The Average-Marginal Relationship and Tractable Equilibrium Forms*

Michal Fabinger† E. Glen Weyl‡

October 2016

Abstract

Economic variables with familiar tractable functional forms (constant-elasticity or linear) are only reweighted in the change from their average to marginal versions. They are also simple, featuring only one or two terms. These properties allow for closed-form solutions. We explicitly characterize all equilibrium systems obeying a generalization of these properties, showing they form a hierarchy of tractability. The resulting forms are more realistic (e.g. bell-shaped demand and U-shaped cost) but highly tractable. These forms have importantly different implications for policy analysis, as we illustrate with applications from innovation, industrial, international, auction and public economics. We discuss close connections to the theory of Laplace transform and completely monotone functions.

Keywords: Laplace transform, tractability, closed-form solutions, innovation incentives, average and marginal variables

*This paper replaces a now-defunct paper "Pass-Through and Demand Forms"/"A Tractable Approach to Pass-Through Patterns". 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, 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, Yale University: glenweyl@microsoft.com.

1 Introduction

> *The things which have the greatest value in use have frequently little or no value in exchange; on the contrary, those which have the greatest value in exchange have frequently little or no value in use. Nothing is more useful than water: but it will purchase scarcely anything; scarcely anything can be had in exchange for it. A diamond, on the contrary, has scarcely any use-value; but a very great quantity of other goods may frequently be had in exchange for it.*

– Adam Smith, 1776, *An Inquiry into the Nature and Causes of the Wealth of Nations*

The "marginal revolution" resolved the diamond-water paradox by deriving the distinction and relationship between marginal (exchange) and average (use) value. Ever since this average-marginal relationship has been central to price theory: average and marginal revenue pervade monopoly theory and mechanism design, average and marginal cost are the central ideas in production analysis and the theory of selection markets, etc. Yet in modern, formal economics these quantities are not just important qualitative concepts, but the basis of systems of equations that define economic equilibrium quantitatively. As a result, the functional forms most frequently used in economic analysis are ones that maintain their tractable (viz. closed-form solvable) structure in both average and marginal analysis, such as linear, exponential and constant elasticity forms.

Despite the centrality of the average-marginal relationship, these forms appear to have been arrived at independently of one another and more or less by chance. The only systematic investigation we are aware of is that of Bulow and Pfleiderer (1983, henceforth BP), who characterize a class of demand functions including those above, but leave out many other equally tractable demand functions and do not allow for non-linear cost. In this paper we build off of BP's insights by systematically deriving the set of all such tractable equilibrium forms that preserve their formal structure between their average and marginal formulations. Within this class we identify a hierarchy of tractable forms with increasing flexibility (expressiveness) and gradually decreasing tractability. The lowest level of this hierarchy includes, as a special case, the BP class. This allows us to propose formulations of demand and cost that are more realistic than but equally or nearly equally tractable as those studied by BP. These forms yield conclusions about policy questions that are more empirically relevant than and which contrast with those under traditional forms in applications such as the relative distortion to incentives for innovation targeted at different income groups.

We do not have any concise characterization of the set of all cases where our approach applies. However, economic models formulated in terms of the relationship between marginal and average variables are pervasive in nearly every field of microeconomics and thus our characterization of the hierarchy of tractable forms applies broadly. To take a few of the examples we discuss, it applies to models of employer-employee bargaining in labor markets (Stole and Zwiebel, 1996a,b), the Rochet and Tirole (2003) model of two-sided markets, the analysis of optimal and symmetric

first-price auctions, competition in selection markets (Einav and Finkelstein, 2011), various types of sequential supply chain models in industrial organization (Salinger, 1988) and organizational economics (Antràs and Chor, 2013) and monopolistically competitive models of international trade (Krugman, 1980; Melitz, 2003).

Some results in these areas can be established analytically without parametric restrictions, while others are considered purely computationally based on forms that are motivated by concerns other than tractability. However in many cases tractability is useful to allow solutions without computers or to reduce computational time (and potentially even yield analytic solutions) when the solutions are aggregated or nested into a broader model, as well as to aid exposition, transparency and pedagogy in cases where general analysis is insufficient to yield sharp and relevant conclusions. In those cases, it is common to parameterize demand and cost using one of the very small number of forms mentioned above (viz. an element of the BP class) that are known to admit linear solutions and to preserve this form in both average and marginal formulations.

Unfortunately the BP class is extremely limited and unsatisfactory in many respects. To take the simplest example that we return to in the next section, the shape of the income distribution and nonparametric econometric analysis both suggest that demand curves usually have bell-shaped rather than constant or constant elasticity derivatives with respect to price (e.g. distributions of willingness-to-pay). Similarly intuition and most teaching in economics suggest costs curves have a U-shape rather than having constant marginal cost. These functional form restrictions have important implications for most policy questions that cannot be settled on general analytic grounds. For example, low-cost mass-market products better reward firms that introduce them (relative to their social value) than do high-cost luxuries under realistic functional forms, while under standard functional forms these quantities are independent of cost level. Thus on many if not most questions where they actually serve a useful purpose (of disambiguating questions that cannot be solved on general theoretical grounds or purely numerically), the BP class is misleading.

Luckily it is not necessary to focus on such misleading examples to obtain the tractability they usefully yield. In fact there is a much larger class of tractable, form-preserving functions that can be used as equilibrium forms. In Section 3 we prove that many other functional forms are linearly tractable, namely any form that can be written as the sum of two constant elasticity functions of quantity; the use of this form of demand (but without the matching cost required for tractability) was first proposed by Mrázová and Neary (2014). Beyond this we define a hierarchy of tractability based on the theory of polynomial equations; in particular, the tractability of a form in this class is characterized by the minimum number of constant elasticity terms *with adjacent terms having evenly-spaced exponents* needed to represent the form.

From a technical perspective, the representation of an arbitrary function by a linear combination of constant elasticity terms is the inverse Laplace transform in the logarithm of quantity, which we refer to as the inverse Laplace-log transform.[1] Tractable functional forms are thus those with

[1] After an extensive literature search of hundreds of articles and talking to numerous economists, including highly

"simple" inverse Laplace-log transforms. Because, as we emphasize in Section 5, essentially any function of economic interest has an inverse Laplace-log transform, arbitrary equilibrium forms may be approximated increasingly accurately with increasingly rich (and thus gradually less tractable) linear combinations of constant-elasticity terms. However, at each additional level of complexity the sacrifice in tractability is relatively small, so that an analyst can make a precise trade-off between tractability and flexibility (expressiveness), rather than simply entering an arbitrary form into the computer and hoping for the best on the one hand or restricting attention to unrealistic forms on the other.

Beyond the illustrative application we develop in the next section, in Section 4 we described a range of applications of our approach we are aware. These span the fields of industrial organization, international trade, auction theory and public economics, with diverse examples within these fields. For each application we describe why it is important to allow for the forms we derive based on a mixture of results from existing literature, simple observations included in the text and, in about half of the cases, detailed, analytical, calibrated or estimated applications that we include in the appendix to the paper.

Because the core results of our paper in Sections 3 and 5 are quite technical in nature, readers more interested in economic applications may wish to skip this material on first reading, though we have reserved the heaviest material for our appendix and the supplementary material at the end of this document. To ease readability, we begin our analysis with a brief example that illustrates the types of forms our analysis leads to and the flexibility it allows.

2 Example: Replacing Constant Elasticity Demand

The most canonical and widely-used demand form in economic analysis is the constant elasticity specification: inverse demand $P(q) = aq^{-b}$. Our study of the literature suggests three reasons this form is so widely used: historically, it was believed to be a good approximation to the income distribution, more recently it has been found to be highly analytically tractable, and it is tightly parameterized and thus easily estimated. To illustrate the use of our approach we propose an alternative form that fits the income distribution as we now know it far better, is nearly as tractable (yielding quadratic solutions in cases where constant elasticity yields linear solutions) and has the same number of parameters. We show that the difference between these forms has important implications for policy through examining the implications of the different forms for the bias of technical progress across products implied by the two forms. In the next section we discuss the general theory we used to derive this form.

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, (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.

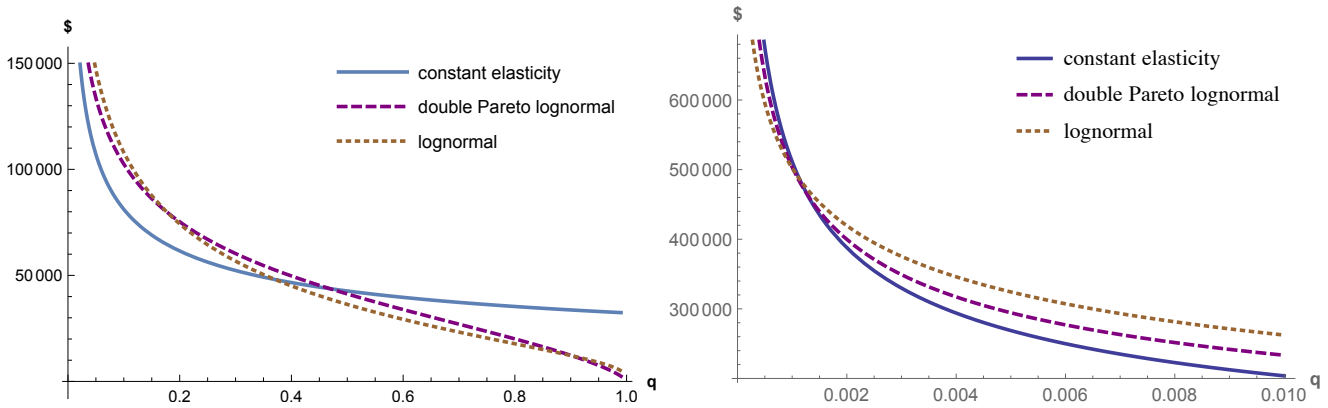


Figure 1: Comparing the fit of the best-fit log-normal to that of the best-fit constant elasticity form to a double-Pareto log-normal estimation of the 2012 US income distribution, represented as a demand (reversed quantile) function. Dollars at any (reversed) quantile represent the income of the individual at that quantile. On the left is the fit for the full income distribution, while the right shows the upper tail. For a functional form with a better fit, see Figure 2.

2.1 Why constant elasticity fails its original intended purpose

The origin of the constant-elasticity demand function 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 has the same shape as the income distribution.[2] Based on extrapolations of early probate measurements of top incomes following power laws (Garnier, 1796; Say, 1828), Dupuit (1844) and Mill (1848) suggested that demand would have a constant elasticity because a power-law distribution of income implies a constant-elasticity form for the reversed income quantile function[3] and thus for demand if it is proportional to income. 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. From here onward we refer to (inverse) demand functions generated from such shapes as themselves being "bell-shaped", even though their characteristic shape is actually that of a sigmoid S curve rotated 90° counter-clockwise. Distributions that accurately match income distributions throughout their full range (Reed and Jorgensen, 2004; Toda, 2012, Forthcoming) have a similar bell shape, but incorporate the Pareto tails measured in the 19th century data.

[2]Here we provide a few clarifying comments. (a) 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. (b) By a discrete-choice product we mean a product such that at most one unit of it can be utilized by any individual. (c) 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.

[3]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 income level.

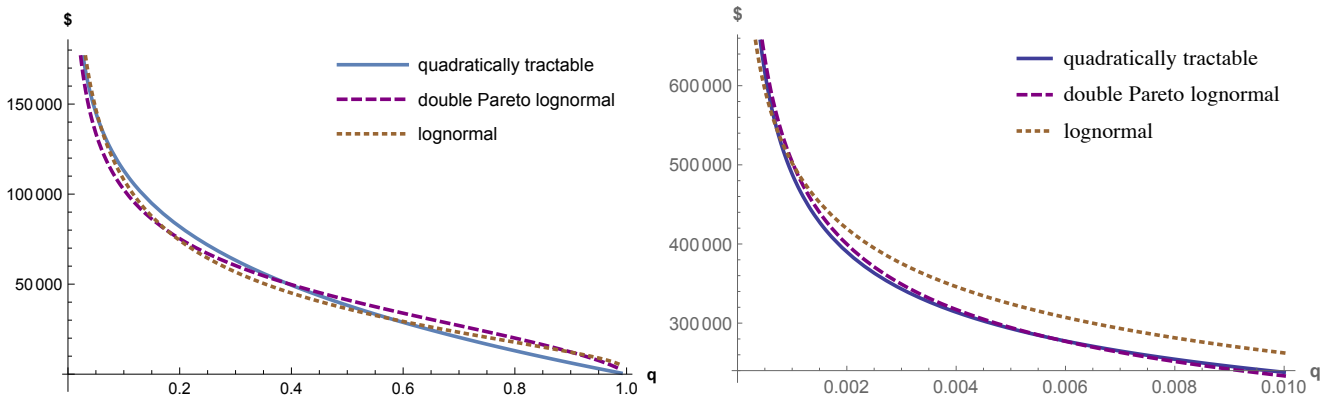


Figure 2: Comparing the fit of the best-fit log-normal to that of the best-fit quadratically solvable form form to a double-Pareto log-normal estimation of the US Income distribution, represented as a demand (reversed quantile) function. Dollars at any (reversed) quantile represent the income of the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

To illustrate what a poor fit the constant elasticity form is relative to these more realistic distributions, we used a standard calibration of a double Pareto log-normal (dPln henceforth) proposed by Reed (2003) to the US income distribution and used the generalized method of moments[4] to find the constant-elasticity demand function that best fits this throughout the full range of the income distribution. To show this we display in Figure 1 the reversed quantile function for these distributions, with the vertical axis representing income and the horizontal axis representing the (reversed) quantile of the distribution. As a benchmark we have included an equally parsimonious approximation, the log-normal distribution. We have compared the performance, on the left, for the full range of quantiles, and on the right only for the upper tail.

In the upper tail the constant elasticity approximation is a bit better of a fit than the log-normal, as suggested by our discussion above. However, in the rest of the income distribution its fit is terrible, while the log-normal fits quite well. Unfortunately, the log-normal distribution is highly intractable for most economic analysis, requiring numerical solutions for even the simplest problems as we will shortly illustrate.

2.2 A nearly-as-tractable replacement of constant-elasticity demand

Luckily, however, there are forms that give an even better fit than the log-normal and are nearly as tractable as constant elasticity. Furthermore while they have more parameters, one can reduce the number of parameters to 2 through a fit to the income distribution. In particular, consider the form $aq^{-b} + c - dq^b$, which strictly generalizes constant elasticity. Figure 2 shows that our best fit of this form is at least as good of a fit to the dPln estimate of the US income distribution as is the log-normal approximation over the full range and that in the upper tail it fits far better.

[4]Technically, we used Mathematica's non-linear fit function.

After obtaining this fit, a natural way to reintroduce two parameters is to allow, as suggested by Weyl and Tirole (2012), one parameter m that scales the fraction of their income individuals are willing to pay for the good and another parameter s that scales the total size of the market for the product. We then obtain with a bit of rounding fractions our proposed parsimonious replacement for the constant-elasticity form:[5]

$$P(q;s,m) \equiv m\left[\frac{1}{2}\left(\frac{q}{s}\right)^{-2/5} + 2 - \frac{5}{2}\left(\frac{q}{s}\right)^{2/5}\right].$$

2.3 Policy implications: the bias of technical progress

The tractability of the constant-elasticity demand form can be illustrated in a simple application that builds off the analysis of Kremer and Snyder (2015a,b). They consider the fraction of the social gains from creating a new product that may be appropriated by a monopolist, 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.[6] 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 when the level of marginal cost of production differs. 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 log-normal distribution form. We then follow Kremer and Snyder (2015b)'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, $P'(q)q = -abq^{-b}$ 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: $MR(q) = a(1-b)q^{-b}$. 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 $q = \left(a(1-b)/c\right)^{1/b}$. From this it follows by substitution that the firm's absolute markup is $\xi = cb/(1-b)$. Furthermore, the average consumer surplus also has the same form as $P(q)$, differing only by a multiplicative constant:

[5] Of course, if we used the form $aq^{-b} + c - dq^b$ in other applications, the appropriate exponents may be different. For example, in applications where the income distribution of some country other than the United States were in question, one would like to choose these to match that income distribution. Luckily income distribution is widely publicly available, making this a relatively simple task. In other applications the income distribution might not be an appropriate calibration target. Thus the particular form here is not a "law of nature" any more than the constant elasticity form is, but we do believe that in most applications of the constant elasticity form this formulation is more useful.

[6] In a different context this problem has also been studied recently by Budish et al. (2015).

$$\overline{CS}(q) \equiv \frac{CS(q)}{q} = \frac{\int_0^q P(x)dx - P(q)q}{q} = \frac{\frac{a}{1-b}q^{1-b} - aq^{1-b}}{q} = \frac{ab}{1-b}q^{-b}.$$

Evaluated at the optimal quantity, the average consumer surplus is $\overline{CS} = cb(1-b)^{-2}$. The appropriability ratio, i.e. the ratio of producer surplus and the total surplus, may be evaluated as $\xi/(\xi + \overline{CS}) = (1-b)/(2-b)$, 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.

This property turns out to hold much more broadly than for the constant elasticity form. In fact it is true of all of the demand functions in the constant pass-through class identified by BP. This class includes all demand forms we are aware of that have been used to provide closed form solutions to problems "of this nature" (we clarify what we mean by this in Section 4 below). Thus all existing tractable forms imply that cost conditions have no impact on the bias of technical progress.

What result would a log-normal distribution of willingness-to-pay yield? While this question can be answered numerically, there is no closed-form solution. To see this, note that for a log-normal distribution characterized by mean μ and standard deviation σ of the exponent, $P(q) = \exp\left(\sigma\Phi^{-1}\left(1-q\right) + \mu\right)$, where Φ is the standard normal cumulative distribution function. Thus

$$P'(q)q = -\sqrt{2\pi}\sigma q \exp\left(\sigma\Phi^{-1}\left(1-q\right) + \mu + \frac{\left[\Phi^{-1}\left(1-q\right)\right]^2}{2}\right) = -P(q)\sqrt{2\pi}\sigma q \exp\left(\frac{\left[\Phi^{-1}\left(1-q\right)\right]^2}{2}\right).$$

There is no analytic closed-form solution, therefore, to the equation $MR = c$, and matters are yet more complicated with the more realistic dPln distribution. These equations can be solved using numerical methods, but this somewhat reduces the transparency, accessibility, compactness and pedagogical value of the exercise. More importantly, in richer models, like those we discuss below in, for example, Subsection 4.2.1, where solutions to these equations must be aggregated into broader models, often with heterogeneity, the repeated calls to numerical operations needed can make the analysis computationally difficult or infeasible. This is presumably an important reason that these more realistic demand shapes are not used in such models.

Our form allows for the basic shape of the dPln distribution, as shown above, but is nearly as tractable as the constant elasticity form. If $P(q) \equiv m\left[\frac{1}{2}\left(\frac{q}{s}\right)^{-2/5} + 2 - \frac{5}{2}\left(\frac{q}{s}\right)^{2/5}\right]$, then $P'(q)q = -m\left[\frac{1}{5}\left(\frac{q}{s}\right)^{-2/5} + \left(\frac{q}{s}\right)^{2/5}\right]$. Note that, as with the constant-elasticity form, the form of $P'(q)q$ is identical to that of $P(q)$, involving only changing each term by a multiplicative constant. This implies that solving $MR(q) = m\left[\frac{3}{10}\left(\frac{q}{s}\right)^{-2/5} + 2 - \frac{7}{2}\left(\frac{q}{s}\right)^{2/5}\right] = c$ for q is, from a mathematical perspective, the same problem as solving $P(q) = c$. Furthermore this problem is just that of a solution to the quadratic equation

$$\frac{3}{10}x^2 + \left(2 - \frac{c}{m}\right)x - \frac{7}{2} = 0,$$

where $x = \left(\frac{q}{s}\right)^{-2/5}$, yielding

$$\frac{q}{s} = \left(\frac{5}{3}\right)^{-5/2} \left(\frac{c}{m} - 2 + \sqrt{\left(\frac{c}{m} - 2\right)^2 + \frac{21}{5}}\right)^{-5/2}.$$

While slightly more elaborate than the linear solution with constant elasticity demand, this is a quite compact closed-form solution. As above the markup and average consumer surplus take the same form and thus we can obtain the appropriability ratio in closed form as a function of cost. However, even easier and perhaps more instructive, we can obtain a very simple closed form expression for the appropriability ratio as a function of the fraction of the population served, $f \equiv \frac{q}{s}$:

$$\frac{21 + 105f^{4/5}}{56 + 180f^{4/5}}.$$

Unlike in the case of constant pass-through demands, this is clearly non-constant. It equals $\frac{21}{56} \approx 37.5\%$ for $f = 0$ (when the product serves a tiny fraction of the population) and monotonically increases in f to $\frac{126}{236} \approx 53.4\%$ for $f = 1$ (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 constant pass-through class so that patents are more desirable and optimal patent protection 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 our form above, that linear combinations of marginal revenue and inverse demand take the same form as inverse demand itself. Within these we then identify *all* forms for equilibrium systems (allowing cost as well as demand to vary) that permit closed-form solutions at each of a hierarchy of levels of tractability, beginning with the linear solutions allowed by BP's constant pass-

through forms (where we find there are many other, more realistic forms that are also tractable at that level), moving through the quadratic forms (an example of which we discussed in the previous section) to cubic and higher-order tractable forms. As was implicit above, the key property of these forms is that they may be written as the sum of constant elasticity terms with the exponents of adjacent terms evenly spaced relative to one another.

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, with a baseline interpretation as the 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 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.*[7]

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.[8]

Theorem 1. (Characterization of Form-Preserving Functions) *Any real finite-dimensional functional form class that is form-preserving under average-marginal transformations must be a set*

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

[8] 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*.

of linear combinations of

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

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

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

where $\{t_j\}_{j=1}^{N_1}$, $\left\{\tilde{t}_j\right\}_{j=1}^{N_2}$, *and* $\left\{\hat{t}_j\right\}_{j=1}^{N_2}$ *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 of this theorem relies on the theory of distributions developed by Laurent Schwartz, which is primarily applied in physics and formalizes the notion of *derivatives* of the Dirac delta function (point mass), and also draws on complex analysis. The theory of distributions and complex analysis do not feature in the standard economics or econometrics toolkit, and for this reason we leave the proof for 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 and to our knowledge none of these claim that, say, some functional form improves tractability relative to some alternative has ever been formally proven.[9]

A potential reason for this is that there is no clear definition within applied mathematics of the notion of tractability of the solution of mathematical equations. For example the classical theory of Galois, on which we will rely below, establishes that generic polynomial equations of degree at most four have solutions in terms of "the method of radical" (roots of different orders) and that generic polynomial equations of higher degree have no such solutions. But this theory does not prove that there is not some other function (other than roots) that can be added to the list of "closed-form" functions to provide solutions to higher order polynomials. In fact 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).[10]

[9]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.

[10]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

For this reason we use a 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 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 $q^{-.5}\left(\log q\right)^2$ it must also include linear combinations $q\log q$, $q^{-.5}\log q$, q and $q^{-.5}$. With a small number of (explicitly enumerable) exceptions classes of functional forms like this can rarely be solved in closed-form because of the mixture of power and logarithmic terms.[11]

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 BP constant pass-through demand with constant marginal cost.[12] As a result we focus on functional form classes composed of linear combinations of power functions q^{-t_j}.

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. When marginal cost is constant it simply adds a constant term to this expression, compatible with the constant p_0. In the special case of linear demand when $t = -1$, linear marginal cost takes the same form as inverse demand. As we showed in the previous section, using the BP demand form with constant marginal cost leads both to tractability (a linear solution) and to an important substantive implication, namely constancy of the pass-through of the constant marginal cost to price. However it is clearly possible to preserve the former property without the latter. For example, consider inverse demand and average cost of the form $P(q) = p_s q^{-s} + p_t q^{-t}$ and $AC(q) = ac_s q^{-s} + ac_t q^{-t}$. Then the monopolist solves

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

software. In this paper we made definite choices to resolve the terminological ambiguity.

[11]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.

[12]The exponential demand of Behrens and Murata (2007, 2012) is not nested in the forms of Theorem 1. In that case closed-form solutions to the firm's problem may be written in terms of the Lambert W function.

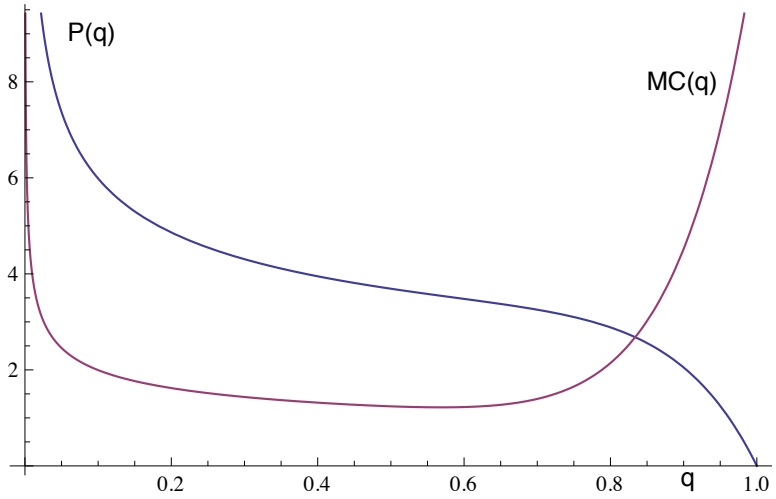


Figure 3: Example of a bell-shaped demand and U-shaped cost curve contributing to equilibrium conditions that can be solved linearly: $P(q) = 3\left(q^{-.3} - q^{10}\right)$ and $MC(q) = q^{-.3} + 10q^{10}$.

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

This more general form may thus also be solved tractably and offers substantially more flexibility. For example, it can accommodate simultaneously bell-shaped demand[13] and U-shaped cost curves; Figure 3 provides an example. A disadvantage of this form, however, is that it does not include a constant term and thus cannot accommodate comparative statics, like the pass-through rate, with respect to constant marginal cost, nor does it typically have an explicit form for the direct demand $Q(p) = P^{-1}(p)$.

It is thus useful to look beyond systems that can be solved linearly. 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. We thus 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 Galois, only for $k \leq 4$ 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 naive conjecture based on the above discussion is that this is simply

[13]We use the term "bell-shaped" as in Subsection 2.1.

Form(s)	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	Bell-shaped demand 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	Bell-shaped demand 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.

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 $(1-0)$ 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.

4 Breadth of Application

Thus far the only application we have discussed is classic monopoly pricing or slight variants thereof. While this is an important problem, the usefulness of our approach is not limited to this setting. We now discuss a range of other applications where tractability under the average and marginal forms is useful. We do not have a comprehensive characterization of all such applications and many are likely left out of our list, but hopefully this gives a sense of the variety of problems where our approach may apply. We divide these models by the fields in which they have been most commonly used, though the models are reasonably abstract and thus potentially more broadly applicable.

In every application we highlight not just the applicability of our techniques but some concrete result from existing literature or our own analysis indicating why more flexible or tractable forms than are used at present are important to that problem and/or how to concretely use our approach to gain insights in these environments. Some of these applications require the development of significant additional analysis inappropriate for the main text of this paper and thus we only briefly discuss these results here. A more thorough analysis in each case appears in Appendix B.

4.1 Industrial organization

Our applications to industrial organization illustrate how the marginal-average relationship appears in a variety of imperfectly competitive settings beyond the simple monopoly model.

4.1.1 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.[14] 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

[14] 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

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 Subappendix B.1 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 demand and U-shaped cost curves).[15]

4.1.2 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

[15] We use the term "bell-shaped" as in Subsection 2.1.

$$P_S(q_S) + P_B(q_B) - c = -\theta P_S'(q_S) q_S = -\theta P_B'(q_B) q_B$$

for some constant $\theta < 1$.[16] On the other hand they show that Ramsey pricing (which nests the 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.

4.2 International trade

More than any other field, international trade economists use models of general equilibrium with imperfectly competitive elements and rely on functional forms selected for tractability, usually from the BP class of demands.

4.2.1 Monopolistic competition models

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 be explicitly solvable even under these forms. Instead, the key property these allow is that the firms' optimization problems may be solved explicitly and aggregation integrals over heterogeneous firms may be expressed in closed form, assuming Pareto-distributed firm productivity.

[16] 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.

Because these models are quite elaborate, we defer a full model set-up to Appendix B.3 and also refer readers to the original sources. However, we now formally state a theorem that shows that demand and cost functions of our form aggregate in closed form just as the constant elasticity (or Pareto) and linear forms do, possibly with a gradual decay of tractability if higher-order forms are used. Then, in the next subsection, we apply this framework to build and learn from an easily-solvable model that allows an important change to the structure of trade costs.

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 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 $q^{\beta}N\left(q^{\alpha}\right)$ 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 share our form and the resulting aggregate function is more complex for higher degrees of the associated polynomial, there are also a few differences worth noting. First, aggregation is still possible when the heterogeneous component of marginal cost is shifted by a uniform multiplicative power factor relative to the other components of demand and supply. 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. We now turn to applications of this framework both to show how it is useful and make the preceding statements clearer.

4.2.2 International trade with marginal cost economies of scale

A vast majority of standard models of international trade assume that the costs of trade are "iceberg": a fraction of all goods (or their value) transported is assumed to be destroyed in transit. While this may be a reasonable model of many forms of tariffs, the vast majority of trade costs modeled this way are not tariffs and it seems implausible they would scale with trade volume and value in this manner. A certain fraction of international trade papers, e.g. Melitz and Ottaviano (2008), allow for constant marginal per-unit costs of trade. 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 international trade observed in data. A last explanation for low international trade volumes employed in typical models, fixed costs of establishing a presence in a foreign country, do a poor job of explaining why trade dies off so rapidly with distance between the exporter and importer; presumably fixed costs do not grow dramatically with distance, as would be required to justify observed trade patterns. It thus seems that most trade costs must arise from other sources.[17]

Recent literature has emphasized the importance of personal aspects of conducting trade, with focus on network effects.[18] If there are important tasks that need to be performed in person and cannot be easily delegated to others, then not just the cost of transporting goods over distance, but also the costs of conducting business travel shape the patterns of international trade. The firm's manager's personal disutility of business travel is likely to sharply increase with distance due to many considerations, such as staying away from home, time-zone adjustment, or the necessity to plan in advance, which would be consistent with trade flows rapidly decreasing with distance. However, managers' costs of coordinating trade seem unlikely to scale linearly with its volume; instead they seem likely to exhibit significant economies of scale.

In fact, despite not appearing in the international trade literature, this coordination cost-based Economic Order Quantity (EOQ) theory of trade costs (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. Thus even if it is not a literally accurate depiction of the physical costs facing firms (though it seems much closer than the iceberg model), it seems plausible that many firms use such a model for making decisions.

In this model, which we treat formally in Appendix B.3, the static cost of firms is viewed as arising from a steady state of dynamic purchases of inventory. Firms face a trade-off between shipping goods frequently in many small shipments to minimize the cost of idle inventory ("just-in-time delivery") and infrequently to minimize the coordination cost of each shipment by maximizing economies of scale in transport. Inventory cost is given by the average time a good lies idle, which is assumed to be inversely proportional to the number of shipments into which the total quantity is divided. In the simplest version of the model there is a constant cost of each shipment. In this case optimizing the number of shipments implies that the total cost of shipping a given quantity is proportional to the square root of the quantity. Essentially as the quantity grows larger fewer shipments suffice to maintain thin inventories, creating economies of scale.

However, the assumption that shipments have purely fixed costs is likely unrealistic; larger shipments seem likely to have at least some additional cost per unit shipped. We represent this by allowing shipments to have a cost that is a general weakly concave power function of the quantity shipped. The power in this function, α then determines the power in the optimized cost of supplying

[17] See Appendix B.3.2 for a more detailed discussion.

[18] See, e.g., Chaney (2014, forthcoming) and references therein.

a given quantity, $1-\beta$, according to the equation $\beta = {}^{(1-\alpha)}/_{(2-\alpha)}$. For any permissible value of α this implies, therefore, smooth power law economies of scale for the overall cost of shipping.

Before discussing our estimation and application of this parameter, it may be helpful to note that allowing coordination costs and the associated smooth economies of scale in trade costs can help explain the relatively low rates of observed trade. The key insight is that once inventory costs and thus the necessity of frequent shipments (and the economies of scale associated with these) are taken into account, the effective marginal cost of shipping may be much greater than it would appear in a static model where transporting the good at any point during the period (say a year) under question is sufficient to supply the market. Furthermore the (smooth) economies of scale associated with this model help explain the absence of exporting to many distant and low-demand locations without resort to implausibly large fixed costs of set-up that seem hard to account for.

Another advantage of this model is that it can be calibrated by data on shipping frequency that is entirely distinct from these sort of macro trade patterns that one might wish to explain. In particular, we show in Appendix B.3 that in a log-log regression of the frequency of shipments against the aggregate annual quantity shipped, the coefficient equals β as defined above under the theory. When shipment costs are fully linear, $\beta = 0$ and when they are fully fixed, $\beta = {}^1/_2$. Thus the theory implies that $\beta \in [0, {}^1/_2]$.

To test this prediction and estimate β we used a dataset on monthly shipments from China to Japan during 2000-2006. We focus on firms in one narrowly-defined product category that export for more than two years. Further details of our estimation appear in Appendix B.3. Our point estimate of β (averaged across industries) is .39 with a confidence interval of $[.36, 42]$. We can thus clearly reject the null of falsity of the model's assumptions, though the costs of shipments appear to be mostly fixed. This is reasonably strong evidence consistent with the (generalized) EOQ model. For tractability, we adopt a rounded version of .39, assuming that the marginal cost for trade decays as $q^{-2/5}$; conveniently this matches precisely the powers we estimated, in a similarly rounded fashion, for our proposed demand function in Subsection 2.2, suggesting a natural pairing of supply and demand sides for tractability and realism.[19] This estimate implies that increasing quantity by 10% reduces (the variable component of) the marginal cost of trade by 4%.

We use this estimate to calibrate a completely standard Melitz model with the only additional element being that we add to the standard iceberg trade costs assumed by Melitz a marginal cost of each unit traded given by our functional form.[20] We calibrate the model to standard moments in the trade literature as we discuss further in Appendix B.3; but it is worth noting that we choose an elasticity of demand of 5 as this leads to a demand form with a power ${}^1/_5$ that matches well with our cost form while being well within the standard estimated range. For brevity we spare

[19] While this coincidence is interesting, it is not necessary for modeling these international trade effects tractably.

[20] Forslid and Okubo (2016) consider iceberg trade costs dependent on the value shipped (see their Equation 7). This dependence is, however, ignored when taking first-order conditions for the firm's profit maximization. In our case the dependence of marginal cost of trade on the volume is taken explicitly into account when solving the firm's problem, since in our framework this does not lead to a loss of tractability.

the reader the resulting equations, which involve a large number of variables. However, they are available in Appendix B.3 and inspection indicates that they are not much more complicated than the corresponding expression in the basic Melitz model. In fact, the generalized model may be solved as explicitly as the original Melitz model: all outcome variables of interest, as well as the exogenous exit rate, may be expressed explicitly in terms of exogenous parameters (excluding the exogenous exit rate) and one endogenous variable. In other words, everything is explicit, up to one implicitly determined variable.[21]

We therefore believe this model is nearly equally tractable to the Melitz model, a better explanation of broad patterns of trade, and founded on a more natural microeconomic model of the source of trade costs. However, its value is also in its implications for empirically relevant patterns, some of which were explained in fairly *ad hoc* manner by modifying the demand away from the traditional constant elasticity assumption in each case to explain particular findings. While there are many implications of this model and its generalizations, here we discuss just two.

The first is the implication that large exporters have more similar domestic and foreign markups over marginal production cost than do small exporters. This follows directly from the fact that in our model there are smooth marginal cost economies of scale in trade costs, so that large exporters face a lower marginal cost of export than do small exporters. This prediction is consistent with intuition: when going abroad on a trip you rarely shop for globally available commodities, but do frequently try to get special deals on niche goods that may be available somewhere in your country but tend to be very expensive as they are sent infrequently in small batches.

The second implication appears in the context of multiproduct firms and is thus a bit subtler. In particular, we consider the elasticity of a multiproduct firm's exports of different products to increases in the demand for those products. All products are assumed to be shipped together and thus jointly share the economies of scale associated with trade. However, different products have different marginal costs of production. For the relatively cheap and thus relatively widely consumed goods, the costs of shipping loom large in the total cost of selling the good. For the relatively expensive and thus narrowly consumed goods, the costs of shipping are a small component.

Because shipping costs exhibit economies of scale while we assume production costs do not, widely consumed products will act more like products with economies of scale than will narrowly consumed products shipped by the same firm. Economies of scale make production react more elastically to a variety of shocks. For example, when there is an upward aggregate demand shock in the foreign market, the shipping costs for both goods will fall, but this will cause a larger proportional increase in the sales of the good where these shipping costs are greater, both because the shipping costs are a larger fraction of cost and because the proportional reduction in cost from the additional sales induced by the reduced price will proportionally reduce the price of the cheaper

[21]Note that the original Melitz model cannot be solved fully explicitly even for Pareto distribution of firm productivity. This is because Equation 12 in Melitz (2003) does not allow for an explicit solution for the productivity cutoff. However, in terms of the productivity cutoff, it is possible to solve explicitly for the exogenous exit rate, as well as for endogenous variables of interest. In our case the situation is closely analogous.

good by more.

We can see this formally in our setting by considering a case where products have very similar marginal costs, but differ slightly in these costs. In that case, our equilibrium formula yields that the difference between the elasticity with respect to a uniform-across-products expansion in demand for two products i and j with marginal costs near a typical value MC is[22]

$$\epsilon_i - \epsilon_j \approx -\frac{10w\kappa_{\text{LT}}\tau}{\sqrt[5]{n}\kappa_R\sqrt{n^{2/5}\kappa_R^2 - 4\text{MC}w\kappa_{\text{LT}}\tau}}\left(\text{MC}_i - \text{MC}_j\right),$$

where n is the number of exported goods. Products that are more popular, that is products that have lower marginal cost of production MC_i, will see their sales increase relatively more in response to an increase in the demand prefactor κ_R. This pattern is observed empirically by Mayer et al. (2014, 2016). They explain this using a model with a particular type of non-CES demand that generates this pattern. While consistent with the observed pattern of elasticities, this demand form has no independent justification, in contrast to the EOQ-based cost explanation we propose, which is micro-founded in a standard model of dynamic shipping and can simultaneously account for a number of other findings in the trade literature.

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

Antràs and Chor (2013) propose a model of vertical supply chains with hold-up problems in the spirit of Grossman and Hart (1986). There is a continuum of stages of production, each of which contributes to the final quality of a product. A firm has to decide which stages of production to insource v. outsource. The advantage of insourcing is that it gives the firm greater bargaining power to appropriate the rents associated with that production stage. The cost of appropriating these rents, however, is that it deters the supplier at that stage from producing high-quality products. This induces a monopsony problem.

Antràs and Chor assume that bargaining at each stage occurs over the marginal revenue generated by the incremental quality at that stage only.[23] They also assume suppliers at all stages have the same convex cost of providing quality. This makes it optimal to equalize the total rents left to each supplier, as this minimizes the total cost of obtaining a given aggregate quantity. Because bargaining is over *marginal* revenue *at a particular stage*, this makes the optimal degree of bargaining power to give to each supplier inversely proportional to the marginal revenue at her production

[22]The formula is straightforward to derive, given our discussion of the single-product case in Appendix B.3, which also introduces the precise notation. The multi-product model considered here is a generalization in which the cost of shipping depends only on the total amount of goods the firm ships, not on the proportions of different types of products. The elasticity is defined as $\epsilon_i \equiv \partial(\log q_{fi})/\partial(\log \kappa_R)$, where q_{fi} is the quantity of product i that arrives in the foreign country.

[23]Our presentation of the model of Antràs and Chor is in terms of different variables than in the original paper. Our version of the model is mathematically isomorphic to theirs if we perform a relabelling of variables, explained in Appendix B.4. It turns out that in terms of our variables the model is simpler to understand and to generalize, so introducing them may be thought of as a separate contribution of this paper.

stage. The total level of quality to choose is given by the equation[24]

$$MR = MC + qMC',$$

where MC is the suppliers' common marginal cost of producing quality. The right-hand side of this expression bears the marginal-average relationship to MC and thus is referred to by Bulow and Roberts (1989) as the "monopsonist's marginal cost curve".

As a result, if inverse demand and average cost are drawn from a form-preserving functional form class, our characterization of tractability applies to this problem as well, in the case when bargaining power can be continuously adjusted for all suppliers, i.e. in the case of the "relaxed" problem Antràs and Chor study in detail. If instead the firm must choose between discretely different levels of bargaining power (insourcing v. outsourcing) then, as we show our supplementary material Subsection I.3, if one of the powers is 0 so that the system has an explicit inverse, the optimal pattern of sourcing can again be solved in closed form.

In their analysis Antràs and Chor focus on a tractable case, namely when both average cost and marginal revenue follow a single-term power law. This implies that marginal revenue is either monotone increasing or monotone decreasing and thus that outsourcing occurs at most at one end of the supply chain. This seems perhaps counterintuitive, as it is often the case that large firms outsource both the production of early inputs and final retailing, insourcing a contiguous production process in between.

In Appendix B.4 we consider an identically tractable and we believe more plausible model in which early production stages have low marginal revenue as the product is just getting started and late stages have low marginal revenue as finishing touches have declining marginal value. In this case outsourcing occurs at both the early and late stages of production, as intuition suggests. Our tractable forms thus allow us to derive more intuitive conclusions from more plausible primitives at no cost to tractability in this case.

4.3 Auction theory

4.3.1 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, $\left[G_\star(x)\right]^{N-1}$.

[24]This equation does not appear in this general form in the original paper, which from the beginning assumes constant-elasticity functional forms.

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

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

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.

4.3.2 Auctions v. posted prices (Einav, Farronato, Levin and Sundaresan, 2016)

Einav et al. (2016) 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 Subsection 4.1.2. 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 Subsection 4.1.2, $\overline{U}$ 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.

4.4 Public economics

4.4.1 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.

Consider this model, as it is usually studied (Helpman and Itskhoki, 2010; Helpman et al., 2010, 2015), for the case of firm with a monopoly in the product market[25] and no market power in the labor market (facing an exogenous wage W_0); our analysis can, as suggested by our treatment of monopsony in Subsection 4.2.3 above, easily be extended to allow monopsony as well. We also assume a linear production technology and for simplicity state all results in terms of the quantity produced. SZ show that in this case if the bargaining power of the worker relative to that of the firm is $\lambda \geq 0$ and the prevailing wage is W_0, optimal production is given (implicitly) by the integral equation

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

This equation, while intractable in general, preserves the functional form of any 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 left-hand side of Equation 1 becomes

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

More generally, for $MR(q)$ a linear combination of power terms, each term of is multiplied by $\frac{1+\lambda}{1+\lambda+t\lambda}$, 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 Appendix B.5 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 Subsection 2.2, we find that over a reasonable business cycle range over-employment should shift by roughly .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.

[25] We use the standard notation $P(q)$ for price, $MR(q)$ for marginal revenue, and $q^\star$ for equilibrium quantity.

4.4.2 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 (2014) extend this framework to nest a variety of models of imperfect competition using a conduct parameter θ as in Subsection 4.1.1 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.

5 General Approximation and the Laplace-Log Transform

So far 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 obviously 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 class. 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.

5.1 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. For the purposes of making 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.[26] 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[27]

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

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 $U(q)$.[28] At the same time, the integral may be thought of as expressing $U(q)$ as a linear combination of form-preserving functions of Theorem 1.

Technical Clarification (Integral Definition).[29] Here we define the integral (2) to be the

[26] $U(q)$ 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 $P(q)$.

[27] 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.

[28] 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.

[29] 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.

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 mixtures 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*[30] *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*[31] *on $[0, \bar{q}]$.*

Theorem 1 also allowed for functions other than mixtures of power functions, e.g. $q^{-\alpha} \log q$, that are useful in some economic context. 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 $\mathbb{C}_{\bar{s}}^{-} \equiv \{s | Re\ s < \bar{s}\}$*

[30] 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).

[31] 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 $\sup_{q \in [0, \bar{q}]} |\tilde{U}(q) - U_j(q)| < \epsilon$.

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).[32] *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*

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

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 Δt as well as by the factorial in the denominator.[33] The last part R_3 may be expressed and bounded using integrals

[32] 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 $\mathbb{C}_{\bar{s}}^{-}$.

[33] 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.

of high derivatives of $h(t)$:

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

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 problem, for which regularization is typically used in the applied mathematics and econometrics literature.[34]

5.2 Complete monotonicity of the demand specification

Many demand curves have economic properties that determine 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.[35] 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)[36]

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

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

[35] 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.

[36] 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.

$n \in \mathbb{N}$,

$$CS^{(n)}_{[s]}(s) \geq 0,$$

or equivalently[37]

$$U^{(n)}_{[s]}(s) - U^{(n+1)}_{[s]}(s) \geq 0.$$

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-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:*[38]
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*[39]*, 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*

[37] 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.

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

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

($\varepsilon > 1$), 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 Section E 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.[40] The log-normal distribution is not quite so well-behaved, but the more realistic income model (the double Pareto log-normal) behaves similarly for calibrated parameter values.

6 Conclusion

This paper makes three contributions. First, it identifies equilibrium systems that are analytically tractable, nest nearly every known tractable equilibrium systems as special cases and allow greater flexibility than existing systems. Second, it shows how these equilibrium systems overcome implausible substantive assumptions imposed by existing tractable systems. Finally, it uses this framework to improve the realism of the predictions of a range of models used in industrial organization, international trade, auction theory and public economics.

Besides direct applications to economic modeling, our work suggests several directions for future research. First, adding additional terms to match features of an equilibrium system closely resembles sieve approximation in non-parametric econometrics. Determining "optimal" procedures for using increasingly less tractable equilibrium systems to approximate empirical equilibrium systems as statistical precision and numerical capacity increase is a potentially powerful way to apply our framework. Similarly determining the loss in approximation accuracy arising from using tractable forms is important to understanding their desirability relative to other approximations.

Second, complete monotonicity of the inverse Laplace-log transform of demand may be a useful property from which to derive substantive economic results. For example, in other work we have conjectured that the complete monotonicity of inverse Laplace-log transforms along with some support conditions on this transform are sufficient conditions for price discrimination to be welfare enhancing.

Finally, we considered only models that can eventually be reduced to a single equation (or a few equations linked by average-marginal relations). Our approach can be extended somewhat beyond this, as we discuss, e.g., in Appendix B.2, to cases where the interactions across equations

[40] 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 $CS(s)$.

occur according to some simple aggregators. However our approach here does not directly apply to models such as imperfectly competitive models with multiple choice dimensions of quality or asymmetric oligopoly, characterized by several, non-aggregable equations. Nonetheless it seems likely that the techniques from applied mathematics we use apply in these richer settings, albeit with an increasingly less favorable trade-off between flexibility and tractability as the number of equations increase.

More broadly, the analysis of imperfect competition has become somewhat divided between approaches that employ simple, explicitly soluble systems and other work that focuses on more complex, realistic systems that require significant computational effort to analyze (Berry et al., 1995). The approximation approach we developed here can bridge between these two extremes by allowing systems that match key policy-relevant features of empirical structures while remaining nearly as tractable as the systems that are usually employed for their convenience rather than their realism. Researchers may then choose more freely which position along the tractability-realism spectrum is most appropriate to their purposes, as well as choosing the features to most closely approximate depending on the policy question of interest.

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.

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.

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.

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.

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," 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.

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 John Roberts, "The Simple Economics of Optimal Auctions," *Journal of Political Economy*, 1989, *97* (5), 1060–1090.

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

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.

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.

Chaney, Thomas, "The Network Structure of International Trade," *The American Economic Review*, 2014, *104* (11), 3600–3634.

_ , "The Gravity Equation in International Trade: An Explanation," *Journal of Political Economy*, forthcoming.

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," 2015. https://goo.gl/bqO5ro.

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.

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

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

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," 2016. http://web.stanford.edu/∼leinav/AFP.pdf.

_ **, Theresa Kuchler, Jonathan Levin, and Neel Sundaresan**, "Assessing Sale Strategies in Online Markets Using Matched Listings," *American Economic Journal: Microeconomics*, Forthcoming.

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," 2013. http://pages.stern.nyu.edu/∼xgabaix/papers/CompetitionEVT.pdf.

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

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.

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 and Oleg Itskhoki, "Labour Market Rigidities, Trade and Unemployment," *Review of Economic Studies*, 2010, *77* (3), 1100–1137.

_ , 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.

_ , _ , Marc-Andreas Muendler, and Stephen Redding, "Trade and Inequality: From Theory to Evidence," 2015. http://www.princeton.edu/∼itskhoki/papers/TradeInequalityEvidence.pdf.

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

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

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," 2015. https://bfi.uchicago.edu/sites/default/files/research/Kremer_Snyder_PriceTheory.pdf.

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," 2014. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2372661.

Mayer, Thierry, Marc J. Melitz, and Gianmarco I. P. Ottaviano, "Market Size, Competition, and the Product Mix of Exporters," *American Economic Review*, 2014, *104* (5), 495–536.

_ , _ , and _ , "Product Mix and Firm Productivity Responses to Trade Competition," 2016. https://goo.gl/cWspK9.

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. http://goo.gl/8CaXPv.

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. http://www.ssc.wisc.edu/∼dquint/papers/quint-complements-substitutes.pdf.

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. http://eml.berkeley.edu/∼saez/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 Manièe 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.

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.

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*, Forthcoming.

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

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). For convenience we express the (infinitely differentiable) functions $F(q)$ on $\mathbb{R}^+$ in terms of functions $G(s)$ defined on $\mathbb{R}$, 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 of 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 non-empty interval $S \equiv (s_1, s_2)$, i.e. with $G_S(s) \equiv G(s) 1_S(s)$. 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 $G_S(s) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} \hat{G}_S(\omega) e^{-i\omega s} d\omega$.[41]

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 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,

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

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

[41] 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.

the expression (6) is a generalized function[42] of s that gives zero when integrated against any test function[43] 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 $T_2(\omega)$. 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_5, 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 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 $F(q)$:

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

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:

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

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

$P_k(x) = 0$, 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 explicity 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 $\int q^{\gamma_8} N_1\left(q^\alpha\right)^{\gamma_9} N_2\left(q^\alpha\right)^{\gamma_{10}} dq$. 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 structues of the polynomials. Due to space limitations, the detailed discussion of these cases is left for the supplementary material Section G. □

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). Due to space limitations the details are provided in supplementary material Section H. □

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). □

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,[44] 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. □

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 Section C, 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). □

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 Section C. □

B Applications

B.1 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

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

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

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

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. (2015) 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 5 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 ESC 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 ESC mass points for tractability. In either case, this approach significantly reduces the complexity of computing and representing the system.

B.2 Monopolistic competition

B.2.1 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 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 we wish to be able to apply Theorem 2, so

we let $u(q_\omega)$ be a linear combination 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 $MC_\omega(q) = \sum_{t\in T} mc_{\omega,t} q_\omega^{-t}$, $\quad AC_\omega(q) = \sum_{t\in T\cup\{1\}} ac_{\omega,t} q_\omega^{-t}$, 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.

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 and in the supplementary material Subsection I.1.

B.2.2 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.[45] The utility has the very general form

$$U_\Omega \equiv F\left(U_\Omega^{(1)}, U_\Omega^{(2)}, ..., U_\Omega^{(m)}\right), \quad U_\Omega^{(i)} \equiv \int_\Omega U^{(i,\omega)}(q_\omega)\, d\omega.$$

In order to preserve tractability, we assume that $U^{(i,\omega)}(q_\omega)$ are linear combinations[46] 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 straightforwad 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 approproate assumptions on the productivity distribution.

[45] 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.

[46] 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.

For more details, see supplementary material Subsection I.1.

B.3 Modeling International Trade

B.3.1 Calibration with variable marginal cost of trade

To complement our Subsection 4.2.2, here we provide an analytic solution to a trade model with heterogeneous firms and variable marginal cost of trade proportional to $q^{-2/5}$, which matches the empirical finding discussed there. For simplicity of exposition, we consider two symmetric countries, but our closed-form aggregation results of Theorem 3 are more general.[47] The setup of the model follows Melitz (2003), but in addition to the usual "iceberg" cost (i.e. damage of goods as they are transported), we allow for a specific cost of trade that varies non-linearly with the traded quantity.

Varieties ω of a differentiated good are produced by monopolistically competitive single-product firms using a single factor of production, for simplicity referred to as labor. The marginal unit labor requirement a is constant, but depends on the firm. In addition to the variable cost of production, there is a fixed cost of operation f and a fixed cost exporting f_x, in units of domestic labor. Entry into the industry is unrestricted, but involves a fixed cost of entry f_e, again in units of domestic labor. Only after the entry cost has been paid does the firm learn its productivity parameter a, drawn from a distribution with cummulative distribution function $G(a)$. For the calibration discussed below we choose Pareto distibution $G(a) = G_0 + \kappa_G \nu_G^{-1} a^{-\nu_G}$ with support $[a_0, a_3]$. When the value of a is revealed, the firm decides whether or not to exit the industry, and if it does not exit, whether to export. In addition to endogenous exit, the firm may be exogenously forced to exit with a probability of δ_e per 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, τ units need to be shipped. The second component requires using an amount of labor given by $L_T(q) = \frac{5}{3}\kappa_{\text{LT}} q^{3/5}$ for storage and coordination tasks, as discussed previously.

Consumers have a CES utility function $U = (\int q_\omega^{1-\frac{1}{\sigma}} d\omega)^{\frac{\sigma}{\sigma-1}}$ that depends on the quantity q_ω of each variety ω consumed. We set elasticity of substitution σ equal to 5, which is consistent with the typical range in the existing empirical literature of about 4 to 8. Each country has an endowment of labor L_E, which is supplied at a competitive wage rate w.

The revenue a firm can earn by selling a quantity q in a given market is $R(q) = \frac{\kappa_R}{\nu_R} q^{\nu_R}$, where $\nu_R = 1 - \frac{1}{\sigma}$. We choose numeraire in such a way that κ_R is an exogenously given constant, and we let the wage rate w, which was set to 1 in Melitz (2003), be endogenously determined.

Let us first discuss the nature of the exporting firms' problem. The first-order condition for choosing the quantity q_f that should reach the foreign market equates the marginal revenue and the comprehensive marginal cost:

$$R'(q_f) = aw\tau - wL_T'(q_f) \Rightarrow \frac{\kappa_R}{\tau \sqrt[5]{q_f}} = aw + \frac{w\kappa_{L_T}}{\tau q_f^{2/5}}$$

[47] In the case of asymmetric countries aggregates may still be computed explicitly in terms of a very small number endogenous variables per country, as in the case of the original Melitz model empirically investigated by Helpman et al. (2008)

and leads to the optimal choice

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

If the marginal cost of production aw exceeds $\kappa_R^2/(4w\tau\kappa_{\mathrm{LT}})$, the first-order condition cannot be satisfied; exporting will not be profitable for this firm even if there is no fixed cost of exporting.[48][49] This implies that our CES-based model can generate an export cutoff without assuming that fixed costs of exporting are extremely large or assuming that advertizing needs to be always untargeted, a point to which we will return later.

The full solution to the model may be constructed as follows. As in the case of the Melitz model, a non-trivial solution features two productivity cutoffs. Cutoff "1" corresponds to firms indifferent between producing for the domestic market and exiting, and cutoff "2" corresponds to firms indifferent between producing only for the domestic market and exporting. We use the corresponding indices for economic variables related to such firms. For notational simplicity, we also introduce index "0" for firms with the lowest possible productivity and "3" for firms with the highest possible productivity.[50]

The firm's first order conditions in the domestic and the foreign market are

$$a = \frac{\kappa_R}{w\sqrt[5]{q_d}} \Rightarrow q_d = \frac{\kappa_R^5}{a^5 w^5},$$

$$a = \frac{\kappa_R}{w\tau\sqrt[5]{q_f}} - \frac{\kappa_{LT}}{\tau q_f^{2/5}} \Rightarrow q_f = \left(\frac{\kappa_R + \sqrt{\kappa_R^2 - 4\tau a \kappa_{\mathrm{LT}} w^2}}{2\tau a w} \right)^5.$$

At special points these explicitly relate a_0 to q_{d0}, a_1 to q_{d1}, a_2 to q_{d2} and q_{f2}, and a_3 to q_{d3} and q_{f3}. The indifference condition at cutoff 1, i.e. the zero cutoff profit condition, allows us to solve for q_{d1}:

$$q_{d1} = 4\sqrt{2}\kappa_R^{-5/4} f^{5/4}.$$

Similarly, the indifference condition at cutoff 2 allows us to solve for the wage in terms of q_{f2}:

$$f_x = \frac{1}{4} q_{f2}^{4/5} \kappa_R - \frac{2}{3} q_{f2}^{3/5} w \kappa_{LT} \Rightarrow w = \frac{3}{8} \frac{q_{f2}^{4/5} \kappa_R - 4 f_x}{q_{f2}^{3/5} \kappa_{LT}}.$$

We see that if we specify q_{f2}, then all the variables previously mentioned in this paragraph may be explicitly computed. What determines q_{f2} in equilibrium is the unrestricted entry condition for

[48] The second-order condition is even slightly more restrictive.

[49] In the model of Arkolakis (2010), firms facing constant-elasticity demand also may not find it profitable to export, even if the fixed cost of exporting is small, due to the model's assumption on advertising costs. Note, however, that such conclusion would not hold if the model allowed for targeted advertising in an environment where it is consumer heteroegenity that is primarily responsible for demand being elastic, as in typical industrial organization models. In our case we do not have to assume the impossibility of targeted advertising even if the demand elasticity comes primarily from the heterogeneity in consumer valuations for the good, as suggested by microeconomic empirical evidence.

[50] Of course, here we assume that the parameter values are such that this is the correct structure of the equilibrium. There exist other parameter values where this would not me true, for example, if the fixed cost of exporting is too high, no firm may choose to export.

firms. Motivated by the empirically observed size distribution of firms, we set the Pareto distribution parameter ν_G to -5. In this case the unrestricted entry condition condition is[51]

$$
\begin{aligned}
wf_e\delta_e = {} & \tfrac{\kappa_G\kappa_R^5}{20w^5}\left(-\tfrac{4fw}{q_{d1}}+\tfrac{4fw}{q_{d3}}+\tfrac{5\kappa_R}{\sqrt[5]{q_{d1}}}-\tfrac{5\kappa_R}{\sqrt[5]{q_{d3}}}\right) \\
& -\tfrac{\kappa_G}{420w^6\tau^5 q_{f2}^2\kappa_{LT}}\left(\sqrt[5]{q_{f2}}\kappa_R-w\kappa_{LT}\right)^5\left(10q_{f2}\kappa_R^2+w\left(84wf_x-55q_{f2}^{4/5}\kappa_R\right)\kappa_{LT}+80w^2q_{f2}^{3/5}\kappa_{LT}^2\right) \\
& +\tfrac{\kappa_G}{420w^6\tau^5 q_{f3}^2\kappa_{LT}}\left(\sqrt[5]{q_{f3}}\kappa_R-w\kappa_{LT}\right)^5\left(10q_{f3}\kappa_R^2+w\left(84wf_x-55q_{f3}^{4/5}\kappa_R\right)\kappa_{LT}+80w^2q_{f3}^{3/5}\kappa_{LT}^2\right)
\end{aligned}
$$

This condition allows us to express δ_e explicitly in terms of q_{f2} and given exogenous parameters. If we wish to find q_{f2} in terms of δ_e, we need to use the inverse of this relation. Similar situation occurs in the original Melitz model.

The relations above represent the desired explicit solution to the model (in terms of q_{f2}). In addition, we may explicitly compute the mass of firms, price indexes, and welfare, since all these lead to the type of integrals covered by the aggregation theorem. For example, the mass M_e of entering firms is given by

$$
\begin{aligned}
\tfrac{L_E\delta_e}{M_e} = {} & f_e\delta_e+\tfrac{\kappa_G\kappa_R^5}{5w^6q_{d1}q_{d3}}\left(fw\left(-q_{d1}+q_{d3}\right)-5q_{n1}q_{d3}^{4/5}\kappa_R+5q_{n1}^{4/5}q_{d3}\kappa_R\right) \\
& -\tfrac{\kappa_G}{105w^7\tau^5 q_{f2}^2\kappa_{LT}}\left(-\sqrt[5]{q_{f2}}\kappa_R+w\kappa_{LT}\right)^5\left(-15q_{f2}\kappa_R^2+21w^2f_x\kappa_{LT}+30wq_{f2}^{4/5}\kappa_R\kappa_{LT}+20w^2q_{f2}^{3/5}\kappa_{LT}^2\right) \\
& +\tfrac{\kappa_G}{105w^7\tau^5 q_{f3}^2\kappa_{LT}}\left(-\sqrt[5]{q_{f3}}\kappa_R+w\kappa_{LT}\right)^5\left(-15q_{f3}\kappa_R^2+21w^2f_x\kappa_{LT}+30wq_{f3}^{4/5}\kappa_R\kappa_{LT}+20w^2q_{f3}^{3/5}\kappa_{LT}^2\right)
\end{aligned}
$$

Similar closed-form formulas for the other variables of interest exist and may be straightforwardly derived by the same method.[52]

B.3.2 Resolving the paradox of trade costs

Here discuss in more detail the question of trade costs raised in Subsection 4.2.2. The Melitz model has been useful for many purposes, but empirical data on costs and prices represent a challenge. In the framework of the original Melitz model it is not easy to explain relatively low international trade flows in a way that is consistent with empirically observed international container shipping charges, costs setting up presence in a foreign country, and prices of imported products. A commonly used technical explanation for low international trade flows is large iceberg trade costs, such as $\tau = 1.33$, which means that each firm wishing to export will face a marginal cost of trade equivalent to destroying 25% of all goods produced for export. This is clearly inconsistent with the fact that in practice only a negligible fraction of goods gets destroyed in international transport and with the fact that international shipping rates for containers are remarkably small. Moreover, such large iceberg trade costs are inconsistent with microeconomic evidence on pricing by major firms, which suggests their prices for the same product in different countries are roughly equal.

The other possible way the Melitz model can be made technically consistent with low international trade flows is to assume very high fixed costs of exporting, such as costs of renting office space in a foreign country or translating product manuals to a different language. This explanation is also problematic. International trade flows decrease rapidly with distance, but it is not clear how such fixed costs of exporting could strongly depend on distance. Moreover, the magnitude of the

[51] The choice $\nu_G = -5$ corresponds to domestic sales Pareto index of 1.25, which is consistent with data from countries such as Italy, see p. 32 of Aoyama et al. (2010). An analogous explicit condition with more general values of the Pareto index is a bit longer and available upon request.

[52] The formulas are available upon request.

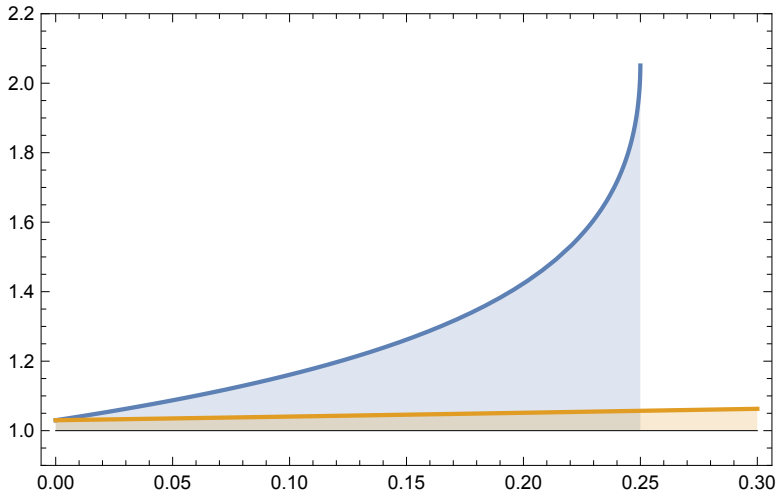


Figure 4: Comparison of foreign-domestic price ratios for small firms (blue line) and large firms (orange line). Here the parameter choices are $\tau = 1.03$ and $a = 10$ (small firms) and $a = 1$ (large firms). The horizontal axis corresponds to rescaled κ_{LT}, namely $\kappa_{LT} \times (\tau w^2 \kappa_R^{-2})$. When this measure is larger than 0.25, exporting is no longer profitable for the small firms.

fixed costs required to enter a foreign market would have to be comparable to the magnitude of the fixed cost of operating the firm itself. It is hard to imagine what could provide a justification for such exessive export fixed costs.

Including coordination costs in the Melitz framework provides a natural solution to this paradox. The calibrated fixed costs of exporting may take realistically low values since coordination costs also enter the decision to export or not. While effective trade costs are small at small quantities, they may still be larger than the corresponding revenue the firm could earn, even with CES demand. (This was discussed previously in connection with the exporting firm's first order conditions.) This phenomenon can account for low international trade flows. Yet major firms will find the effective trade costs negligible, and set similar prices in different countries, in agreement with microeconomic evidence on their pricing; see Figure 4. Similarly, low international container shipping rates and low damage of goods during transport do not pose a problem for the model; costs of this kind do not have to be high to be consistent with low international trade flows. The rapid falloff of trade flows with distance also has a natural interpretation in the model, since business travel is certainly more challenging for longer distances.

This resolution of the paradox may be interpretted also more broadly. While for simplicity we assumed that the shipments are coordinated within firm, one can also consider the case in which the shipments are coordinated between an exporter and an importer. In this case it is even more natural to expect that business travel and face-to-face contact play an important role for determining international trade flows.

B.3.3 Generalized Economic Order Quantity model applied to international shipping

To gain an empirical insight into the scale economies of international trade of Subsection 4.2.2, 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 (1913). Consider a firm that produces a single good 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 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: $C_s = \kappa_t {q_s}^{\alpha}$, $\alpha \in [0,1]$. (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 $C_t = C_s q/q_s = \kappa_t q {q_s}^{\alpha-1}$. Minimizing the sum of the inventory cost and the coordination cost leads to the optimal choice $q_s = (q(1-\alpha)\kappa_t {\kappa_i}^{-1})^{\frac{1}{2-\alpha}}$, the minimized value $(2-\alpha)(1-\alpha)^{-\frac{1-\alpha}{2-\alpha}} \kappa_i^{\frac{1-\alpha}{2-\alpha}} \kappa_t^{\frac{1}{2-\alpha}} q^{\frac{1}{2-\alpha}}$, and the optimal frequency of shipping $f_s = q/q_s$ equal to

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

This result implies that we can infer the coordination cost exponent α 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 $\beta \equiv (1-\alpha)/(2-\alpha)$.[53] The model predicts that this coefficient always lies between 0 and $\frac{1}{2}$, since $\alpha \in (0,1)$.

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 $\alpha = 0$ and $\beta = 1/2$, implying effective cost of trade (here inventory and coordination) proportional to $\sqrt{q}$. The other extreme case has $\alpha = 1$ and $\beta = 0$ 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 obtain empirical estimates of β, we used a Chinese customs dataset on firm-level monthly shipment data on exports from China to Japan in years 2000 to 2006. We selected firms by requiring that they specialize on one narrowly defined 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 performed the industry-level regressions and calculated overall statistics. The mean of industry-level estimates of β was 0.39 with a 95% confidence interval of [0.36, 0.42]. Remarkably for all industries, even those with very precise estimates of β, the lower bound of the 95% confidence interval for β was smaller than 0.47, meaning that the model's nontrivial prediction that for any industry β does not exceed 1/2 is consistent with the data.

The value of mean β of 0.39 translates into $\alpha = 0.36$, which is closer to 0 than to 1. In this sense we can say that the original Economic Order Quantity model ($\alpha = 0$) matches the data better than the model with constant marginal cost of trade ($\alpha = 1$). Both of these extreme parameter values are rejected, however, since the confidence interval for β of [0.36, 0.42] translates into a confidence interval of α of [0.28, 0.44].

The resulting marginal cost of trade (inventory and coordination)

$$\left(\frac{\beta}{1-\beta}\right)^{-\beta} \kappa_i^{\beta} \kappa_t^{1-\beta} q^{-\beta}$$

is proportional to $q^{-\beta}$, $\beta \approx 0.39$, which after rounding becomes $q^{-2/5}$. In other words, increasing the quantity by 10% reduces the variable part of the marginal cost of trade by 4%.[54]

[53] Note that this definition implies that the exponent of quantity in the formula for minimized cost above is consistent with the text of Subsection 4.2.2.

[54] In addition, there can be a constant contribution to marginal cost of trade, arising e.g. from per-container

The supplementary material Subsection I.2 provides more details on our estimation. As we discuss there, our estimates are robust to changes in specification (nonlinear storage technology) and are not driven by seasonality patterns that certain industries exhibit or by the choice of the cutoff on the number of firms per industry.

B.4 Supply chains with hold-up

Antràs and Chor (2013, henceforth AC) model the decisions of firms that manufacture complex, multistage products about vertical structure (insourcing v. outsourcing) to address hold-up problems as in, e.g., Grossman and Hart (1986). Firms contributing to critical stages of the production process, where the marginal revenue associated with their contributions is very high, should be insourced to avoid hold-up, while those at more marginal stages of the production process should be outsourced to avoid hold-up by the main firm that discourages quality production. For tractability, they assume a Dixit and Stiglitz (1977) structure, implying that marginal revenue is monotone and thus that either the early or the late stages of production are outsourced, but not both. However, a more natural assumption may be that while marginal revenue rises at early production stages, as the product is first taking shape, it falls at later stages once it is nearly finished and thus its quality is reaching saturation, causing standard downward sloping demand to kick in. This would lead to outsourcing of both early and late stages, an arguably more plausible conclusion. In this application we show how a model exhibiting these features can be formulated and solved as simply as that studied by

A firm produces a final good using a continuum of customized inputs each provided by a different supplier indexed by $j \in [0,1]$. If production proceeds smoothly the *effective* (quality-adjusted) quantity q of the final good is the integral of the quality 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 quality.[55] The lower is j, the further *upstream* a supplier is; that is, the more basic inputs to the good she supplies. However, if production is "disrupted" by the failure of some supplier, $\overline{j} \in [0,1)$, to cooperate, then only the effective quantity accumulated to that point in the chain is available, with all further quality-enhancement impossible. The firm faces an inverse demand function $P(q)$, which need not be decreasing as is a standard inverse demand is, because, for example, consumers may have little willingness-to-pay for a very early stage 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, $MR\left(q(j)\right) q_s(j)$ is therefore split between the firm and supplier j, where $MR = P + P'q$.[56] In particular, the supplier receives a fraction $1 - \beta(j)$ (its bargaining power).

international sea shipping charges. For many industries such as apparel or footwear, such charges are of trivial magnitude compared to the value of the goods.

[55] Here we use notation compatible with the rest of this paper both for consistency and because it simplifies the exposition. The relation to AC's notation 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 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.

[56] 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.

The cost of producing quality $q_s(j)$ is homogeneous across supplies and equal to $C(q_s(j))$, which is assumed convex.[57] Thus the first-order condition of supplier j equates the share of 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 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 $qMC(q)$. Thus 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

[57] 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 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 $q_s\left(j\right) = \theta^\alpha (x\left(j\right))^\alpha$.

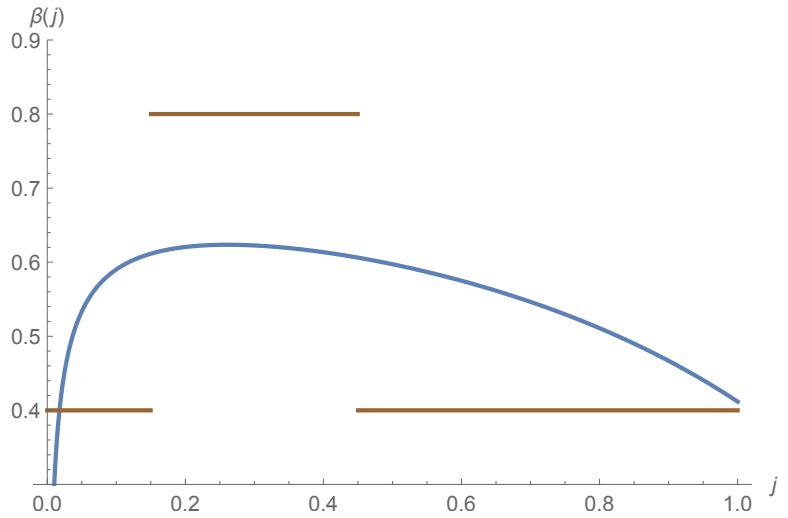


Figure 5: Optimal relaxed and restricted $\beta^\star$ in the AC model when $t = .35, u = .7, \frac{p_{-u}}{mc_{-u}} = -4$.

specification as the special case when $p_{-t} = 0$ and $mc_{-u} = 0$ so that each has constant elasticity.[58] 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 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 numerator and first denominator term in the ratio are positive and the second denominator term is negative. This implies that at small q, where j^t dominates, $\beta^\star$ increases in j, while at large j it decreasing 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 the joint level of p_{-u} and mc_{-u}.

However, for many parameters an inverted U-shape emerges. For example, Figure 5 shows the case when $t = .35, u = .7, p_{-t} = 1.8, \frac{p_{-u}}{mc_{-u}} = -4$. 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 Subsection I.3 of our supplementary material we study the constrained problem using largely closed-form methods for the case when outsourcing gives $\beta_O = .8$ and insourcing gives $\beta_I = .4$. This is illustrated by the lines in Figure 5, which show the constrained optimum. This gives the same qualitative answer as the relaxed problem, unsurprisingly. In Subsection I.3 of our supplementary material we show we can get an even tighter closed-form solution if we use a tractable form with an explicit inverse, which requires a quadratic solution to obtain our intuitive non-monotone cont

B.5 Labor bargaining without commitment

Stole and Zwiebel (1996a,b, henceforth SZ) study a model of wage bargaining where firms employ

[58] 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.

workers mutually at-will and where hiring new workers happens with delay. This leads to "labor hoarding" (viz. over-employment relative to standard monopoly under-employment) in order make each worker more expendable and thus weaken their bargaining position. However, the model has been primarily applied (Helpman et al., 2010; Helpman and Itskhoki, 2010; Helpman et al., 2015) assuming constant elasticity demand and power law technology, which implies that the relative degree of labor hoarding is unaffected by the prevailing state of the economic cycle (viz. the strength of demand relative to the outside option). Because the model is governed by a complex differential equation, most intuitions about the model arise from these special cases. We show that stepping just slightly outside this particular class yields the qualitatively different result that labor hoarding is *counter-cyclical*. We derive this result by mechanically applying our tractable forms and do not fully understand the intuition behind this result. Such insights unguided by intuition are possible because the fairly elaborate equilibrium of the SZ model is simple when framed in terms of the Laplace first-order conditions of a standard monopoly model.

In SZ a firm hires workers, each of whom supplies one unit of labor if employed. 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 the excess of his wage over the outside option $S_w = W(l) - W_0$, which depends on labor l supplied. We assume linear production and thus $l = q$.

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

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

where $MR \equiv P + P'q$. Integrating both of the sides of the equation, imposing the boundary condition that the wage bill shrinks to 0 at $q = 0$ yields an integral equation for the firm's profit that we state explicitly in Subsection I.4 of our supplementary material. The firm's optimal q solves its first-order condition, $\Pi'(q) = 0$.

Let us define (relative) labor hoarding as $h \equiv \frac{q^\star - q^{\star\star}}{q^{\star\star}}$, 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 the first-order condition alluded to above gives a useful condition for h in terms of the equilibrium employment level $q^\star$:

$$MR\left(\frac{q^\star}{1+h}\right) = \frac{(1+\lambda)\int_0^{q^\star} x^{\frac{1}{\lambda}} MR(x)dx}{\lambda\,(q^\star)^{1+\frac{1}{\lambda}}}. \tag{13}$$

Note that this equation involves only a) marginal revenue and b) integrals of it multiplied by a power of q and then divided by one power higher of q. It can easily be shown that the support of the Laplace marginal revenue is preserved by this transformation using arguments related to those we use in the text to illustrate for the extreme cases of large and small λ. Thus Equation 28 has

precisely the same tractability characterization as does the basic monopoly model.

Given the complexity of Equation 28 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 case when $p_0 = 0$. Solving Equation 28 for h yields

$$h = \left(\frac{1+\lambda}{1+\lambda-t\lambda} \right)^{\frac{1}{t}} - 1. \tag{14}$$

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.[59] Instead it appears that the reverse is the case. This shows one advantage of considering an 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 Subsection 2.2 above, 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):

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

Plugging this into Equation (28) 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:

$$h = 1.6 \left(\frac{1 + \sqrt{1 + \frac{1.2 \cdot 10^9}{(100000 - W_0)^2}}}{1 + \sqrt{1 + \frac{1.1 \cdot 10^8}{(100000 - W_0)^2}}}. \right)^{5/2} - 1, \tag{15}$$

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 15 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

[59] 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.

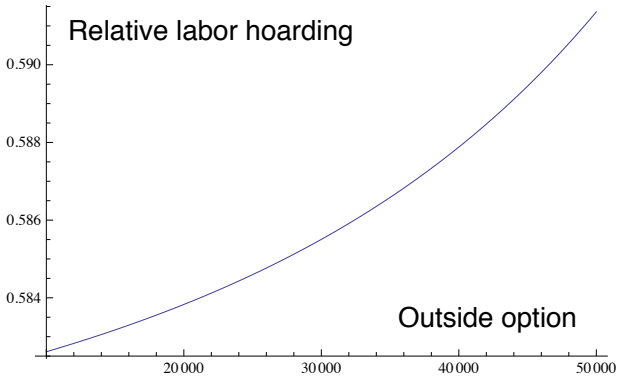


Figure 6: Relative labor hoarding in the SZ model with $\lambda = 1$ and demand given by the approximation for $W_0 \in 10^4 \cdot [1, 5]$

collective bargaining does. Figure 6 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 may of a similar magnitude to cyclic shifts in employment and are thus worth considering.

Supplementary Material

C 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: $q(P) = \left(\frac{P}{\beta}\right)^{-\epsilon} \quad P(q) = \beta q^{-1/\epsilon}$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 a 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 D.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.[60] 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 log-normal 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,

[60] 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 our paper's Section 5. 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 E of this supplementary material.

D 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.

D.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).

D.2 Analytic Evaluation of Inverse Laplace Transform

Mathematical software allows for symbolic inversion of Laplace transform.[61] However, it is often more convenient to evaluate the inverse Laplace transform using more direct methods.

D.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 Section C, which corresponds to inverse demand $P(q) = \sigma(\frac{q}{1-q})^{-1/\gamma}$. This may be written as $P(q) = \sigma q^{-1/\gamma} \times (1-q)^{1/\gamma}$, i.e. a product of a power function and a function that has a well-defined Taylor expansion at $q = 0$:

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

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

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

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 Section C.

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

D.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:[62]

$$f(t) = \frac{1}{2\pi i} \lim_{T\to\infty} \int_{\gamma-iT}^{\gamma+iT} e^{st} f_{\mathcal{L}}(s)\, ds, \text{ where } f_{\mathcal{L}}(s) \equiv \int_0^\infty e^{-st} f(t)\, dt. \tag{16}$$

For the purposes of this paper we did not actually need it. We obtained the Laplace inverse demand function listed in supplementary material Section C by simpler methods.

D.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 16, 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 16 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.

E Demand Forms

E.1 Curvature properties

Table 2 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 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.

[62] 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).

	$\kappa < 1$	$\kappa > 1$	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 (Fr?chet) with shape $\alpha > 1$	
Price-dependent				
Parameter-dependent				
Does not globally satisfy $\kappa < 2$		Type II Extreme Value (Fr?chet) with shape $\alpha < 1$ Weibull with shape $\alpha < 1$ Gamma with shape $\alpha < 1$		

Table 2: A taxonomy of some common demand functions

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

$$\kappa(p) = -\frac{\frac{s^2}{m^2}F''\left(\frac{p-\mu}{m}\right)\left(1 - F\left(\frac{p-\mu}{m}\right)\right)}{\frac{s^2}{m^2}\left[F'\left(\frac{p-\mu}{m}\right)\right]^2} = -\frac{F''\left(\frac{p-\mu}{m}\right)\left(1 - F\left(\frac{p-\mu}{m}\right)\right)}{\left[F'\left(\frac{p-\mu}{m}\right)\right]^2}.$$

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. (Forthcoming), 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., 2013; Quint, 2014). Which conclusion is most appropriate will obviously depend on the empirical context and the views of the analyst.

Proposition 1. *Table 2 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.[63] 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 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 Fr?chet, 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 (Fr?chet) 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:

$$\kappa(x) = \frac{e^{x}(1 - \alpha + x)\Gamma(\alpha, x)}{x^{\alpha}}. \tag{17}$$

   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 $\kappa > 2$.

[63] 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.

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 (Fr?chet) distribution with shape $\alpha > 1$: From the formula above it is easy to show that the derivative of the pass-through rate is

$$\kappa'(x) = x^{-(1+\alpha)}\Big([1+\alpha]\big[x^{2\alpha}(e^{x^{-\alpha}} - 1) - e^{x^{-\alpha}}x^{\alpha}\big] + \alpha e^{x^{-\alpha}}\Big) > 0,$$

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

$$\kappa'(x) = (-x)^{\alpha-1}\alpha^2\left[1 - \alpha + e^{(-x)^{\alpha}}\Big([1-\alpha]\big[(-x)^{\alpha} - 1\big] + [-x]^{2\alpha}\alpha\Big)\right],$$

which has the same sign as

$$1 - \alpha + e^{(-x)^{\alpha}}\Big([1-\alpha]\big[(-x)^{\alpha} - 1\big] + [-x]^{2\alpha}\alpha\Big). \tag{18}$$

Note that the limit of this expression as $x \to 0$ is

$$1 - \alpha - (1 - \alpha) = 0$$

and its derivative is

$$\frac{e^{(-x)^{\alpha}}(-x)^{2\alpha}\alpha\big(1 + \alpha + [-x]^{\alpha}\alpha\big)}{x},$$

which is clearly strictly negative for $x < 0$. Thus Expression 18 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 Fr?chet 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 17 yields:

$$\kappa'(x) = \frac{\alpha - 1 - x + \frac{e^x}{x^\alpha}\big(x^2 - 2x[\alpha-1] + [\alpha-1]\alpha\big)\Gamma(\alpha, x)}{x},$$

which has the same sign as

$$\alpha - 1 - x + \frac{e^x}{x^\alpha}\big(x^2 - 2x[\alpha-1] + [\alpha-1]\alpha\big)\Gamma(\alpha, x), \tag{19}$$

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

$$q(x, \alpha) \equiv \frac{x - (\alpha - 1)}{x^2 + (\alpha - 2x)(\alpha - 1)},$$

$$T(x, \alpha) = \frac{2(\alpha - 1)x}{\big(\alpha^2 + x[2 + x] - \alpha[1 + 2x]\big)^2} > 0$$

for $\alpha > 1, x > 0$. So $\Gamma(\alpha, x) > x^\alpha e^{-x} q(x, \alpha)$. Thus Expression 19 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 Fr?chet distribution is

$$\kappa(x) = \frac{\alpha - e^{x^{-\alpha}}\big(\alpha - x^\alpha[1 + \alpha]\big) - x^\alpha(1 + \alpha)}{\alpha} = \frac{\left(1 - e^{x^{-\alpha}}\right)\left[\alpha - x^\alpha\,(1 + \alpha)\right]}{\alpha}.$$

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

$$\left(1 - e^{x^{-\alpha}}\right)\left[\alpha - x^\alpha\,(1 + \alpha)\right].$$

By the same argument as above with the Fr?chet 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 $p \leq e^{-\frac{a}{b}} < e^{2-\frac{a}{b}}$. 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*

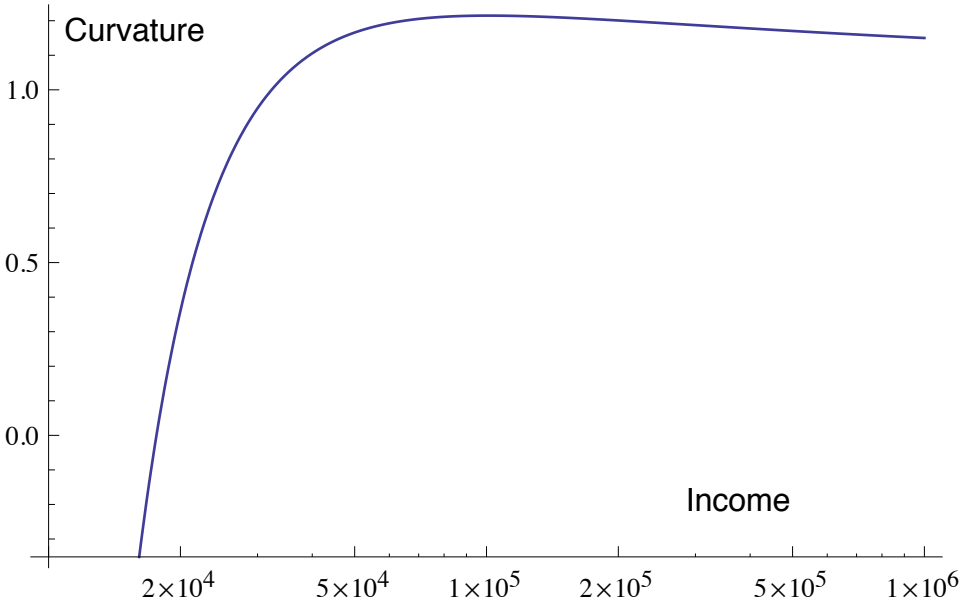


Figure 7: Curvature of a lognormal distribution calibrated to the US income distribution: parameters are $\mu = 10.5$ and $\sigma = .85$.

1. *If* $y \geq \overline{y}(\sigma)$ *then* $\kappa' \leq 0$*, and similarly with strict inequalities or if the directions of the inequalities both reverse.*

2. *If* $y \geq \underline{y}(\sigma)$ *then* $\kappa \geq 1$*, and similarly with strict inequalities or if the directions of the inequalities both reverse.*

Both $\overline{y}$ *and* $\underline{y}$ *are strictly decreasing in* σ*.*

Under the lognormal distribution, 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 7), 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.[64] Thus, in practice,

[64]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.

curvature is closer to flat at the top than significantly declining.

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

$$\kappa\left(p(y)\right) = \frac{(y+\sigma)\left[1-\Phi(y)\right]}{\phi(y)}. \tag{20}$$

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'H?spital'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'H?spital'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 eventual 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

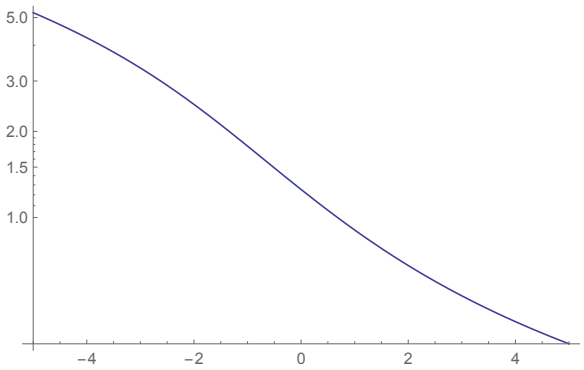


Figure 8: The figure shows the value, in logarithmic scale, of the left-hand side of inequality 21.

threshold structure for κ'. Furthermore, the same logic implies that that the region where $\kappa > 1$ must be 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.

$$\kappa' = \frac{(1-\Phi)\,\phi - (y+\sigma)\phi^2 - (y+\sigma)(1-\Phi)\phi'}{\phi^2} = \frac{1-\Phi-(y+\sigma)\left[\phi - y(1-\Phi)\right]}{\phi} \propto$$

$$1-\Phi-(y+\sigma)\left[\phi - y(1-\Phi)\right] \propto \frac{1-\Phi}{\phi - y(1-\Phi)} - y - \sigma.$$

where the last sign relationship follows by the common inequality that $\phi(y) > y\left[1-\Phi(y)\right]$. Thus $\kappa' > 0$ if and only if

$$\frac{1-\Phi}{\phi - y(1-\Phi)} - y > \sigma. \tag{21}$$

Figure 8 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 21 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 21 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'H?spital's rule yields

$$\lim_{y\to\infty} \frac{(1-\Phi)\,(1+y^2) - y\phi}{\phi - y\,(1-\Phi)} = \lim_{y\to\infty} \frac{-\phi\,(1+y^2) + 2y\,(1-\Phi) - \phi - y\phi'}{\phi' - (1-\Phi) + y\phi} =$$

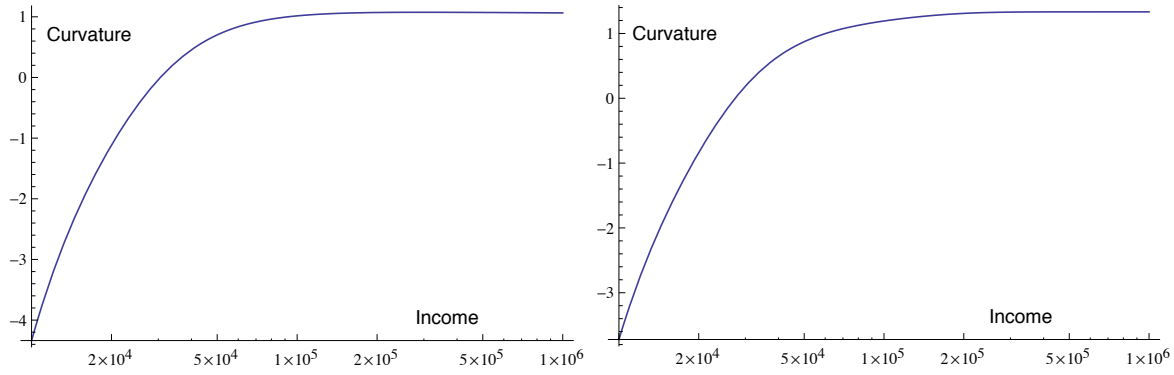


Figure 9: Curvature of the double Pareto-lognormal distribution lognormal under parameters estimated by (Reed, 2003) (left) and by updated by us (right); parameters in the former case are $\alpha = 22.43, \beta = 1.43, \mu = 10.9, \sigma = .45$ ad in the latter case are $\alpha = 3, \beta = 1.43, \mu = 10.9, \sigma = .5$. 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'H?spital'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 9'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 log-normal calibration, except at exorbinate incomes, where the lognormal distribution has thin tails. Thus it actually has a *thinner* tail, except at the very extreme tail, than the log-normal 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 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 9. 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 9 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.

F 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.[65]

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 solutions 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.

G Details of the Proof of the Aggregation Theorem

Details of the Proof of Theorem 3 Here we provide additional details of the proof of Theorem 3 in Appendix A. To fully complete the proof, we need to examine the results of the integral evaluation. 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.

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

(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 $y \to x_0/x, x_1 \to u_1x$ and $x_n \to u_nx$ then leads to the desired result for the integral:

$$\int_0^x x_0^{b-1}\left(1-u_1x_0\right)^{-b_1}\ldots\left(1-u_nx_0\right)^{-b_n} dx_0 = \frac{x^b}{b} {F_D}^{(n)}\left(b, b_1, \ldots, b_n, b+1; u_1x, \ldots, u_nx\right)$$

(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. □

H Details of the Proof of the Discrete Approximaton Theorem

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

$$\sum_{k=1}^{n_T} F(k) = \int_1^{n_T} F(x)\,dx + \mathcal{C}(F) + E_F\left(n_T\right),$$

$$\mathcal{C}(F) = \frac{1}{2}F(1) - \sum_{r=1}^{m} \frac{B_{2r}}{(2r)!}F^{(2r-1)}(1) + \frac{1}{(2m+1)!}\int_1^{\infty} P_{2m+1}(x)F^{(2m-1)}(x)\,dx,$$

$$E_F\left(n_T\right) = \frac{1}{2}F\left(n_T\right) - \sum_{r=1}^{m} \frac{B_{2r}}{(2r)!}F^{(2r-1)}\left(n_T\right) + \frac{1}{(2m+1)!}\int_{n_T}^{\infty} P_{2m+1}(x)F^{(2m-1)}(x)\,dx.$$

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

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

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,

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

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

$$\sum_{t\in T} q^{-t} f(t) = \sum_{k=1}^{n_T} F(k),$$

$$\frac{1}{2}q^{-t_{\min}} f\left(t_{\min}\right) + \frac{1}{2}q^{-t_{\max}} f\left(t_{\max}\right) = \frac{F(1)}{2} + \frac{F\left(n_T\right)}{2},$$

$$\frac{R_2}{\Delta \mathrm{t}} = \sum_{r=1}^{m} \frac{B_{2r}}{(2r)!}\left(F^{(-1+2r)}(1) - F^{(-1+2r)}\left(n_T\right)\right),$$

$$\frac{R_3}{\Delta \mathrm{t}} = -\frac{\Delta \mathrm{t}^{2m}}{(1+2m)!}\int_{t_{\min}}^{t_{\max}} P_{1+2m}(t)h^{(1+2m)}(t)\,dt = -\frac{1}{(2m+1)!}\int_1^n P_{1+2m}(x)F^{(1+2m)}(x)\,dx.$$

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). □

I Details of Applications

To make this appendix self-contained we largely repeat our analysis from the text, but in greater detail and omitting some final results when they are unnecessarily repetitive with the text.

I.1 Monopolistic competition

I.1.1 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:

$$U_\Omega = \int_\Omega u_\omega\left(q_\omega\right)\,d\omega. \tag{22}$$

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 $u\left(q_\omega\right)$ is assumed to be a function of a combination different powers of q_ω . More explicitly, consumer optimization requires that marginal utility of extra spending is equalized across varieties: $u'_\omega\left(q_\omega\right) = \lambda P_\omega$, 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 $P_\omega\left(q_\omega\right) = {}^{u'_\omega(q_\omega)}/_{\lambda}$ and the corresponding revenue $R_\omega\left(q_\omega\right)$ be linear combinations of equally-spaced powers of q_ω:

$$P_\omega\left(q_\omega\right) = \sum_{t\in T} p_{\omega,t}q_\omega^{-t}, \quad R_\omega\left(q_\omega\right) = \sum_{t\in T} p_{\omega,t}q_\omega^{1-t}$$

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 num?raire in a way that keeps $P_\omega\left(q_\omega\right)$ 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.1.2 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 22 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 $MC^\star$.[66] 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).[67] 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.[68] 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.[69] The export FOC is now

[66] 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: $(L(q)+f)R'(q) = R(q)L'(q)$.

[67] As mentioned in the paper, a convenient choice of numéraire allows us to keep the revenue function $R(.)$ independent of economic circumstances.

[68] 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 $MC^\star$.

[69] In a symmetric equilibrium it does not matter how this labor is split between the countries, as long as 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.

$R'(q_x) - wL'_T(q_x) = \tau MC^\star$, 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.1.3 Flexible Melitz model

The Melitz (2003) model is again based on monopolistic competition and assumes 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.[70]

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.[71]

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.[72] In terms of the labor requirement function $L(q;a)$, the firm profit maximization condition and the z*ero cutoff profit condition* are $R'(q) = wL'(q;a)$

[70] For simplicity, consumers do not discount 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.

[71] 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.

[72] 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 $R(q) = r_t q^{-t} + r_u q^{-u}$.

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

$$R'(q) = wL'(q; a), \tag{23}$$

$$R(q_c) = wL(q_c; a_c) + wf. \tag{24}$$

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 (23) and (24) 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 (24): $w = R(q_c)/(L(q_c; a_c) + f)$.

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:

$$w\delta_e f_e = \bar{\Pi} = \int_{q \geq q_c} (R(q) - wL(q; a) - wf)\, dG(a(q)).$$

Here $a(q)$ is the firm's productivity parameter as an explicit function of the optimally chosen quantity q that results from using (23). For Pareto G, and L and R tractable from the point of view of q (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 enters 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 fashion. In a future draft of this paper we will provide a more detailed discussion.

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 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:

$$MC_n = R'(q_n) \tag{25a}$$
$$MC_n = wL'(q_n;a) \tag{25b}$$
$$MC_x = R'(q_d) \tag{25c}$$
$$MC_x = \tfrac{1}{\tau}R'(q_f) - \tfrac{1}{\tau}wL'_T(q_f) \tag{25d}$$
$$MC_x = wL'(q_d+\tau q_f;a) \tag{25e}$$
$$R(q_{1n}) - wL(q_{1n};a_1) = f \tag{25f}$$
$$R(q_{2d}) + R(q_{2f}) - wL(q_{2d}+\tau q_{2f};a_2) - wL_T(q_{2f}) = f + f_x \tag{25g}$$

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 (25a). The corresponding a_1 may be found by solving a linear equation that results from combining (25b) and (25f) in a way that eliminates wages. Wages then may be recovered by substituting back to (25b).

For a chosen MC_{2x}, we can derive q_{2d} from (25c) and q_{2f} from (25d). The value of a_2 is then determined by (25e). We find q_{2n} by solving (25a) and (25b) 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. Fixed cost of exporting f_e is then identified from (25g).

For a given marginal cost, we can find the corresponding quantities and productivity parameters a by a similar method from (25a-25e), 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:

$$w\delta_e f_e = \bar{\Pi} = \int_{y\in S_n} \Pi\left(q_n(y);a_n(y)\right) dG\left(a_n(y)\right) + \int_{y\in S_x} \Pi\left(q_x(y);a_x(y)\right) dG\left(a_x(y)\right),$$

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 $M_e f_e + Mf + M_x f_x + (M - M_x)\bar{L}_n + M_x \bar{L}_x = L_E$, where

$$\bar{L}_n \equiv \tfrac{1}{G(a_1)-G(a_2)} \int_{y\in S_n} L\left(q_n(y);a_n(y)\right) dG\left(a_n(y)\right),\ \bar{L}_x \equiv \tfrac{1}{G(a_2)} \int_{y\in S_x} L\left(q_x(y);a_x(y)\right) dG\left(a_x(y)\right).$$

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.1.4 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 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.[73] The utility is of the form

$$U_\Omega \equiv F\left(U_\Omega^{(1)}, U_\Omega^{(2)}, ..., U_\Omega^{(m)}\right), \quad U_\Omega^{(i)} \equiv \int_\Omega U^{(i,\omega)}\left(q_\omega\right) d\omega.$$

In order to preserve tractability, we assume that $U^{(i,\omega)}\left(q_\omega\right)$ are linear combinations[74] 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 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.

I.2 Details of the Generalized Economic Order Quantity model applied to international shipping

Here we provide additional details of the estimation of the cost parameter $\beta = (1-\alpha)/(2-\alpha)$. To obtain empirical estimates of β, we used a Chinese customs dataset on firm-level monthly shipment data on exports from China to Japan in years 2000 to 2006. We selected firms by requiring that they specialize in one narrowly defined product category (one 8-digit HS code). The exporting firm

[73] 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.

[74] Of course, without loss of generality we could assume that $U^{(i,\omega)}\left(q_\omega\right)$ 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.

$N_{f,\min}$	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.

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. The resulting sample had seventy industries.

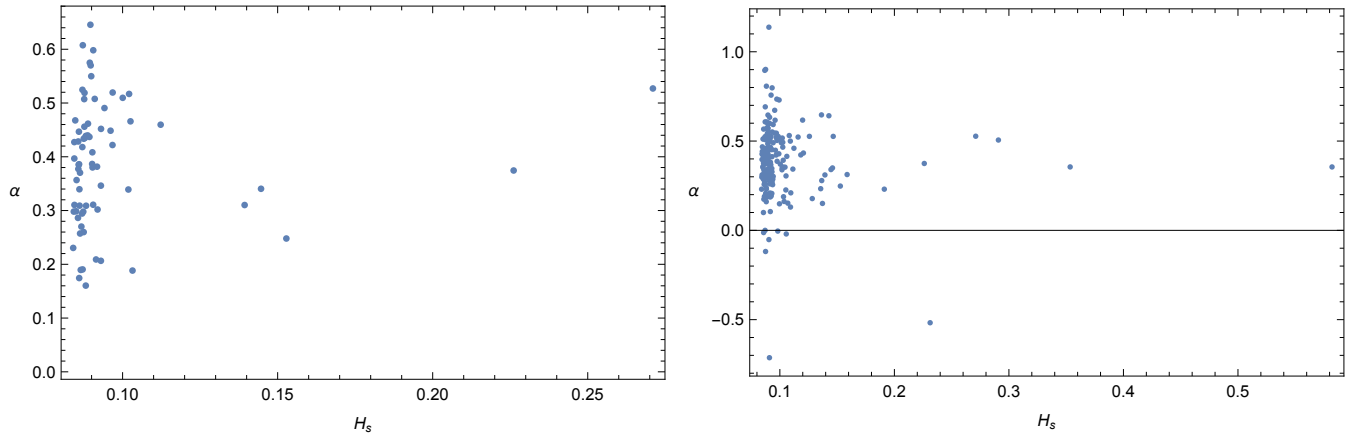
The value of average estimated β could in principle be sensitive to the cutoff on the number of firms per industry. Table 3 summarizes the dependence of the resulting average β on the choice of the cutoff. It turns out that the average β remains roughly the same even for large changes of the cutoff on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

To investigate whether the estimated values of α could be influenced by seasonality patterns, we construct a measure of seasonality of individual industries. We calculate a Herfindahl-like seasonality index based on the shares of trade in individual months of the year, defined as $H_s = \sum_{i=1}^{12} v_i$, 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 α 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 10.

It is also possible to consider a generalized version of our model in which the storage cost is not linearly proportional to the time a typical unit needs to remain in storage, but corresponds to a certain power of it. The model may be solved along the lines of our baseline model and leads to the same conclusion for the power of quantity that describes the non-constant component of the marginal cost as a function of the average quantity shipped per year.

I.3 Antràs-Chor

In this subsection 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 quantity $\hat{q}$. 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 $q(j;\beta) \in [0,\hat{q}]$ and then invert the resulting $q(j;\beta^{\star\star})$ 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



(a) (b)

Figure 10: 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.

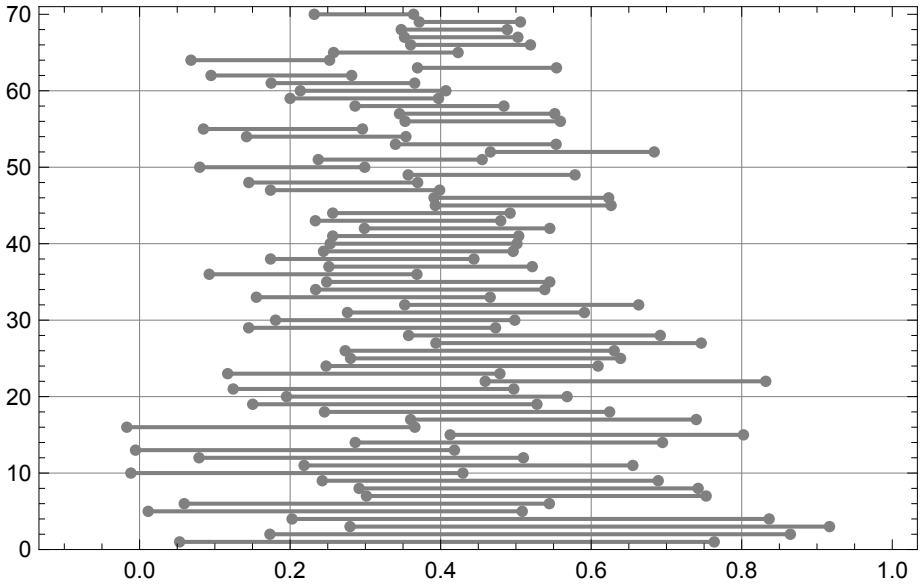


Figure 11: Confidence intervals for the cost exponent α for individual industries at the 95% level. For visualization purposes, the industries are ordered by the standard deviation of α and stacked vertically.

preserve notational economy.

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

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

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

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

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:

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

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

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

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

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

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

$$MR(q) > \frac{\lambda\left[S([1-\beta_O]\, MR(q)) - S([1-\beta_I]\, MR(q))\right]}{[1-\beta_O]\, S([1-\beta_O]\, MR(q)) - [1-\beta_I]\, S([1-\beta_I]\, MR(q))}. \tag{26}$$

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

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

This implies the differential equation $q'(j) = S([1-\beta^{\star\star}(q(j))]\, MR(q^{\star\star}(j)))$ 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 26 is satisfied $B_I(\lambda)$ and on which it is reversed $B_O(\lambda)$.[75] 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 26 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 26 becomes

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

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

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.

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

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

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

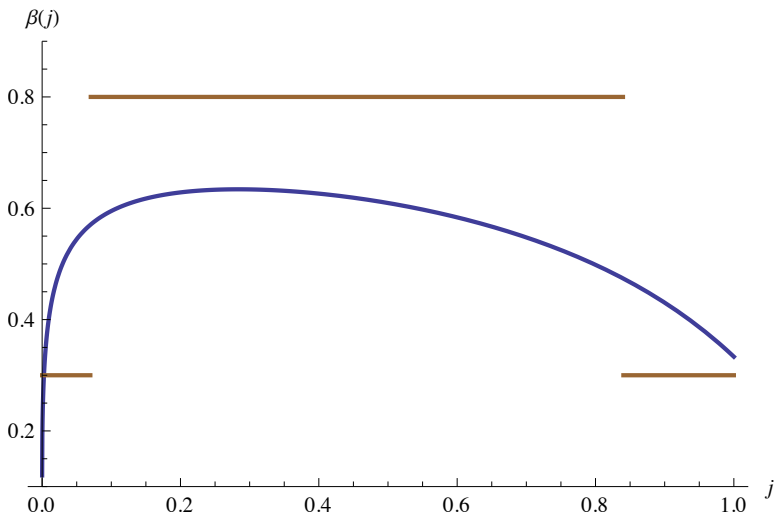


Figure 12: Relaxed and restricted solutions to the AC model when $P(q) = .2+2q^{\frac{1}{2}}-4q$, $MC(q) = \frac{q^{\frac{1}{2}}}{2}$, $\beta_O = .3$ and $\beta_I = .8$.

Between these two roots, in-sourcing is optimal; outside them, outsourcing is optimal.[76]

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 $MR(q_\lambda) = \lambda$ 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 12, 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.

I.4 Stole-Zweibel

In SZ, at the beginning of a period, a firm hires workers, each of whom supplies one unit of labor if employed.[77] 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 relation-

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

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

ship 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 differential equation

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

where $MR \equiv P + P'q$ 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

$$W(q) = q^{-\left(1+\frac{1}{\lambda}\right)} \int_0^q x^{\frac{1}{\lambda}} MR(x)dx + \frac{W_0}{1+\lambda}$$

and thus profits

$$\Pi(q) = P(q)q - q^{-\frac{1}{\lambda}} \int_0^q x^{\frac{1}{\lambda}} MR(x)dx - \frac{W_0}{1+\lambda}.$$

The firm's optimal q solves its first-order condition, $\Pi'(q) = 0$, which, after some algebraic manipulations, is

$$\frac{(1+\lambda) \int_0^q x^{\frac{1}{\lambda}} MR(x)dx}{\lambda q^{1+\frac{1}{\lambda}}} = W_0. \tag{27}$$

Let us define (relative) labor hoarding as $h \equiv \frac{q^\star - q^{\star\star}}{q^{\star\star}}$, 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 (27) gives a useful condition for h in terms of the equilibrium employment level $q^\star$:

$$MR\left(\frac{q^\star}{1+h}\right) = \frac{(1+\lambda) \int_0^{q^\star} x^{\frac{1}{\lambda}} MR(x)dx}{\lambda (q^\star)^{1+\frac{1}{\lambda}}}. \tag{28}$$

Note that this equation, and Equation 27, involves only a) marginal revenue and b) integrals of it multiplied by a power of q and then divided by one power higher of q. 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 in when consumer surplus is calculated. This implies that Equations 27 and 28 have precisely the same tractability characterization as does the basic monopoly model we studied in Section 2 of the paper.[78]

[78]Note that Equation 27 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.