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

Michal Fabinger† E. Glen Weyl‡

August 2018

Abstract

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

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

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

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

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

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

We then find all functions that have such property.

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

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

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

---

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

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

3 A simple economic interpretation would be to identify $F(q)$ with the price $P(q)$ of a good sold by a monopolist, i.e. with the average revenue the firm receives per unit sold, in which case $F(q) + q F'(q) = P(q) + q P'(q)$ is the marginal revenue. The name of the transformation is chosen to be consistent with this and similar examples. The Bulow-Pfeiferer demand class (Bulow and Pfeiferer, 1983) discussed later is also invariant under average-marginal transformations.
requires parameter restrictions involving parameters both from the supply side and the demand side. These restrictions may or may not be approximately satisfied in a particular market. If they are not satisfied, one may be tempted to conclude that our method is not applicable. Most likely this is the reason why economists have not found (or have not attempted to find) the kind of solutions we discuss in our paper.

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

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

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

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

4 Yet way of extending the usefulness of the solutions is to use Taylor series expansions around them, which may be useful for certain types of models.
5 Even though we do need considerable computational power to fit our model to the data, without the tractability of our functional forms the computations would be significantly harder and we would not have attempted to obtain a calibration of world trade flows that we discuss below.
6 This puzzle in various forms has been discussed by many authors; see Disdier and Head (2008) for an overview and Head and Mayer (2013) for an in-depth discussion of the problem.
quantities due to increasing marginal costs of production. At the same time, they find it beneficial to concentrate their exports to a few destinations due to economies of scale in shipping. With this cost structure, even a small cost advantage of a particular destination will be enough to make the firm export there instead of other destinations. If trade costs are slightly smaller for closer destinations, this cost advantage will lead to substantially larger trade flows at smaller distances and substantially smaller trade flows at larger distances.

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

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

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

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

\textsuperscript{7}We discuss alternative mechanisms in Section 4.

\textsuperscript{8}In economics there are many combinatorially difficult problems, and we expect our methods to be useful there.
cyclical patterns. We also discuss applications to imperfectly competitive supply chains, two-sided platforms, selection markets, auctions, and, extensively, monopolistic competition.

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

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

2 Example: Replacing Constant-Elasticity Demand

2.1 Constant-elasticity demand and its flexible replacement

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

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

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

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

\(^11\)Similarly, this flexibility could allow us to get a better match to a distribution of valuations in cases where it differs from the exact income distribution.
Figure 1: Comparing the fit of the best-fit lognormal and the best-fit constant-elasticity form to a double-Pareto lognormal estimation of the 2012 US income distribution, represented as a demand function (reversed quantile function). Dollars at any reversed quantile represent the income of the corresponding individual. On the left is the fit for the full income distribution, while the right shows the upper tail. We used a standard calibration of a double Pareto lognormal proposed by Reed (2003) and used the generalized method of moments to find the constant-elasticity demand function that best fits this throughout the full range of the income distribution. In the upper tail the constant elasticity approximation is a bit better of a fit than the lognormal. However, in the rest of the income distribution its fit is terrible, while the lognormal fits quite well (although in economic models it is harder to work with).

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

Consider the task of representing the empirical income distribution using a corresponding demand function. Constant-elasticity demand fails this purpose, as illustrated in Figure 1. This is because the income distribution is not Pareto but approximately double Pareto lognormal (Reed and Jorgensen, 2004; Toda, 2012, 2017). Working with the double Pareto lognormal distribution (or with the more loosely fitting lognormal distribution) in economic models would be quite difficult.\footnote{See Footnote 14.}

To overcome this difficulty, we propose a functional form that allows for the same basic shape, but leads to calculations almost as easy as for the constant-elasticity form:

\[
P(q) = m - ma_- \left( \frac{q}{q_0} \right)^{-b} - ma_+ \left( \frac{q}{q_0} \right)^b, \quad a_- \equiv 1 - a_+ .
\]

A set of parameter values that matches the income distribution (for the US in 2012) very well is \( a_- = -1/2, a_+ = 5/2, b = 2/5 \). We obtained these values by a generalized-method-of-moments fit and rounding the results. The match is illustrated in Figure 2.
Figure 2: Comparing the fit of the best-fit lognormal and the best-fit quadratically solvable form to a double-Pareto lognormal estimation of the US Income distribution, represented as a demand function (reversed quantile function). Dollars at any reversed quantile represent the income of the corresponding individual. On the left is the fit for the full income distribution, while the right shows the upper tail.

2.2 Bias in technological progress

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

When the private sector decides which products to develop, it chooses profit-making products, not necessarily those products that create the greatest social value. This bias in technological progress depends on the discrepancy between private and social gains. Kremer and Snyder (2015, 2017) consider the fraction of the social gains from creating a new product that may be appropriated by a monopolist\(^1\), referred to as the appropriability ratio, and show that the maximal fraction of potential surplus that may be lost due to imperfect appropriability is equal to one minus this appropriability ratio. They compare different demand functions since they lead to different bias in research and development, but always assume no costs. Here we assume a fixed demand function and consider biases at different levels of marginal production cost. We walk quite didactically through the process of solving the model in order to illustrate the source of the tractability of the constant-elasticity form and why it carries over to our proposed generalized form but not to the lognormal distribution form. We then follow Kremer and Snyder (2017)’s argument that a sensible demand function is one matching the world income distribution and use this as motivation for using our form to study the impact of cost on the appropriability ratio, which is very different under our form than under constant elasticity.

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

---

\(^1\)Budish, Roin and Williams (2015) studied this problem recently in a different context.
\(-abq^{-b}\), which has the same form as \(P(q)\), just a different multiplicative constant out front. For this reason, the marginal revenue has the same form as well: \(\text{MR}(q) = a(1 - b)q^{-b}\). The monopolist optimally equates it to the marginal cost, so the optimal quantity may be determined by solving the linear equation \(a(1 - b)x = c\) with \(x \equiv q^{-b}\), yielding \(q = (a(1 - b)/c)^{1/b}\). From this it follows by substitution that the firm’s absolute markup is \(\overline{PS} \equiv PS/q = cb/(1 - b)\), where PS stands for the producer surplus. Furthermore, the average consumer surplus also has the same form as \(P(q)\), differing only by a multiplicative constant:

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

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 \(\overline{PS}/(\overline{PS} + \overline{CS}) = (1 - b)/(2 - b)\), 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.

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

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

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

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

If we introduce the notation

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

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

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

This leads to the closed-form solution

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

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

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

The appropriability ratio is then

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

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

\[
\frac{\overline{PS}}{\overline{PS} + \overline{CS}} = \frac{21 + 105 (q/q_0)^{4/5}}{56 + 180 (q/q_0)^{4/5}}.
\]

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

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

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

3.1 Form preservation under the average-marginal transformation

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

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

Obviously, if $C$ 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 $(0, \bar{q})$ for some $\bar{q}$ 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 $F \rightarrow aF + bqF'$. 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 $a_1e^{-a_2q}$, where $a_1$ and $a_2$ are continuously varying real numbers is two-dimensional, while $a_1e^{-a_2q^2}q^{-a_3}$ with continuously varying real $a_1$, $a_2$, and $a_3$ is three-dimensional.\(^{(16)}\)

---

\(^{(15)}\)Note that any form-preserving class is also form-preserving under multiple applications of operators of this type.

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

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

\[
\begin{align*}
\text{log } q & \quad a_{jk} q^{-t_j}, \quad a_{jk} = 0, 1, \ldots, n^{(1)}_j, \quad j = 1, 2, \ldots, N_1, \\
\text{log } q \cos (\tilde{t}_j \text{ log } q) & \quad b_{jk} q^{-\tilde{t}_j}, \quad b_{jk} = 0, 1, \ldots, n^{(2)}_j, \quad j = 1, 2, \ldots, N_2, \\
\text{log } q \sin (\tilde{t}_j \text{ log } q) & \quad c_{jk} q^{-\tilde{t}_j}, \quad c_{jk} = 0, 1, \ldots, n^{(2)}_j, \quad j = 1, 2, \ldots, N_2,
\end{align*}
\]

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

\[
q^{-t_j}, \quad q^{-t_j} \text{ log } q, \quad q^{-t_j} (\text{log } q)^2, \quad \ldots, q^{-t_j} (\text{log } q)^{n_j}, \quad j = 1, 2, \ldots, N_1.
\]

The proof is provided in Appendix A.

3.2 Tractability

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

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

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

\(^{17}\)Of course, in other contexts the word “tractability” may have other meanings that are also useful. We specify below what we mean by “tractability” in this paper.

\(^{18}\)Of course, the notion of “tractability” and “closed-form solutions” is subjective to some extent. Equations whose solutions may be expressed in terms of functions that are familiar enough are often said to have closed-form solutions. That does not imply, however, that such notion is meaningless. Familiar functions are easier to work with for researchers thanks to existing intuition, as well as thanks to their implementation in symbolic or numerical software. In this paper we made definite choices to resolve the terminological ambiguity.
using power functions and a solution to a polynomial equation of degree no greater than \( k \). While this definition eliminates many other functions with known solutions, it does a good job capturing existing forms that are widely considered tractable while allowing an extension to richer forms in a pragmatic manner given the ease with which polynomial equations can be solved both analytically and computationally (Kubler and Schmedders, 2010).

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

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

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

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

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

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

\(^{21}\)The BP demand, defined below, gives constant pass-through rate of specific taxes to monopolist’s prices only in the case of a constant marginal cost. For this reason, we prefer to use the term Bulow-Pfleiderer (BP) demand, instead of the frequently used term “constant pass-through demand”.

\(^{22}\)There are a few cases not nested in the forms of Theorem 1 for which the firm’s first-order condition may be solved. Hyperbolic demand curves used by Simonovska (2015) are one of them. Cases where the solution is in terms of the Lambert W function are the exponential utility function of Behrens and Murata (2007, 2012) and single-product versions of the Almost Ideal Demand System and of translog demand.
Figure 3: Example of a bell-shaped-distribution-generated demand and U-shaped cost curve contributing to equilibrium conditions that can be solved linearly: \( P(q) = 3(q^{-0.3} - q^{10}) \) and \( MC(q) = q^{-0.3} + 10q^{10} \).

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

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

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

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

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

---

23Mrázová and Neary (2014) studied the properties such bi-power form applied to inverse demand functions in combination with constant marginal cost. Their goal was not to obtain closed-form solutions.
Form \( F(q) = f_0 + f_{-1}q \) | Tractability properties | Flexibility | Special cases | Historical notes
--- | --- | --- | --- | ---
Linearly tractable | Linear<br>MC | Constant<br>MC | Farrell and Shapiro (1990)
Linearly tractable | Any constant<br>pass-through | Linear<br>Constant<br>elasticity<br>Exponential | BP<br>constant pass-through demand
Linearly tractable | Demand generated by<br>bell-shaped distribution<br>Constant<br>pass-through<br>cost | BP | Mrázová and Neary (2014)<br>bi-power demand
Quadratically tractable | Demand generated by<br>bell-shaped distribution<br>U-shaped<br>cost | BP | Fabinger and Weyl (2012)<br>APT demand

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

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

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

\[ q + 1 + q^{-1/2} = 0. \]

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

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

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

\[ q^{1/2} + 1 + q^{-1/2} = 0 \]

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

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

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

3.3 Aggregation over heterogeneous firms

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

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

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

While this result is closely related to our general theory and our other applications (in particular, because this aggregation is possible when the relevant variables have our proposed forms), there are also a few differences worth noting. First, aggregation is still possible when the heterogeneous component $MC_1(q)$ of marginal cost is “shifted” (in the exponent space) by a uniform multiplicative power factor relative to $MR(q) - MC_0(q)$. This corresponds to the “possibly differing $\beta$” in the statement of the theorem. Second, our results here are about aggregation, not solution, and the resulting functions are not, therefore, solutions to polynomial equations but rather various functions that may be exotic to some economists, but are widely used in mathematics and related applied fields. Finally, as the complexity of the forms rises, it is the complexity of these functions that rises.
Closed-form aggregation is useful for at least three reasons. First of all, in the simplest cases the resulting aggregation integrals are just polynomial functions, which means that at the aggregate level the economic equations are relatively simple. Second, when the aggregation integrals lead to commonly used special functions, these are likely to be implemented in numerical software of the researcher’s choice. The researcher gets a fast and numerically reliable implementation of these functions and their derivatives without spending time on approximation methods. Third, it is possible to take advantage of the properties of these functions that have been studied in the mathematical literature.

3.4 Interpolation between solutions

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

Let us illustrate this approach with a toy example that is not computationally intensive. Consider a monopolistic firm with marginal revenue \( MR(q) = MR_0 q^{b_R}, \) \( b_R > 0 \) and marginal cost \( MC(q) = MC_0 + MC_1 q^{-bc}, \) \( MC_1 > 0. \) After the substitution \( x = q^{-b_R}, \) the firm’s first-order condition becomes \( MC_0 + MC_1 x^b = MR_0 x, \) with \( b \equiv bC/bR. \) This equation admits closed-form solutions by the method of radicals for \( b \in \{-3, -2, -1, -\frac{1}{2}, -\frac{1}{3}, 0, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1, 2, 3, 4\}. \) The second-order condition is satisfied only for \( b < 1, \) so we restrict our attention to the first 9 values. For illustration, consider the simple goal of finding numerical values of \( q \) for \( b \) between these points. Instead of solving the first-order conditions by usual numerical methods, we may interpolate between the closed-form solutions, say, using cubic splines. Figure 4 shows the result of such interpolation, as well as the true solutions to the first-order condition, for \( MC_0 = MC_1 = MR_0 = 1. \) The agreement is extremely good, with average absolute value of relative deviations equal to 0.00013 and maximum absolute value of relative deviations equal to 0.00056.\(^{24}\)

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

\(^{24}\)The corresponding values for Mathematica’s Hermite polynomial interpolation were 0.00069 and 0.0021.
Figure 4: Comparison of an interpolation between analytic solutions and the correct values. The blue dots represent analytic solutions. The blue solid line corresponds to an interpolation using cubic splines. The green dashed line represents correct values.

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

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

4 World Trade

4.1 Overview

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

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

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

4.2 Computational considerations

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

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

25PyTorch is an open source software framework developed by Facebook primarily for deep learning in artificial neural networks. Its first version was released in 2017.
4.3 Firm-level economies of scale in shipping: a generalized Economic Order Quantity model

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

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

To gain empirical insight into the scale economies of international trade, we estimate a model of optimal shipping frequency using monthly international shipment data. Our approach generalizes the classic Economic Order Quantity model of Ford W. Harris, which is widely taught in operations management courses in business schools and applied by logistics planning managers in corporations.\textsuperscript{28} Consider a firm that produces a single good in one country and wishes to ship to a different country quantity \( q \) per year, on average. The firm faces a trade-off 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

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

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

\textsuperscript{28}Despite not appearing in the international trade literature, the Economic Order Quantity (EOQ) model (Harris, 1913) is perhaps the most classical model of trade costs in the operations research literature and is regularly taught to business students as a method of optimizing their inventory decisions; see e.g. Cárdenas-Barrón et al. (2014) for highlights of its importance. Judging from the absence of citations, the academic international trade community is largely unaware of Harris’ publication. When fixed costs per shipment are included in international trade models, they are incorporated in theoretical models with different structures. Those models are similar in spirit, but do not strictly speaking contain the EOQ model or its generalizations (Kropf and Sauré, 2014; Hornok and Koren, 2015a). These papers provide very useful insights into shipment frequency issues, and so does the purely empirical paper Hornok and Koren (2015b). Note also that economies of scale in shipping were studied by Anderson et al. (2014) and Forslid and Okubo (2016), but those approaches were not based on shipping frequency and in the former case involved external (i.e., not within-firm) economies of scale.
storage. If the size of each shipment is $q_s$, then $T$, in turn, is linearly proportional to $q_s/q$, implying $C_i = \kappa_i q_s$, for some constant $\kappa_i$. The coordination cost $C_s$ of each shipment is proportional to its size: $C_s = \kappa_s q_s^2$, $\gamma \in [0,1)$. (In addition, we could assume an additional term proportional to $q_s$, but this would not affect the optimal choice of $q_s$ for given $q$.) The resulting average annual coordination cost is $C_i = C_s q / q_s = \kappa_i q q_s^{2-\gamma}$. Minimizing the sum of the inventory cost and the coordination cost leads to the optimal choice $q_s = \left(q(1-\alpha)\kappa_i\kappa_s^{-1}\right)^{1/2}$, the minimized value (2 $\gamma$)$(1-\gamma)^{-1/2} \kappa_i^{-1/2} \kappa_s^{-1/2} q^{-1/2} \gamma$, and the optimal frequency of shipping $f_s = q / q_s$ equal to $f_s = (1-\alpha)^{-1/2} \kappa_i^{1/2} \kappa_s^{-1/2} q^{1-\gamma} / q^{1-\gamma}$.

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

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

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

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

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

\textsuperscript{30} The confidence interval corresponds to a simple statistical model in which $\beta$ for different industries is drawn from a normal distribution.
arises from per-shipment coordination costs and from inventory costs with optimally chosen shipping frequency. In the rest of this section, we use the notation \( \nu_{LT} \) for what was \( 1 - \beta = 1/(2 - \gamma) \) here, and set \( \nu_{LT} = 0.6 \).

### 4.4 Export quantity determination

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

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

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

The firm’s first-order condition is

\[
R'(q_f) - \tau MC - wL'_T(q_f) = 0 \implies -\frac{w \kappa_{LT}}{\tau} q_f^{-\frac{2}{5}} + \frac{\kappa_R}{\tau} q_f^{-\frac{1}{5}} - MC = 0.
\]

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

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

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

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

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

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

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

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

### 4.5 Increasing marginal cost of production

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

\textsuperscript{31}If exporting leads to zero profit just like not exporting, we specify that the firm chooses not to export.

\textsuperscript{32}The restaurant industry is an obvious example: few people would associate chain restaurants with outstanding culinary experience. Another fairly obvious example is the automobile industry: there are many automakers in the world, each having a relatively small market share, very stable over time, even though cars produced by different automakers are highly substitutable from customers’ perspective. With constant marginal costs of production this would require a remarkably small dispersion of marginal costs across firms, which is especially hard to rationalize given the large observed fluctuations of currency exchange rates. Also, the increasing nature of marginal costs of production was one of the reasons why socialist economies were unsuccessful: state-controlled monopolies avoid duplication of effort in product design and other fixed costs of production, yet they suffer from severe agency problems that private sector competition can mitigate. Although here we emphasize increasing marginal costs of production in the long term, they are also interesting at short time scales; see Almunia, Antràs, Lopez-Rodriguez and Morales (2018) and references therein.
Issues of this kind are the subject of interest to vast literature within organizational economics, which includes Williamson (1967), Calvo and Wellisz (1978), and Tirole (1986).\footnote{Oliver E. Williamson’s Nobel lecture (Williamson (2009)) provides an excellent, compact discussion of the many things that may go wrong in a large organization. For a related discussion, see Tirole (1988).}

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

4.6 Model setup

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

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

Consider a firm located in country \( k \) and identified by an index \( i \). In order to produce a quantity \( q_i \), the firm needs to pay a variable cost of
\[
\frac{1}{1+\alpha}\kappa_{C,i}w_k^{1+\alpha},
\]
where \( w_k \) is the competitive wage the firm’s country \( k \) and \( \kappa_{C,i} \) is a positive constant that depends on the firm. Importantly, the constant \( \alpha \) determines how quickly marginal costs increase when any firm decides to scale up production; it is the elasticity of the marginal cost of production with respect to quantity. In addition to the variable cost, there is a fixed cost \( f_o \) of operation and a fixed cost \( f_x \) of exporting to a destination country \( k_d \), expressed in units of domestic labor.\footnote{In general, we can allow for country-dependence of these costs: \( f_{o,k} \) and \( f_{x,k,k_d} \). We chose to make them country-independent for simplicity, not for tractability or computational feasibility reasons.}

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

Trade costs have two components. The first corresponds to standard iceberg trade costs: in order of one unit of the good to arrive in the destination country \( k_d \), \( \tau_{k,k_d} \) units need to be shipped.\footnote{Including also per-unit trade costs would not affect the computational feasibility of the model.}

The second component requires using an amount of labor given by
\[
L_{T,k,k_d}(q) = \nu_{LT}^{1/\nu_{LT}}K_{LT,k,k_d}^{\nu_{LT}},
\]
where we set \( \nu_{LT} = \frac{3}{5} \) to be consistent with the empirical value, as in Subsection 4.4.\footnote{The cost \( L_{T,k,k_d}(q) \) is associated with coordination/shipment preparation tasks and with inventory costs. Its form is motivated by the empirical results of Subsection 4.3.}

Consumers in each country have a CES utility function
\[
U = \left( \frac{1}{\sigma} \int q^{1-\frac{1}{\sigma}} d\omega \right)^{\frac{1}{\sigma-1}}
\]
that depends on the
quantity $q_\omega$ of each variety $\omega$ consumed. As in Subsection 4.4, we set the elasticity of substitution $\sigma$ equal to 5, which is consistent with the typical range in the existing empirical literature of about 4 to 8. This exact choice is motivated by analytic tractability. Each country $k$ has an endowment of labor $L_{E,k}$, which is supplied at a country-specific competitive wage rate $w_k$ mentioned above.

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

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

4.7 The exporting firm’s problem

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

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

$$R_{k_d}'(q_{f,i,k_d}) = \tau_{k,k_d}MC_i(q_i) + w_kL'_{T,k,k_d}(q_{f,i,k_d}) \Rightarrow \frac{\kappa_{R,k_d}}{\tau_{k,k_d}} q_{f,i,k_d}^{\frac{1}{2}} = MC_i(q_i) + \frac{w_k\kappa_{LT,k,k_d}}{\tau_{k,k_d}} q_{f,i,k_d}^{\frac{2}{5}},$$

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

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

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

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

---

37 The model has no explicit discounting of future utility, but $\delta_e$ plays a role similar to a discount rate.

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

4.8 Solution strategy

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

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

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

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

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

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

\(^{43}\)We tried several accelerated and non-accelerated gradient descent algorithms. Adam performed the best.
case) and backpropagation.\textsuperscript{44, 45}

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

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

4.9 Fitting the model

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

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

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

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

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

\textsuperscript{46}The value of 1.05 for the Pareto index of the firm size distribution has empirical support in Aoyama et al. (2010), at least for the advanced economies studied there. In our open-economy model, there is no simple closed-form expression for the firm size (revenue) as a function of firm productivity. For this reason we use a formula that would hold exactly for closed economies, as well as in the absence of trade costs for the world. Simple algebra shows that the required value of the productivity Pareto index is $\mu_R(\sigma - 1)/(1 + \sigma \alpha)$ and we use this value. The calibration results
the model; any real-world overall firm productivity differences across countries are represented by adjustments to the countries’ effective labor endowments.

For computational purposes we discretize the productivity distribution to \( N_p = 20 \) discrete values, each representing the same probability mass.\(^{47}\)

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

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

### 4.10 Elasticity of the marginal cost of production

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

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

---

\( \begin{array}{|c|c|c|c|c|c|c|}
\hline
N_c & 100 & N_p & 20 & N_v & 10 & \alpha & \text{multiple values} \\
\hline
\sigma & 5 & \nu_R & 0.8 & \nu_L & 0.6 & \mu_R & 1.05 \\
\hline
\delta_e & 0.05 & f_e & 0.1 & f_x & 10^{-5} & \tau & 1.05 \\
\hline
\end{array} \)

Table 2: Calibration parameter values

---

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

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

\(^{49}\)More precisely, we use the generalized method of moments to fit functions of the form \( c_0 (\text{rank})^{-c_1} + c_2 \).
firms in the data that choose to export to the given destination. We also computed such best-fit curve for the data. The results are intuitive: For smaller \( \alpha \), the difference between the median (log) size of firms exporting to unpopular destinations and to popular destinations is larger because in this case the most productive firms will dominate world trade, and if a less productive firm decides to export at all, it will choose a few of the popular destinations.

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

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

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

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

matches well the gravity equation implied by the fitted model\(^{52}\)

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

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

\(^{51}\)This is for 30 largest economies. For 100 economies the results would be more noisy.

\(^{52}\)Here we used \( \alpha = 0.225 \).
Figure 5: Median revenue of Chinese exporting firm (base-10 log scale) by destination country for different values of $\alpha$ and for the observed values. For the first export destination (United States), the median log revenue is normalized to 0 to make visual comparisons easier. The top figure corresponds to all 99 export destinations in the model, while the bottom figure corresponds to the top third by export popularity.

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

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

We see that trade flows decrease rapidly with distance despite only a very mild increase of trade cost prefactors $\kappa_{LT}$ with distance. Although this may look surprising at first sight, there is clear intuition for this phenomenon: Due to increasing marginal costs of production, firms effectively have only a limited output to sell and due to economies of scale in shipping, they need to concentrate their exports to only a few countries. They choose close countries because the trade costs are slightly lower, which leads to strong effects for the decrease of trade with distance.\footnote{We briefly discuss related literature and mechanisms in Appendix C.2.}
From the histogram in Figure 6 we see that the dispersion of $\kappa_{LT}$ is very small, which is only possible with a very small dependence on distance. This small dispersion is very much consistent with sea shipping over large distances being only mildly more expensive than over short distances. This provides a very natural resolution to the trade cost puzzle in the international trade literature.

### 4.12 Choice of export destinations

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

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

---

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

\(^{55}\)Although for identical firms this would be impossible if we introduce differences between the firms, there are other phenomena that may play a role. We briefly discuss them in Appendix C.3.
otherwise. With increasing marginal costs of production, a destination that is profitable for one firm may not be profitable for another identical firm, if that firm already serves other locations. Of course, if there were no economies of scale in shipping (and no significant export entry fixed costs), firms would dilute their exports over more destinations and would not face a combinatorial discrete choice problem. In that case, two identical firms would serve the same destinations, unless, again, there were firm-destination specific shocks. For this reason, both increasing marginal costs of production and economies of scale in shipping are crucial for our model’s ability to resolve the export destination choice puzzle.

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

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

4.13 Implications for modeling international trade

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

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

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

56 Similarly, the mechanism may help explain why in the long run trade liberalization can have dramatic effects on trade flows, as for example in the case of the 2001 US-Vietnam trade liberalization; see McCaig and Pavcnik (2018).
Figure 7: The top map highlights export destinations of two Chinese firms in the fitted model. Specifically, for two identical firms the map shows destinations to which both firms export (yellow), destinations to which only firm 1 exports (red), destinations to which only firm 2 exports (purple), and the country of origin (green). The bottom map shows similar information for two firms from the Czech Republic.

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

\textsuperscript{57}For example, for welfare consequences we can no longer use simple, elegant formulas such as those derived by Arkolakis, Costinot and Rodríguez-Clare (2012).
5 Breadth of Application

5.1 Overview of applications

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

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

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

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

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

5.1.2 Labor bargaining without commitment (Stole and Zwiebel)

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

58 Using our transformed variables would have saved at least 10 pages of the original paper Antràs and Chor (2013). But of course, relative to these authors we have the benefit of hindsight.
While this model appears to differ from previous examples we considered, as it is not a straightforward monopsony model, we show that behavior under the Stole-Zwiebel model corresponds to a “partial” application of the marginal-average transformation (“partial monopolization”) and thus remains tractable under our forms. Thus it is common to use standard, form-preserving tractable forms to analyze this model, especially constant-elasticity. The downside of the assumption is that interesting effects are suppressed: the overemployment ratio (ratio of actual employment and employment under competitive labor markets) is a constant independent of economic conditions.

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

5.1.3 Imperfectly competitive supply chains

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

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

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

5.1.5 Auction Theory

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

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

5.1.6 Selection markets

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

5.1.7 Monopolistic competition

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

6 General Approximation and the Laplace-Log Transform

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

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

6.1 The Laplace-log transform and arbitrary approximation

Under quite general conditions, univariate equilibrium conditions may be expressed as linear combinations of average-marginal form-preserving functions. To make this statement precise, we focus on the demand side here and write an inverse demand curve of interest as \( P(q) = U'(q) \), where \( U(q) \) is a function primitive to \( P(q) \). We assume that \( P(q) \) is non-increasing, which implies that such primitive function exists. Depending on the model of choice, \( U(q) \) may or may not be proportional to the utility of an agent, but to keep the terminology simple, here we refer to \( U(q) \) as the utility.\(^{59}\)

Even though we explicitly discuss the demand side here, the mathematical theorems below apply to the cost side as well, with a straightforward reinterpretation.

We observe that virtually all shapes of demand functions that are useful in economics may be associated with a utility function of the form\(^{60}\)

\[
U(q) = \int_{-\infty}^{0} u(t) q^{-t} dt,
\]

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 \( s = \log(q) \), and for this reason, we refer to \( u(t) \) as the inverse Laplace-log transform of \( U(q) \).\(^{61,62}\) At the same time, the integral may

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

\(^{60}\)The Laplace-log representation (2) of a given utility function \( U(q) \) exists under various conditions. Theorem 18b in Section VII.18 of Widder (1941) states general necessary and sufficient conditions on \( U(q) \) for the existence of \( u_1(t) \) 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 \( u_1(t) \) exists and has certain properties, such as being of bounded variation, nondecreasing, or belonging to the functional space \( L^p \). Additional conditions may be found in Chapter 2 of the book by Arendt et al. (2011), which contains recent developments in the theory. In situations when utility unbounded below is desired, e.g. for constant demand elasticity smaller than 1, we can depart from (2) and instead use the bilateral specification \( U(q) = \int_{-\infty}^{\infty} u(t) q^{-t} dt\). However this generalization requires the use of more technically involved bilateral Laplace transforms and thus we do not discuss it in greater detail here, though analogous results are available on request.

\(^{61}\)Our use of \( t \) for exponents throughout the text and our use of \( s = \log(q) \) here match the standard notation in the literature on Laplace transforms.

\(^{62}\)After an extensive literature search of hundreds of articles and talking to numerous economists, including highly accomplished econometricians, we concluded that this is almost certainly the first time (inverse) Laplace transform in log quantity is used in the economic literature. Note, however, that a different transform, namely (inverse) Laplace transform in quantity, as opposed to log quantity, has been used in economics. These transforms have different properties and should not be confused. Note also that the way we use Laplace transform is different from, say, engineering fields in the sense that, because of the additional logarithm, functions of main interest for us in economics typically would not be of interest in engineering, and vice versa. For this reason, books containing detailed tables of Laplace transform were not of help to us. Except for trivial cases, we needed to derive the transforms listed in Supplementary Material E by ourselves.
be thought of as expressing $U(q)$ as a linear combination of form-preserving functions of Theorem 1.

**Technical Clarification (Integral Definition).**\(^{63}\) Here we define the integral (2) to be the Riemann-Stieltjes integral

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

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

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

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

**Theorem 4. (Laplace-log Transform with Riemann-Stieltjes Integrals)**

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

Note that part D of this theorem is a simple consequence of the Weierstrass approximation theorem.\(^{66}\) The reader may ask why we do not instead work simply with polynomials in $q$ and use

---

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

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

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

\(^{66}\)There is also a related, more powerful theorem, the Müntz-Szász theorem. Barnett and Jonas (1983) use it to propose to write direct demand as Müntz-Szász polynomials of prices. Here we write inverse demand as polynomials of powers of quantities (times possibly another power of quantity), but the same logic would apply here: we could use Müntz-Szász polynomials.
them as approximations. Even though this would be possible in principle, it would not be practical. This is because in economics we often need flexibility in the $q \to 0_+$ limit behavior of the inverse demand function. With any (finite-order) polynomial, we would always get finite $\lim_{q \to 0_+} P(q)$, i.e., a choke price; to allow for $\lim_{q \to 0_+} P(q) = \infty$, we could not stay within a finite-order approximation.

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

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

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

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

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

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

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

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

38
where $T \equiv \{t_{\text{min}}, t_{\text{min}} + \Delta t, \ldots, t_{\text{max}}\}$ is an evenly spaced grid with at least two points, $m$ is an integer such that $f$ is $(2m+1)$-times continuously differentiable on $[t_{\text{min}}, t_{\text{max}}]$ 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 $\int_{-\infty}^{t_{\text{max}}} q^{-t} f(t) \, dt$ and $\int_{t_{\text{min}}}^{t_{\text{max}}} q^{-t} f(t) \, dt$, and can be made very small, since $\int_{-\infty}^{t_{\text{min}}} q^{-t} f(t) \, dt = \int_{t_{\text{min}}}^{0} q^{-t} f(t + t_{\text{min}}) \, dt$, which is exponentially suppressed for $t_{\text{min}}$ 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 $t_{\text{min}}$ and $t_{\text{max}}$:

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

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

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

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

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

6.2 Complete monotonicity and pass-through behavior

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

\footnote{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.}

\footnote{As mentioned above, the validity of such approximations may be proved along the lines of the proof given here.}

39
demand function \( p(t) \) is positive or negative. Choosing a few evenly spaced mass points with a similar positivity/negativity pattern is then likely to lead to a tractable approximation to the original inverse demand function that has similar qualitative properties.

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

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

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

**Definition 5. (Complete Monotonicity of the Demand Specification)**\(^72\)

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

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

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

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

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

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

**Theorem 9. (Complete Monotonicity of Demand Specification)** The following demand functions satisfy the complete monotonicity criterion:

1. 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, 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:

1. 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 ($\epsilon > 1$), gamma distribution ($\alpha < 1$), Singh-Maddala distribution ($a < 1$), Tukey lambda distribution ($\lambda > 1$), Wakeby distribution ($\beta < 1$), generalized Pareto distribution ($\gamma > 1$).

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

---

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

---

41
7 Conclusion

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

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

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

References

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

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

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

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

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

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

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

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

_ and Elhanan Helpman, “Global Sourcing,” Journal of Political Economy, 2004, 112 (3), 552–580.
and _, “Contractual Frictions and Global Sourcing,” in Elhanan Helpman, Dalia Marin, and Thierry Verdier, eds., The Organization of Firms in a Global Economy, Cambridge, MA: Harvard University Press, 2008.

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

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

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

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

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

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

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

Bagnoli, Mark and Ted Bergstrom, “Log-Convex Probability and its Applications,” Economic Theory, 2005, 26 (2), 445–469.

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

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

_ and _, “Trade, Competition, and Efficiency,” Journal of International Economics, 2012, 87 (1), 1–17.

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

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

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

Buchbinder, Niv, Moran Feldman, Joseph Seffi, and Roy Schwartz, “A Tight Linear Time (1/2)-Approximation for Unconstrained Submodular Maximization,” SIAM Journal on Computing, 2015, 44 (5), 1384–1402.
Budish, Eric, Benjamin Roin, and Heidi Williams, “Do Firms Underinvest in Long-Term Research? Evidence from Cancer Clinical Trials,” American Economic Review, 2015, 105 (7), 2044–2085.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Einav, Liran, Amy Finkelstein, and Mark R. Cullen, “Estimating Welfare in Insurance Markets Using Variation in Prices,” Quarterly Journal of Economics, 2010, 125 (3), 877–921.
and, “Selection in Insurance Markets: Theory and Empirics in Pictures,” *Journal of Economic Perspectives*, 2011, 25 (1), 115–138.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Hart, Oliver and John Moore, “Property Rights and the Nature of the Firm,” *Journal of Political Economy*, 1990, 98 (6), 1119–1158.
Head, Keith and Thierry Mayer, “What Separates Us? Sources of Resistance to Globalization,” *Canadian Journal of Economics/Revue canadienne d’économique*, 2013, 46 (4), 1196–1231.

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

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

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

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

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

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

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

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

Kingma, Diederik P. and Jimmy Ba, “Adam: A Method for Stochastic Optimization,” 2014. [http://arxiv.org/abs/1412.6980](http://arxiv.org/abs/1412.6980).

Klenow, Peter and Jon Willis, “Real Rigidities and Nominal Price Changes,” 2006. [http://klenow.com/RealRigidities.pdf](http://klenow.com/RealRigidities.pdf).

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

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

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

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

Kubler, Felix and Karl Schmedders, “Competitive Equilibria in Semi-Algebraic Economies,” *Journal of Economic Theory*, 2010, 145 (1), 301–330.
_ , Philipp Renner, and Karl Schmedders, “Computing All Solution to Polynomial Equatoins in Economics,” in Karl Schmedders and Kenneth L. Judd, eds., Handbook of Computational Economics, Vol. 3, Amsterdam, Holland: Elsevier, 2014.

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

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

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

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

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

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

_ and Gianmarco I. P. Ottaviano, “Market Size, Trade and Productivity,” Review of Economic Studies, 2008, 75 (1), 295–316.

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

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

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

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

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

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

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

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

_ and Murray Jorgensen, “The Double Pareto-Lognormal Distribution – A New Parametric Model for Size Distributions,” Communications in Statistics – Theory and Methods, 2004, 33 (8), 1733–1753.
Rochet, Jean-Charles and Jean Tirole, “Platform Competition in Two-Sided Markets,” *Journal of the European Economic Association*, 2003, 1 (4), 990–1029.

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

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

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

_ , *Cours Complet D’Économie Politique Practique*, Paris: Rapilly, 1828.

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

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

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

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

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

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

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

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

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

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

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

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

_ , “Monopoly, Ramsey and Lindahl in the Rochet and Tirole (2003),” *Economics Letters*, 2009, 103 (2), 99–100.
and Jean Tirole, “Market Power Screens Willingness-to-Pay,” Quarterly Journal of Economics, 2012, 127 (4), 1971–2003.

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

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

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

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

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

Appendix

A Proofs of Theorems

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

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

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

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

\(^{78}\) If we worked with infinite intervals, the convergence of the integrals below would not be always guaranteed.

\(^{79}\) The Fourier transform used in the proof is equivalent to the Laplace transform with imaginary $s$. Both transforms may be thought of as parts of the holomorphic Fourier-Laplace transform.
addition to $G(s)$. For $n = 1, 2, ..., m$, we denote by $G^{(n)}_S(s)$ the truncation of $G^{(n)}(s)$ to the interval $S$, i.e. $G^{(n)}_S(s) \equiv G^{(n)}(s) 1_{s \in S}$. Inside the interval $S$,

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

The $m + 1$ functions $G_S(s), G^{(1)}_S(s), G^{(2)}_S(s), ..., G^{(m)}_S(s)$ 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 $G_S(s), G^{(1)}_S(s), G^{(2)}_S(s), ..., G^{(m)}_S(s)$ must be linearly dependent on the interval $S$. As a result, there must exist a polynomial $T_0(.)$ (with real coefficients), such that

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

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 $n$th 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 $n$th derivative. Given that the expression (6) is a generalized function\(^{80}\) of $s$ that gives zero when integrated against any test function\(^{81}\) supported on $(-\infty, s_1 - \epsilon) \cup [s_1 + \epsilon, s_2 - \epsilon] \cup [s_2 + \epsilon, \infty)$ for any $\epsilon > 0$, we may write it as a linear combination of Dirac delta functions and a finite number of their derivatives located at $s_1$ and $s_2$. By computing its Fourier transform we find that $T_0(-i\omega) \hat{G}_S(\omega)$ must be of the form $T_1(\omega) e^{i\omega s_1} + T_2(\omega) e^{i\omega s_2}$ with some polynomials $T_1(\omega)$ and $T_2(\omega)$, with complex coefficients in general. Consequently, $\hat{G}_S(\omega)$ may be written as

$$\hat{G}_S(\omega) = \frac{T_1(\omega)}{T_0(-i\omega)} e^{i\omega s_1} + \frac{T_2(\omega)}{T_0(-i\omega)} e^{i\omega s_2}. \quad \text{for } T_0(-i\omega) \neq 0$$

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^{i\omega s_1} + \frac{T_4(\omega)}{T_6(\omega)} e^{i\omega s_2}, \quad \text{for } T_3, T_4, T_5, \text{ and } T_6, \text{ such that } T_3 \text{ has no common divisors with } T_5 \text{ and similarly for } T_4 \text{ with } T_6. \quad \text{Let us compute the inverse Fourier transform of the last expression for } \hat{G}_S(\omega) \text{ using the residue theorem. To perform the integration, we consider each of the two terms in (7) separately and specialize to } s \in S. \text{ 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 \text{ multiplied by polynomials of } \omega. \text{ We see that for } s \in S, G_S(s) = \sum_{j=1}^{N} D_j(s) e^{-i\omega t_j}, \text{ for some integer } N, \text{ complex numbers } t_j \text{ and polynomials } D_j(s). \text{ Since the interval } S \text{ was chosen arbitrarily, not just } G_S(s), \text{ but also } G(s) \text{ 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 \text{ numbers } t_j \text{ are real and the remaining ones have an imaginary part. By combining individual terms into real contributions so}

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

\(^{81}\)A test function here refers to an element of the space $S(\mathbb{R})$ of space of rapidly decreasing functions.
that $G(s)$ is real, we get

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

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 $F(q)$:

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

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

**Proof of Theorem 2 (Closed-Form Solutions).** The proof is straightforward. By assumption, there exists some definite power $b$ such that $x \equiv q^b$ satisfies an algebraic equation of order $k$: $P_k(x) = 0$, where $P_k(x)$ is a polynomial of order at most $k$. For this to be true, all elements of the functional form class must factorize as $q^b P_k(q^b)$ 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 $\int MC(q) dq$, and profit are all linear combinations of powers of $q$. For this reason, it suffices to show that it is possible to perform explicitly aggregation integrals $\mathcal{I}$ for powers of $q$ (the quantity optimally chosen by a firm with productivity parameter $a$): $\mathcal{I} \equiv \int q(a)^\gamma dq$. Changing the integration variable to $q$ gives: $\mathcal{I} = \int q^\gamma G'(a(q)) a'(q) \, dq$. The firm’s first-order condition equates the marginal revenue $R'(q) = P(q) + qP'(q)$ to the marginal cost $MC_0(q) + aMC_1(q)$ and implies

$$a = \frac{R'(q) - MC_0(q)}{MC_1(q)} \Rightarrow a'(q) = \frac{R''(q) - MC_0'(q)}{MC_1(q)} - \frac{R'(q) - MC_0(q)}{MC_1(q)^2} MC_1'(q).$$

Substituting these expressions into the integral gives

$$\mathcal{I} = \int q^\gamma \left( \frac{R''(q) - MC_0'(q)}{MC_1(q)} - \frac{R'(q) - MC_0(q)}{MC_1(q)^2} MC_1'(q) \right) G' \left( \frac{R'(q) - MC_0(q)}{MC_1(q)} \right) dq.$$

Since $G'(a)$ is a mixture of powers of $a$, and $(R'(q) - MC_0(q))MC_1'(q)$ and $R''(q) - MC_0'(q)$ 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

$$\int q^\gamma MC_1(q)^\gamma (-MC_0(q) + R'(q)) \, dq,$$

where $\gamma$ equals $-\gamma_6 - 1$ or $-\gamma_6 - 2$. Given our assumptions, up to a known multiplicative constant this integral equals $\int q^{\gamma_6} N_1(q^a)^\gamma N_2(q^a)^{\gamma_10} dq$. If we change the integration variable to $x \equiv q^a$, the problem reduces to computing the integral $\int x^{\gamma_6} N_1(x)^{\gamma_12} N_2(x)^{\gamma_13} dx$. To complete the proof, it
suffices to examine the structure of this integral for different structures of the polynomials.

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

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

2. If either $N_1$ or $N_2$ is trivial and the other polynomial is linear, the integral leads to the standard hypergeometric function, denoted $2F_1$, since up to an additive constant

$$
\int x^{\gamma_1} (1 + \gamma_1 x) \gamma_2 \, dx = \frac{x^{1+\gamma_1}}{1+\gamma_1} 2F_1(1 + \gamma_1, -\gamma_2; 2 + \gamma_1; -x\gamma_1)
$$

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

$$
\int x^{\gamma_1} (1 + \gamma_1 x + \gamma_2 x^2) \gamma_3 \, dx =
\frac{\gamma_3 x^{1+\gamma_1}}{1+\gamma_1} \left( \frac{1 + x\gamma_1 + x^2\gamma_2}{\gamma_1 + x\gamma_2 \gamma_3 + x^2\gamma_3} \right) F_1(1 + \gamma_1, -\gamma_2, -\gamma_3; 2 + \gamma_1, \gamma_1 x, \gamma_2 x, \gamma_3 x)
$$

where $\gamma_1 = -2\gamma_15 \left( \gamma_14 + \sqrt{\gamma_14^2 - 4\gamma_15} \right)^{-1}$, and $\gamma_17 = 2\gamma_15 \left( -\gamma_14 + \sqrt{\gamma_14^2 - 4\gamma_15} \right)^{-1}$.

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 $F_1$:

$$
\int x^{\gamma_1} (1 + \gamma_1 x + \gamma_2 x^2) \gamma_3 \, dx =
\frac{\gamma_3 x^{1+\gamma_1}}{1+\gamma_1} \left( \frac{1 + x\gamma_1 + x^2\gamma_2}{\gamma_1 + x\gamma_2 \gamma_3 + x^2\gamma_3} \right) F_1(1 + \gamma_1, -\gamma_2, -\gamma_3; 2 + \gamma_1, \gamma_1 x, \gamma_2 x, \gamma_3 x)
$$

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, $x^{\gamma_1} (1 + \gamma_1 x) \gamma_2 (1 + \gamma_1 x) \gamma_3$ equals $x^{b-1} (1 - u_1 x)^{-b_1} \ldots (1 - u_n x)^{-b_n}$, where $u_i$ represent the reciprocals of the roots of the polynomials. These roots, as well the constants $b_1, b_2, \ldots, b_n$ 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

$$
\int x^{b-1} (1 - u_1 x)^{-b_1} \ldots (1 - u_n x)^{-b_n} \, dx = \frac{x^b}{b} F_D^{(n)}(b, b_1, \ldots, b_n, b + 1; u_1 x, \ldots, u_n x)
$$

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

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

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

$$
F_D^{(n)}(b, b_1, \ldots, b_n, b + 1; x_1, \ldots, x_n) = b \int_0^1 y^{b-1} (1 - x_1 y)^{-b_1} \ldots (1 - x_n y)^{-b_n} \, dy
$$
Substituting \( y \to x_0 / x, \) \( x_1 \to u_1 x \) and \( x_n \to u_n x \) then leads to the desired result for the integral:

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

(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. \( \square \)

**Proof of Theorem 4 (Laplace-log Transform with Riemann-Stieltjes Integrals).** (A) This follows from Theorem I.6.3 of Widder (1941). (B) If we choose \( u(t) \) appearing in Equation 3 from the paper to be piecewise constant with a finite number \( N \) of points of discontinuity \( \{ t_j, j = 1, 2, \ldots, N \} \), the integral becomes 

\[
U(q) = \sum_{j=1}^{N} a_j q^{-t_j}, \quad \text{where } a_j \text{ is the (signed) magnitude of the discontinuity at point } t_j, \text{ i.e. the magnitude of the mass that } u(t) \text{ has at point } t_j.
\]

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). \( \square \)

**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. \( \square \)

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

\[
\sum_{k=1}^{n_T} F(k) = \int_T^T F(x) \, dx + \mathcal{C}(F) + E_F(n_T),
\]

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

\[
E_F(n_T) = \frac{1}{2} F(n_T) - \sum_{m=1}^{m} \frac{B_{2m}}{(2m)!} F^{(2m-1)}(n_T) + \frac{1}{(2m+1)!} \int_{n_T}^T \frac{P_{2m+1}}{F(2m-1)}(x) \, dx.
\]

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

\[
\sum_{t \in T} q^{-t} f(t) = \frac{1}{\Delta t} \int_{t_{\min}}^{t_{\max}} q^{-t} f(t) \, dt + \frac{1}{2} q^{-t_{\min}} f(t_{\min}) + \frac{1}{2} q^{-t_{\max}} f(t_{\max}) - \frac{R_1 + R_2 + R_3}{\Delta t},
\]

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

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

If we use the notation

\[
F(k) \equiv q^{-t_{\min} - k \Delta t} f(t_{\min} + (k - 1) \Delta t)
\]

we can rewrite the individual terms in the desired formula as

\[
\sum_{t \in T} q^{-t} f(t) = \sum_{k=1}^{n_T} F(k), \quad \frac{1}{\Delta t} \int_{t_{\min}}^{t_{\max}} P_{2m}(t) f^{(1+2m)}(t) \, dt = -\frac{1}{(2m+1)!} \int_1^T P_{2m+1}(x) F^{(1+2m)}(x) \, dx,
\]

\[
\frac{R_3}{\Delta t} = -\frac{\Delta t^{2m}}{(1+2m)!} \int_{t_{\min}}^{t_{\max}} P_{1+2m}(t) h^{(1+2m)}(t) \, dt
\]
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).

**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 $\rho$ 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 $t$, and $1/\rho$ decreases. We conclude that $\rho$ is decreasing in $q$. Only if $t cs(t)$ is supported at one point will there be no shift in weight and $\rho$ 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 $tp(t)$ is nonnegative and supported on $(-\infty, 1)$, with the corresponding Laplace inverse demand functions $p(t)$ listed in our Supplementary Material E, which also contains additional discussion. Note that for most of the inverse demand functions listed in the theorem, it is also possible to prove complete monotonicity using Theorems 1–6 of Miller and Samko (2001).
Table 3: Sensitivity to the cutoff $N_{f,\text{min}}$ 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 $\beta$ and the corresponding standard deviation $\sigma_\beta$.

| $N_{f,\text{min}}$ | $N_I$ | $\beta$ | $\sigma_\beta$ |
|-------------------|------|--------|--------------|
| 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         |

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

B Details of the Generalized EOQ Model Estimation

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

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

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

C World Trade

C.1 Details of data construction

Here we provide details of the data construction for Section 4. The economies used to fit our model are, in descending order by 2006 GDP, United States, Japan, Germany, China, United Kingdom, France, Italy, Canada, Spain, Brazil, Russia, South Korea, Mexico, India, Australia, Netherlands, Turkey, Switzerland, Sweden, Belgium, Saudi Arabia, Norway, Poland, Austria, Denmark,
Figure 9: The relationship of the cost exponent $\alpha$ 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 $\alpha$ and $H_s$. Figure (b) corresponds to a cutoff set to 5 firms as a robustness check. Also, in this case the values of $\alpha$ do not seem to be influenced by $H_s$.

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

*In the model, all imports are consumed domestically, which implies that exports cannot be larger than tradable GDP. However, such situation might arise for small, highly open economies. To avoid this discrepancy, when the calculated portion of tradable GDP that goes to domestic consumption is smaller than five percent of the tradable GDP in the data, we increment it so that it reaches that level. This is done by correspondingly increasing both the (adjusted) tradable GDP and the (adjusted) consumption in the economy. This criterion was satisfied for just one economy, Hong Kong. Of course, a more realistic way of modeling this situation is to include multi-stage production and/or multi-stage transportation in the model. This will require some additional research work, but it is a clear direction to pursue.*
C.2 Related literature

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

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

C.3 Firm export patterns

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

D Applications

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

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

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

Following the property rights theory of the firm (Grossman and Hart, 1986; Hart and Moore, 1990; Antràs, 2003), input production requires relationship-specific investments. The marginal revenue from additional quality brought by supplier $j$, $MR(q(j)) q_s(j)$ is therefore split between the firm and supplier $j$, where $MR = P + P'q$.\footnote{See AC’s Subsection 3.1 for a discussion of why only marginal revenue, and not the full-downstream revenue, is effective quality-adjusted quantity $q$. This effective quantity represents both the quantity of the good and its quality level. But we will refer to it simply as “quality”, since this will make the discussion sound more natural. If production is “disrupted” by the failure of some supplier $j \in [0, 1)$ to cooperate, then only the quality accumulated to that point in the chain is available, with all further quality-enhancement impossible: $q = \int_0^j q_s(j) \, dj$. The firm faces an inverse demand function $P(q)$, which does not necessarily have to be decreasing because, for example, consumers may have little willingness-to-pay for an improperly finished product. If there is no disruption in production, $q = q(1)$.

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

The cost of producing quality $q_s(j)$ is homogeneous across suppliers and equal to $C(q_s(j))$, which is assumed strictly convex.\footnote{The AC model corresponds to $C(q_s) = (q_s)^{1/\alpha} c / \theta$, where $c$ and $\theta$ 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

\begin{align*}
\text{D.1 Supply chains with hold-up (Antràs and Chor, 2013)}
\end{align*}
marginal revenue she bargains for with her marginal cost:

\[ MC(qs(j)) = C'(qs(j)) = [1 - \beta(j)] MR(q(j)). \]  

(9)

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

The firm chooses \( \beta(j) \) 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 \( \beta(j) \) 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 \( \beta \) corresponding to outsourcing (low \( \beta \)) and insourcing (high \( \beta \)) 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 \( q^* \),

\[ \beta^*(j) = 1 - \frac{MC(q^*)}{MR(jq^*)}. \]

(10)

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

Equation 10 implies that the surplus left to each supplier is \( q_sMC(q) \). The problem reduces to choosing \( q \) to maximize revenue \( qP(q) \) less cost \( qMC(q) \), giving first-order condition

\[ MR(q) = MC(q) + qMC'(q). \]

(11)

This differs from the familiar neoclassical first-order condition \( MR(q) = MC(q) \) only by the presence of the (positive) term \( qMC'(q) \). 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 monopolsony 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 \( P(q) = p_uq^u + p_oq^o \) and \( MC(q) = mc_{-t}q^t + mc_{-u}q^u \). This includes AC’s specification as the special case when \( p_{-t} = 0 \) and \( mc_{-u} = 0 \) so that each has constant elasticity.\(^{87}\) However, let us focus instead on the case when \( t, u, mc_{-u}, p_{-t} > 0 = mc_{-t} > p_{-u} \) and \( u > t \) so that...

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

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

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

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

### D.2 Imperfectly competitive supply chains

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

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

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

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

\[
P + P'q = MC + \hat{P},
\]

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

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

\^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.