon}{\theta}} = \frac{e^{\frac{1}{\epsilon}}}{6\theta\epsilon^3} - \frac{e^{\frac{1}{\epsilon}}}{6\epsilon^3}, \quad \ldots$

$p(t): \quad \left(\frac{e^{\frac{1}{\epsilon}}}{2\epsilon^2} - \frac{e^{\frac{1}{\epsilon}}}{2\theta\epsilon^2}\right)\delta\left(t+\frac{2\epsilon}{\theta}\right) + \delta(t)\left(e^{\frac{1}{\epsilon}} - \frac{e^{\frac{1}{\epsilon}}}{\theta}\right) + \left(\frac{e^{\frac{1}{\epsilon}}}{\theta\epsilon} - \frac{e^{\frac{1}{\epsilon}}}{\epsilon}\right)\delta\left(t+\frac{\epsilon}{\theta}\right) + \ldots$

Cauchy distribution: $q(P) = \frac{\tan^{-1}\left(\frac{a-P}{b}\right)}{\pi} + \frac{1}{2} \quad P(q) = a + b\tan\left(\pi\left(\frac{1}{2}-q\right)\right)$

$p(t): \quad p_1 = \frac{b}{\pi}, \quad p_0 = a, \quad p_{-1} = -\frac{\pi b}{3}, \quad p_{-3} = -\frac{\pi^3 b}{45}, \quad p_{-5} = -\frac{2\pi^5 b}{945}, \quad p_{-7} = -\frac{\pi^7 b}{4725}, \quad \ldots$

$p(t): \quad a\delta(t) + \frac{b\delta(t-1)}{\pi} - \frac{1}{3}\pi b\delta(t+1) - \frac{1}{45}\pi^3 b\delta(t+3) - \frac{2}{945}\pi^5 b\delta(t+5) - \frac{\pi^7 b\delta(t+7)}{4725} + \ldots$

Singh Maddala distribution: $q(P) = \left(\left(\frac{P}{b}\right)^a + 1\right)^{-\tilde{q}} \quad P(q) = b\left(q^{-\frac{1}{\tilde{q}}} - 1\right)^{\frac{1}{a}}$

$p(t): \quad p_{\frac{1}{a\tilde{q}}} = b, \quad p_{-\frac{a-1}{a\tilde{q}}} = -\frac{b}{a}, \quad p_{-\frac{2a-1}{a\tilde{q}}} = \frac{b}{2a^2} - \frac{b}{2a}, \quad p_{-\frac{3a-1}{a\tilde{q}}} = -\frac{b}{6a^3} + \frac{b}{2a^2} - \frac{b}{3a}, \quad \ldots$

$p(t): \quad \left(\frac{b}{2a^2} - \frac{b}{2a}\right)\delta\left(\frac{2a-1}{a\tilde{q}} + t\right) + b\delta\left(t - \frac{1}{a\tilde{q}}\right) - \frac{b\delta\left(\frac{a-1}{a\tilde{q}}+t\right)}{a} + \ldots$

Tukey lambda distribution: $q(P) = P^{(-1)}(P) \quad P(q) = \frac{(1-q)^\lambda - q^\lambda}{\lambda}$

$p(t): \quad p_{-\lambda} = -\frac{1}{\lambda}, \quad p_0 = \frac{1}{\lambda}, \quad p_{-1} = -1, \quad p_{-2} = \frac{\lambda}{2} - \frac{1}{2}, \quad p_{-3} = -\frac{\lambda^2}{6} + \frac{\lambda}{2} - \frac{1}{3}, \quad \ldots$

$p(t): \quad \left(-\frac{\lambda^2}{6} + \frac{\lambda}{2} - \frac{1}{3}\right)\delta(t+3) + \frac{\delta(t)}{\lambda} + \left(\frac{\lambda}{2} - \frac{1}{2}\right)\delta(t+2) - \frac{\delta(t+\lambda)}{\lambda} - \delta(t+1) + \ldots$

Wakeby distribution: $q(P) = P^{(-1)}(P) \quad P(q) = \mu - \frac{\gamma\left(1-q^{-\delta}\right)}{\delta} + \frac{\alpha\left(1-q^\beta\right)}{\beta}$

$p(t): \quad p_0 = \frac{\alpha}{\beta} - \frac{\gamma}{\delta} + \mu, \quad p_{-\beta} = -\frac{\alpha}{\beta}, \quad p_\delta = \frac{\gamma}{\delta}$

$p(t): \quad \delta(t)\left(\frac{\alpha}{\beta} - \frac{\gamma}{\delta} + \mu\right) - \frac{\alpha\delta(t+\beta)}{\beta} + \frac{\gamma\delta(t-\delta)}{\delta} + \ldots$

Here we provide clarification of some of the expressions in the table above. In many cases the terms in the Laplace inverse demand were obtained utilizing series expansions, as discussed in Subsection F.2.1 of this supplementary material. More explicitly, we utilized the following series representations of the Laplace inverse demand.

Fréchet distribution:	$P(q) = \mu + \beta q^{-1/\alpha}\sum_{k=0}^{\infty}\binom{-\frac{1}{\alpha}}{k}\left(\sum_{j=2}^{\infty} j^{-1}q^{j-1}\right)^k$
Log-logistic distribution:	$P(q) = \sigma q^{1/\gamma}\sum_{n=0}^{\infty}(-1)^n\binom{\frac{1}{\gamma}}{n}q^n$
Almost Ideal Demand System:	$P(q) = \beta\sum_{n=0}^{\infty}\frac{(-1-n)^n}{(1+n)!}\left(-\frac{e^{-\frac{\alpha}{\beta}}}{\beta}\right)^{1+n}q^n$
Constant superelasticity:	$P(q) = \sum_{n=0}^{\infty}\frac{e^{1/\epsilon}(-\epsilon)^{-n}(\theta-1)}{\theta n!}q^{\frac{n\epsilon}{\theta}}$
Cauchy distribution:	$P(q) = a + \frac{b}{\pi q} + b\sum_{k=1}^{\infty}\frac{(-1)^k 2^{2k}\pi^{-1+2k}B_{2k}}{(2k)!}q^{-1+2k}$
Singh-Maddala distribution:	$P(q) = bq^{-\frac{1}{a\tilde{q}}}\sum_{n=0}^{\infty}(-1)^n\binom{\frac{1}{a}}{n}q^{\frac{n}{\tilde{q}}}$

We used the standard notation for generalized binomial coefficients and denoted Bernoulli numbers by B_{2k}. Constant superelasticity refers to the inverse demand function introduced by Klenow and Willis (2006). The Gumbel distribution is also known as the type I extreme value distribution, the Fréchet distribution as type II extreme value, and the Weibull distribution as type III extreme value.

In the case of the Gumbel distribution, the Laplace inverse demand is not an ordinary function, but a distribution (generalized function) in the sense of the distribution theory by Laurent Schwartz. For this reason, we use regularization to give a precise meaning to integrals involving the Laplace inverse demand function that was schematically written in the previous table. We provide three regularization prescriptions and illustrate them for the case of Laplace inverse demand itself. The first prescription is

$$P(q) = \lim_{a\to 0^+}\left(\int_{-\infty}^{\infty}(\alpha - \beta(\gamma + \log(a)))\delta(t)dt + \int_{-\infty}^{a}\frac{\beta q^{-t}}{t}\,dt\right).$$

Here we moved the upper bound in the second integral beyond zero and added the regularization term proportional to $\log(a)$. The second integral is to be interpreted in the sense of principal value.[89] It is straightforward to verify that this expression leads to the correct expression for $P(q)$. First, we evaluate the integrals to get

$$P(q) = \lim_{a\to 0^+}(\alpha - \gamma\beta + \beta\mathrm{Ei}(-a\log(q)) - \beta\log(a)).$$

Here γ is the Euler gamma, $\gamma \approx 0.577216$, and Ei stands for the special function called exponential integral. Evaluating the limit then leads to the correct expression

$$P(q) = \alpha + \beta\log(-\log(q)).$$

The second prescription is analogous and shifts the upper bound of the second integral to negative numbers:

$$P(q) = \lim_{a\to 0^+}\left(\int_{-\infty}^{\infty}(\alpha - \beta(\gamma + \log(a)))\delta(t)dt + \int_{-\infty}^{-a}\frac{\beta q^{-t}}{t}\,dt\right).$$

Evaluating the integral gives

$$P(q) = \lim_{a\to 0^+}(\alpha - \gamma\beta - \beta\Gamma(0, -a\log(q)) - \beta\log(a)),$$

and taking the limit leads again to the correct expression. Here Γ is the incomplete gamma function.

The third prescription is computationally most convenient because it does not involve taking a limit. The regularizing term is expressed in the form of an integral

$$P(q) = \int_{-\infty}^{\infty}(\alpha - \beta\gamma)\delta(t)dt + \beta\int_{-\infty}^{0}\frac{q^{-t} - 1_{t>-1}}{t}\,dt.$$

Here $1_{t>-1}$ is an indicator function. The integral may then be computed directly, again leading to the correct expression. The same methods may be used when interpreting other integrals involving generalized functions that behave as $1/t$ close to $t = 0$.

For the normal and lognormal distributions the Laplace inverse demand functions are again not ordinary functions. They were obtained using Taylor series expansions of the error function. Expressions involving these Laplace inverse demand functions may need to be summed using the Euler summation method to ensure proper convergence.

The expressions above may be used to straightforwardly derive Theorems 9 and 10 of the main paper. An alternative, but sometimes less straightforward way is to utilize Theorems 1–6 of Miller and Samko (2001). If we are interested in the monotonicity properties of the pass-through rate,

[89] In Mathematica, principal value integrals may be computed by choosing the option *PrincipalValue* $\to$ *True* for the *Integrate* function.

we can use the corollary in Section 6. However, we also identified more direct ways to prove monotonicity properties of the pass-through rate for certain demand functions. These proofs are included in Section J of this supplementary material.

F Evaluation of Inverse Laplace Transform

In the main text of the paper we used the term Laplace-log transform to emphasize that this is Laplace transform in terms of the logarithm of an economic quantity. In this section, which focuses on mathematical issues, we use the term Laplace transform, keeping the economic interpretation implicit.

F.1 Numerical Evaluation of Inverse Laplace Transform

There exist many methods for numerical evaluation of inverse Laplace transform, now usually integrated into mathematical and statistical software. For a summary and important references, see, e.g., Chapter 6 of Egonmwan (2012). Note that just like other types of non-parametric methods, numerical inversion of Laplace transform requires some regularization, such as the Tikhonov (1963) regularization. This is because Laplace transform inversion is a so-called *ill-posed* problem, which means that there exist large changes in the inverse Laplace transform that lead to only small changes in the original function in the domain of interest. For a classic discussion, see Bellman et al. (1966).

F.2 Analytic Evaluation of Inverse Laplace Transform

Mathematical software allows for symbolic inversion of Laplace transform.[90] However, it is often more convenient to evaluate the inverse Laplace transform using more direct methods.

F.2.1 Using Taylor series expansion

We would like to emphasize that finding analytic expressions for Laplace inverse demand is often much simpler than it seems since it many cases it only requires finding a Taylor series expansion of a definite function. Consider, for example, the case of log-logistic distribution of valuations included in the Supplementary Material E, which corresponds to inverse demand $P\left(q\right)=\sigma(\frac{q}{1-q})^{-1/\gamma}$. This may be written as $P\left(q\right)=\sigma q^{-1/\gamma}\times\left(1-q\right)^{1/\gamma}$, i.e. a product of a power function and a function that has a well-defined Taylor expansion at $q=0$:

$$\left(1-q\right)^{1/\gamma}=1-\frac{1}{\gamma}\;q+\frac{1-\gamma}{2\gamma^{2}}\;q^{2}+...=\sum_{n=0}^{\infty}(-1)^{n}\binom{\frac{1}{\gamma}}{n}q^{n},$$

where the nth term contains a generalized binomial coefficient. This immediately translates into

$$P\left(q\right)=\sigma q^{-\frac{1}{\gamma}}-\frac{\sigma}{\gamma}\;q^{1-\frac{1}{\gamma}}+\frac{\left(1-\gamma\right)\sigma}{2\gamma^{2}}\;q^{2-\frac{1}{\gamma}}+...=\sum_{n=0}^{\infty}(-1)^{n}\binom{\frac{1}{\gamma}}{n}q^{n-\frac{1}{\gamma}},$$

and from here we can read off the masses at points $t=\frac{1}{\gamma},\frac{1}{\gamma}-1,\frac{1}{\gamma}-2,...$ that together constitute the Laplace inverse demand included in Supplementary Material E.

[90] The corresponding functions are *InverseLaplaceTransform* in Mathematica, *ilaplace* in MATLAB, or *inverse_laplace_transform* in Python (SymPy).

F.2.2 Using the traditional inverse Laplace transform formula

The readers may be familiar with the traditional inverse Laplace transform formula based on the Bromwich integral in the complex plane:[91]

$$f(t) = \frac{1}{2\pi i} \lim_{T\to\infty} \int_{\gamma - iT}^{\gamma + iT} e^{st} f_{\mathcal{L}}(s)\, ds, \text{ where } f_{\mathcal{L}}(s) \equiv \int_0^\infty e^{-st} f(t)\, dt. \tag{13}$$

For the purposes of this paper we did not actually need it. We obtained the Laplace inverse demand function listed in Supplementary Material E by simpler methods.

F.2.3 Piecewise inverse Laplace transform

Readers familiar with Fourier transform but not with Laplace transform may potentially be concerned about applicability of our approach to the case of linear demand. Our prescription is simple: If the inverse demand takes the form $P(q) = a - bq$, we restrict our attention to $q \in \left(0, \frac{a}{b}\right)$, without affecting the form of Laplace inverse demand $p(t)$. The reason why this is possible is that to evaluate $p(t)$ using, say, Equation 13, we do not need the values of $P(q) \equiv P(e^s)$ for $s \in (-\infty, \infty)$, as a superficial analogy with Fourier transform might suggest. Instead, the integral in Equation 13 is in the imaginary direction. Writing the inverse demand as $P(e^s) = a - be^s$ for Re $s < \log \frac{a}{b}$ and $P(e^s) = 0$ for Re $s > \frac{a}{b}$ and working with each piece separately will not make the Laplace inverse demand complicated. We will just have two different Laplace inverse demand functions, each valid for a range of q.

G Solving Cubic and Quartic Equations Simply

The readers may have seen general formulas for solutions to cubic and quartic equations that looked very complicated. It turns out that the intimidating look is caused just by shifts and rescalings of variables. Solving these equations is actually very straightforward:

Cubic equations. To solve the equation $x^3 + 3ax + 2 = 0$, we substitute $x \equiv y^{1/3} - ay^{-1/3}$, which leads to the quadratic equation $y^2 + 2y - a^3 = 0$ with solutions $y = \pm\sqrt{1 + a^3} - 1$. Given this result, the solutions to any other cubic equation may be obtained by rescaling and shifting of x.[92]

Quartic equations. A quartic equation of the form $x^4 + ax^2 + bx + 1 = 0$ is equivalent to $(x^2 + \sqrt{\alpha}x + \beta)(x^2 - \sqrt{\alpha}x + \beta^{-1}) = 0$ with $\alpha \geq 0$ and β chosen to satisfy $\beta + \beta^{-1} = a + \alpha$ and $\beta^{-1} - \beta = \frac{b}{\sqrt{\alpha}}$ so that the coefficients of different powers of x match. If we substitute the right-hand-side expressions into the trivial identity $(\beta + \beta^{-1})^2 - (\beta - \beta^{-1})^2 = 4$, we get a cubic equation for α, which we know how to solve. With the help of the quadratic formula, a solution for α then translates into a solution for β, and consequently for x. Given these results, the solutions to any other quartic equation may be obtained simply by rescaling and shifting of x.

[91] Here i is the imaginary unit and γ is a real number large enough to ensure that $F(s)$ is holomorphic in the half-plane Re $s > \gamma$ (or has a holomorphic analytic continuation to this half-plane).

[92] What we described here is a version of Vieta's substitution that we customized to avoid cluttering of various rational factors.

H Discussion of the Characterization Theorem

Here we present a discussion of the logic behind Theorem 1 that only requires knowledge of very elementary calculus, i.e. simple differentiation and integration, instead of assuming knowledge of complex analysis and functional transforms.[93]

H.1 One-dimensional functional form classes

Consider a real function $P(q)$ (on an open interval of positive q) that belongs to a one-dimensional functional form class invariant under (i.e., preserved by) average-marginal transformations. It must be the case that $aP(q) + bqP'(q)$ also belongs to this class. For this reason, if $qP'(q)$ were not a multiple of $P(q)$, the class would not be one-dimensional, which would be a contradiction. Denoting the coefficient of proportionality as A, we get the differential equation

$$qP'(q) = AP(q),$$

which implies

$$\frac{dP}{P} = A\frac{dq}{q} \quad \Longrightarrow \quad \log|P| = A\log q + \text{const.} \quad \Longrightarrow \quad P(q) = c_1 q^A,$$

with some real constant c_1. We conclude that the one-dimensional functional form classes invariant under the average-marginal transformations are the classes of power functions with a fixed exponent.

In the next subsection we work in terms of the variable $s = \log q$. For clarity, let us repeat the computation above using s, with the identification $H(s) \equiv P(q)$ for $s = \log q$. The differential equation becomes $H'(s) = AH(s)$ and leads to the same result as above:

$$\frac{dH}{H} = A\,ds \quad \Longrightarrow \quad \log|H| = As + \text{const.} \quad \Longrightarrow \quad H(s) = c_1 e^{As}.$$

H.2 Two-dimensional functional form classes

Any member $H(\log(q)) = P(q)$ of a two-dimensional class invariant under average-marginal transformations must satisfy the following differential equation:

$$H''(s) = AH(s) + BH'(s).$$

The logic is analogous to the one-dimensional case in the previous subsection. If $H'(s)$ and $H(s)$ are proportional to each other, the equation is automatically satisfied. Consider the case where $H'(s)$ and $H(s)$ are not proportional to each other. By the invariance property, any linear combination of $H(s)$ and $H'(s)$ must belong to the class. Consequently, any linear combination of $H(q)$, $H'(s)$, and $H''(s)$ must belong to the class. These linear combinations span a three-dimensional space, unless $H''(s)$ is a linear combination of $H(q)$ and $H'(s)$, i.e. unless the differential equation is satisfied for some A and B. This establishes the validity of the differential equation.

Denote by r_1 and r_2 the two roots of the quadratic equation

$$x^2 = A + B\ x,$$

[93] We are grateful to an anonymous referee for the excellent suggestion that this is possible and worth doing. In the original version of the paper we only included the proof of Theorem 1 based on functional transforms. That proof is quick and easy, but less pedagogical because it requires knowledge of more advanced mathematics.

namely

$$r_1 = \frac{1}{2}\left(B + \sqrt{4A + B^2}\right), \quad r_2 = \frac{1}{2}\left(B - \sqrt{4A + B^2}\right).$$

It is straightforward to check that the differential equation $H''(s) = AH(s) + BH'(s)$ may be alternatively written as

$$\left(1 - r_1 \frac{d}{ds}\right)\left(\left(1 - r_2 \frac{d}{ds}\right) H(s)\right) = 0.$$

The function $\left(1 - r_2 \frac{d}{ds}\right) H(s)$, which we denote $f(s)$, therefore satisfies the differential equation

$$\left(1 - r_1 \frac{d}{ds}\right) f(s) = 0.$$

This differential equation gives the solution

$$f(s) - r_1 f'(s) = 0 \Longrightarrow ds = r_1 \frac{df}{f} \Longrightarrow \log|f| = \frac{s}{r_1} + \text{const.} \Longrightarrow f(s) = \tilde{c}_1 e^{\frac{s}{r_1}},$$

which means

$$\left(1 - r_2 \frac{d}{ds}\right) H(s) = \tilde{c}_1 e^{\frac{s}{r_1}}.$$

To get the final expression for $H(q)$, we will solve this differential equation in two alternative cases.

H.2.1 The case of two distinct roots

Let us consider the case where the two roots, r_1 and r_2, are not equal. To solve this last differential equation, let us perform the substitution $H(s) = e^{\frac{s}{r_1}} g(s) + \frac{1}{r_1 - r_2} \tilde{c}_1 r_1 e^{\frac{s}{r_1}}$. The differential equation then becomes

$$\left(1 - r_2 \frac{d}{ds}\right)\left(e^{\frac{s}{r_1}} g(s) + \frac{e^{\frac{s}{r_1}} \tilde{c}_1 r_1}{-r_1 + r_2}\right) = \tilde{c}_1 e^{\frac{s}{r_1}},$$

or after canceling terms proportional to $\tilde{c}_1$,

$$\left(1 - r_2 \frac{d}{ds}\right)\left(e^{\frac{s}{r_1}} g(s)\right) = 0.$$

This differential equation has the following solution:

$$\left(1 - \frac{r_2}{r_1} - r_2 \frac{d}{ds}\right) g(s) = 0 \quad \Longrightarrow \quad g(s) = c_2 e^{\left(1 - \frac{r_2}{r_1}\right)\frac{s}{r_2}} \quad \Longrightarrow \quad g(s) = c_2 e^{s\left(\frac{1}{r_2} - \frac{1}{r_1}\right)}.$$

For the function $H(s)$ this implies

$$H(s) = c_2 e^{\frac{s}{r_2}} + \frac{\tilde{c}_1 r_1}{r_1 - r_2} e^{\frac{s}{r_1}}.$$

After introducting the notation $c_1 = \frac{1}{r_1 - r_2} \tilde{c}_1 r_1$, the expression for $H(s)$ becomes

$$H(s) = c_1 e^{\frac{s}{r_1}} + c_2 e^{\frac{s}{r_2}}.$$

If both roots r_1 and r_2 are real, then this is the desired final form. If they have a non-zero imaginary part, we would like to manipulate the expression further. In this case $\frac{1}{r_1}$ and $\frac{1}{r_2}$ are complex conjugates, which means we may write them as

$$\frac{1}{r_1} = a_R + a_I i, \quad \frac{1}{r_2} = a_R - a_I i.$$

For $H(s)$ to be real, we also need c_1 and c_2 to be complex conjugates:

$$c_1 = c_R - c_I i, \quad c_2 = c_R + c_I i.$$

$H(s)$ then becomes

$$H(s) = (c_R - c_I i)\, e^{a_R s} \left(\cos(a_I s) + i \sin(a_I s)\right) + (c_R + c_I i) \left(\cos(a_I s) - i \sin(a_I s)\right),$$

which after canceling terms gives the desired final form

$$H(s) = c_R e^{a_R s} \cos(a_I s) + c_I e^{a_R s} \sin(a_I s).$$

H.2.2 The case of a double root

Now let us consider the case where the two roots are equal: $r_1 = r_2$. In this case, we need to solve the equation

$$\left(1 - r_1 \frac{d}{ds}\right) H(s) = \tilde{c}_1 e^{\frac{s}{r_1}}$$

We perform the substitution $H(s) = e^{\frac{s}{r_1}} g(s)$, which leads to

$$\left(1 - r_1 \frac{d}{ds}\right)\left(e^{\frac{s}{r_1}} g(s)\right) = \tilde{c}_1 e^{\frac{s}{r_1}} \quad \Longrightarrow \quad -r_1 \frac{d}{ds} g(s) = \tilde{c}_1 \quad \Longrightarrow \quad g(s) = -\frac{\tilde{c}_1}{r_1} s + \text{const.}$$

The result for $H(s)$ is $H(s) = e^{\frac{s}{r_1}}(-\frac{\tilde{c}_1}{r_1} s + \text{const.})$, which, after renaming the constants, gives the desired final form

$$H(s) = e^{\frac{s}{r_1}} \left(c_2 + c_1 s\right).$$

We see that overall, the resulting characterization of two-dimensional classes of functional forms invariant under average-marginal transformations is consistent with the statement of Theorem 1.

H.3 Higher-dimensional functional form classes

The same method may be used to derive higher-dimensional form-preserving classes of functional forms. The differential equations one needs to solve are standard and have known solutions that utilize the properties of the *characteristic equations* of the differential equations, which are analogous to the equation $x^2 = A + B\, x$ we used above. The proof of Theorem 1 described in Appendix A may be thought of as using such differential equations, just represented in a different, transformed way as the vanishing of the expression in Equation 6 inside the interval S. Solving the differential equations using transforms is much quicker and more convenient.

I Applications

I.1 Supply chains with hold-up: the restricted problem

Extending our analysis of the Antràs and Chor model in Appendix D.1, we consider the solution of the restricted AC model in the case where the firm is restricted to two discrete levels of bargaining power corresponding to "out-sourcing" and "in-sourcing". As in the relaxed solution, consider the optimal choice of a path for β subject to producing a total quality $\hat{q}$. Note that $q(j;\beta)$ is a strictly increasing function of j for any path of β achieving $\hat{q}$ by definition. Thus it is equivalent, instead of solving for the optimal restricted β for each j, to solve for the optimal $\beta^{\star\star}$ for each $q\left(j;\beta\right)\in[0,\hat{q}]$ and then invert the resulting $q\left(j;\beta^{\star\star}\right)$ function to recover the value optimal β at each j. This method preserves the separability we used in the relaxed problem and thus greatly simplifies the restricted problem. Wherever it does not create confusion we suppress as many arguments as possible, especially the dependence on β, to preserve notational economy.

By the same arguments as in the restricted case, the cost of production $\hat{q}$ is $C\left(\hat{q};\beta\right)$ where

$$C\left(\hat{q};\beta\right)=\int_0^{\hat{q}}\left[1-\beta(q)\right]MR(q)dq,$$

where $\beta(q)$ is a notationally-abusive contraction of $\beta\left(j\left(q;\beta\right)\right)$. However, to actually produce $\hat{q}$, we need

$$\int_0^1 S\left(\left[1-\beta\left(q(j)\right)\right]MR\left(q(j)\right)\right)dj=\hat{q},$$

where $S=MC^{-1}$, the supply curve, exists because of our assumption that MC is strictly monotone increasing. Changing variables so that both integrals are taken over j:

$$C\left(\beta\right)=\int_0^1\left[1-\beta\left(q(j)\right)\right]MR\left(q(j)\right)S\left(\left[1-\beta\left(q(j)\right)\right]MR\left(q(j)\right)\right)dj.$$

Thus the firm solves a Lagrangian version of this problem that is separable in each j, or equivalently q:

$$\max_{\beta}\int_0^1\lambda S\left(\left[1-\beta\left(q(j)\right)\right]MR\left(q(j)\right)\right)-\left(\left[1-\beta(q)\right]MR(q)S\left(\left[1-\beta\left(q(j)\right)\right]MR\left(q(j)\right)\right)\right)dj-\lambda\hat{q}.$$

At each q this is a simple maximization problem. The firm chooses the value of β maximizing

$$\lambda S\left(\left[1-\beta\left(q\right)\right]MR\left(q\right)\right)-\left[1-\beta(q)\right]MR(q)S\left(\left[1-\beta\left(q\right)\right]MR\left(q\right)\right),$$

the difference between the total value of the production by that firm and the total cost of that production. Clearly, both terms are decreasing in β given that $MR>0$ in any range where the firm would consider producing, so given that the firm chooses between only two values of β, $\beta_I>\beta_O$, the firm will strictly choose in-sourcing if and only if

$$MR(q)>\frac{\lambda\left[S\left(\left[1-\beta_O\right]MR\left(q\right)\right)-S\left(\left[1-\beta_I\right]MR\left(q\right)\right)\right]}{\left[1-\beta_O\right]S\left(\left[1-\beta_O\right]MR\left(q\right)\right)-\left[1-\beta_I\right]S\left(\left[1-\beta_I\right]MR\left(q\right)\right)}.\tag{14}$$

If the sign here is equality (which generically occurs on a set of measure 0 so long as the functions are nowhere constant relative to one another) then the firm is indifferent and if the inequality is reversed the firm strictly chooses in-sourcing. As λ rises, the firm will in-source less and produce more; thus varying λ over all positive numbers traces out all potentially optimal solutions. Note that this could easily be extended to a situation where the firm has any simple restricted choice of β, not just two values.

Furthermore, once $\beta(q)$ is set, we can easily recover the optimal $\beta^{\star\star}$ for each j by noting that the optimal value of $\beta^{\star\star}$ at $\tilde{j}$ is the optimal value at $\tilde{q}$ satisfying the production equation

$$\int_0^{\tilde{j}} S\left(\left[1-\beta^{\star\star}\left(q(j)\right)\right] MR\left(q^{\star\star}(j)\right)\right) dj = \tilde{q}.$$

This implies the differential equation $q'(j) = S\left(\left[1-\beta^{\star\star}\left(q(j)\right)\right] MR\left(q^{\star\star}(j)\right)\right)$ and thus the inverse differential equation $j'(q) = \frac{1}{S([1-\beta^{\star\star}(q)]MR(q^{\star\star}))}$ which together with the boundary condition $j(0)=0$ yields $j(q)$ and thus $\beta^{\star\star}$ at each j.

It remains only to pin down the optimal value of λ. To do this, denote the set of q on which Inequality 14 is satisfied $B_I(\lambda)$ and on which it is reversed $B_O(\lambda)$.[94] Total production is

$$q_\lambda = \int_{j\in(0,1):q(j)\in B_I(\lambda)} S\left(\left(1-\beta_I\right) MR\left(q(j)\right)\right) dj + \int_{j\in(0,1):q(j)\in B_O(\lambda)} S\left(\left(1-\beta_O\right) MR\left(q(j)\right)\right) dj,$$

while total cost $C_\lambda =$

$$\int_{B_I(\lambda)\cap(0,q_\lambda)} \left[1-\beta_I\right] MR\left(q\right) dq + \int_{B_O(\lambda)\cap(0,q_\lambda)} \left[1-\beta_O\right] MR\left(q\right).$$

Profit is

$$R\left(q_\lambda\right) - C_\lambda$$

and the first-order condition for its maximization is

$$MR\left(q_\lambda\right)\frac{\partial q_\lambda}{\partial \lambda} - \frac{\partial C_\lambda}{\partial \lambda} = 0 \implies MR\left(q_\lambda\right) = \frac{\frac{\partial C_\lambda}{\partial \lambda}}{\frac{\partial q_\lambda}{\partial \lambda}} = \lambda,$$

because λ is defined as the shadow cost of relaxing the constraint on production.

Now we consider obtaining as close as possible to an explicit solution. Note that, to do so, we must be able to characterize S, B_O and B_I explicitly. S is the inverse of MC and thus MC must admit an explicit inverse. To characterize B_O and B_I explicitly requires solving Inequality 14 with equality to determine the relevant thresholds, which, as we will see, requires marginal revenue to have an explicit inverse.

One of the simplest forms satisfying these conditions and yet yielding our desired non-monotonicity is $P(q) = p_0 + p_{-t}q^t + p_{-2t}q^{2t}$ and $MC(q) = mc_{-t}q^t$, where $t, p_0, p_{-t}, mc_{-t} > 0 > p_{-2t}$. In this case $S(p) = \left(\frac{p}{mc_{-t}}\right)^{\frac{1}{t}}$. Thus the equality version of Inequality 14 becomes

$$MR(q) = \frac{\lambda\left(\left[\frac{(1-\beta_O)MR(q)}{mc_{-t}}\right]^{\frac{1}{t}} - \left[\frac{(1-\beta_I)MR(q)}{mc_{-t}}\right]^{\frac{1}{t}}\right)}{(1-\beta_O)\left[\frac{(1-\beta_O)MR(q)}{mc_{-t}}\right]^{\frac{1}{t}} - (1-\beta_I)\left[\frac{(1-\beta_I)MR(q)}{mc_{-t}}\right]^{\frac{1}{t}}} \implies$$

[94] We ignore the generically 0-measure set on which it is an equality.

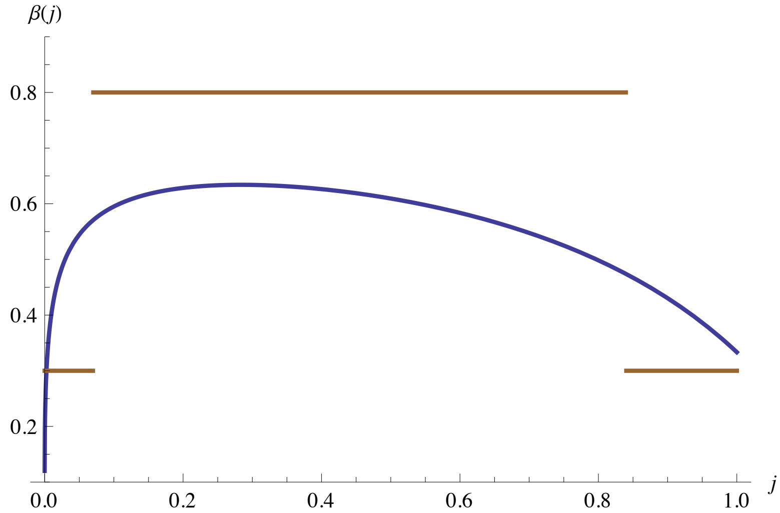


Figure 11: Relaxed and restricted solutions to the AC model when $P(q) = 0.2+2q^{\frac{1}{2}} - 4q$, $MC(q) = \frac{q^{\frac{1}{2}}}{2}$, $\beta_O = 0.3$ and $\beta_I = 0.8$.

$$\implies MR(q) = \frac{\lambda\left[(1-\beta_O)^{\frac{1}{t}} - (1-\beta_I)^{\frac{1}{t}}\right]}{(1-\beta_O)^{\frac{1+t}{t}} - (1-\beta_I)^{\frac{1+t}{t}}} \equiv \lambda k,$$

where k is the relevant collection of constants. Note that this is an extremely simple threshold rule in terms of marginal revenue. Given that we have chosen a form of marginal revenue that admits an inverse, it is simple to solve out for the threshold rule in terms of quantities; this is why we needed marginal revenue to have an inverse solution.

$$p_0 + (1+t)p_{-t}q^t + (1+2t)p_{-2t}q^{2t} = \lambda k \implies$$

$$q = \left(\frac{-p_{-t}(1+t) \pm \sqrt{p_{-t}(1+t)^2 + 4\left(p_0 - k\lambda\right)p_{-2t}(1+2t)}}{2p_{-2t}(1+2t)}\right)^{\frac{1}{t}}.$$

Between these two roots, in-sourcing is optimal; outside them, outsourcing is optimal.[95]

This provides closed-form solutions as a function of λ, but λ remains to be determined. This is, unfortunately, where things start to get a bit messier. The integral determining q_λ can be explicitly taken, but only in terms of the less-standard Appell Hypergeometric function. The equation for $MR\left(q_\lambda\right) = \lambda$ therefore cannot be solved explicitly for λ. However, it is a single explicit equation. Once λ has been determined, optimal sourcing is determined in closed-form as described above. We plot this and the relaxed optimal β, in Figure 11, in the same format as in the paper for the case when $p_0 = .2, p_{-t} = 2, p_{-2t} = -4, mc_{-t} = .5, t = .5, \beta_I = .8, \beta_O = .3$. Clearly, we obtain similar, non-monotone results, but now these require only a single call of Newton's method to solve an otherwise explicit equation, as opposed to the two-dimension search we required to solve the case presented in the paper.

We do not discuss second-order conditions here, but they can easily be derived and checked to hold for this example as well as for the example in the paper. A grossly sufficient condition is that marginal revenue is declining over the solution range, as is the case in both of these examples.

[95]Actually if $\lambda k < p_0$ then the lower root should be interpreted as 0.

I.2 Labor bargaining without commitment (Stole and Zwiebel, 1996a,b)

Stole and Zwiebel (1996a,b, henceforth SZ) consider a model of labor market bargaining where contracts cannot commit workers. Each worker is, therefore, able to extract a share of the surplus the firm gains from a marginal worker. However, that surplus is determined by the profits the firm would earn if that worker were to leave, in which case the firm would bargain with other workers for a share of the remaining surplus. This causes (a) wages to depend on infra-marginal profits and (b) firms to over-employ workers relative to a standard labor market since having reserve workers decreases the marginal value of any given employee, lowering equilibrium wages and raising profits.

The setup of the SZ model is as follows. At the beginning of a period, a firm hires workers, each of whom supplies one unit of labor if employed.[96] When this process has been completed but before production takes place, the workers are free to bargain over their wages for this period. At that time the firm cannot hire any additional workers, so if any bargaining is not successful and any worker leaves the firm, fewer workers will be available for production in this period. Moreover, after the worker's departure, the remaining employees are free to renegotiate their wages, and in principle the process may continue until the firm loses all its employees. Assuming its revenues are concave in labor employed, this gives the firm an incentive to "over-employ" or *hoard* workers as hiring more workers makes holding a marginal worker less valuable to the firm and thus reduces workers' bargaining power.

If the bargaining weight of the worker relative to that of the firm's owner is λ, then the relationship surplus splitting condition is $S_w{=}\lambda S_f$. The worker's surplus is simply the equilibrium wage corresponding to the current employment level minus the outside option: $S_w = W(l) - W_0$, where W is the wage as a function of l, the labor supplied. For expositional simplicity, we assume the firm transforms labor into output one-for-one, though analytic solutions also exist for any power law production technology when $\lambda = 1$ and in other cases. Thus we assume $q = l$ and henceforth use q as our primary variable analysis for consistency with previous sections.

The firm faces inverse demand $P(q)$ and thus its profits are $\Pi(q) = [P(q) - W(q)]\, q$. The firm's surplus from hiring an additional worker is then $\Pi'(q)$. This gives the differential equation

$$W(q) - W_0 = \lambda MR(q) + \lambda \left(W(q)\, q\right)' \Rightarrow \lambda(W(q)\, q^{1+\frac{1}{\lambda}})' = q^{\frac{1}{\lambda}} \left(\lambda MR(q) + W_0\right), \qquad (15)$$

where $MR(q) \equiv P(q) + P'(q)q$ and the implication can be verified by simple algebra and is a standard transformation for an ordinary differential equation of this class. Integrating both of the sides of the equation, imposing the boundary condition that the wage bill shrinks to 0 at $q = 0$, and solving out yields wages and profits

$$W(q) = q^{-\left(1+\frac{1}{\lambda}\right)} \int_0^q x^{\frac{1}{\lambda}} MR(x)dx + \tfrac{W_0}{1+\lambda}, \quad \Pi(q) = P(q)q - q^{-\frac{1}{\lambda}} \int_0^q x^{\frac{1}{\lambda}} MR(x)dx - \tfrac{W_0}{1+\lambda}.$$

While the wage equation is intractable in general, the operation on the right-hand side does not change the functional form of any element of the average-marginal form-preserving class. To gain intuition for this, note that as $\lambda \to 0$, the model converges to the neoclassical model because the worker has no bargaining power; thus the equation becomes $MR(q) = W_0$. On the other hand as $\lambda \to \infty$, the equation converges to $P(q) = W_0$ as workers capture all revenue and divide it equally. Thus for intermediate λ the marginal-average transformation is effectively applied "partially" to $P(q)$. To see this mathematically, suppose $MR(q) = aq^{-b}$. Then the integral term in Equation 15 becomes

[96] The model is formally dynamic but is usually studied in its steady state as described here.

$$\frac{(1+\lambda)a\int_0^{q^\star} x^{\frac{1}{\lambda}}x^{-b}dx}{\lambda(q^\star)^{1+\frac{1}{\lambda}}} = \frac{(1+\lambda)a}{(q^\star)^{1+\frac{1}{\lambda}}}\ \frac{(q^\star)^{\frac{1+\lambda-b\lambda}{\lambda}}}{1+\lambda-b\lambda} = \frac{1+\lambda}{1+\lambda-b\lambda}\ a(q^\star)^{-b}.$$

More generally, for $MR(q)$ a linear combination of power terms, each term of becomes multiplied by $(1+\lambda)/(1+\lambda+t\lambda)$, where t is the power on the term. This tractability under form-preserving classes, but general intractability, has led researchers to study the SZ model almost exclusively under linear and constant elasticity demand.

While this class can yield important insights, it also has significant limitations. In particular, in the rest of this subsection we show that under this class the percentage over-employment relative to the neoclassical benchmark is constant as a function of the prevailing wage and multiplicative demand shifters. Thus proportional over-employment does not vary, for example, over the business cycle as consumers become richer and employment grows overall. By contrast in a calibrated model with equal bargaining weights ($\lambda = 1$), using demand derived from the US income distribution as in Section 2, Equation 1, we find that over a reasonable business cycle range over-employment should shift by roughly 0.4% of total employment. While quite small in absolute terms, this could account for a non-trivial fraction of cyclic variation in employment and is ruled out by the standard model. Furthermore, this model is quadratically tractable, nearly as tractable as the standard constant elasticity or linear specifications that are linearly tractable. It thus seems a natural alternative to make future analysis of labor bargaining more realistic without losing significant tractability.

To carry out this calculation, we note that the firm's optimal q solves its first-order condition, $\Pi'(q) = 0$, which, after some algebraic manipulations, is

$$\frac{(1+\lambda)\int_0^{q} x^{\frac{1}{\lambda}}MR(x)dx}{\lambda q^{1+\frac{1}{\lambda}}} = W_0. \tag{16}$$

Let us define (relative) labor hoarding as $h \equiv (q^\star - q^{\star\star})/q^{\star\star}$, where $q^\star$ is SZ employment and $q^{\star\star}$ is the employment level that a neoclassical firm with identical technology would choose: $MR(q^{\star\star}) = W_0$. Combining these definitions with Equation 16 gives a useful condition for h in terms of the equilibrium employment level $q^\star$:

$$MR\left(\frac{q^\star}{1+h}\right) = \frac{(1+\lambda)\int_0^{q^\star} x^{\frac{1}{\lambda}}MR(x)dx}{\lambda (q^\star)^{1+\frac{1}{\lambda}}}. \tag{17}$$

Note that this equation, and Equation 16, involves only (a) marginal revenue and (b) integrals of it multiplied by a power of q and then divided by a power of q higher by 1. It can easily be shown that the support of the Laplace marginal revenue is preserved by this transformation using essentially the same argument we used in the paper to show this support was shifted by exactly one unit when consumer surplus is calculated. This implies that Equations 16 and 17 have precisely the same tractability characterization as does the basic monopoly model we studied in Section 2 of the paper.[97]

Given the complexity of Equation 17 from any perspective other than our tractable forms, we investigate it using these forms, following Helpman et al. (2010) who study the model under constant elasticity demand. First consider the BP class, $P(q) = p_0 + p_t q^{-t}$, which nests the constant elasticity

[97] Note that Equation 16 also involves a constant and thus only our tractable forms with a constant term will maintain their tractability in this model. This is why we focus on this class below.

case when $p_0 = 0$. Solving Equation 17 for h yields

$$h = \left(\frac{1+\lambda}{1+\lambda-t\lambda}\right)^{\frac{1}{t}} - 1. \tag{18}$$

Therefore hoarding is constant in $q^\star$ and consequently in W_0. Thus under the BP class of demand, including constant elasticity, the economic cycle (the nominal outside option) has *no effect* on relative hoarding. It can easily be shown that h monotonically increases in t, so that the less concave demand (and thus profits) are, the more hoarding occurs. We found this counterintuitive, as we believed, building off the intuition supplied by SZ about the relationship between the "front-loading" that drives hoarding and concavity, that labor hoarding was driven by concavity in the firm's profit function.[98] Instead it appears that the reverse is the case. This shows one advantage of considering explicit functional forms: they help correct false intuitions. In particular, because t clearly parameterizes concavity the comparative static has a natural interpretation.

This new intuition suggests that the hoarding may not be constant over the economic cycle if, during that cycle, the curvature of firm profits change. For example, if during booms broad parts of the population are served and during recessions only wealthier individuals are served, then labor hoarding should be counter-cyclical as the distribution of income among the wealthy is more convex than among the middle-class and poor.

To analyze this we used our proposed functional form from Equation 1 in Section 2, in the version where it actually represents the income distribution as this is appropriately normalized to our assumption of $q = l$ (willingness-to-pay for a unit of labor):

$$P(q) = 50000\left(\frac{1}{2}q^{-\frac{2}{5}} + 2 - \frac{5}{2}q^{\frac{2}{5}}\right).$$

(in US dollars). Plugging this into Equation 17 and $MR(q^{\star\star}) = W_0$, assuming to match the convention in the literature that $\lambda = 1$ (though this plays no role in the simple form of our solution) and rounding to the second significant digit yields:

$$h = 1.6\left(\frac{1+\sqrt{1+\frac{1.2\cdot 10^9}{(100000-W_0)^2}}}{1+\sqrt{1+\frac{1.1\cdot 10^8}{(100000-W_0)^2}}}.\right)^{5/2} - 1, \tag{19}$$

We interpret a reduction in W_0, or equivalently a multiplicative scaling up of P, to be a boom (as it leads to higher production) and a rise in W_0 to be a recession. The expression on the right-hand side of Equation 19 can easily be shown to be increasing in W_0 (in fact, this is true quite broadly beyond this particular calibration). Thus a recession hoarding rises, contrasting with the standard intuition that unions exacerbate recessions by creating nominal wage rigidity and suggesting the effects of individual workers' bargaining may have qualitatively different comparative statics than collective bargaining does. Figure 12 shows the results quantitatively. Hoarding is large ($\approx 59\%$), but its comparative statics are less pronounced. It rises by a bit less than one percentage point when the outside option rises from \$30k to \$50k, a reasonable range of variation over the economic cycle. Thus, while the BP approximation of constancy appears not to be very far off these effects

[98] However, it is worth noting that another source of profit convexity, fixed costs, has an opposite effect and is a natural element to include in the model. This can be done in a straightforward way using our technology given our previous discussion, but we omit it here for brevity.

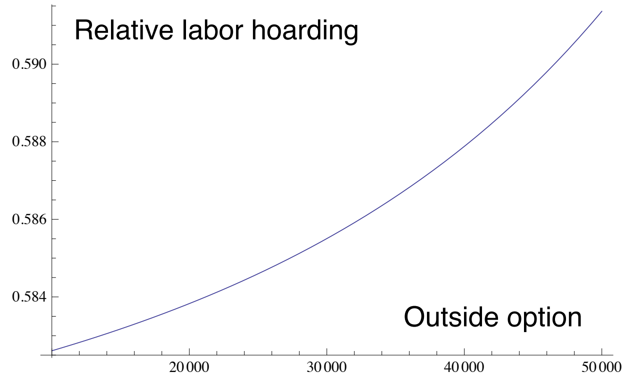


Figure 12: Relative labor hoarding in the Stole-Zwiebel model with $\lambda = 1$ and demand given by the approximation for $W_0 \in [10000{,}50000]$ (in USD).

may of a similar magnitude to cyclic shifts in employment and are thus worth considering.

I.3 Imperfectly competitive supply chains

Consider the model of imperfectly competitive supply chains where each stage of production strategically anticipates the reactions of the subsequent stage proposed by Salinger (1988). There are m stages of production interacting via linear pricing. Producers at each stage act simultaneously and the stages act in sequence. We solve by backwards induction.

Producers at stage m take an input from producers at stage $m-1$ and sell it to final consumers, facing inverse demand P_m. The n_m firms at stage m are symmetric Cournot competitors with average cost AC_m. The linear price clearing the market between stage $m-1$ and m is $\hat{P}_{m-1}$. Using the standard first-order condition for Cournot competition and dropping arguments, the first-order equilibrium conditions are

$$P_m + \frac{1}{n_m} P'_m q = \hat{P}_{m-1} + AC_m + \frac{1}{n_m} AC'_m q \iff$$

$$\hat{P}_{m-1} = P_m + \frac{1}{n_m} P'_m q - AC_m - \frac{1}{n_m} AC'_m q.$$

Thus the effective inverse demand facing the firms at stage $m-1$ is

$$P_{m-1} \equiv P_m + \frac{1}{n_m} P'_m q - AC_m - \frac{1}{n_m} AC'_m q,$$

as all output produced at stage $m-1$ is used as an input at stage m. Effectively the inverse demand at stage $m-1$ is the (competition-adjusted) marginal profit (competition-adjusted marginal revenue less marginal cost) at stage m.

This analysis may be back-propagated up the supply chain to obtain a first-order condition at the first stage determining the quantity in the industry. However, at each stage one higher derivative of P_m, at least and also of some of the cost curves, enters the first-order conditions. Thus the implicit equation for the first-order conditions characterizing the supply chain is usually quite elaborate and is both difficult to analyze in general and highly intractable, even computationally, for many functional forms. For example, Crawford et al. (2018) use this computational tractability concern to justify their focus on simultaneous decisions upstream and downstream in a related vertical contracting model.

However, we now derive a simple explicit transformation of the Laplace inverse demand and average cost characterizing the supply chain and discuss how this can be used to overcome these difficulties. Note that

$$P_m + \frac{1}{n_m} P'_m q = \left(1 - \frac{1}{n_m}\right) P_m + \frac{1}{n_m} MR_m,$$

where $MR_m = P_m + P'_m q$. Let p_m be the Laplace inverse demand. From Section 6 we have that the Laplace marginal revenue is $(1-t)p_m$ and thus that the inverse Laplace-log transform of $\left(1 - \frac{1}{n_m}\right) P_m + \frac{1}{n_m} MR_m$ is just $\left(1 - \frac{t}{n_m}\right) p_m$. By the same logic, if we denote the Laplace average cost by ac_m the inverse Laplace-log transform of $AC_m + \frac{1}{n_m} AC'_m q$ is $\left(1 - \frac{t}{n_m}\right) ac_m$.

Iterating this process, one obtains that the Laplace first-order condition at the initial stage, which we denote f_1, is

$$p_m \prod_{i=1}^{m} \left(1 - \frac{t}{n_i}\right) - \sum_{i=1}^{m} \left[ac_i \prod_{j=1}^{i} \left(1 - \frac{t}{n_j}\right) \right].$$

This obviously differs only in its (trivially computed) coefficients and not in its support from the ac_i's and p_m that make it up. Thus if all ac_i's and p_m are chosen to have the same tractable support (with the desired number of evenly spaced mass points to achieve desired tractability) then the full will be equally tractable. Beyond this, even if p_m and the ac_i's are specified in an arbitrary manner, the resulting Laplace first-order condition can be trivially computed from the inverse Laplace-log transforms of each of these inputs and then either solved directly by applying the Laplace transform or approximated using a small number of evenly spaced mass points for tractability. In either case, this approach significantly reduces the complexity of computing and representing the system.

I.4 Two-sided platforms à la Rochet and Tirole (2003)

Rochet and Tirole (2003) propose a model of a two-sided platform motivated by the credit card industry. Sellers and buyers are randomly matched and independently decide whether they want to accept credit cards and whether they want to use them conditional on cards being accepted. These decisions are driven by the price charged (or subsidy paid) to each side. In particular, in order for a fraction of sellers q_S to wish to accept cards, the price that must be charged to sellers is $P_S(q_S)$, and similarly for buyers.

Let $U_I(q_I) \equiv \int_0^{q_I} P_I(x)dx$ be the gross utility on side I. Because $U'_I(q_I) = P_I(q_I)$, the average gross utility $\overline{U}_I(q_I) \equiv {}^{U_I(q_I)}/_{q_I}$ has the average-marginal relationship to inverse demand. Thus average consumer surplus $\overline{V}_I(q_I) = \overline{U}_I(q_I) - P_I(q_I)$ has the same functional form as $P'_I q_I$ for a form-preserving functional form class.

Rochet and Tirole show that, when there is a constant and symmetric marginal cost of clearing transactions c, imperfectly competitive equilibrium between symmetric firms is characterized by

$$P_S(q_S) + P_B(q_B) - c = -\theta P'_S(q_S)\, q_S = -\theta P'_B(q_B)\, q_B$$

for some constant $\theta < 1$.[99] On the other hand, they show that Ramsey pricing (which nests the

[99] Weyl (2008) extends this characterization to the case of complements when $\theta > 1$. For analogous reasons to the previous applications all results here may be extended to arbitrary imperfectly competitive supply chains.

unconstrained social planner's problem as a special case) is characterized by

$$P_S(q_S) + P_B(q_B) - c = -\theta \overline{V}_S(q_S) = -\theta \overline{V}_B(q_B)$$

for some constant θ, equal to unity in the case of the unconstrained social optimum and approaching 0 as Ramsey pricing is required to break even. Thus if inverse demand on both sides of the market is specified within the same form-preserving class (Rochet and Tirole assume linear demand in their example) then our characterization of tractability applies here as well.

Again the added flexibility of our forms is important in this context. For example, Weyl (2009) considers how platforms would choose an "interchange fee" between two sides of the market, holding fixed the overall level of prices. He demonstrates that if both sides have BP demand, then users on both sides of the market and profit maximization *all* agree on the same optimal interchange fee. However, this is generally false and thus assuming BP demand trivializes the wide-ranging regulatory debate over interchange fees. In fact under plausible (bell-shaped) demand forms, perhaps surprisingly, both sides in aggregate prefer to face higher prices (consumers prefer lower interchange, merchants prefer higher) to subsidize use on the other side of the market. From a social perspective, the more heterogeneous side and/or the side which has more complete adoption should be taxed to subsidize the other side more than will be in the interest of a profit-maximizing platform, even for fixed aggregate prices.

I.5 Auction theory

I.5.1 Symmetric independent private values first-price auctions

Consider N symmetric bidders with privately-known values v_i for a single object drawn independently and identically from a distribution with differentiable CDF F. Let $V(q) \equiv F^{-1}(q)$ be the quantile function of F. Let $b_\star$ be a symmetric-equilibrium bid function mapping values to bids in a first-price auction in which the highest bidder wins and pays her bid value; any such equilibrium bid function can be shown to be strictly monotone increasing under weak conditions. The probability that the bid of any individual bidder is below x is then $G_\star(x) \equiv F\left(b_\star^{-1}(x)\right)$. Thus the probability that bidder i wins if she submits a bid of x is, by symmetry, $[G_\star(x)]^{N-1}$.

The expected utility a bidder with value v thus earns from a bid of x is

$$(v - x)\left[G_\star(x)\right]^{N-1} = (v - B_\star(q))\, q^{N-1},$$

where q is the fraction of other bidders with (weakly) lower bids and $B_\star(q) \equiv G_\star^{-1}(q)$ is the quantile function of the equilibrium bid distribution. A necessary condition for her optimization is therefore

$$(v - B_\star(q))\,(N-1)q^{N-2} + B_\star'(q)q^{N-1} = 0 \iff v = B_\star(q) + \frac{1}{N-1} q B_\star'(q).$$

For this to be a symmetric, monotone equilibrium for the posited bid distribution, it must be that a bidder with value at reversed quantile q of the value distribution chooses to bid (weakly) higher than precisely a fraction q of her rivals. Thus a necessary condition for a symmetric equilibrium is

$$V(q) = B_\star(q) + \tfrac{1}{N-1} q B_\star'(q).$$

Sufficient conditions, which we omit here, are well-known in the literature. Note that the right-hand side of this expression involves the marginal and average forms of $B_\star$. Thus, by simple coefficient matching, if V is chosen to be from a form-preserving class then there is always an equilibrium $B_\star$

from the same class. This may be used directly to analytically relate the values and bids at various quantiles, which is all that is necessary for many analytic problems.

However if one wishes to obtain a closed form for $b_\star$ itself, then one must choose the class to be tractable at the level of complexity of the desired closed form and include a constant (a power of 0) in the class. By definition, $G_\star = F \circ b_\star^{-1}$, so $b_\star^{-1} = V \circ G_\star$ and consequently $b_\star = B_\star \circ F$. Thus if F and V have forms tractable at level k, then so does $b_\star$. Evidently uniform and exponential distributions, which have linear and logarithmic V respectively, are linearly tractable, explaining why they are ubiquitously used for examples in symmetric first-price auction models.

However these forms are quite restrictive in that they cannot, for example, have the bell shape usually found in empirical studies of valuation distributions in auctions (Haile and Tamer, 2003; Cassola et al., 2013). Our forms can easily generate such shapes and thus allow tractable examples with realistic value distributions.

I.5.2 Auctions v. posted prices (Einav, Farronato, Levin and Sundaresan, 2018)

Einav et al. (2018) consider the trade-off a seller faces between using an auction and setting a posted price in an online retail market. They assume sellers of goods know the common (positive) hassle cost λ for buyers to participate in an auction, but may still use an auction because they do not know their common value v of the good. The seller has an opportunity cost of selling c, and v is drawn from a distribution F that the seller knows. Assuming, as the authors do, that at least two bidders participate, the auction guarantees that the seller gets value $v - \lambda$ as long as $v - \lambda \geq r$, where r is the reserve price the seller sets. Alternatively, the seller may set a posted price p, in which case she will sell the good if $v \geq p$.

Let $P(q) \equiv F^{-1}(1-q)$. If a seller sells the good with probability q, then in an auction with the reserve price set to $P(q) - \lambda$ she will receive an average price $\overline{U}(q) - \lambda$, where $\overline{U}(q) \equiv {\int_0^q P(x)dx}/{q}$ by the same logic as in Supplementary Material I.4. If the seller sells the good with probability q with a posted price by setting price $P(q)$, she will receive price $P(q)$ with certainty. Thus the region in which she wishes to use an auction rather than a posted price is when she wishes to sell with probability q such that $\overline{U}(q) > P(q) + \lambda$. As noted in Supplementary Material I.4, $\overline{U}$ has the average-marginal relationship to P. For this reason, if P is specified according to an average-marginal form-preserving class including a constant term (power 0 term), then the resulting optimal cut-off rules for using a definite mechanism are tractable at the level of tractability of the class (in terms of both the cost and the desired probability of sale, which is more directly observed in the authors' data).

Einav et al. present such an example, by assuming a uniform distribution and thus a linear form for P. In this case $\overline{U} - P$ uniformly grows in q. This implies that sellers with a low cost (low opportunity cost of sale), such as impatient private individuals clearing old property out of the house, who wish to achieve sale with high probability (quickly) will use auctions. On the other hand, those who have a high cost, such as professional vendors, who want to achieve a sale with low probability (slowly) to wait will set a high posted price. However this is not generally true. If P takes a constant elasticity form, for example, the reverse pattern holds: low-cost sellers set a low posted price and sell quickly while high cost (patient) sellers run an auction.

For the bell-shaped demands that appear to fit Einav et al.'s data best, the gap between $\overline{U}$ and P is actually non-monotone, first declining an then rising. This suggests auctions should be polarized into goods that sell with very low and very high probability; that is among those clearing out their houses and among the most professional sellers. This is in fact what the authors find; they cannot even measure the posted-price demand curve at very low sale probabilities as they do not observe sufficiently many items selling that infrequently with posted prices, while the same is true at very high probabilities. This suggests richer classes of tractable, form-preserving demand may be more useful in modeling this trade-off than is the uniform distribution.

I.6 Selection markets

Akerlof (1970) analyzed markets where the cost of providing a service differs by the identity of the consumer to whom it is provided. He studies a case that he labels "adverse selection" in which consumers differ in only a single characteristic and in which raising this one dimension increases both consumers' willingness-to-pay for the product and the cost of serving them. Einav et al. (2010) and Einav and Finkelstein (2011) maintain Akerlof's assumption of a single product but allow consumers to differ along multiple dimensions that may impact their willingness to pay and cost in potentially rich ways.

Einav et al. (2010) define an inverse demand curve $P(q)$ for $q \in (0,1)$ as the willingness to pay of the individual in the $(1-q)$th quantile of the willingness-to-pay distribution. They define average cost $AC(q)$ as the average cost of individuals who are in the quantiles above $1-q$ of the willingness-to-pay distribution. They argue that perfectly competitive equilibrium requires $AC(q) = P(q)$ while social optimization requires $MC(q) = P(q)$, where MC has the average-marginal relationship to AC. Mahoney and Weyl (2017) extend this framework to nest a variety of models of imperfect competition using a conduct parameter θ as in Subsection D.2 above and show that equilibrium is characterized by $\theta MC(q) + (1-\theta)AC(q) = (1-\theta)P(q) + \theta MR(q)$.

As is clear by now, both sides of this equation are tractable for any value of θ at whatever level the cost and demand side are specified if these are chosen to be part of a form-preserving class. Many analyses have assumed linear forms on both the cost and demand side (Cutler and Reber, 1998; Einav et al., 2010; Einav and Finkelstein, 2011), partly for tractability. As Scheuer and Smetters (2014) highlight, this assumption rules out many interesting phenomena, such as selection that is "advantageous" (higher willingness to pay correlating with lower cost) over some range but adverse over other ranges or multiple local competitive equilibria that Scheuer and Smetters argue may have challenged the introduction of the Affordable Care and Patient Protection Act in the United States. Broader tractable form-preserving classes, especially those with bell-shaped demand and cost curves, allow these possibilities and appear to fit existing empirical evidence more closely.

I.7 Monopolistic competition

I.7.1 Tractable generalizations of the Dixit-Stiglitz framework with separable utility

In the simplest monopolistic competition model, consumers derive their utility from a continuum of varieties $\omega \in \Omega$ of a single heterogeneous good in a separable way: $U_\Omega = \int_\Omega u_\omega(q_\omega)\, d\omega$. In the original Dixit-Stiglitz model specialized to the case of constant elasticity of substitution σ, $u_\omega(q_\omega)$ is a power of the consumed quantities q_ω: $u_\omega(q_\omega) \propto q_\omega^{1-1/\sigma}$. In our generalization we wish to be able to apply Theorem 2, so we let $u(q_\omega)$ be a linear combination of different powers of q_ω. More explicitly, consumer optimization requires that marginal utility of extra spending is equalized across varieties: $u'_\omega(q_\omega) = \lambda P_\omega$, where P_ω is the price of variety ω and λ is a Lagrange multiplier related to consumers' wealth. To ensure tractability, we let the residual inverse demand $P_\omega(q_\omega) = u'_\omega(q_\omega)/\lambda$ and the corresponding revenue $R_\omega(q_\omega)$ be linear combinations of equally-spaced powers of q_ω: $P_\omega(q_\omega) = \sum_{t\in T} p_{\omega,t} q_\omega^{-t}, \quad R_\omega(q_\omega) = \sum_{t\in T} p_{\omega,t} q_\omega^{1-t}$ for some finite and evenly-spaced set T, with the number of elements of T determining the precise degree of tractability.

For convenience of notation, we choose a numéraire in a way that keeps $P_\omega(q_\omega)$ for a given q_ω independent of macroeconomic circumstances.

Each variety of the differentiated good is produced by a single firm. We assume that the marginal cost and average cost of production can be written as $MC_\omega(q) = \sum_{t\in T} mc_{\omega,t} q_\omega^{-t}$, $AC_\omega(q) = \sum_{t\in T\cup\{1\}} ac_{\omega,t} q_\omega^{-t}$, where $mc_{\omega,t} = (1-t)\, ac_{\omega,t}$. A constant component of average cost (and marginal cost) would correspond to $ac_{\omega,0}$ and a fixed cost would correspond to $ac_{\omega,1}$. However, given the generality possible here we do not necessarily have to assume that these components are present in all models under consideration.

With this specification, Theorem 2 applies and the firm's problem has closed-form solutions unless T has six elements or more. Moreover, if firms are heterogeneous in their productivity, then 3 leads to closed-form aggregation integrals for suitable choices of the productivity distribution, as in the case of a generalized Melitz model discussed below.

I.7.2 Tractable generalizations of the D-S framework with non-separable utility

Here we briefly discuss tractable monopolistic competition in the case of non-separable utility.[100] The utility has the very general form

$$U_\Omega \equiv F\left(U_\Omega^{(1)}, U_\Omega^{(2)}, ..., U_\Omega^{(m)}\right), \quad U_\Omega^{(i)} \equiv \int_\Omega U^{(i,\omega)}\left(q_\omega\right) d\omega.$$

In order to preserve tractability, we assume that $U^{(i,\omega)}(q_\omega)$ are linear combinations[101] of equally-spaced powers of q_ω and that the set of exponents does not depend on i or ω. For example, we could specify $U_\Omega \equiv U_\Omega^{(1)} + \kappa_1 (U_\Omega^{(1)})^{\xi_1} + \kappa_2 (U_\Omega^{(2)})^{\xi_2}$, $U_\Omega^{(1)} \equiv \int_\Omega q_\omega^{\gamma_1} d\omega$, and $U_\Omega^{(2)} \equiv \int_\Omega q_\omega^{\gamma_2} d\omega$, with $(\gamma_1+1)/(\gamma_2+1)$ equal to the ratio of two small integers. In the language of heterogeneous-firm models, the choice $\kappa_1 = \kappa_2 = 0$ corresponds to the Melitz model, while the choice $\xi_1 = 2$, $\xi_2 = 1$, $\gamma_1 = 1$, and $\gamma_2 = 2$ gives the Melitz and Ottaviano model, which is based on a non-homothetic quadratic utility. Our general specification allows also for homothetic non-separable utility functions that feature market toughness effects analogous to those in the Melitz and Ottaviano model.

It is straightforward to verify that as in the separable-utility case, Theorems 2 and 3 still apply and lead to closed-form solutions to the firm's problem and closed-form aggregation. This is because the structure of the firm's problem is unchanged. Non-separability only makes the resulting system of equations for macroeconomic aggregates more complex. The system itself may still be written in closed form due to Theorem 3, under appropriate assumptions on the productivity distribution.

I.7.3 Tractable generalizations of the Dixit-Stiglitz framework

In the baseline monopolistic competition model consumers derive their utility from a continuum of varieties $\omega \in \Omega$ of a single heterogeneous good:

$$U_\Omega = \int_\Omega u_\omega\left(q_\omega\right) d\omega. \tag{20}$$

[100] In the case of heterogeneous firms, this generalization contains as special cases both the Melitz model and the Melitz and Ottaviano model. To be more precise, let us note that in addition to the heterogeneous-good varieties explicitly considered here, the Melitz and Ottaviano model includes a homogeneous good. In our discussion, the homogeneous good is absent, but adding it to the model is straightforward.

[101] Of course, without loss of generality we could assume that $U^{(i,\omega)}(q_\omega)$ are power functions and let the function F combine them into any desired linear combinations. However, for clarity of notation it is preferable to keep the number m of different expressions $U_\Omega^{(i)}$ small.

In the original Dixit-Stiglitz model with constant elasticity of substitution σ, $u_\omega(q_\omega)$ is a power of the consumed quantities q_ω: $u_\omega(q_\omega) \propto q_\omega^{1-1/\sigma}$. In our generalization $u(q_\omega)$ is assumed to be a function of a combination of different powers of q_ω . More explicitly, consumer optimization requires that marginal utility of extra spending is equalized across varieties: $u'_\omega(q_\omega) = \lambda P_\omega$, where P_ω is the price of variety ω and λ is a Lagrange multiplier related to consumers' wealth. To ensure tractability, we let the residual inverse demand $P_\omega(q_\omega) = {u'_\omega(q_\omega)}/{\lambda}$ and the corresponding revenue $R_\omega(q_\omega)$ be linear combinations of equally-spaced powers of q_ω:

$$P_\omega(q_\omega) = \sum_{t\in T} p_{\omega,t} q_\omega^{-t}, \quad R_\omega(q_\omega) = \sum_{t\in T} p_{\omega,t} q_\omega^{1-t}$$

for some finite and evenly-spaced set T, with the number of elements of T determining the precise degree of tractability. For the convenience of notation, we choose a numéraire in a way that keeps $P_\omega(q_\omega)$ for a given q_ω independent of macroeconomic circumstances.

Each variety of the differentiated good is produced by a single firm. We assume that the marginal cost and average cost of production can be written as

$$MC_\omega(q) = \sum_{t\in T} mc_{\omega,t} q_\omega^{-t}, \quad AC_\omega(q) = \sum_{t\in T\cup\{1\}} ac_{\omega,t} q_\omega^{-t},$$

where $mc_{\omega,t} = (1-t)\, ac_{\omega,t}$. A constant component of average cost (and marginal cost) would correspond to $ac_{\omega,0}$ and a fixed cost would correspond to $ac_{\omega,1}$. However given the generality possible here we do not necessarily have to assume that these components are present in all models under consideration.

I.7.4 Flexible Krugman model

The Krugman (1980) model of trade, featuring monopolistic competition and free entry of identical single-product firms, may be solved explicitly for the tractable demand and cost functions mentioned above, not just constant-elasticity demand and constant marginal cost specified in the original paper. Here we consider these solutions in the case of two symmetric countries, which leads to a symmetric equilibrium.

There is a continuum of identical consumers with preferences as in Equation 20 who earn labor income. The amount of labor a firm needs to hire in order to produce quantity q may be split into a fixed part f and a variable part $L(q)$ that vanishes at zero quantity. Both $L(q)$ and the revenue function $R(q)$ are assumed to allow for a linear term. The firm only uses labor for production, so its total cost is $w(L(q)+f)$, where w is the competitive wage rate. Having produced quantity q, the firm splits it into q_d to be sold domestically, and τq_x to be shipped abroad. Due to iceberg-type trade costs ($\tau \geq 1$), a fixed fraction of the shipped good is lost during transport, and only quantity q_x is received in the other country. (Non-iceberg trade costs are considered in the appendix.) Let us denote the equilibrium level of marginal cost, measure of firms, international trade flows, and welfare by $MC^\star$, $N^\star$, $X^\star$, and $W^\star$, respectively, and similarly for other variables. The total labor endowment of one of the two symmetric economies is L_E.

Observation 1. *There exists an explicit map $MC^\star \to (f, q_d^\star, q_x^\star, w^\star)$ and an explicit map $(MC^\star, L_E) \to (N^\star, X^\star, W^\star)$. These relationships represent a closed-form solution to the model in terms of $MC^\star$ and exogenous parameters.*

To see briefly why this is the case, it is convenient to express the model's equations in terms of the equilibrium level of marginal cost $MC^{\star}$.[102] Output optimally designated for the domestic market and the export market will satisfy $R'(q_d) = MC^{\star}$ and $R'(q_x) = \tau\, MC^{\star}$, respectively, and therefore may be solved for in closed form in terms of $MC^{\star}$ for tractable specifications of the revenue function (or consumer preferences).[103] The same is true for wages since $w = MC^{\star}/L'(q_d + \tau q_x)$.

For a chosen $MC^{\star}$ we may compute the level of fixed cost f consistent with it using the free-entry condition: $R(q_d) + R(q_x) = wL(q_d + \tau q_x) + wf$. The equilibrium number (measure) $N^{\star}$ of firms in each economy then satisfies $N^{\star} = L_E/(L\,(q_d + q_x) + f)$, where L_E is the labor labor endowment one of the two economies.[104] Other variables of interest, e.g. trade flows or welfare, are then simply functions of the ones discussed above.

Krugman model with non-iceberg and iceberg international trade costs. Although the Krugman model with non-iceberg trade costs is not our main focus here, we mention it for completeness. Let us assume the presence of non-iceberg international trade costs that require hiring labor $L_T\,(q_x)$ in order for q_x to reach its destination in the other country.[105] The export FOC is now $R'(q_x) - wL_T'(q_x) = \tau MC^{\star}$, while the free entry condition becomes $R(q_d) + R(q_x) = wL(q_d + \tau q_x) + wL_T(q_x) + wf$. The resulting number (measure) of firms is $N^{\star} = L_E/((L\,(q_d + q_x) + f) + L_T(q_x))$. The model may be solved explicitly along the same lines in terms of chosen $MC^{\star}$ and w, with f and τ treated as derived quantities.

I.7.5 Flexible Melitz model

The Melitz (2003) model is again based on monopolistic competition and assumes a constant elasticity of substitution between heterogeneous-good varieties. Relative to the Krugman (1980) model, it introduced a novel channel for welfare gains from trade, namely increased average firm productivity resulting from trade liberalization or analogous decreases in trade costs. Here we generalize the model to allow for more flexible demand functions, non-constant marginal costs of production, and trade costs that may have components that are neither iceberg-type nor constant per unit.

Single country. For clarity of exposition, we first describe the flexible and tractable version of the Melitz model in the case of a single country and later discuss its generalization. Just like the Krugman model, it involves two types of agents: monopolistic single-product firms and identical consumers, who supply their labor in a competitive labor market and consume the firms' products.[106]

[102] The case of a single country corresponds to the Dixit-Stiglitz model. It may be obtained from our two-country discussion by setting $\tau \to \infty$ and $q_x = 0$. In this case one does not have to express the model's equations in terms of the equilibrium level of marginal cost $MC^{\star}$ as we do below. Instead, for tractable functions $R\,(q)$ and $L\,(q)$ one can solve for equilibrium quantity $q^{\star}$ in closed form (in terms of the fixed cost of production f) from an equation that combines profit maximization and free entry: $(L(q) + f)R'(q) = R(q)L'(q)$.

[103] As mentioned in the paper, a convenient choice of numéraire allows us to keep the revenue function $R\,(.)$ independent of economic circumstances.

[104] L_E may be exogenous, as in the original Krugman model, but even for endogenous labor supply, it is possible to obtain fully explicit solutions to the model in terms of the parameter $MC^{\star}$.

[105] In a symmetric equilibrium it does not matter how this labor is split between the countries, as long as the symmetry of the model is maintained. For asymmetric countries, we could assume that the transport requires labor from both countries. The model may be solved in terms of marginal costs of serving each market.

[106] For simplicity, consumers do not discount the future, although it would be easy to incorporate an explicit discount factor. Formally, the model includes an infinite number of periods, but it may be thought of as a static model because the equilibrium is independent of time.

Labor is the only factor of production: all costs have the interpretation of labor costs and are proportional to a competitive wage rate w. Each heterogeneous-good variety is produced by a unique single-product firm, which uses its monopolistic market power to set marginal revenue equal to marginal cost. Demand and costs are specified tractably as discussed above; this time we do not need to assume that variable cost and revenue functions allow for a linear term.

If the firm is not able to make positive profits, it is free to exit the industry. In situations of main interest, this endogenous channel of exit is active: there exist firms that are indifferent between production and exit. There is also an exogenous channel of exit: in every period with probability δ_e the firm is forced to permanently shut down.

Entry into the industry is unrestricted but comes at a fixed one-time cost wf_e. Only after paying this fixed cost, the entering firm observes a characteristic a, drawn from a distribution with cumulative distribution function $G(a)$, that influences the firm's cost function. In the original Melitz model the constant marginal cost of production is equal to wa. Here we leave the specification more general while maintaining the convention that increasing a increases the firm's cost at any positive quantity q. In expectation, the stream of the firm's profits must exactly compensate the (risk-neutral) owner for the entry cost, which implies the *unrestricted entry condition* $wf_e = \mathbb{E}\Pi(q;a)/\delta_e$, with the profit $\Pi(q;a)$ evaluated at the optimal quantity.[107]

The amount of labor needed to produce quantity q is $L(q;a) + f$, where $L(q;a)$ corresponds to variable cost ($L(0;a) = 0$) and f to a fixed cost. $L(q;a)$ is assumed to be tractable with respect to q, but also with respect to a.[108] In terms of the labor requirement function $L(q;a)$, the firm profit maximization condition and the *zero cutoff profit condition* are $R'(q) = wL'(q;a)$ and $R(q_c) = wL(q_c;a_c) + wf$, where q_c and a_c correspond to a *cutoff firm*, i.e. a firm that is in equilibrium indifferent between exiting and staying in the industry. We denote by L_E, $M^\star$,$M_e^\star$, $W^\star$ the labor endowment, and the equilibrium measure of firms, measure of entering firms, and level of welfare, respectively.

The firm profit maximization condition and the free entry condition are

$$R'(q) = wL'(q;a), \tag{21}$$

$$R(q_c) = wL(q_c;a_c) + wf. \tag{22}$$

A convenient solution strategy is to choose q_c and then calculate f_e as a derived quantity. For a chosen q_c we can find a_c explicitly by combining (21) and (22) into $R'(q_c)(L(q_c;a_c)+f)= R(q_c)L'(q_c;a_c)$, since $L(q;a)$ is assumed to be tractable also with respect to a. Wages are then given recovered from (22): $w = R(q_c)/(L(q_c;a_c)+f)$.

Now we need to show how to calculate the fixed cost of entry f_e and the measure of firms. The fixed cost of entry consistent with the chosen cutoff quantity is given simply by the unrestricted entry condition:

$$w\delta_e f_e = \bar{\Pi} = \int\limits_{q \geq q_c} (R(q) - wL(q;a) - wf)\, dG(a(q)).$$

Here $a(q)$ is the firm's productivity parameter as an explicit function of the optimally chosen quantity q that results from using (21). For Pareto G, and L and R tractable from the point of view of q

[107]In the case of a single country, the profit is simply $\Pi(q;a) = q[P(q;a) - AC(q;a)]$. Also, note that the unrestricted entry condition is often referred to as the *free entry condition*, but here we avoid this term since there is a positive entry cost.

[108]For example, the function $L(q;a)$ could be linear in a, as would be the case in the original Melitz model. A simple example of a tractable choice of functional forms is $L(q) = \tilde{L}(q) + a\hat{L}(q)$, $\hat{L}(q) \equiv q^t$, $\tilde{L}(q) \equiv \tilde{\ell}_t q^{-t} + \tilde{\ell}_u q^{-u}$, and $R(q) = r_t q^{-t} + r_u q^{-u}$.

(but not necessarily having a linear term) and $L(q;a)$ linear in a, there exist closed-form expressions for this integral in terms of special functions, which are straightforward to derive, especially if one uses symbolic manipulation software such as Mathematica. If the shape parameter of the Pareto distribution is a negative integer, the integrals actually reduce to simple power functions.

If M_e denotes the measure of firms that enter each period (in one country), then the measure of operating firms is $M = G(a_c) M_e/\delta_e$. The total labor used in the economy is given by $L_E = M_e f_e + Mf + M\bar{L}$, where $\bar{L} = G(a_c)^{-1} \int_{q \geq q_c} L(q;a)\, dG(a(q))$ is the labor on average hired for the variable cost of production. Under the same assumptions, the integral again has an explicit form in terms of special functions. We see that in these cases we can get fully explicit expressions for f_e and M in terms of chosen q_c and L_E.

Other quantities of interest, such as trade flows or welfare, may be found in an analogous, straightforward fashion.

Two countries with non-iceberg and iceberg international trade costs. Just like in the case of the flexible Krugman model, it is convenient to write the model in terms of equilibrium marginal cost, which this time is firm-specific and also depends on the firm's chosen export status. For tractability we will need the revenue function $R(q)$ and the production labor requirement function $L(q;a)$ to allow for a linear term. The same is true for labor corresponding to the non-iceberg trade costs, here denoted by $L_T(q_x)$. As in the original Melitz (2003) paper, we consider equilibria characterized by two cutoffs, here denoted a_1 and a_2, such that least productive firms with $a > a_1$ exit, more productive firms with $a \in (a_2, a_1]$ serve only their domestic market, and most productive firms with $a \leq a_2$ serve both countries. In general, we denote the equilibrium marginal cost of a non-exporting firm as $MC_n^\star$ and that of an exporting firm as $MC_x^\star$. Variables corresponding to the two cutoffs are distinguished by subscripts 1 and 2, so for example $MC_{1n}^\star$ is the optimal marginal cost of a firm with $a = a_1$, and $MC_{2x}^\star$ and $MC_{2n}^\star$ are optimal marginal costs of a firm with $a = a_2$ that decides to export or not to export, respectively. We denote by $M_x^\star$ and $X^\star$ the equilibrium measure of exporting firms and international trade flows.

Our solution strategy is to treat MC_{1n} and MC_{2x} as given and to express other variables of the model in terms to these two chosen parameters. In particular, we will show how to derive explicit expressions for the fixed cost of exporting f_x and cost of entry f_e. The (variable-cost) labor requirement $L(q;a)$ is assumed to be a tractable combination of equidistant powers of a, with coefficients that in general depend on q. Firms' profit maximization leads to the set of equations:

$$MC_n = R'(q_n) \tag{23a}$$
$$MC_n = wL'(q_n;a) \tag{23b}$$
$$MC_x = R'(q_d) \tag{23c}$$
$$MC_x = \tfrac{1}{\tau} R'(q_f) - \tfrac{1}{\tau} wL_T'(q_f) \tag{23d}$$
$$MC_x = wL'(q_d + \tau q_f; a) \tag{23e}$$
$$R(q_{1n}) - wL(q_{1n};a_1) = f \tag{23f}$$
$$R(q_{2d}) + R(q_{2f}) - wL(q_{2d} + \tau q_{2f}; a_2) - wL_T(q_{2f}) = f + f_x \tag{23g}$$

Here q_n is the quantity sold by a non-exporting firm, while q_d and q_f represent quantities that reach domestic and foreign customers of an exporting firm, respectively. In addition to exporting cost $wL_T(q_f)$, we allow for an iceberg trade cost factor $\tau \geq 1$.

For a chosen MC_{1n}, we can calculate q_{1n} from (23a). The corresponding a_1 may be found by solving a linear equation that results from combining (23b) and (23f) in a way that eliminates wages.

Wages then may be recovered by substituting back to (23b).

For a chosen MC_{2x}, we can derive q_{2d} from (23c) and q_{2f} from (23d). The value of a_2 is then determined by (23e). We find q_{2n} by solving (23a) and (23b) with MC_{2n} eliminated, and then in turn use one of these to find MC_{2n}. This means that we know the marginal cost at the cutoffs. The fixed cost of exporting f_e is then identified from (23g).

For a given marginal cost, we can find the corresponding quantities and productivity parameters a by a similar method from (23a-23e), this time treating w as known. We denote the resulting functions $q_n(MC_n)$, $q_d(MC_x)$, $q_x(MC_x)$, $a_n(MC_n)$, and $a_x(MC_x)$. Using these functions we can now determine the entry labor requirement f_e from the unrestricted entry condition:

$$w\delta_e f_e = \bar{\Pi} = \int\limits_{y\in S_n} \Pi\left(q_n(y); a_n(y)\right) dG\left(a_n(y)\right) + \int\limits_{y\in S_x} \Pi\left(q_x(y); a_x(y)\right) dG\left(a_x(y)\right),$$

where Π is the profit function (revenue minus cost), $G(a)$ is the cumulative distribution function of a, and the integration ranges are $S_n \equiv (MC_{2n}, MC_{1n})$ and $S_x \equiv (0, MC_{1n})$. Under various assumptions these integrals may be evaluated in closed form, often involving special functions. If a measure M_e of firms enters each period (in one of the countries), then the equilibrium measure of operating firms is $M = M_e G(a_1)/\delta_e$ and that of exporting firms is $M_x = M_e G(a_2)/\delta_e$. These measures may be calculated from the labor market clearing condition $M_e f_e + M f + M_x f_x + (M - M_x)\bar{L}_n + M_x \bar{L}_x = L_E$, where

$$\bar{L}_n \equiv \tfrac{1}{G(a_1)-G(a_2)} \int\limits_{y\in S_n} L\left(q_n(y); a_n(y)\right) dG\left(a_n(y)\right),\ \bar{L}_x \equiv \tfrac{1}{G(a_2)} \int\limits_{y\in S_x} L\left(q_x(y); a_x(y)\right) dG\left(a_x(y)\right).$$

Under the same assumptions as before, these integrals may be evaluated in closed form. Again, other variables of interest, such as trade flows or welfare, may be obtained in a similar way.

I.7.6 Flexible Melitz/Melitz-Ottaviano model with non-separable utility

While a significant part of the international trade literature relies on separable utility functions, there exist realistic economic phenomena what are more easily modeled with non-separable utility. An instantly classic alternative to the Melitz model that uses non-separable utility is the model of Melitz and Ottaviano, which assumes that with a greater selection of heterogeneous-good varieties available to consumers, the marginal gain from an additional variety decreases relative to the gains from increased quantity. Trade liberalization leads to tougher competition, which results not only in higher productivity but also in the decrease of markups charged by a given firm.

Here we briefly discuss a generalization of the flexible Melitz model where the utility function is allowed to be non-separable. This generalized model contains as special cases both the Melitz model and the Melitz and Ottaviano model.[109] The utility is of the form

$$U_\Omega \equiv F\left(U_\Omega^{(1)}, U_\Omega^{(2)}, ..., U_\Omega^{(m)}\right), \quad U_\Omega^{(i)} \equiv \int_\Omega U^{(i,\omega)}\left(q_\omega\right) d\omega.$$

[109]In addition to the heterogeneous-good varieties explicitly considered here, the Melitz and Ottaviano model includes a homogeneous good. In our discussion, the homogeneous good is absent but adding it to the model is straightforward.

In order to preserve tractability, we assume that $U^{(i,\omega)}(q_\omega)$ are linear combinations[110] of equally-spaced powers of q_ω and that the set of exponents does not depend on i or ω. For example, we could specify $U_\Omega \equiv U_\Omega^{(1)} + \kappa_1 (U_\Omega^{(1)})^{\xi_1} + \kappa_2 (U_\Omega^{(2)})^{\xi_2}$, $U_\Omega^{(1)} \equiv \int_\Omega q_\omega^{\gamma_1}\, d\omega$, and $U_\Omega^{(2)} \equiv \int_\Omega q_\omega^{\gamma_2}\, d\omega$, with $(\gamma_1+1)/(\gamma_2+1)$ equal to the ratio of two small integers. The choice $\kappa_1 = \kappa_2 = 0$ corresponds to the Melitz model, while the choice $\xi_1 = 2$, $\xi_2 = 1$, $\gamma_1 = 1$, and $\gamma_2 = 2$ gives the Melitz and Ottaviano model, which is based on a non-homothetic quadratic utility. Our general specification allows also for homothetic non-separable utility functions that feature market toughness effects analogous to those in the Melitz and Ottaviano model.

It is straightforward to verify that just like the flexible Melitz model with separable utility, this more general version leads to tractable optimization by individual firms, as well as for tractable aggregation under the same conditions. The reason for the tractability of the firm's problem is simple: the firm's first-order condition will have the same structure as previously, a linear combination of equidistant powers (with an additional dependence of the coefficients of the linear combination on *aggregate* variables of the type $\int_\Omega q_\omega^\gamma\, d\omega$ for some constants γ). Given that the nature of the firm's problem is unchanged, it follows that being able to explicitly aggregate over heterogeneous firms does not require any additional functional form assumptions relative to the separable utility case.

J Demand Forms

J.1 Curvature properties

Table 4 provides a taxonomy of the curvature properties of demand functions generated by common statistical distributions and the single-product version of the Almost Ideal Demand System. Following Caplin and Nalebuff (1991b,a), we define the curvature of demand as

$$\kappa(p) \equiv \frac{Q''(p)Q(p)}{[Q'(p)]^2}.$$

Cournot (1838) showed that the pass-through rate of a constant marginal cost monopolist is

$$\frac{1}{2-\kappa}$$

and thus that a) that the comparison of κ to unity determines the comparison of pass-through to unity in this case and b) that if $\kappa'(p) > 0$ that pass-through rises with price (falls with quantity), and conversely if κ declines with price (rises with quantity). The comparison of κ to unity also determines whether a demand is log-convex and its sign whether demand is convex. The comparison of κ to 2 determines whether demand has declining marginal revenue, a condition also known as Myerson (1981)'s regularity condition.

For probability distribution F, the corresponding demand function $Q(p) = s\left(1 - F\left(\frac{p-\mu}{m}\right)\right)$ where s and m are stretch parameters (Weyl and Tirole, 2012) and μ is a position parameter. Note that in this case

$$\kappa(p) = -\frac{\frac{s^2}{m^2} F''\left(\frac{p-\mu}{m}\right)\left(1 - F\left(\frac{p-\mu}{m}\right)\right)}{\frac{s^2}{m^2}\left[F'\left(\frac{p-\mu}{m}\right)\right]^2} = -\frac{F''\left(\frac{p-\mu}{m}\right)\left(1 - F\left(\frac{p-\mu}{m}\right)\right)}{\left[F'\left(\frac{p-\mu}{m}\right)\right]^2}.$$

[110] Of course, without loss of generality we could assume that $U^{(i,\omega)}(q_\omega)$ are power functions and let the function F combine them into any desired linear combinations. However, for clarity of notation it is preferable to keep the number m of different expressions $U_\Omega^{(i)}$ small.

	$\kappa < 1$	$\kappa > 1$	Price-dependent	Parameter-dependent
$\kappa' < 0$			AIDS with $b < 0$	
$\kappa' > 0$	Normal (Gaussian) Logistic Type I Extreme Value (Gumbel) Laplace Type III Extreme Value (Reverse Weibull) Weibull with shape $\alpha > 1$ Gamma with shape $\alpha > 1$		Type II Extreme Value (Fréchet) with shape $\alpha > 1$	
Price-dependent				
Parameter-dependent				
Does not globally satisfy $\kappa < 2$		Type II Extreme Value (Fréchet) with shape $\alpha < 1$ Weibull with shape $\alpha < 1$ Gamma with shape $\alpha < 1$		

Table 4: A taxonomy of some common demand functions

Note, thus, that neither global level nor slope properties of κ are affected by s, m or μ. We can thus analyze the properties of relevant distributions independently of their values, as represented in the table and the following proposition.

The most prominent conclusion emerging from this taxonomy is that the vast majority of forms used in practice in computational, statistical models such as Berry et al. (1995) have monotonically increasing curvature and most have curvature below unity. This suggests two conclusions. The first, highlighted in the paper, is that, to the extent we believe these forms are more realistic than tractable forms, they have properties systematically differing from the BP class and thus it is important to derive tractable forms capable of matching their central property of monotonically increasing in price/decreasing in quantity curvature.

A second possible conclusion is that, to the extent that in some cases these properties are *not* empirically relevant, such as in the data of Einav et al. (2015), standard forms rule out observed behavior and thus analysts may wish to consider more flexible forms along these dimensions, such as those we derive in the paper. To the extent there are not strong theoretical reasons to believe in the restrictions imposed by standard statistically based forms (which, in many cases, there are) allowing such relaxation is important because in many contexts the properties of firm demand and equilibrium are inherited directly from the demand function, at least with constant marginal cost (Weyl and Fabinger, 2013; Gabaix et al., 2016; Quint, 2014). Which conclusion is most appropriate will obviously depend on the empirical context and the views of the analyst.

Proposition 1. *Table 4 summarizes global properties of the listed statistical distributions generating demand functions. α is the standard shape parameter in distributions that call for it.*

Proof. Characterization of the curvature level (comparisons of κ to unity) follow from classic classifications of distributions as log-concave or log-convex as in Bagnoli and Bergstrom (2005), except in the case of AIDS in which the results are novel.[111] Note that our discussion of stretch parameters in the paper implies we can ignore the scale parameter of distributions, normalizing this to 1 for any distributions which has one. A similar argument applies to position parameter: because this

[111]We do not classify the slope of pass-through for demand functions violating declining marginal revenue as this is such a common assumption that we think such forms would be unlikely to be widely used and because it is hard to classify the slope of pass-through when it is infinite over some ranges.

only shifts the values where properties apply by a constant, it cannot affect global curvature or higher-order properties. This is useful because many of the probability distributions we consider below have scale and position parameters that this fact allows us to neglect. We will denote this normalization by *Up to Scale and Position* (USP).

We begin by considering the first part of the proof, that for any shape parameter $\alpha < 1$ the Fréchet, Weibull and Gamma distributions with shape α violate DMR at some price. We show this for each distribution in turn:

1. Type II Extreme Value (Fréchet) distribution: USP, this distribution is $F(x) = e^{-x^{-\alpha}}$ with domain $x > 0$. Simple algebra shows that

$$\kappa(x) = \frac{(e^{x^{-\alpha}} - 1)x^{\alpha}(1+\alpha) + \left(1 - e^{x^{-\alpha}}\right)\alpha}{\alpha}.$$

   As $x \to \infty$ and therefore $x^{-\alpha} \to 0$ (as shape is always positive), $e^{x^{-\alpha}}$ is well-approximated by its first-order approximation about 0, $1 + x^{-\alpha}$. Therefore the limit of the above expression is the same as that of

$$\frac{x^{-\alpha}x^{\alpha}(1+\alpha) - x^{-\alpha}\alpha}{\alpha} = \frac{1+\alpha+x^{-\alpha}\alpha}{\alpha} \to \frac{1}{\alpha} + 1$$

   as $x \to \infty$. Clearly, this is greater than 2 for $0 < \alpha < 1$ so that for sufficiently large x, $\kappa > 2$.

2. Weibull distribution: USP, this distribution is $F(x) = 1 - e^{-x^{\alpha}}$. Again algebra yields:

$$\kappa(x) = \frac{1-\alpha}{\alpha x^{\alpha}} + 1.$$

   Clearly, for any $\alpha < 1$ as $x \to 0$ this expression goes to infinity, so that for sufficiently small x, $\kappa > 2$.

3. Gamma distribution: USP, this distribution is $F(x) = \frac{\gamma(\alpha,x)}{\Gamma(\alpha)}$ where $\gamma(\cdot,\cdot)$ is the lower incomplete Gamma function, $\Gamma(\cdot,\cdot)$ is the upper incomplete Gamma function and $\Gamma(\cdot)$ is the (complete) Gamma function:

$$\kappa(x) = \frac{e^{x}(1-\alpha+x)\Gamma(\alpha,x)}{x^{\alpha}}. \tag{24}$$

   By definition, $\lim_{x\to 0}\Gamma(\alpha,x) = \Gamma(\alpha) > 0$ so

$$\lim_{x\to 0}\kappa(x) = +\infty$$

   as $1-\alpha > 0$ for $\alpha < 1$. Thus clearly for small enough x, the Gamma distribution with shape $\alpha < 1$ has $\kappa > 2$.

We now turn to the categorization of demand functions as having increasing or decreasing pass-through. As price always increases in cost, this can be viewed as either pass-through as a function of price or pass-through as a function of cost.

1. Normal (Gaussian) distribution: USP, this distribution is given by $F(x) = \Phi(x)$, where Φ is the cumulative normal distribution function; we let ϕ denote the corresponding density. It is

well-known that $\Phi''(x) = -x\phi(x)$. Thus

$$\kappa(x) = \frac{x\left[1 - \Phi(x)\right]}{\phi(x)}.$$

Taking the derivative and simplifying yields

$$\kappa'(x) = \frac{\left[1 - \Phi(x)\right](1 + x^2) - x\phi(x)}{\phi(x)},$$

which clearly has the same sign as its numerator, as ϕ is a density and thus everywhere positive. But a classic strict lower bound for $\Phi(x)$ is $\frac{x}{1+x^2}\phi(x)$, implying $\kappa' > 0$.

2. Logistic distribution: USP, this distribution is $F(x) = \frac{e^x}{1+e^x}$. Again algebra yields

$$\kappa'(x) = e^{-x} > 0.$$

Thus the logistic distribution has $\kappa' > 0$.

3. Type I Extreme Value (Gumbel) distribution: USP, this distribution has two forms. For the minimum version it is $F(x) = 1 - e^{-e^x}$. Algebra shows that for this distribution

$$\kappa'(x) = e^{-x}.$$

Note that this is the same as for the logistic distribution; in fact κ for the Gumbel minimum distribution is identical to the logistic distribution. This is not surprising given the close connection between these distributions (McFadden, 1974).

For the maximum version it is $F(x) = e^{-e^{-x}}$. Again algebra yields

$$\kappa'(x) = e^{-x}\big(e^{2x}[e^{e^{-x}} - 1] - e^{e^{-x}}[e^x - 1]\big).$$

For $x < 0$ this is clearly positive as both terms are strictly positive: $1 > e^x$ and because $e^{-x} > 0$, $e^{e^{-x}} > 1$. For $x > 0$ we can rewrite κ' as

$$e^{e^{-x}}\left(e^x - 1\right) + e^{-x}\left(e^{e^{-x}} - 1\right),$$

which again is positive as $e^x > 1$ for $x > 0$ and $e^{e^{-x}} > 1$ by our argument above.

4. Laplace distribution: USP, this distribution is

$$F(x) = \begin{cases} 1 - \frac{e^{-x}}{2} & x \geq 0, \\ \frac{e^x}{2} & x < 0. \end{cases}$$

For $x > 0$, $\rho = 1$ (so in this range pass-through is not strictly increasing). For $x < 0$

$$\kappa'(x) = 2e^{-x} > 0.$$

So the Laplace distribution exhibits globally weakly increasing pass-through, strictly increasing for prices below the mode. The curvature for this distribution is $1 - 2e^{-x}$ as opposed to $1-e^{-x}$ for Gumbel and Logistic. However, these are very similar, again pointing out the similarities among curvature properties of common demand forms.

5. Type II Extreme Value (Fréchet) distribution with shape $\alpha > 1$: From the formula above it is easy to show that the derivative of the pass-through rate is

$$\kappa'(x) = x^{-(1+\alpha)}\left([1+\alpha]\left[x^{2\alpha}(e^{x^{-\alpha}}-1) - e^{x^{-\alpha}}x^{\alpha}\right] + \alpha e^{x^{-\alpha}}\right) > 0,$$

which can easily be shown to be positive as follows. Let us multiply the inequality by the positive factor $\frac{e^{-x^{-\alpha}}}{\alpha+1}$. Denoting $X \equiv x^{-\alpha}$, the inequality becomes

$$\left(\frac{\alpha}{\alpha+1} - \frac{1}{2}\right) + \left(\frac{1}{X^2} - \frac{e^{-X}}{X^2} - \frac{1}{X} + \frac{1}{2}\right) > 0.$$

The first term is positive because $\alpha > 1$. The second term is positive because $e^{-X} < 1 - X + \frac{1}{2}X^2$ for any $X > 0$. Thus this distribution, as well, has $\kappa' > 0$.

6. Type III Extreme Value (Reverse Weibull) distribution: USP, this distribution is $F(x) = e^{-(-x)^{\alpha}}$ and has support $x < 0$. Algebra shows

$$\kappa'(x) = (-x)^{\alpha-1}\alpha^2\left[1-\alpha+e^{(-x)^{\alpha}}\left([1-\alpha]\left[(-x)^{\alpha}-1\right] + [-x]^{2\alpha}\alpha\right)\right],$$

which has the same sign as

$$1-\alpha+e^{(-x)^{\alpha}}\left([1-\alpha]\left[(-x)^{\alpha}-1\right] + [-x]^{2\alpha}\alpha\right). \tag{25}$$

Note that the limit of this expression as $x \to 0$ is

$$1-\alpha-(1-\alpha) = 0$$

and its derivative is

$$\frac{e^{(-x)^{\alpha}}(-x)^{2\alpha}\alpha\left(1+\alpha+[-x]^{\alpha}\alpha\right)}{x},$$

which is clearly strictly negative for $x < 0$. Thus Expression 25 is strictly decreasing and approaches 0 as x approaches 0. It is therefore positive for all negative x, showing that again in this case $\kappa' > 0$.

7. Weibull distribution with shape $\alpha > 1$: As with the Fréchet distribution algebra from the earlier formula shows

$$\kappa'(x) = x^{\alpha-1}(\alpha-1)\alpha^2,$$

which is clearly positive for $\alpha > 1$ as the range of this distribution is positive x. Thus the Weibull distribution with $\alpha > 1$ has $\kappa' > 0$.

8. Gamma distribution with shape $\alpha > 1$: Taking the derivative of Expression 24 yields:

$$\kappa'(x) = \frac{\alpha-1-x+\frac{e^x}{x^{\alpha}}\left(x^2-2x[\alpha-1]+[\alpha-1]\alpha\right)\Gamma(\alpha,x)}{x},$$

which has the same sign as

$$\alpha - 1 - x + \frac{e^x}{x^\alpha}\left(x^2 - 2x[\alpha - 1] + [\alpha - 1]\alpha\right)\Gamma(\alpha, x), \tag{26}$$

given that $x > 0$. Note that as long as $\alpha > 1$

$$x^2 + (\alpha - 2x)(\alpha - 1) = x^2 - 2(\alpha - 1)x + \alpha(\alpha - 1) > x^2 - 2(\alpha - 1)x + (\alpha - 1)^2 = (x + 1 - \alpha)^2 > 0.$$

Therefore so long as $x \leq \alpha - 1$ this is clearly positive. On the other hand when $x > \alpha - 1$ the proof depends on the following result of Natalini and Palumbo (2000):

Theorem (Natalini and Palumbo, 2000). *Let a be a positive parameter, and let q(x) be a function, differentiable on* $(0, \infty)$ *, such that* $lim_{x \to \infty} x^\alpha e^{-x} q(x, \alpha) = 0$. *Let*

$$T(x, \alpha) = 1 + (\alpha - x)q(x, \alpha) + x\frac{\partial q}{\partial x}(x, \alpha).$$

If $T(x, \alpha) > 0$ *for all* $x > 0$ *then* $\Gamma(\alpha, x) > x^\alpha e^{-x} q(x, \alpha)$.

Letting

$$q(x, \alpha) \equiv \frac{x - (\alpha - 1)}{x^2 + (\alpha - 2x)(\alpha - 1)},$$

$$T(x, \alpha) = \frac{2(\alpha - 1)x}{\left(\alpha^2 + x[2 + x] - \alpha[1 + 2x]\right)^2} > 0$$

for $\alpha > 1, x > 0$. So $\Gamma(\alpha, x) > x^\alpha e^{-x} q(x, \alpha)$. Thus Expression 26 is strictly greater than

$$\alpha - 1 - x + x - (\alpha - 1) = 0$$

as, again, $x^2 + (\alpha - 2x)(\alpha - 1) > 0$. Thus again $\kappa' > 0$.

This establishes the second part of the proposition. Turning to our final two claims, algebra shows that the curvature for the Fréchet distribution is

$$\kappa(x) = \frac{\alpha - e^{x^{-\alpha}}\left(\alpha - x^\alpha[1 + \alpha]\right) - x^\alpha(1 + \alpha)}{\alpha} = \frac{\left(1 - e^{x^{-\alpha}}\right)\left[\alpha - x^\alpha\left(1 + \alpha\right)\right]}{\alpha}.$$

Note for any $\alpha > 1$ this is clearly continuous in $x > 0$. Now consider the first version of the expression. Clearly, as $x \to 0$, $x^\alpha \to 0$ and $e^{x^{-\alpha}} \to \infty$ so the expression goes to $-\infty$. So for sufficiently small $x > 0$, $\kappa(x) < 1$. On the other hand, consider the second version of the expression. Its numerator is

$$\left(1 - e^{x^{-\alpha}}\right)\left[\alpha - x^\alpha\left(1 + \alpha\right)\right].$$

By the same argument as above with the Fréchet distribution the limit of the above expression as $x \to \infty$ is the same as that of

$$\left(-x^{-\alpha}\right)\left(-x^\alpha\left(1 + \alpha\right)\right)$$

as $x \to \infty$. Thus

$$\lim_{x\to\infty} \kappa(x) = \frac{1+\alpha}{\alpha} > 1$$

and thus for sufficiently large x and any $\alpha > 1$, this distribution has $\kappa > 1$.

Finally, consider our claim about AIDS. First note that for this demand function

$$\kappa(p) = 2 + \frac{b\big(a - 2b + b\log p\big)}{\big(a - b + b\log p\big)^2} < 1$$

as $b < 0$ and $p \le e^{-\frac{a}{b}} < e^{2-\frac{a}{b}}$. This is less than 1 if and only if

$$a^2 + 2ab\big(\log p - 2\big) + b^2\Big(1 + \big[\log(p) - 2\big]\log p\Big) < b^2\big(2 - \log p\big) - ab$$

or

$$\big(a + b\log p\big)^2 - b^2\big(\log p + 1\big) < 0.$$

Clearly, as $p \to 0$ the second term is positive; therefore there is always a price at which $\kappa(p) > 1$. On the other hand as $p \to e^{-\frac{a}{b}}$ this expression goes to

$$0 - b^2\left(1 - \frac{a}{b}\right) = b(a - b) < 0.$$

Thus there is always a price at which $\kappa(p) < 1$.

$$\kappa'(p) = b^2 - \big(a - 2b + b\log p\big)^2,$$

which has the same sign as

$$b^2 - \big(a - 2b + b\log p\big)^2 < b^2 - (2b)^2 = -3b^2 < 0.$$

Thus $\kappa' < 0$.

□

We now turn to two important distributions, which are typically used to model the income distribution, whose behavior is more complex and which, to our knowledge, have not been analyzed for their curvature properties. We focus only on the two that we believe to be most common (the first), best theoretically founded (both) and to provide the most accurate match to the income distribution (the second). Namely, we analyze the lognormal and double Pareto-lognormal (dPln) distributions, the latter of which was proposed by Reed (2003) and Reed and Jorgensen (2004). Other common, accurate models of income distributions which we have analyzed in less detail, appear to behave in a similar fashion.

We begin with the lognormal distribution, which is much more commonly used, and for which we have detailed, analytic results. However, while most of the arguments for the below proposition are proven analytically, some simple points are made by computational inspection.

Proposition 2. *For every value σ, there exist finite thresholds $\overline{y}(\sigma) > \underline{y}(\sigma)$ such that*

1. *If $y \ge \overline{y}(\sigma)$ then $\kappa' \le 0$, and similarly with strict inequalities or if the directions of the inequalities both reverse.*

2. *If $y \ge \underline{y}(\sigma)$ then $\kappa \ge 1$, and similarly with strict inequalities or if the directions of the inequalities both reverse.*

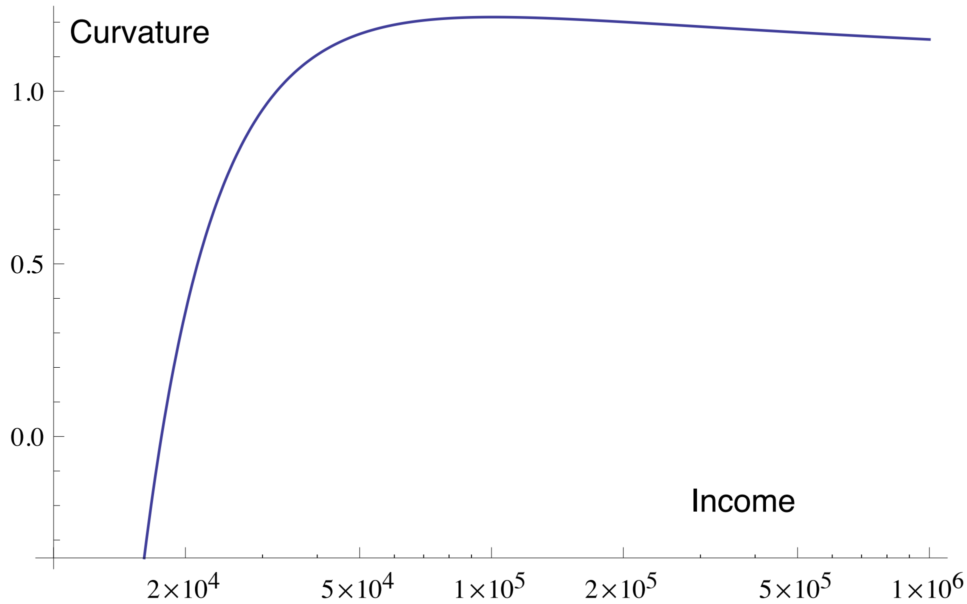


Figure 13: Curvature of a lognormal distribution calibrated to the US income distribution: parameters are $\mu = 10.5$ and $\sigma = 0.85$.

Both $\overline{y}$ and $\underline{y}$ are strictly decreasing in σ.

Under the lognormal distribution, the behavior depends critically on the amount of inequality or equivalently the standard deviation of the logarithm of the distribution: there is famously a one-to-one relationship between the Gini coefficient associated with a lognormal distribution and its logarithmic standard deviation. If inequality is not high, the behavior of curvature like a normal distribution occurs except at fairly high incomes levels; for a Gini of .34, for example, monotonicity of κ is preserved until the top 1% of the income distribution and log-concavity outside of the top 30%. However, if inequality is sufficiently high, in particular if the Gini coefficient is above about .72, then the lognormal distribution has $\kappa > 2$ over some range and then κ converges back to 1 for very large incomes. This result is not discussed in the proposition but can easily be seen by inspecting a graph of the expression for κ given in the proof of the proposition for various values of σ yielding Gini coefficients of various magnitudes around .72.

For intermediate levels of inequality between these, like that seen in nearly every country, the lognormal distribution has curvature that rises from $-\infty$ to above unity before gradually returning towards unity. For an example calibrated to the US income distribution (Figure 13), the crossing to above unity occurs at an income of about \$33k, between the mode and the median and the downward slope begins at about \$100k. Despite this, curvature never falls below unity again and in fact is at each quantile increasing in σ (again, not discussed in the proposition). Again taking the example of the US income-calibrated distribution, curvature peaks at about 1.21 and only falls to 1.20 by \$200k, eventually leveling out to about 1.1 for the extremely wealthy.[112] Thus, in practice, curvature is closer to flat at the top than significantly declining.

[112]Note, however, that in the true limit as $y \to \infty$, $\kappa \to 1$. However, in practice this occurs at such high income levels that the asymptote to a bit above 1 is a more realistic representation.

Proof. For a lognormal distribution with parameters (μ, σ), $F(x) = \Phi\left(\frac{\log(x)-\mu}{\sigma}\right)$, so that

$$Q(p) = 1 - \Phi\left(\frac{\log(p) - \mu}{\sigma}\right), Q'(p) = -\frac{\phi\left(\frac{\log(p)-\mu}{\sigma}\right)}{\sigma p}$$

and

$$Q''(p) = -\frac{\phi'\left(\frac{\log(p)-\mu}{\sigma}\right)}{\sigma^2 p^2} + \frac{\phi\left(\frac{\log(p)-\mu}{\sigma}\right)}{\sigma p^2} = -\frac{\phi\left(\frac{\log(x)-\mu}{\sigma}\right)}{\sigma^2 p^2}\left(\sigma + \frac{\log(x) - \mu}{\sigma}\right).$$

where the second equality follows from the identities regarding the normal distribution from the previous proof and $y \equiv \frac{\log(p)-\mu}{\sigma}$. Thus

$$\kappa\left(p(y)\right) = \frac{(y+\sigma)\left[1 - \Phi(y)\right]}{\phi(y)}. \tag{27}$$

Note that we immediately see, as discussed above, that κ increases in σ at each quantile as the inverse hazard rate $\frac{1-\Phi}{\phi} > 0$; similarly, for any quantile associated with y, $\kappa \to \infty$ as $\sigma \to \infty$ so it must be that the set of y for which $\kappa > 1$ a) exists for sufficiently large σ and b) expands monotonically in σ. This implies that, if point 2) of the proposition is true, $\underline{y}$ must strictly decrease in σ. This also implies that for sufficiently large σ, $\kappa > 2$ for some y.

Now note that $\lim_{y\to\infty} \frac{y[1-\Phi(y)]}{\phi(y)} = 1$. To see this, note that both the numerator and denominator converge to 0 as $1 - \Phi$ dies super-exponentially in y. Applying l'Hospital's rule:

$$\lim_{y\to\infty} \frac{y\left[1 - \Phi(y)\right]}{\phi(y)} = \lim_{y\to\infty} \frac{1 - \Phi(y) - \phi(y)y}{\phi'(y)} = \frac{y\phi(y) - \left[1 - \Phi(y)\right]}{y\phi(y)} = \frac{0}{0}.$$

where the first equality follows from the identity for ϕ' we have repeatedly been using, and from here on we no longer note the use of. Again applying l'Hospital's rule:

$$\lim_{y\to\infty} \frac{y\left[1 - \Phi(y)\right]}{\phi(y)} = \lim_{y\to\infty} \frac{\phi(y) + y\phi'(y) + \phi(y)}{\phi(y) + y\phi'(y)} = \lim_{y\to\infty} \frac{2\phi(y) - y^2\phi(y)}{\phi(y) - y^2\phi(y)} = \lim_{y\to\infty} \frac{2 - y^2}{1 - y^2} = 1.$$

The same argument, but one step less deep, shows that $\lim_{y\to\infty} \frac{\sigma[1-\Phi(y)]}{\phi(y)} = 0$. Together these imply that $\lim_{y\to\infty} \kappa\left(p(y)\right) = 1$ and thus that, if $\kappa > 1$ at some point, it must eventually decrease to reach 1.

Similar methods may be used to show, as discussed in the paper, that $\kappa \to -\infty$ as $y \to -\infty$. Furthermore, we know from the proof for the normal distribution above that $\frac{y[1-\Phi(y)]}{\phi(y)}$ is monotone increasing and that $\frac{\sigma[1-\Phi(y)]}{\phi(y)}$ is monotone decreasing. The latter point implies that the set of y for which κ is decreasing must be strictly increasing in σ and thus that, if point 1) of the proposition is true, then $\overline{y}$ must strictly decrease in σ.

All that remains to be shown is that κ's comparison to unity and the sign of κ' obey the threshold structure posited. Note that we only need to show the cut-off structure for κ' and that this immediately implies the structure for κ, given the smoothness of all functions involved, because if κ increases up to some threshold and then decreases monotonically while reaching an asymptote of unity, it must lie above unity above some threshold. Otherwise, if it ever crossed below unity, it would have to be increasing in some region to asymptote to unity at very large p, violating the threshold structure for κ'. Furthermore, the same logic implies that the region where $\kappa > 1$ must be

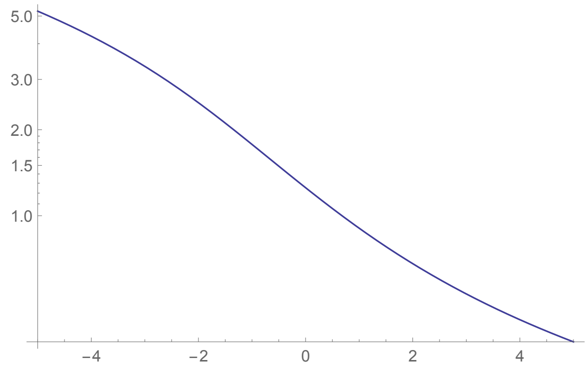


Figure 14: The figure shows the value, in logarithmic scale, of the left-hand side of Inequality 28.

strictly larger than the region where $\kappa' < 0$ (that $\overline{y} > \underline{y}$) as κ must rise strictly above unity before sloping strictly down towards it.

We drop arguments wherever possible in what follows to ease readability. We use the symbol $\propto$ to denote expressions having the same sign, not proportionality as is typical.

$$\kappa' = \frac{(1-\Phi)\,\phi - (y+\sigma)\phi^2 - (y+\sigma)(1-\Phi)\phi'}{\phi^2} = \frac{1-\Phi-(y+\sigma)\,[\phi - y(1-\Phi)]}{\phi} \propto$$

$$1-\Phi-(y+\sigma)\,[\phi - y(1-\Phi)] \propto \frac{1-\Phi}{\phi - y(1-\Phi)} - y - \sigma.$$

where the last sign relationship follows by the common inequality that $\phi(y) > y\,[1-\Phi(y)]$. Thus $\kappa' > 0$ if and only if

$$\frac{1-\Phi}{\phi - y(1-\Phi)} - y > \sigma. \tag{28}$$

Figure 14 shows that the left-hand side of this inequality is strictly decreasing. We have not found a simple means to prove this formally, but it is clearly true by inspection of the figure. Thus the left-hand side of Inequality 28 must cross σ at most once, and this must be from above to below.

It only remains to show that this expression does, in fact, make such as single crossing for all values of σ. It suffices to show that the small y limit of the left-hand side of inequality 28 is ∞ and that its large y limit is 0. We show these in turn.

The first claim is easy: clearly $-y\,(1-\Phi) \to \infty$, while $1-\Phi$ is finite, as $y \to -\infty$. Thus the first term approaches 0 and the second ∞ as $y \to -\infty$.

The second claim is more delicate. The expression is the same as

$$\frac{(1-\Phi)\,(1+y^2) - y\phi}{\phi - y\,(1-\Phi)}.$$

This asymptotes to the indefinite expression $\frac{0}{0}$ as $y \to \infty$ as it is well-known that $\lim_{y\to\infty} \frac{\phi}{y(1-\Phi)} = 1$. Applying l'Hospital's rule yields

$$\lim_{y\to\infty} \frac{(1-\Phi)\,(1+y^2) - y\phi}{\phi - y\,(1-\Phi)} = \lim_{y\to\infty} \frac{-\phi\,(1+y^2) + 2y\,(1-\Phi) - \phi - y\phi'}{\phi' - (1-\Phi) + y\phi} =$$

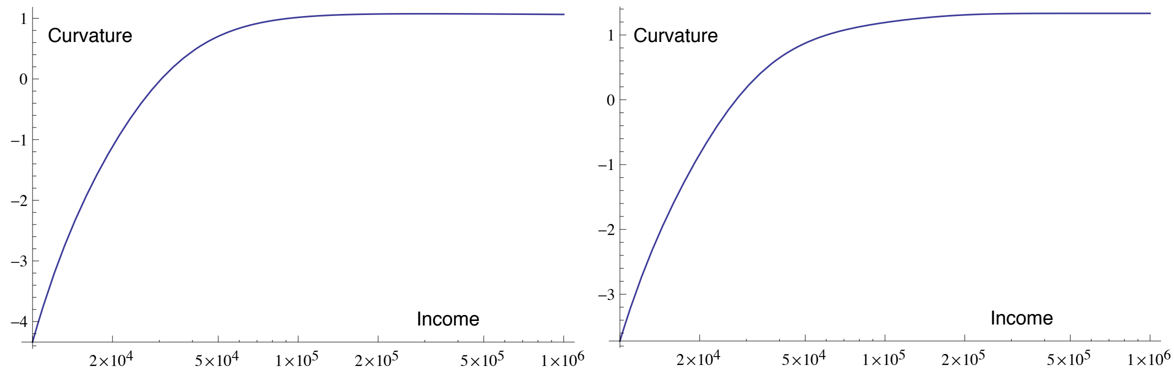


Figure 15: Curvature of the double-Pareto lognormal distribution lognormal under parameters estimated by (Reed, 2003) (left) and by updated by us (right); parameters in the former case are $\alpha = 22.43, \beta = 1.43, \mu = 10.9, \sigma = 0.45$ ad in the latter case are $\alpha = 3, \beta = 1.43, \mu = 10.9, \sigma = 0.5$. The x-axis has a logarithmic scale in income.

(applying now-familiar tricks)

$$\lim_{y\to\infty} 2\frac{\phi - y(1-\Phi)}{1-\Phi} = \frac{0}{0}.$$

Again, we apply l'Hospital's rule:

$$\lim_{y\to\infty} 2\frac{\phi - y(1-\Phi)}{1-\Phi} = \lim_{y\to\infty} 2\frac{\phi' - (1-\Phi) + y\phi}{\phi} = \lim_{y\to\infty} -\frac{1-\Phi}{\phi} = 0.$$

□

Even the slight decline in the lognormal distribution's curvature at very high incomes is an artifact of its poor fit to incomes distributions at very high incomes. It is well-known that at very high incomes the lognormal distribution fits poorly; much better fit is achieved by distributions with fatter (Pareto) tails, especially in countries with high top-income shares like the contemporary United States (Atkinson et al., 2011). A much better fit is achieved by the dPln distribution (Reed, 2003). Figure 15's left panel shows curvature as a function of income for the parameters Reed estimates (for the 1997 US income distribution). Curvature monotonically increases up the income distribution.

However, it levels off at quite moderate income (it is essentially flat beyond \$100k) and at a lower level (≈ 1.04) than under the lognormal calibration, except at exorbitant incomes, where the lognormal distribution has thin tails. Thus it actually has a *thinner* tail, except at the very extreme tail, than the lognormal calibration, paradoxically. This is because Reed calibrated only to the mid-section of the US income distribution, given that the survey he used is notoriously thin and inaccurate at higher incomes; this led him to estimate a very high (thin-tailed) Pareto coefficient in the upper tail of 22.43. Consensus economic estimates, for example Diamond and Saez (2011), suggest that 1.5-3 is the correct range for the Pareto coefficient of the upper tail of the income distribution in the 2000's.

We therefore construct our own calibration consistent with that finding. To be conservative we set the upper tail Pareto coefficient to 3, maintain $\beta = 1.43$ to be consistent with Reed and because the lower-tail is both well-measured in his data and has not changed dramatically in the last decade and a half (Saez, 2013). We then adjust μ and σ in the unique way, given these coefficients, to match the latest US post-tax Gini estimates (.42), using a formula derived by Hajargasht and

Griffiths (2013), and average income ($53k). This yields the plot in the right panel of Figure 15. There curvature continues to monotonically increase at a significant rate up to quite high incomes: at $50k it is .87, at $100k it is 1.19 and by $200k it has leveled off at 1.31, near its asymptotic value of $1 + \frac{1}{\alpha} = \frac{4}{3}$. It is this last calibration that we use to represent the dPln calibration US income distribution in the paper.

Moreover, the monotone increasing nature of curvature is not only true in the US data. While we have not been able to prove any general results about this four-parameter class, we have calculated similar plots to Figure 15 for every country for which a dPln income distribution has been estimated, as collected by Hajargasht and Griffiths. In every case curvature is monotone increasing in income, though in some cases it levels off at a quite low level of income (typically when the Gini is high relative to the upper tail estimate). Even this leveling off seems to us likely to be a bit of an artifact, arising from the lack of reliable top incomes tax data in many of the developing countries on which Hajargasht and Griffiths focus. In any case, it appears that a "stylized fact" is that a reasonable model of most country's income distributions has curvature that is significantly below unity among the poor, rises above unity for the rich and monotone increasing over the full range so long as top income inequality is significant relative to overall inequality.