Worst-Case Bounds on R&D and Pricing Distortions: Theory and Disturbing Conclusions if Consumer Values Follow the World Income Distribution

Michael Kremer

Harvard University
Center for Global Development
National Bureau of Economic Research

Christopher M. Snyder

Dartmouth College
National Bureau of Economic Research

September 2018

Abstract: We prove that, for general demand and cost conditions and market structures, the fraction of first-best surplus that a monopolist is unable to extract in a market provides a tight upper bound on the relative distortions arising from firms' equilibrium decisions at all margins (entry and pricing). Continuing with this worst-case perspective, we show that a symmetrically truncated Zipf (STRZ) distribution of consumer values generates the lowest producer surplus among those with a given mean and maximum value. This allows us to relate potential deadweight loss from all margins in a market to the Zipf-similarity of its demand curve. The STRZ distribution also bounds deadweight loss at just the pricing margin. We leverage existing results from industrial organization (e.g., on demand curvature) and statistics (e.g., on the relation between means and medians) to bound producer surplus in an array of important special cases. Calibrations based on the world distribution of income generate extremely Zipf-similar demand curves, with disturbing consequences for potential deadweight loss in global markets. We gauge the extent to which various policies—such as progressive taxation or price discrimination—can ameliorate potential deadweight loss.

JEL codes: D21, D42, O31, L11

Contact information: Kremer: Department of Economics, Harvard University, Littauer Center 207, Cambridge MA 02138; email: mkremer@fas.harvard.edu. Snyder: Department of Economics, Dartmouth College, 301 Rockefeller Hall, Hanover NH 03755; email: chris.snyder@dartmouth.edu.

Acknowledgments: The authors thank Dan Björkegren, Severin Borenstein, Eric Edmonds, James Feyrer, Berhnard Ganglmair, Jerry Green, Douglas Irwin, Sonia Jaffe, Justin Johnson, Anup Malani, Scott Pauls, Nina Pavcnik, Robert Staiger, Michael Waldman, Glen Weyl, Jonathan Zinman, and seminar participants at Cornell University, Duesseldorf Institute for Competition Economics, Indiana University, the Advances in Price Theory Workshop at the Becker-Friedman Institute, and the International Industrial Organization Conference at Drexel University for insightful comments. The authors are grateful to Maxim Pinkovskiy and Xavier Sala-i-Martin for sharing the data behind their estimates of the world distribution of income used in our calibrations and for detailed advice on constructing these and alternative distributions. The authors thank Branco Milanovic for use of the Lakner-Milanovic (2015) data, publicly posted on the CUNY Graduate Center website. The authors thank Matthew Goodkin-Gold for excellent research assistance. Snyder gratefully acknowledges the support and hospitality of the Becker-Friedman Institute at the University of Chicago where he was a visiting scholar while undertaking some of this research.

1. Introduction

Economists have an ambivalent view of producer surplus. On the one hand, high producer surplus may be indicative of high markups above marginal cost, resulting in large distortions measured by the Harberger triangle. On the other hand, producer surplus provides the incentives for investment in entry, capacity, and innovation that allow markets to exist and thrive. If firms cannot easily appropriate their investment returns, this may result in a different source of distortion—underinvestment relative to the first best.

This paper contributes to an understanding of markets by studying them through the lens of the worst-case scenarios. How large can the distortions possibly be in a given market? Can their magnitude be related to a simple sufficient statistic? What is the source of this maximal distortion: above-cost prices, inefficient investment, or both? We show that the greatest potential deadweight loss is at the investment margin, determining the very existence of the market. We show quite generally that the amount of total surplus that a monopolist is unable to extract provides a tight upper bound on the maximum possible distortion in any given market. The monopoly bound also serves as a bound on distortions arising in a wide range of other, oligopolistic industrial structures.

These results may be best understood in a simple numerical example, drawing on Romer (1994). Suppose a monopoly serves a market with linear demand $Q(p) = 1-p$ and constant marginal cost normalized to zero, shown in Figure 1. Equilibrium quantity is $1/2$, price is $1/2$, and producer surplus is $1/4$ (the area of square B). First-best surplus is $1/2$ (the areas of A, B, and C). Let k denote the fixed cost that the firm needs to expend in order to enter the market. Consider the thought experiment of letting k vary, tracing out the total distortion resulting from departures from both first-best product and pricing decisions. If $k < 1/4$, then the product is produced both in equilibrium and in the first best. The only distortion is the Harberger triangle C resulting from supra-competitive prices, the area of which is $1/8$. If $k > 1/2$, then there is no distortion in the market because nothing is produced either in the first best or in equilibrium. If $k = 1/4+\epsilon$ for small, positive ϵ, the product is produced in the first best but not in equilibrium. The distortion then amounts to all of first-best welfare net of k, i.e., $1/2-k = 1/4-\epsilon$. Deadweight loss increases as ϵ decreases, coming arbitrarily close to $1/4$ in the limit as $\epsilon \downarrow 0$. We conclude that up to half of first-best surplus can be lost because of inefficient investment incentives. While the "half" result is specific to this numerical example, a general principle is being illustrated here. Let ρ^* denote the ratio of the monopolist's to first-best surplus. Theorem 1 shows that the relative surplus the monopolist cannot extract, $1-\rho^*$, provides a tight upper bound on relative deadweight loss for a general class of markets.

The worst case for potential distortions has a variety of policy implications. For example, the worst case for distortions can be converted into an upper bound on the social gains from offering an optimal subsidy. In the numerical example, the government can obtain the first best by offering a subsidy of $1/2$ for each unit sold at marginal cost, 0. If the firm accepts, enters, and sells to all consumers at 0, it ends up selling 1 unit, earning profit net of R&D cost of $1/2-k = 1/4-\epsilon$ for $k = 1/4+\epsilon$. One can see that this policy

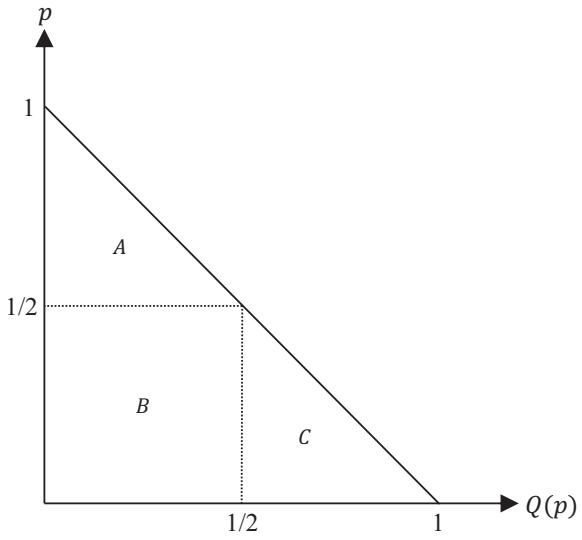


Figure 1: Numerical example.

leads the firm to make the first-best entry decision—entering if $\epsilon < 1/4$ and not if $\epsilon > 1/4$—and the first-best pricing decision—charging price equal to marginal cost 0. The potential gains from this policy can come arbitrarily close to $1/4$ assuming, as one would to attain an upper bound on gains, that there is no social cost of public funds beyond the dollar-for-dollar transfer from taxpayers. This $1/4$ also bounds the potential social loss from banning price discrimination. At best, the firm can engage in first-degree price discrimination, generating social welfare $1/2-k$, exceeding social welfare from linear pricing by an amount approaching $1/4$. These policy implications generalize beyond this numerical example: Theorem 2 shows that the tight bound on deadweight loss, $1-\rho^*$, also serves as a tight bound on the social gain from a subsidy and the social loss from a ban on price discrimination for a general class of markets.

A series of additional theorems provide further generalizations. Theorem 3 generalizes the nature of the product-development cost k, allowing a fraction β of it to be a socially neutral transfer (say a licensing fee or a bribe). The theorem shows that the potential for deadweight loss expands in proportion to β. Theorems 4 and 5 show that the bound derived for a monopoly is relevant to a quite broad range of oligopoly market structures.

Our analysis of the numerical example so far has fixed the demand curve and asked how high the distortion could be in that market. In the spirit of worst-case analyses, it is natural to ask whether other demand curves could generate even greater distortions. Among demands with the same vertical intercept and total surplus as the given linear one, the one minimizing the producer-surplus ratio ρ^* (and thus maximizing potential distortions) is shown in Figure 2 as the black curve superimposed over the grey linear demand. The difference in producer surplus is shown as the difference in the area of the shaded rectangles. The black demand curve is everywhere unit elastic, yielding the special property that no price is particularly attractive for the monopolist because they all generate the same revenue (also the same producer surplus given

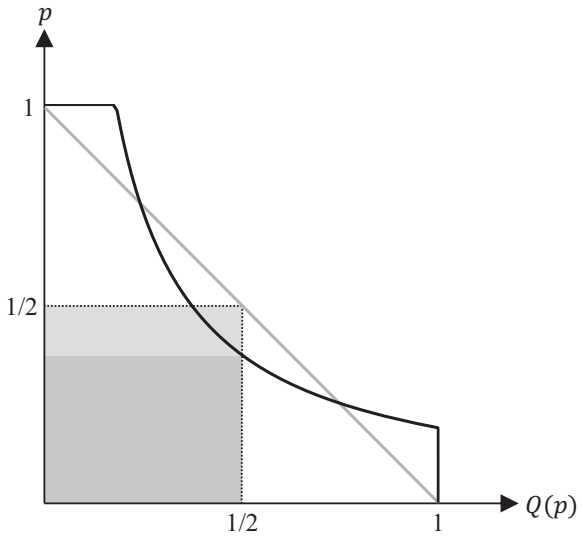


Figure 2: STRZ demand curve.

marginal cost is zero). The consumer values underlying the black demand curve follow a Zipf distribution, with the property that each doubling of consumer value cuts the number of consumers with at least that value in half. It is truncated so that the ends match, hence the label—symmetrically truncated Zipf (STRZ) distribution—given to it in our previous work (Kremer and Snyder 2015), where it also played a key role. [1] The numerical example is admittedly special, involving unit intercepts and zero marginal cost, but for any constant marginal cost and any demand with finite intercepts, we derive a suitable rescaling such that the demand can be compared to its STRZ counterpart having the same intercepts and mean value. In this way, producer surplus for a large class of markets can be decomposed into the Zipf-similarity of the rescaled demand curve and the lower bound on producer surplus attained by the STRZ demand with that mean. As the rescaled mean (denoted μ) shrinks to 0, the producer-surplus ratio from a STRZ demand curve approaches 0 and the bound on deadweight loss approaches 100% of total surplus. Thus the worst of the worst cases is a perfectly Zipf-similar demand curve with vanishingly small mean (relative to peak) consumer values.

Given that the distribution of a wide range of economic variables—city size, firm size, income, wealth, CEO compensation, stock price changes, various international-trade indexes—is approximately Zipf (Gabaix 2009), demand in many real-world markets may have a STRZ-like shape. For example, the Zipf distribution

[1] A growing economics literature proves a wide range of important results based on demand curves that, like STRZ demand in this paper, generate equal producer surplus whatever price is charged (alternately labeled truncated Pareto, unit-elasticity, equal-revenue, or extremal demand). The earliest published article in this literature appears to be Neeman (2003), who uses the construction to bound the effectiveness of English auctions in extracting bidder surplus. The idea first appears in our work in an early working-paper version (Kremer and Snyder 2003) of Kremer and Snyder (2015). Malueg and Snyder (2006) use a sequence of linear demands with the equal-revenue property to bound the ratio of profits from discriminatory to uniform pricing by a function of the number of markets. The demand curve that Brooks (2013) derives to generate the worst case for his belief survey auction is precisely the STRZ form. Bergemann, Brooks, and Morris (2014) show that any market can be represented as a convex combination of segmented markets with equal-revenue demands, allowing them to construct segmentations attaining arbitrary divisions of first-best surplus across the firm, consumers, and deadweight loss. A growing literature in computer science uses the construction to bound worst cases for approximately optimal mechanisms in various settings. See Hartline and Roughgarden (2009) for an early such reference.

of income may lead individuals' demand for cancer cures, pay-per-view television series, electronic gadgets, or professional sports tickets to have a STRZ shape. The Zipf distribution of city size may lead municipal demand for management software to have a STRZ shape. Facing a STRZ demand curve, a firm lacking the ability to price discriminate may be able to extract only a small share of consumer surplus, raising the possibility that in many different contexts socially valuable investments will not be undertaken. A natural policy response is to identify markets with Zipf-similar demands and target R&D subsidies there, or at least refrain from imposing policies such as banning price discrimination that would impair firm profits.

As a proof of concept, in Section 7 we calibrate demand for some representative product based on the world distribution of income using the estimates from Pinkovskiy and Sala-i-Martin (2009). Calibrated demands indeed closely resemble STRZ worst cases; for the most recent year of income estimates, the Zipf-similarity of calibrated demand is $Z = 83\%$. Because it is so Zipf similar, and because the ratio of the mean to peak consumer value is so low, implying that the Zipf demand yields very little producer relative to total surplus, potential distortions are enormous with this demand curve. In the case of a good that can be produced at little or no marginal cost (software, digital media, some small-molecule drugs), deadweight loss can be as high as 72% of total surplus. A series of robustness checks (varying income elasticities, marginal costs, price indexes, data sources, and market regions) confirms that potential deadweight loss remains high for a broad range of cases. We also run a series of policy counterfactuals, analyzing the effect on potential deadweight loss of a ban on price discrimination, redistribution via a proportional tax, and redistribution from just the top 1% (or other precentiles) of consumers.

While Zipf-similarity provides a sufficient statistic for potential deadweight loss that has the virtue of broad applicability, it can require numerical methods to compute. Section 6 leverages existing results from microtheory and statistics to provide analytical bounds in a rich set of specific cases. Anderson and Renault's (2003) upper bound on the ratio of producer to total surplus as a function of the generalized curvature (c-concavity) of demand yields a bound on deadweight loss as an immediate corollary. Based on the simple observation that a monopolist can guarantee sales to at least half of the consumers by pricing at the median valuation, we provide a bound on the ratio of producer to total surplus that is related to the ratio of the median to the mean of a distribution. This allows us to leverage a rich statistics literature on the mean-median-mode inequality (including Runnenburg 1978, van Zwet 1978, Dharmadhikari and Joag-dev 1988, and Basu and DasGupta 1997) to provide a series of bounds on deadweight loss that applies to a variety of unimodal distributions. We also provide bounds for the beta, Pareto, and general discrete distributions.

Turning the focus from distortions on the entry margin to distortions on the pricing margin conditional on entry, (we will label this Harberger deadweight loss because it is equal to the area of the Harberger triangle), the STRZ demand curve has useful implications for this source of deadweight loss as well. As we show in Section 5, Harberger deadweight loss is chaotic for a STRZ demand. There are multiple equilibria, one of which (the equilibrium in which only the highest-value consumers are served) attains the upper bound

on Harberger deadweight loss for any market with that mean consumer value. However, a tiny shift in mass from high to low consumer values is enough to tip the market to a unique equilibrium with no Harberger deadweight loss. The policy implication is that Harberger deadweight loss is potentially huge in markets with Zipf-similar demands, but this inefficiency can be corrected at small cost to the policymaker with a small subsidy.

This paper is most closely related to our own previous work on which it builds, Kremer and Snyder (2015). Our previous paper focused on pharmaceutical markets. In some simple cases, we were able to bound the deadweight loss from the firm's bias to produce a drug rather than a vaccine by $1-\rho^*$, where ρ^* is the ratio of monopoly producer surplus to first-best surplus on the vaccine market. Here we show that the bound $1-\rho^*$ applies much more generally, beyond pharmaceuticals to any product market, and does not require the availability of two substitute products to serve the same market. In the previous paper, the bound on potential deadweight loss was much fuzzier outside of the simple case of a monopoly serving consumers who differ only in disease risk. The present paper shows that the sharp bound $1-\rho^*$ applies regardless of the factors determining consumer values and applies to a wide range of market structures beyond monopoly. We build on the derivation of STRZ demand in the previous paper, applying Zipf similarity to decompose changes in potential deadweight loss over time and applying the STRZ demand to bound Harberger deadweight loss. The computation of ρ^* for the beta and Pareto distributions is new here, as are the bounds on ρ^* leveraging results for c-concave demands and leveraging the mean-median-mode inequality. While the previous paper provided demand calibrations for HIV pharmaceuticals, the present paper calibrates global demand for a general product and uses the calibrations to conduct a variety of policy counterfactuals. [2]

Our paper is close in spirit to Romer's (1994) agenda-setting article. Although his focus is on the contribution of trade to developing economies, he makes the broader point, which we share, that product-market distortions on the extensive margin can overwhelm those on the intensive margin. In a sense our paper can be viewed as taking up his call for more formal modeling of basic ideas related to the market for new goods: "Formal theoretical analysis can contribute to our use of evidence largely by forcing us to make explicit the habits of mind that we take for granted" (Romer 1994, p. 36). Also philosophically related is the work of Makowski and Ostroy (1995, 2001). They provide perhaps the clearest statement of the idea that efficiency requires each agent to appropriate the returns from his investment. They show that price taking is not necessary for the efficiency of general equilibrium entailed by the first welfare theorem. Price taking is just one setting in which which agents appropriate their marginal contributions to social welfare; any other such setting—a single producer with small capacity auctioning goods to competing suppliers, etc.—will likewise generate the first best. Rather than looking at conditions for first-best efficiency, our

[2] Each of the relevant section of results starts with a footnote detailing the precise connection between the propositions in Kremer and Snyder (2015) and this paper. Another of our related papers is Kremer and Snyder (2018), which uses some of the results in the present paper to bound deadweight loss in a calibration of the global market for an HIV pharmaceutical. It should be emphasized that Kremer and Snyder (2018) is the successor to rather than the antecedent of the present paper.

paper contributes by looking at the flip side, conditions for the greatest distortions to arise. Consistent with Makowski and Ostroy's perspective, we find that distortions are greatest when conditions for appropriability are worst; we characterize which demand shapes are worst for appropriability.

We already acknowledged our debt to Anderson and Renault (2003), which led to a burgeoning literature linking the shape of the demand curve to producer surplus. Weyl and Fabinger (2013) (see also follow-on work by Fabinger and Weyl 2014) provide bounds on the surplus ratio for arbitrary demand and cost curves and oligopoly models; the bounds are tied to the elasticity of the marginal-producer-surplus function and oligopoly conduct parameters. Johnson and Myatt (2006) construct several orderings on demand curves including clockwise rotations. They provide a rich set of applications in which a firm's strategy—e.g., advertising or product design—is isomorphic to a choice of an ordered demand curve.[3] They show profit is typically quasiconvex in the demand ordering, rationalizing all-or-nothing strategy choices observed in the applications. Our decomposition formula provides a different way to order demand curves based on Zipf similarity and mean consumer values, which can be used to compare any two demands from a quite general class. Garber, Jones, and Romer (2006) relate deadweight loss to the shape of the demand curve as we do. The distortion in their pharmaceutical application arises comes from co-insurance: by defraying a fraction of the pharmaceutical price, co-insurance can induce overconsumption and excess entry.

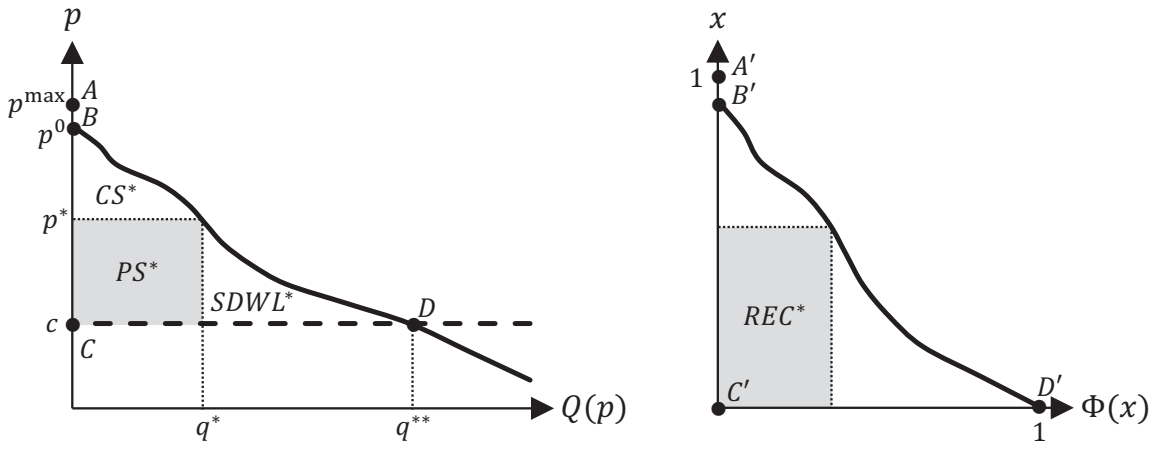
Our emphasis on incentives for market creation based on surplus appropriation and our bound on deadweight loss from that source contribute to the broader literature on incentives for innovation from a general microeconomics perspective (see, e.g., Dosi 1988, Freeman 1994, Weyl and Tirole 2012) as well as to particular R&D-intensive industries (see, e.g., Newell, Jaffee, and Stavins 1999; Acemoglu and Linn 2004; Finkelstein 2004; Furman and Stern 2011; and Budish, Roin, and Williams 2013). Our result bounding the loss from banning price discrimination is related to welfare bounds on price discrimination in Malueg (1993) and Bergemann, Brooks, and Morris (2014)

2. Model

Consider a market m for a product. Demand is given by $Q(p)$, a nonincreasing and left-continuous function of price p.[4] Let $P(q) = \sup\{p|Q(p) > q\}$ be the inverse demand curve. Let $p^0 = P(0)$ denote the choke price. It will be useful to distinguish between the choke price p^0 associated with a particular demand curve and the maximum conceivable valuation for the product, denoted $p^{\max} \geq p^0$. For example, suppose a disease causes

[3]The idea that the firm will choose the most profitable shape for consumer demand has been applied to a diverse set of phenomena in industrial organization. For example, DeGraba (1995) explains buying frenzies as a response to supply limitations, inducing consumers to race to buy before acquiring more information about their true valuations. Biehl (2001) applies the idea to the sell versus lease decision.

[4]Relaxing full continuity accommodates the case of discrete consumer types. Left-continuity amounts to assuming that consumers who are indifferent between purchasing and not choose to purchase the product. This assumption avoids an existence problem. If indifferent consumers choose not to purchase, the optimal monopoly price may fail to exist with discrete consumer types. In particular, if marginal consumers at a given price are of a discrete type, then any price in the open interval between the given price and the marginal consumers' maximum willingness to pay is more profitable.



(a) Original demand (b) Rescaled demand

Figure 3: Demand and its rescaling.

\$1 of harm to affected individuals. The maximum conceivable valuation for a perfectly effective vaccine would be $p^{\max} = 1$. On the other hand, in a market consisting of a country in which no consumer has greater than a 50% risk of contracting the disease, the choke price for a vaccine may be only $p^0 = 0.5$. Assume $p^0 > c$, ensuring the market is non-trivial. The first panel of Figure 3 illustrates the inverse demand curve and some of the other notation introduced in this subsection (we will come back to the second panel in a later section).

We will assume throughout most of the analysis that $p^{\max} < \infty$. The importance of this assumption may not be fully apparent until Section 6.3, where we consider a demand curve arising from a Pareto distribution of consumer values, which does not have a finite maximum. The assumption can be justified in several ways. One is to realize that continuous distributions are just convenient approximations for values in a population of discrete consumers. There will be some highest valuer with some well-defined finite value for the discrete product. Another justification is that buyers with unbounded consumption values, well before paying an infinite amount to the monopolist, could save money by integrating backward into the production of the good themselves (Katz 1987). The option of backward integrating could then be related back to a finite $p^{\max}$.

The product is supplied by a monopolist. The firm decides whether or not to enter the market, reflected by the indicator function E. Entry requires a fixed R&D expenditure k. Most of the analysis takes all of k to be a social as well as a private cost, although we will briefly consider an extension in which part of k is not a social cost—possibly a licensing fee, tax, or other transfer of surplus to another party, or an investment with some external social benefit. After entering, the monopolist produces at constant marginal cost c and sells to consumers at a linear price p.[5] Let $PS(p) = (p-c)Q(p)$ be the resulting producer

[5]Harris and Raviv (1984) show that a linear price is the optimal mechanism for a monopolist who cannot engage in third-degree price discrimination and who serves consumers with unit demand.

surplus, $CS(p) = \int_p^{p^0} Q(x)dx$ be consumer surplus, and $TS(p) = PS(p) + CS(p)$ be total surplus. Note $PS(p)$ and $TS(p)$ are surpluses from an an ex post perspective, i.e., treating k as a sunk cost and thus ignoring it. Profit from an ex ante perspective—treating k as an economic cost—is denoted $\Pi(p) = PS(p) - k$. Ex ante social welfare is denoted $W(p) = TS(p) - k$. Denote equilibrium values of variables with stars. Thus $p^* = \text{argmax} PS(p)$ is the profit-maximizing price conditional on entry; and $q^* = Q(p^*)$, $PS^* = PS(p^*)$, $CS^* = CS(p^*)$, $TS^* = TS(p^*)$, $\Pi^* = \Pi(p^*) - k$, and $W^* = W(p^*)$. The profit-maximizing entry decision is denoted E^*, i.e., $E^* = 1$ if $\Pi^* > 0$ and $E^* = 0$ if $\Pi^* < 0$. Denote first-best values of variables with double stars. We have $p^{**} = c$, $q^{**} = Q(c)$, $PS^{**} = PS(c)$, $CS^{**} = CS(c)$, $TS^{**} = TS(c)$, and $W^{**} = W(c)$. The first-best entry decision is E^{**}, i.e., $E^{**} = 1$ if $W^{**} > 0$ and $E^{**} = 0$ if $W^{**} < 0$.

Deadweight loss captures equilibrium distortion compared to the first best. We will distinguish between two deadweight-loss concepts. Harberger deadweight loss is $HDWL(p) = TS^{**} - TS(p)$. This is the area of the Harberger triangle, reflecting just the distortion at the intensive margin of charging a supra-competitive price $p \geq c$, but taking the decision to develop the product as given. The equilibrium value of Harberger deadweight loss is $HDWL^* = HDWL(p^*)$. Deadweight loss without without the "Harberger" modifier is a more comprehensive concept, capturing distortions at all margins, both the intensive margin (pricing) and the extensive margin (entry). Denote the equilibrium value of this deadweight-loss concept by $DWL^* = E^{**}W^{**} - E^*W^*$.

3. Bounding Deadweight Loss

Much of the second half of the paper is devoted to characterizing when markets are lucrative for a monopolist and when not. This section motivates that analysis, providing a series of theorems connecting the monopoly producer-surplus ratio, $\rho^* = PS^*/TS^{**}$, to potential deadweight loss, and in turn connecting potential deadweight loss to bounds on the gains and losses from policy interventions. Several theorems show that the monopoly bound is not special to that market structure but under some conditions serves as a bound for arbitrary oligopoly models.[6]

3.1. Baseline Bounds for Monopoly

Much of the paper is concerned with factors affecting a monopolist's ability to extract surplus in a market. The results will have a number of useful positive implications, allowing one to predict which markets firms will enter and to predict the level of monopoly profits conditional on entry. The next theorem, a core result of the paper, implies that the results also have normative implications, which may be even more important.

[6]To connect the results in this section to related results in our earlier paper, Kremer and Snyder (2015), Theorem 1 here is a generalization of Proposition 2 from our earlier paper, showing that the previous bound on potential deadweight loss in a pharmaceutical market owing to the firm's bias toward producing a drug rather than a vaccine applies to arbitrary product markets and does not require the firm to have two available products. Theorem 4 is likewise a generalization of our previous Proposition 13. The remaining results in this section (Theorems 2, 3, and 5) have no antecedent in our previous paper.

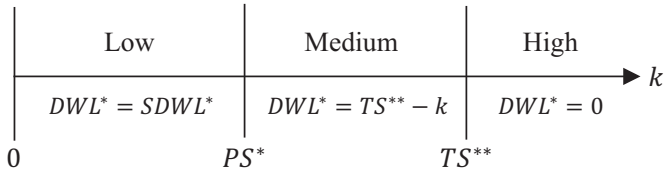


Figure 4: Intervals of k in the Proof of Theorem 1.

Intuitively, the more surplus the monopolist can extract, the greater its R&D incentives, and the more likely that there is any supply on the market at all. Thus the ability of the monopolist to extract surplus, while not a social good in itself, limits potential deadweight loss.

Theorem 1. *In a monopoly market, the total surplus that cannot be extracted by the firm provides a tight upper bound on the level of deadweight loss, i.e.,*

$$\sup_{k\geq 0}(DWL^*) = TS^{**} - PS^*. \tag{1}$$

The tight upper bound can equivalently be expressed as a proportion of total surplus:

$$\sup_{k\geq 0}\left(\frac{DWL^*}{TS^{**}}\right) = 1-\rho^*. \tag{2}$$

Proof. The set of k over which the supremum is taken can be partitioned into the three subintervals shown in Figure 4: low, $(0, PS^*)$; moderate, (PS^*, TS^{**}), and high, (TS^{**}, ∞). For k in the low subinterval, the firm develops the product both in equilibrium and the first best. Hence the only source of deadweight loss is Harberger deadweight loss: $DWL^* = HDWL^* = TS^{**} - TS^*$. For k in the high subinterval, the firm does not enter either in equilibrium or the first best, implying $DWL^* = 0$. Deadweight loss at the extensive margin arises only for k in the moderate subinterval. The firm does not enter in equilibrium because $PS^* < k$ but the first best involves entry because $TS^{**} > k$. In this case, all of first-best welfare $W^{**} = TS^{**} - k$ is deadweight loss. Because welfare is decreasing in k, the supremum is achieved at the lower boundary of this subinterval, $k = PS^*$, where the firm is indifferent between entering and not in equilibrium. The supremum, $TS^{**} - PS^*$, over the moderate subinterval of k exceeds that, $TS^{**} - TS^*$, over the low subinterval of k by $CS^* \geq 0$. This establishes equation (1). Dividing both sides of (1) by TS^{**} establishes (2). *Q.E.D.*

Theorem 1 indicates that that the extensive margin can be a substantial source of deadweight loss. If the monopolist decides not to enter when it is socially efficient to do so, the consumer surplus that would have been generated, CS^*, is lost to society. The deadweight loss at the extensive margin can approach all of CS^* if k is such that the monopolist is close to being indifferent between entering and not. The deadweight loss from the extensive margin adds to Harberger deadweight $HDWL^*$ for a total potential loss (in levels) of $HDWL^* + CS^* = HDWL^* + TS^* - PS^* = TS^{**} - PS^*$, as stated in (1).

The theorem could be generalized in a more general model of monopoly behavior. The model described in Section 2 assumes linear pricing, and the optimal linear price p^* figured into the equilibrium producer surplus PS^*. More generally, PS^* could represent the producer surplus from optimally choosing more complicated pricing strategies, say involving second-degree price discrimination (i.e., nonlinear pric-

ing, bundling) or third-degree price discrimination (i.e. pricing on observable consumer characteristics). The theorem would still hold interpreting PS^* in this way.

The normative implications of the previously derived upper bounds on deadweight loss can be further developed. The next theorem links DWL^* to bounds on social gains and losses from a variety of policies. This allows us to translate bounds on DWL^* directly into bounds on the welfare effects of these policies.

Theorem 2. *The social loss from banning price discrimination in a monopoly market is tightly bounded above by DWL^*, as is the social gain from subsidy policies.*

The proof is provided in Appendix A. Intuitively, the bound on the social loss from banning price discrimination is approached by taking k to be arbitrarily close to, but greater than, PS^* and assuming the firm can perfectly price discriminate in the absence of a ban. Then the monopolist chooses not to produce if price discrimination is banned, but could generate social welfare TS^*-k if allowed to discriminate. The loss in welfare equals $DWL^* = TS^*-PS^*$ in the limit as $k \downarrow PS^*$. The bound on the social gain from a subsidy is approached similarly, again taking k to be arbitrarily close to, but greater than, PS^*. A small subsidy would shift the firm from not producing to producing, and a sufficiently large subsidy could induce first-best output and social welfare TS^*-k. Assuming the subsidy is funded by a non-distortionary tax, the social gain from the subsidy approaches $DWL^* = TS^*-PS^*$ in the limit as $k \downarrow PS^*$.

Theorem 2 provides additional motivation for quantifying ρ^*, which will occupy much of the analysis in subsequent sections. We already saw from Theorem 1 that $1-\rho^*$ is a tight upper bound on relative deadweight loss. Putting the two theorems together, we see that $1-\rho^*$ also provides a tight upper bound (in relative terms) on the social loss from banning price discrimination and the social gain from subsidy policies.

3.2. Private Versus Social Entry Costs

The analysis so far has taken k to be a social cost, using up socially valuable resources, as would be the case in constructing labs and hiring scientists to invent a new product. In some cases, the monopolist's entry expenditures include transfers to other parties (for example, bribes, licensing fees, corporate taxes, tariffs) or investments with some external social value (for example, mandated clinical trials that expand medical knowledge). To accommodate these cases, we extend the model to allow a fraction β of k to be a socially-neutral transfer from the monopolist. Thus, while the whole k is internalized by the monopolist as an entry cost, only the residual $(1-\beta)k$ is a social cost. The following theorem, proved in Appendix A, generalizes the bound on relative deadweight loss from equation (2) in this extension.

Theorem 3. *Suppose that a fraction $\beta \in [0,1]$ of the monopoly's entry cost k is a private cost to the monopolist but not a social cost. Then a tight upper bound on relative deadweight loss is given by*

$$\sup_{k\geq 0}\left(\frac{DWL^*}{TS^{**}}\right) = 1-(1-\beta)\rho^*. \tag{3}$$

The theorem implies that imposing costs on the monopolist over and above the true social cost exacerbates the potential for deadweight loss. In the $\beta = 1$ extreme in which k is a pure transfer to other parties (or in any event involves no expenditure of real social resources), deadweight loss can dissipate all of first-best surplus. One way to view the relationship between Theorems 1 and 3 is to see the former as a bound on deadweight loss arising in a solely private market whereas the latter is a bound accounting for additional distortions due to government and other outside-party activities. The increase in potential deadweight loss from (2) to (3) provides a concrete measure of the potential social loss from corporate taxes, corruption in the form of bribe taking, and other externally imposed distortions. For the remainder of the paper, we will return to the model in which all of k is a social cost, keeping in mind that this assumption leads (2) to be a conservative bound on potential deadweight loss.

3.3. Bounds for General Models of Competition

The analysis so far has assumed that the market is served by a monopolist. Although this is a fairly special market structure, the results are more general than they first appear, providing what is again a conservative bound on potential deadweight loss applicable to a broad range of other industrial structures. Letting C be the model of competition under consideration—Bertrand, Cournot, perfect or imperfect cartel, etc.—with few restrictions on C, these models will be able to generate at least as much deadweight loss as monopoly.

Formally, consider a model of competition C involving $N \in \mathbb{N}$ potential entrants, where $\mathbb{N}$ is the set of natural numbers. Let $i = 1, \ldots, N$ index firms. At the start of the game, each i obtain a draw $k_i \geq 0$ of the fixed cost of developing the product. For simplicity, we will return to the assumption maintained prior to Theorem 3 that k_i is a social as well as a private cost. After observing the vector of draws, firms decide whether or not to enter by sinking the fixed cost. We will focus on pure strategies for the entry decision.

Let $\overline{PS}^*(C,n)$ be the most any single firm earns in equilibrium given $n \in \mathbb{N}$ firms have entered. To allow for general forms of competition, we will put few constraints on this function. Assume

$$\overline{PS}^*(C,1) = PS^*, \tag{4}$$

meaning that a single entrant can achieve the monopoly outcome and thus earn the monopoly producer surplus PS^*. Among other things, this assumption rules out the possibility that potential entrants who do not materialize as actual entrants does not constrain the surplus a monopoly can earn, thus ruling out forms of contestability along the lines of Baumol, Panzar, and Willig (1982). Further, assume

$$\overline{PS}^*(C,n) \leq \overline{PS}^*(C,1) \tag{5}$$

for all numbers of actual entrants $n \geq 1$. This inequality captures the idea that competition destroys industry profits, so a monopoly generates weakly more producer surplus than any other market structure. Finally

assume

$$k_i \leq \overline{PS}^*(C,n) \tag{6}$$

for each firm i among the n actual entrants. This inequality is a minimal assumption on the rationality of the entry decision: if it is violated for some i, that firm would have gained by staying out of the market. We have the following theorem.

Theorem 4. *Consider any model of competition C satisfying conditions* (4)–(6) *and any number of potential entrants $N \in \mathbb{N}$. The upper bound on relative deadweight loss is weakly higher than under monopoly:*

$$\sup_{\{k_i \geq 0 | i=1,\ldots,N\}} \left[\frac{DWL^*(C,N)}{TS^{**}}\right] \geq 1-\rho^*, \tag{7}$$

where $DWL^*(C,N)$ is the deadweight loss in model C with N firms and $\rho^* = PS^*/TS^{**}$ is the monopoly surplus-extraction ratio.

The proof is intuitive. With N firms, the entry costs may be sufficiently high for all but one firm that the only feasible outcome involves monopoly. Thus the monopoly distortion is always a possibility with any of the models under consideration. The $N-1$ additional entrants and entry costs just add "degrees of freedom" that can create even greater distortions. A formal proof is provided in Appendix A.

If one is willing to impose further assumptions regarding symmetry and free entry, much more precise bounds can be obtained. To this end, assume there are a potentially unlimited number of symmetric potential entrants (i.e., $N \uparrow \infty$), each facing fixed cost $k \geq 0$. Assume that for any number of actual entrants $n \in \mathbb{N}$ the model of competition C has a symmetric equilibrium. If the symmetric equilibrium is not unique, we will isolate a focal one that is always played in that situation. Let $PS^*(C,n)$ denote one firm's producer surplus and $CS^*(C,n)$ denote consumer surplus in this equilibrium.

Analogous to conditions (4)–(6), assume

$$PS^*(C,1) = PS^* \tag{8}$$

$$(n+1)PS^*(C,n+1) \leq nPS^*(C,n) \tag{9}$$

$$PS^*(C,n^*+1) \leq k \leq PS^*(C,n^*), \tag{10}$$

where n^* is the equilibrium number of entrants in the two-stage game of entry followed by competition governed by model C. Symmetry allows us to replace the maximum $\overline{PS}(C,n)$ among firms' idiosyncratic producers surpluses by the common producer surplus $PS(C,n)$ and to replace the idiosyncratic fixed cost k_i by k. Condition (9) is stronger than (5) in that it involves a strict inequality and assumes each of any number of market participants, not just a monopoly, is harmed by further entry. Condition (10) is stronger than (6). By omitting a lower bound on k_i, (6) allows for possible barriers to entry. The lower bound on k in (10) embodies the free-entry condition that no strictly profitable entry opportunities remain unexploited in

equilibrium. We have the following theorem, proved in Appendix A.

Theorem 5. *Consider any model of competition C with symmetric firms and a symmetric equilibrium for all $n \in \mathbb{N}$ satisfying conditions* (8)–(10)*. The upper bound on relative deadweight loss is the same as in Theorem 1 for a monopoly market, i.e.,*

$$\sup_{k \geq 0} \left[\frac{DWL^*(C, n^*)}{TS^{**}} \right] = 1 - \rho^*, \tag{11}$$

if and only if

$$PS^* \leq CS^*(C, n) + PS^*(C, n) \tag{12}$$

for all $n \in \mathbb{N}$. Otherwise

$$\sup_{k \geq 0} \left[\frac{DWL^*(C, n^*)}{TS^{**}} \right] = 1 - \frac{CS^*(C, \hat{n}) + PS^*(C, \hat{n})}{TS^{**}}, \tag{13}$$

where $\hat{n}$ solves $\min_{n \in \mathbb{N}} [CS^*(C, n) + PS^*(C, n)]$.

With the additional assumptions of symmetry and free entry, Theorem 5 is able to replace the weak inequality bounding relative deadweight loss in Theorem 4 with a tight bound. The tight bound is the lower of two values in (11) and (13), depending on the extensive margin at which deadweight loss is worse: insufficient or excess entry. If condition (12) holds for all $n \in \mathbb{N}$, the supremum on relative deadweight loss is generated by insufficient entry, i.e., by a socially valuable product that ends up not being produced. In this case, the bound $1 - \rho^*$ provided in Theorem 1 for the monopoly case carries over to the general model of competition C.

On the other hand, if condition (12) does not hold for some $n \in \mathbb{N}$, the greatest source of deadweight loss moves from insufficient to excess entry. The proof shows the problem of excess entry is worst when $\hat{n}$ firms enter the market, for $\hat{n}$ defined in the statement of the theorem. This amount of entry can be induced by setting the fixed cost equal to $k = PS^*(C, \hat{n}) - \epsilon$ for sufficiently small $\epsilon > 0$, generating equilibrium welfare $W^* = CS^*(C, \hat{n}) + \hat{n} PS^*(C, \hat{n}) - \hat{n} k = CS^*(C, \hat{n}) - \hat{n}\epsilon$. The social planner would prefer to obtain first-best welfare $W^{**} = TS^{**} - k = TS^{**} - PS^*(C, \hat{n}) + \epsilon$ by having one firm enter and price at marginal cost. Deadweight loss is thus $W^{**} - W^* = TS^{**} - CS^*(C, \hat{n}) - PS^*(C, \hat{n}) + (\hat{n} - 1)\epsilon$, which gives the bound in equation (13) in the limit as $\epsilon \downarrow 0$.

Whether or not condition (12) holds is *a priori* ambiguous. The right-hand side adds consumer and producer surplus, while the left-hand side involves just producer surplus. However, the producer surplus on the left-hand side is that for a monopoly, exceeding the producer surplus on the right-hand side earned by one of $\hat{n}$ competitors.

Condition (12) holds in a wide range of familiar cases. Suppose C is given by Cournot competition among homogeneous firms facing linear inverse demand $P = a - bQ$ and constant marginal cost c. One can show $PS^* = (a-c)^2/4$, $CS^*(C, n) = 2PS^* n^2/(n+1)^2$, and $PS^*(C, n) = 4PS^*/(n+1)^2$. Dividing through by PS^*, we see that condition (12) holds if and only if $n^2 - 2n + 3 \geq 0$, which is true for all $n \in \mathbb{N}$. Hence, the

bound in Theorem 1 for the monopoly case carries over to free entry in a symmetric Cournot model with linear demand and cost.

The bound in Theorem 1 for the monopoly case also carries over to free entry in a symmetric Bertrand model with linear cost. Then, $CS^*(C,n)+PS^*(C,n) = TS^{**} \geq PS^*$ for all $n > 1$ because Bertrand competition among multiple firms leads to marginal-cost pricing. Further, $CS^*(C,1)+PS^*(C,1) \geq PS(C,1) = PS^*$. These facts together imply that (12) holds for all $n \in \mathbb{N}$.

For a contrasting case in which excess entry is the source of the most deadweight loss and equation (13) is the relevant bound, consider a model of a perfect cartel, maintaining the monopoly price regardless of how many firms enter. Then $CS^*(C,n) = CS^*$ and $PS^*(C,n) = PS^*/n$. It follows that (12) holds for all $n \in \mathbb{N}$ if and only if it holds in the limit as $n \uparrow \infty$, i.e., $PS^* \leq \lim_{n\in\mathbb{N}}(CS^* + PS^*/n) = CS^*$. We conclude that excess entry is the source of the most deadweight loss and equation (13) is the relevant bound in a cartel model if $PS^* > CS^*$, i.e., if producer surplus exceeds consumer surplus when a monopoly serves the market. This will be the case, for example, in a market with linear demand and cost, since it can be shown using the formulas from a previous paragraph that producer surplus is twice consumer surplus in the monopoly equilibrium. That said, condition (12) can hold in the perfect-cartel case, in particular for any demand and cost specifications for which $PS^* \leq CS^*$.

4. Characterizing the Producer-Surplus Ratio, ρ^*

The theorems presented in the previous section highlight the relevance of the monopoly surplus-extraction ratio, ρ^*, for the computation of bounds on relative deadweight loss and bounds on gains and losses from policy interventions. The theorems were general, applying to different types of investment and to a broad range of market structures beyond monopoly including the familiar oligopoly models such as Cournot, Bertrand, and perfect cartels. These general results motivate the analysis of the determinants of ρ^* undertaken in this section. The analysis is broken into three subsections. The first subsection presents a rescaling that places demands for diverse products on the same footing. The second introduces the STRZ demand curve. The third presents a formula for decomposing equilibrium changes into changes in how much the demand curve resembles a STRZ one and other factors.[7]

[7] We can connect the results in each subsection of Section 4 to related results in our earlier paper, Kremer and Snyder (2015). In Section 4.1, the demand rescaling is new. Our previous paper worked with a vaccine that was costless to produce and that was sold to a unit mass of consumers who differed only in disease risk, so no rescaling was needed. The lemmas below, though related to the same-numbered lemmas in the previous paper, are more general and thus require new proofs. Proposition 1 is identical to Proposition 3 of Kremer and Snyder (2015). In Section 4.2, the STRZ demand was derived in our previous paper. We provide a more elegant and rigorous proof that it is the producer-surplus minimizer in its class here. The analytical expression for the producer surplus from the STRZ demand in Proposition 3 is new. In Section 4.3, our previous paper proposed the index of Zipf similarity Z, but the decomposition of a market change in terms of Z in equations (26) and 27 are new.

4.1. Rescaling Demand

Rather than working directly with demand $Q(p)$, we will employ a convenient change of variables to allow us to work with net consumer values, further rescaled so that relevant function's argument and value are unitless. Define

$$x = \frac{p-c}{p^{\max}-c}, \tag{14}$$

interpreted as the marginal consumer's net value for the product relative to the maximum conceivable. We will call x the *rescaled consumer value* for short. For the domain of feasible equilibrium prices $p \in [c, p^{\max}]$, the range of x is $[0,1]$. Let $\Phi : [0,1] \to [0,1]$ be the function satisfying

$$\Phi(x) = \Phi\left(\frac{p-c}{p^{\max}-c}\right) = \frac{Q(p)}{q^{**}}, \tag{15}$$

Intuitively, $\Phi(x)$ is the share of consumers in the first best whose relative net values are at least x. Dividing by first-best quantity q^{**}—which the monopoly quantity never exceeds—ensures $\Phi(x) \leq 1$ for all $x \in [0,1]$ and thus that the range of Φ is $[0,1]$. Figure 3 illustrates the derivation of $\Phi(x)$ in panel (b) from the original demand curve in panel (a). We will call Φ the *rescaled demand curve*. Points A through D in panel (a) map to corresponding points A' through D' in panel (b). As the figure shows, rescaling preserves the basic shape of the original demand curve.[8]

The key assumption needed for this rescaling to be valid are that q^{**} and $p^{\max}$ are finite. If $c > 0$, q^{**} will be finite as long as demand is finite at any positive price, a relatively weak assumption. If $c = 0$, then for q^{**} to be finite, a stronger assumption that consumers reach a satiation point with the free good. On the finiteness of $p^{\max}$, our results would be similar if we replaced the assumption of finite $p^{\max}$ with finite p^0.[9]

An equivalent, distributional interpretation of Φ, will be particularly convenient in the subsequent discussion. Letting X be the rescaled value of a consumer randomly selected from those purchasing in the first best, then $\Phi(x) = \Pr(X \geq x)$. Following along with this distributional interpretation, we can define the cumulative distribution function $F(x) = \Pr(X \leq x)$ and its complement $\bar{F}(x) = \Pr(X > x) = \Phi(x) - \Pr(X = x)$. Let $\mu = \int_0^1 x\,dF(x)$ be the mean of X. The next lemma provides several equivalent expressions for μ. The proof, provided in Appendix A, uses integration by parts and facts about Riemann integrals.

Lemma 1. $\mu = \int_0^1 \Phi(x)dx = TS^{**}/q^{**}(p^{\max}-c)$.

By definition, μ is the mean of rescaled consumer values. The first equality in Lemma 1 provides an equivalent interpretation of μ as the area under the rescaled demand curve, Φ. The second equality in

[8] Our rescaling is similar to the stretch parameterization used in Weyl and Tirole's (2012) study of optimal rewards for innovation. In both, the horizontal axis is rescaled/stretched so that the first-best quantity is 1. We rescale the vertical axis so the maximum conceivable value is 1 whereas they scale it so the monopoly price is 1.

[9] The only formal difference would be that p^0 rather than $p^{\max}$ would appear in the the rescaling formula in equation (14). Conceptually, any result comparing a given demand curve with intercept p^0 to other demand curves would be conditional on the others not having an intercept greater than p^0. The virtue of rescaling by the maximum conceivable value $p^{\max}$ when this is finite is that no demand curve can have a higher intercept and thus no additional constraint about intercepts need be considered.

Lemma 1 provides yet another equivalent interpretation of μ in terms of the original (unscaled) consumer values and demand curve. According to this interpretation, μ is the mean surplus in the first best per unit of output, TS^{**}/q^{**}, as a proportion of the maximum conceivable surplus, $p^{\max}-c$. Borrowing from engineering terminology, this can be called the *mean-to-peak surplus ratio*.

The next lemma provides a simple formula for ρ^* that can be read off a graph, as illustrated in the second panel of Figure 3. Note that $x\Phi(x)$ is the area of the rectangle of height x inscribed under Φ. Let $REC^* = \max_{x\in[0,1]}[x\Phi(x)]$ denote the area of the largest such rectangle, shown in the figure as the shaded region. Lemma 1 showed that μ equals the whole area under Φ. The next lemma, proved in Appendix A, states that $\rho^* = REC^*/\mu$, implying that ρ^* is the ratio of the area of the shaded rectangle to the area under the whole curve.

Lemma 2. *The surplus-extraction ratio in a market satisfies $\rho^* = REC^*/\mu$. Under the maintained assumptions, $\rho^* \in [0,1]$.*

Much of the subsequent analysis will be devoted to examining the conditions on the distribution of X leading to high or low values of ρ^*. The next proposition covers one extreme: a necessary and sufficient condition for $\rho^* = 1$ is that positive values of X are homogeneous. The proposition, restated here for reference, is identical (except for some minor notational differences) to Proposition 3 of Kremer and Snyder (2015); see that paper for a proof.

Proposition 1. *$\rho^* = 1$ in a market if and only if there exists some $x' \in (0,1]$ such that $\Pr(X = x') = 1$. Otherwise, $\rho^* < 1$.*

An implication of Proposition 1 is that, starting from a homogeneous X, introducing variance causes ρ^* to fall from 1 to some lower value. This leads one to hope that ρ^* is monotonically decreasing in the variance of X, allowing variance to be used as a sufficient statistic for ρ^*. As Proposition 4 of Kremer and Snyder (2015) shows, this is not the case: not only is variance not a sufficient statistic for ρ^* but neither is skewness, kurtosis, or any other higher moment of the distribution of X.[10] The next section provides a sufficient statistic based not on moments but on the the resemblance of the distribution of X to the worst-case distribution, which has the lowest possible ρ^* for a given μ. As indicated in the introduction, this is the STRZ distribution, which we turn to next.

4.2. STRZ Demand

Let $\underline{\Phi}(x,\mu)$ denote the producer-surplus minimizer among rescaled demand curves having mean (equivalently area underneath) equal to μ. Brooks (2013) and Kremer and Snyder (2015) independently derived its

[10] The result holds whether the moment is raw, central, or standardized. Kremer and Snyder (2015) also show that other standard measures of heterogeneity besides variance (mean-preserving spreads and increases in the Gini mean difference) are not monotonically related to ρ^*.

functional form:

$$\underline{\Phi}(x,\mu) = \min\left\{\frac{A(\mu)}{x}, 1\right\}, \tag{16}$$

where $A(\mu)$ is the implicit solution to

$$\mu = A(\mu)[1 - \ln A(\mu)]. \tag{17}$$

The $A(\mu)$ solving equation (17) is the unique value preserving the property $\mu = \int_0^1 \underline{\Phi}(x,\mu)dx$.

We will not repeat the derivation here, but instead provide intuition with the help of Figure 2 from the introduction. Begin by considering the grey demand curve, which happens to be linear but, because it has unit intercepts, can represent any rescaled demand curve $\Phi(x)$. The shaded rectangle (which happens to be a square) is the largest rectangle that can be inscribed under it. Now consider transforming the grey demand curve by moving some mass away from the corner of the shaded square to other parts of the curve, maintaining the same area under the whole curve—thus maintaining μ. This transformation will reduce REC^* and, because μ is constant, reduce ρ^*, which equals $\rho^* = REC^*/\mu$ by Lemma 2. The process can be repeated, at each step moving mass away from the corner of the largest rectangle inscribed under the transformed curve elsewhere, again reducing REC^* and thus ρ^*. The limit of this process is the demand curve drawn in black in the figure: no further transformations are possible because inscribed rectangles have the same area. For all $x \geq A(\mu)$, $x\underline{\Phi}(x,\mu) = A(\mu)$, a constant independent of x, verifying that the demand curve in equation (16) has the equal-area property and is thus the unique producer-surplus minimizer.

Proposition 2. *Consider the demand curve $\underline{\Phi}(x,\mu)$ defined in* (16)*, where the associated value of $A(\mu)$ is the implicit solution to* (17)*. This is the unique (almost everywhere) minimizer of producer surplus among the set of rescaled demand curves having a mean rescaled value of at least $\mu \in (0,1)$.*

Proposition 2 is a restatement of Proposition 5 from Kremer and Snyder (2015) for general products beyond vaccines. We were able to derive a more rigorous and elegant proof, which we provide in Appendix A.

Kremer and Snyder (2015) call $\underline{\Phi}(x,\mu)$ the symmetrically truncated Zipf (STRZ) demand. The origin of the "Zipf" part of the name can be better understood by interpreting the demand curve as a complementary cdf over consumer values. The function in equation (16) has a power-law form with power-law exponent equal to 1 (see Gabaix 2009), thus satisfying the definition of Zipf's law. Students will recognize $\underline{\Phi}(x,\mu)$ as a globally unit-elastic demand curve—as expected from the property that where demand is unit elastic, revenue is unchanging in price. It is a special form, though, with truncated endpoints making it symmetric. Figure 5 shows how the shape of $\underline{\Phi}(x,\mu)$ varies with μ. The black curve is the same STRZ demand curve from Figure 2, having a mean value of $\mu = 0.5$. The grey curve, which has a lower mean of $\mu = 0.25$, hugs the axes more closely.

We can use the STRZ demand to construct a lower bound on the producer-surplus ratio. Let $\underline{REC}(\mu) = \max_{x\in[0,1]} x\underline{\Phi}(x,\mu)$ denote producer surplus in the market with STRZ demand and $\underline{\rho}(\mu)$ denote the associated surplus-extraction ratio, i.e., $\underline{\rho}(\mu) = \underline{REC}(\mu)/\mu$. It is immediate from Proposition 2 that $\underline{\rho}(\mu)$ must be a

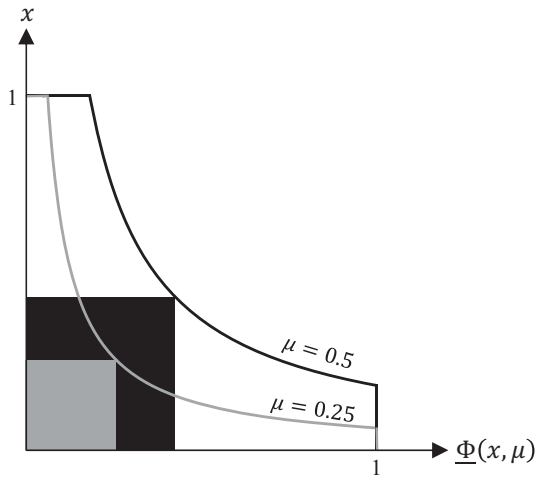


Figure 5: STRZ demands for various μ.

lower bound on the surplus-extraction ratio. To see this, let ρ^* be the surplus-extraction ratio associated with rescaled demand $\Phi(x)$ in some market m. Then $\rho^* = REC^*/\mu \geq \underline{REC}(\mu)/\mu = \underline{\rho}(\mu)$, with strict inequality as long as $\Phi(x)$ and $\underline{\Phi}(x,\mu)$ are not the same curves almost everywhere. We can derive a formula for $\underline{\rho}(\mu)$ by first noting

$$\underline{REC}(\mu) = \max_{x\in[0,1]} [x\underline{\Phi}(x,\mu)] = \max_{x\in[0,1]} [\min\{A(\mu),x\}] = \min\{A(\mu),1\} = A(\mu), \tag{18}$$

implying $\underline{\rho}(\mu) = A(\mu)/\mu$. Substituting $A(\mu) = \mu\underline{\rho}(\mu)$ into equation (17) and simplifying yields

$$1 = \underline{\rho}(\mu)[1 - \ln\mu - \ln\underline{\rho}(\mu)]. \tag{19}$$

While this equation does not have a closed-form solution for $\underline{\rho}(\mu)$ in terms of elementary functions, it does have an analytic solution in terms of the lower branch of the Lambert W function, $W_{-1}(z)$.[11]

Proposition 3. *The producer-surplus ratio for a market with demand curve $\underline{\Phi}(x,\mu)$ is*

$$\underline{\rho}(\mu) = \frac{-1}{W_{-1}(-\mu/e)}, \tag{20}$$

providing a lower bound on ρ^* for any market with mean rescaled value μ. Furthermore, $\underline{\rho}'(\mu) > 0$, $\lim_{\mu\downarrow 0}\underline{\rho}(\mu) = 0$, and $\underline{\rho}(1) = 1$.

The proof provided in Appendix A verifies that the formula for $\underline{\rho}(\mu)$ in equation (20) satisfies (19) and

[11] The Lambert W function, also called the product log function, is the inverse relation $W(z)$ of the function $z = We^W$. Its branches are built-in functions in standard mathematical software packages including Mathematica, Matlab, and R.

establishes the derivative and limit results. For reference, the proof includes a tabulation of $\underline{\rho}(\mu)$ for a grid of values of μ.

We can apply the formula to compute the producer-surplus ratios for the example STRZ demand curves in Figure 5. Considering the black demand curve, for which $\mu = 0.5$, we have $\underline{\rho}(0.5) = 0.373$; considering the grey curve, for which $\mu = 0.25$, we have $\underline{\rho}(0.25) = 0.271$. The figure illustrates that the producer-surplus rectangles inscribed under the associated STRZ demands cover a smaller proportion of area the lower is μ. In the limit, as the proposition states, $\underline{\rho}(\mu)$ approaches 0 as μ approaches 0. This limiting result has a crucial implication for our analysis that deserves highlighting. The result implies that cases can be constructed in which ρ^* is arbitrarily close to 0. But this means, by Theorem 1, that cases can be constructed such that deadweight loss is arbitrarily close to 1.

4.3. Decomposition

Kremer and Snyder (2015) define the Zipf-similarity, Z, of demand in a given market by the ratio of unextracted surplus in that market relative to that in the worst (STRZ) case with the same μ:

$$Z = \frac{1-\rho^*}{1-\underline{\rho}(\mu)}. \tag{21}$$

The fact that $\underline{\rho}(\mu) \leq \rho^* \leq 1$ ensures $Z \in [0,1]$, with $Z = 0$ for homogeneous consumers and $Z = 1$ for a market with STRZ demand.

With this definition of Z, a simple rearrangement allows potential deadweight loss to be expressed as the product of two factors:

$$1-\rho^* = Z[1-\underline{\rho}(\mu)]. \tag{22}$$

In words, potential deadweight loss is the product of the Zipf-similarity of demand and $1-\underline{\rho}(\mu)$, which can be interpreted as the difficulty in capturing surplus from a STRZ demand curve. Since $1-\underline{\rho}(\mu)$ is a monotonic function of the parameter μ, we see that potential deadweight loss $1-\rho^*$ is completely determined by two variables, Z and μ and moreover is monotonic in them.

According to Gabaix's (2009) survey, power laws—Zipf distributions in particular—characterize the distribution of many natural and social phenomena, including earthquakes and word frequency, but also including economic phenomena that may translate into distributions of values for products. For example, the upper tail of city size, firm size, income, wealth, CEO compensation, stock price changes and volatility, and various international-trade indexes are Zipfian. Equation (22) implies that R&D incentives will be quite low in markets in which demand is proportional to these variables. For example, the demand for lifesaving pharmaceuticals may be proportional to income or wealth; the demand for software used by enterprises or municipal governments might be proportional to firm or city size. As highlighted by our careful treatment of demand rescaling, the relevant distribution to analyze for the Zipf shape is not the raw distribution of

consumer values but net values above marginal production cost (normalizing values by the peak value and quantity by first-best quantity to ensure the resulting demand has unit intercepts). The focus on net values means that markets in which only the upper tail is Zipfian may nonetheless generate highly Zipf-similar distributions of net values, since consumers in the lower tail of gross values may not have positive net values and may thus not be part of the relevant distribution. The case of goods produced at little or no marginal cost—software, digital media, certain small-molecule drugs—is straightforward because net and gross consumer values are similar for them.

Equation (22) provides a starting point toward a useful decomposition of comparative-statics changes. Consider a change in parameters from some ex ante constellation indicated by subscript 0 to some ex post constellation indicated by subscript 1. The resulting percentage changes in $1-\rho^*$, Z, and $1-\underline{\rho}(\mu)$ (sometimes called semi-elasticities) are

$$\widehat{1-\rho^*} = \frac{(1-\rho_1^*)-(1-\rho_0^*)}{1-\rho_0^*} \tag{23}$$

$$\hat{Z} = \frac{Z_1-Z_0}{Z_0} \tag{24}$$

$$\widehat{1-\underline{\rho}(\mu)} = \frac{[1-\underline{\rho}(\mu_1)]-[1-\underline{\rho}(\mu_0)]}{1-\underline{\rho}(\mu_0)}. \tag{25}$$

Differencing (22) and rearranging leads to the following decomposition of the change in potential deadweight loss:

$$\widehat{1-\rho^*} = \hat{Z} + \left(\frac{Z_1}{Z_0}\right) \widehat{1-\underline{\rho}(\mu)}. \tag{26}$$

Equation (26) tells us how much of the change in potential deadweight loss is due to changes in Z and μ in percentage-point terms. That is, the first term on the right-hand side is the percentage-point change in potential deadweight loss due to the change in Z and the second term the percentage-point change due to the change in μ.

Perhaps more useful is to express the decomposition in percentages (i.e., as a percent of the total change) rather than percentage points. Dividing (26) through by its left-hand side yields, after some manipulation,

$$100\% = \left(\frac{1-\rho_0^*}{\rho_1^*-\rho_0^*}\right) \hat{Z} + \left(\frac{Z_1}{\rho_1^*-\rho_0^*}\right) [\underline{\rho}(\mu_1)-\underline{\rho}(\mu_0)]. \tag{27}$$

The first term on the right-hand side tells us what percent of the total (100%) change in potential deadweight loss is due to the change in Z and the second, residual term what percent is due to the change in μ. We will apply this decomposition repeatedly in the calibrations in Section 7.

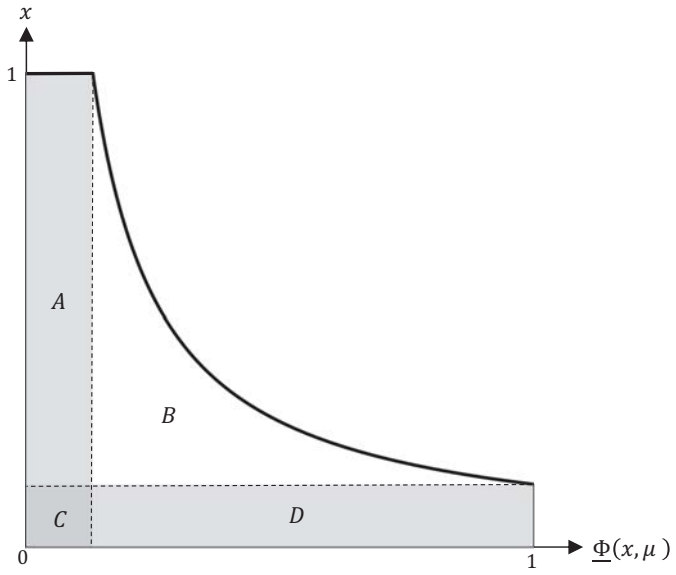


Figure 6: Static deadweight loss with Zipf-similar demand.

5. Harberger Deadweight Loss

The previous subsection emphasized the implications of Zipf-similarity for dynamic R&D incentives and overall deadweight loss *DWL*. Zipf-similarity also has implications for Harberger deadweight loss *HDWL* *.[12] A monopolist in a market with rescaled demand $\underline{\Phi}$ is indifferent over a wide range of prices. Thus *HDWL* * is indeterminate, depending on which equilibrium price the monopolist selects. For example, in Figure 6, the monopolist is indifferent between selling to all consumers with positive net values (earning producer surplus proportional to the area of *C* plus *D*) or just to the mass of the highest-value consumers (earning producer surplus proportional to the area of *A* plus *C*). In the first, there is no Harberger deadweight loss; in the second, Harberger deadweight loss is proportional to the areas of regions *B* plus *D*. This turns out to be the greatest Harberger deadweight loss possible for any market with mean rescaled value μ. We have the following proposition, proved in Appendix A.

Proposition 4. *Consider a market with rescaled demand* $\underline{\Phi}(x,\mu)$. *There exists an equilibrium in this market in which relative Harberger deadweight loss* $HDWL^*/TS^{**}$ *equals* $1-\underline{\rho}(\mu)$, *which is strictly greater than in any other market with a rescaled demand that is not almost everywhere identical but with the same mean rescaled value* μ.

Let $\underline{HDWL}^*$ denote the worst-case Harberger deadweight loss associated with the STRZ demand curve $\underline{\Phi}$. While $\underline{HDWL}^*$ is only realized in one among a continuum of equilibria, a unique equilibrium with Hargerger deadweight loss arbitrarily close to $\underline{HDWL}^*$ can be generated by perturbing demand by adding tiny mass to the highest support point of $\underline{\Phi}$. Combining Proposition 4 with the result from Section 4.2 that $\lim_{\mu\downarrow 0}\underline{\rho}(\mu)=0$ implies that Hargerger deadweight loss can come close to fully dissipating total surplus in

[12] Our previous related paper, Kremer and Snyder (2015), did not analyze Harberger deadweight loss.

a market with a Zipf-similar demand and low mean rescaled value.

The next proposition, proved in Appendix A, turns from conditions under which Harberger deadweight loss is large to those under which it is small.

Proposition 5. *In a given market, HDWL*/TS** is bounded above by $Z[1-\underline{\rho}(\mu)]$. Harberger deadweight loss is vanishingly small in the limit as $Z \downarrow 0$ or $\mu \uparrow 1$.*

The limiting results are intuitive: in both limits $Z \downarrow 0$ and $\mu \uparrow 1$, the population of consumers with positive demand becomes perfectly homogeneous, causing the Harberger triangle to disappear.

We mentioned that adding a tiny mass to the highest support point of $\underline{\Phi}$ can generate a unique equilibrium with close to the maximum possible Harberger deadweight loss. Conversely, adding tiny mass to the lowest support point of $\underline{\Phi}$ would generate a unique equilibrium in which Harberger deadweight loss is eliminated. Thus welfare in a market with Zipf-similar demand is chaotic. This leaves a powerful role for policy in such a market. Price ceilings or mandatory licensing could swing the equilibrium to the one without Harberger deadweight loss while generating negligible reductions in dynamic R&D incentives. This can be seen in Figure 6, where the imposition of such policies could ensure that the equilibrium in which all consumers with positive values purchase while keeping producer surplus arbitrarily close to proportional to the original areas of *C* plus *D*. A tiny subsidy would likewise eliminate Harberger deadweight loss at negligible fiscal cost.

The results in this section, though derived for the monopoly market structure, have implications for more general market structures. Price under monopoly is typically higher than under more competitive market structures. Thus Harberger deadweight loss is also typically higher under monopoly than under other market structures, implying that the upper bound on Harberger deadweight loss from a monopoly will also bound Harberger deadweight loss from other market structures. This is the logic of the following proposition, proved formally in Appendix A.

Proposition 6. *Consider any model of competition C among n homogeneous firms having a symmetric equilibrium resulting in market price $x^*(C,n)$. Assuming this price is weakly lower than under monopoly, i.e., $x^*(C,n) \leq x^*$, then relative Harberger deadweight loss from this model of competition, HDWL $^*(C,n)$/TS**, is bounded above by $Z[1-\underline{\rho}(\mu)]$.*

Tighter bounds can be obtained in specific competition models. Consider, for example, a Cournot model. To construct the demand curve maximizing relative Harberger deadweight loss under this model, we can use the same approach as we did with a monopoly, ensuring firms only serve the highest-demand consumers, leaving all the rest of the consumers to constitute the deadweight loss triangle. For this outcome to be an equilibrium, each firm must weakly prefer the revenue obtained from serving a $1/n$ share of highest-demand consumers to any higher output. The worst case is generated by distributing the given mass μ under the demand curve such that each firm is indifferent among all these quantities. Appendix B works out the precise formula for the resulting worst-case demand curve, denoted $\underline{\Phi}(x,\mu,n)$. The appendix provides a figure

graphing several examples of $\underline{\Phi}(x,\mu,n)$, showing that they are roughly a clockwise rotation of the STRZ demand through a point, with a larger rotation the higher is n. Since the equilibrium price extracts all the consumer surplus of the subset of consumers served in this Cournot construction, the Harberger-deadweight-loss ratio is $HDWL^*/TS^{**} = 1-\underline{\rho}(\mu,n)$, where $\underline{\rho}(\mu,n)$ is the ratio of these n firms' worst-case producer surpluses to first-best surplus. For reference, the appendix provides a table of $\underline{\rho}(\mu,n)$ for different values of μ and n. For example, fixing $\mu = 0.2$, one can read from the table that $\underline{\rho}(0.2,1) = 0.25$, $\underline{\rho}(0.2,4) = 0.53$, and $\underline{\rho}(0.2,16) = 0.78$, implying an upper bound on Harberger deadweight loss of $1-\underline{\rho}(0.2,1) = 0.75$ for the monopoly case ($n = 1$), $1-\underline{\rho}(0.2,4) = 0.47$ for $n = 4$ Cournot firms, and $1-\underline{\rho}(0.2,16) = 0.22$ for $n = 16$ Cournot firms. Evidently, with as many as $n = 16$ Cournot competitors, the potential for Harberger deadweight loss, even in the worst case, is quite limited.

6. Specific Distributions

The analysis so far has provided general results for unrestricted distributions of consumer values X. This subsection derives additional results in specific cases in which more structure can be put on demand. We first consider discrete distributions of consumer values, bounding ρ^* as a function of the number of consumer types. Then we move to continuous distributions, first beta distributions, then general distributions having global curvature properties. The last subsection builds on the simple insight that the monopolist can earn at least as much as by pricing so that the median consumer is marginal. This allows us to derive a quite general lower bound on ρ^* as a function of the ratio of the median to the mean of X. This simple insight has powerful consequences, allowing us to leverage a rich set of results from the statistics literature on the "mean-median-mode inequality" to provide alternative bounds on ρ^* in a variety of cases.[13]

6.1. Discrete Distributions

The following proposition provides a lower bound on ρ^* when X is discrete.

Proposition 7. *Consider the set of markets in which the distribution of consumer values X has T discrete types. Then $1/T$ is a tight lower bound on ρ^*.*

The proof is similar to that of a related result (Proposition 8) in Kremer and Snyder (2015); the reader is referred there for a proof. Here we will provide intuition for the proposition, illustrated by Figure 7 in an example with $T = 3$ discrete types. Let $A-F$ represent the areas of the indicated rectangles. By construction, the rescaled demand curve is a discrete analogue to the symmetrically truncated Zipf distribution, with equal producer surplus whichever of the three types the firm targets as the marginal consumer: $A+B+C$ from

[13] To connect the results in this section to our earlier paper, the results for discrete distributions appears as Proposition 8 in Kremer and Snyder (2015), restated here for reference. Proposition 13 generalizes the results for concavity in Proposition 9 and log-concavity in Proposition 10 to c-concavity. The rest of the results—for beta, Pareto, and the mean-median-mode inequality—are new here.

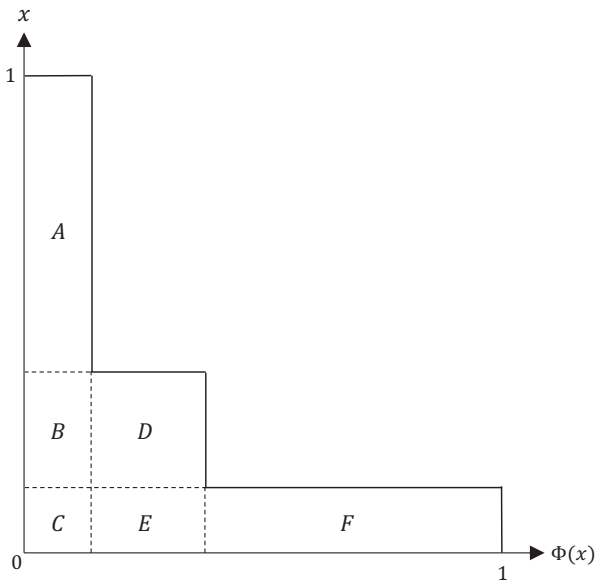


Figure 7: Example with three discrete types.

targeting the high type equals $B+C+D+E$ from targeting the middle type equals $C+E+F$ from targeting the low type. Then

$$\rho = \frac{A+B+C}{A+B+C+D+E+F} = \frac{A+B+C}{3(A+B+C)-(B+2C+E)}. \tag{28}$$

We see that ρ^* exceeds $1/3$ in proportion to the overlap among the inscribed rectangles captured by the $B+2C+E$ term. In the limit as μ approaches 0, the demand curve increasingly hugs the axes, and the area of these overlapping rectangles becomes second order.

6.2. Beta Distribution

The next special case moves from discrete to continuous distributions. We consider the beta distribution, a useful one in our context because it has the $[0,1]$ domain suitable for rescaled consumer values but with just two parameters can fit many distribution shapes fairly closely. The beta density is

$$f(x,a,b) = \frac{x^{a-1}(1-x)^{b-1}}{\int_0^1 t^{a-1}(1-t)^{b-1}dt}. \tag{29}$$

It turns out that considerable insight can be obtained in our setting by re-parameterizing the beta in terms of its mean μ and variance σ^2. Standard results for the beta give

$$\mu = \frac{a}{a+b} \tag{30}$$

$$\sigma^2 = \frac{ab}{(a+b)^2(a+b+1)}. \tag{31}$$

This system of two equations can be solved for a and b in terms of μ and σ^2. Substituting into the beta

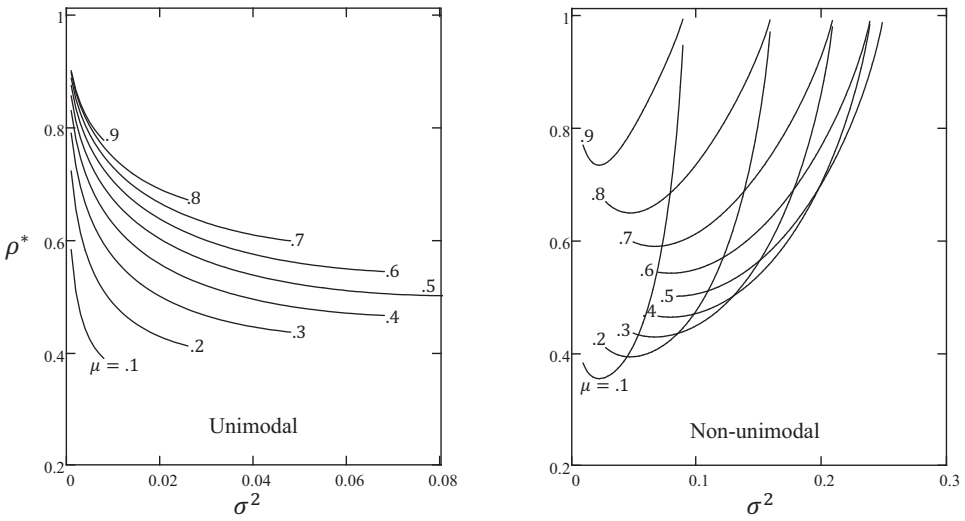


Figure 8: Surplus-extraction ratio ρ^* for beta distribution. First panel shows only parameters for which beta is unimodal; second for non-unimodal. Each curve fixes a different indicated value of μ.

density gives the re-parameterized density

$$\tilde{f}(x,\mu,\sigma^2) = f(x,\mu k(\mu,\sigma^2),(1-\mu)k(\mu,\sigma^2)), \tag{32}$$

where $k(\mu,\sigma^2) = \mu(1-\mu)/\sigma^2 - 1$.

Our results for the beta turn out to depend on whether or not it is unimodal. In the standard parameterization, beta is unimodal if and only if $a,b > 1$. (It reduces to the uniform distribution if $a = b = 1$. The beta density is U-shaped if $a,b < 1$.) Using equations (30) and (31), it can be shown that an equivalent condition for $a,b > 1$ is

$$\sigma^2 < \min\left\{\frac{\mu^2(1-\mu)}{1+\mu}, \frac{\mu(1-\mu)^2}{2-\mu}\right\}. \tag{33}$$

Thus the re-parametrized beta is unimodal if and only if (33) holds. Intuitively, holding its mean constant, a beta is unimodal if and only if its variance is sufficiently small.

Figure 8 graphs ρ^* over a fine grid of parameters μ,σ^2. Each point on a curve, representing a unique (μ,σ^2) parameter configuration, was generated by numerically optimizing $REC^* = \max_{x\in[0,1]}[x\Phi(x)]$ using standard quasi-Newton methods and dividing by μ to obtain ρ^*. The first panel shows the range of parameters for which the beta is unimodal, in particular, shows all values of σ^2 satisfying equation (33) for each μ. For each μ, ρ^* approaches 1 as σ^2 approaches 0; intuitively, the firm can extract all surplus by pricing just below the mode as the mass of values piles up at the mode. As σ^2 increases holding μ constant, ρ^* falls in the unimodal case. The second panel shows values of the parameters for which the beta is non-unimodal. Here, ρ^* is non-monotonic, increasing for sufficiently high σ^2. What is happening for these high values of σ^2 is that the density asymptotes to infinity at $x = 1$, leading the firm to be able to extract nearly all surplus by setting x^* close to 1.

As is obvious from the second panel of the figure, it is not generally true that ρ^* is monotonic in σ^2 holding μ constant. However, in the unimodal region, ρ^* is strictly decreasing in σ^2 for each μ shown. This pattern holds for all μ we have tried over a fine grid, leading us to conjecture that it is generally true for a unimodal beta that ρ^* is strictly decreasing in σ^2 holding μ, although we have not been able to prove this result.

6.3. Pareto Distribution

We next consider the Pareto distribution, a useful special case because, besides being widely used in economics, it is a generalization of the Zipf distribution featured in our analysis. We will use this special case to better understand the role played by the maintained assumption of an upper bound on consumer values.

According to the textbook definition, a Pareto random variable has distribution function $F(x) = 1 - (x_0/x)^\alpha$ on support $[x_0, \infty)$, where $x_0, \alpha > 0$. With no upper bound on the support, the textbook Pareto does not fit our model, which posited a maximum conceivable consumer value, $p^{\max}$. We relax that feature of the model and proceed with the analysis of the Pareto.

Suppose a monopolist with costless production sells to a continuum of consumers whose unit demands for the good follow the textbook Pareto distribution. Demand then is $Q(p) = 1 - F(p) = \min\{1, (x_0/p)^\alpha\}$, implying $PS(p) = \min\{p, x_0^\alpha p^{1-\alpha}\}$. The equilibrium monopoly price is generically a corner solution: $p^* = x_0$ for $\alpha > 1$ and $p^* = \infty$ for $\alpha < 1$. In the knife-edged case of $\alpha = 1$, the monopolist is indifferent among all $p^* \in [x_0, \infty)$.

We will proceed with the analysis by considering the case $\alpha > 1$. Substituting the equilibrium price $p^* = x_0$ into the expression for $PS(p)$ in the previous paragraph yields $PS^* = x_0$. One can further show $TS^{**} = \int_{x_0}^{\infty} (x_0/p)^\alpha dp = \alpha x_0/(\alpha - 1)$, implying $\rho^* = 1 - 1/\alpha$. The lower bound on ρ^* is reached in the limit as $\alpha \downarrow 1$, when we have $\lim_{\alpha \downarrow 1} \rho^* = 0$. Note that the limit as $\alpha \downarrow 1$ is equivalent to the limit as a general Pareto approaches a Zipf. Hence we have recovered the result from Proposition 2 that the worst case on surplus extraction is achieved by a Zipf distribution. However, we now appear to have punctured the lower bound $\underline{\rho}(\mu)$ on ρ^*, shown in Section 4.2 to be a positive number depending on μ, whereas for the textbook Pareto here the lower bound is 0, independent of the mean consumer value. The apparent discrepancy arises because consumer values are unbounded above with the textbook Pareto. Scaling by the infinite peak value produces a mean-to-peak ratio μ approaching 0 for the textbook Pareto. Proposition 3 states $\lim_{\mu \downarrow 0} \underline{\rho}(\mu) = 0$. Thus the textbook Pareto does not contradict Proposition 2 and other earlier results. In equilibrium in the remaining ($\alpha < 1$) case, the good is sold at an infinite price to an infinitesimal segment of highest-value consumers, straining credibility. As noted in Section 2, the consumer would backward integrate into production before paying an infinite price, motivating an upper bound on Pareto values.

For the rest of the subsection we will consider a modified version of the Pareto with three parameters: upper bound on the support x_1 in addition to lower bound x_0 and shape parameter α. With the upper

bound, this now fits our maintained model and allows for the required rescaling. Rescaling reduces the modified Pareto to a two parameter distribution. Different parameterizations are possible; perhaps the most natural specifies shape parameter $\alpha > 0$ and lower bound x_0, now interpreted as the lower bound on *rescaled* consumer values, so restricted to the interval $x_0 \in (0,1)$. The following rescaled demand captures this parameterization:

$$\Phi(x,\alpha,x_0) = \begin{cases} \min\{1,(x_0/x)^\alpha\} & x \in [0,1] \\ 0 & x > 1. \end{cases} \tag{34}$$

The reader can check that equation (34) nests the STRZ demand $\underline{\Phi}(\mu)$ when $\alpha = 1$ and $x_0 = A(\mu)$, recalling the definition of $A(\mu)$ as the implicit solution to (17). Proposition 2 guarantees that the STRZ demand minimizes ρ^* among demands with fixed μ. We would like to be able to conclude that $\alpha = 1$ is the shape parameter minimizing ρ^* among rescaled Pareto demands with fixed x_0. Unfortunately, we cannot conclude this directly from Proposition 2 because Proposition 2 holds μ not x_0 constant. One can show by integrating equation (34) that fixing x_0 and varying α causes μ to vary for the rescaled Pareto demand:

$$\mu = \int_0^1 \Phi(x,\alpha,x_0)dx = \begin{cases} \dfrac{x_0^\alpha - \alpha x_0}{1-\alpha} & \alpha \neq 1 \\ x_0(1-\ln x_0) & \alpha = 1. \end{cases} \tag{35}$$

Thus Proposition 2 does not directly apply to the exercise of varying α holding x_0 constant. The desired result is supplied by the next proposition, proved in Appendix A.

Proposition 8. *For the family of Pareto rescaled demands in equation* (34)*, if x_0 is held constant and α varied, ρ^* is strictly quasiconvex in α, reaching a minimum at $\alpha = 1$.*

6.4. Means, Medians, and Modes

There exists a large literature in statistics determining when measures of centrality mean, median, and mode can be ordered alphabetically. That is, letting μ denote the mean, m the median, and M the mode of a distribution, the literature seeks conditions under which $\mu \leq m \leq M$, the so-called mean-median-mode inequality. We begin with a theorem that will allow us to leverage such results to help us bound ρ^*.

The theorem is based on the simple observation that the firm can always ensure that half the consumers purchase by charging the median m rescaled consumer value. By so doing, the monopolist can ensure producer surplus of at least $m\Phi(m) = m/2$. Letting $x^* = \text{argmax}_{x\in[0,1]}[x\Phi(x)]$, we have $REC^* = x^*\Phi(x^*) \geq m\Phi(m) = m/2$. Dividing by μ and using Lemma 2, $\rho^* = REC^*/\mu \geq m/2\mu$. This proves the next theorem.

Theorem 6. *Letting μ be the mean and m the median of the distribution of rescaled consumer values, the monopoly surplus-extraction ratio has the following lower bound:*

$$\rho^* \geq \frac{1}{2}\left(\frac{m}{\mu}\right). \tag{36}$$

Theorem 6 has immediate consequences for symmetric distributions, which have the property that $\mu = m$. Substituting this equality into equation (36) yields $\rho^* \geq 1/2$.

More generally, Theorem 6 has consequences for distributions satisfying the mean-median-mode inequality. The mean-median-mode inequality implies $\mu \leq m$. Substituting this inequality into (36) again yields $\rho^* \geq 1/2$. Hence any sufficient condition for the mean-median-mode inequality is sufficient for $\rho^* \geq 1/2$.

The next propositions leverage a variety of results from the statistics literature on the mean-median-mode inequality for unimodal distributions.

Proposition 9. *Assume that X has a unimodal density function f; i.e., $f(x)$ is strictly increasing in x for $x < M$ and strictly decreasing in x for $x > M$. Then $\rho^* \geq 1/2$ if any one of the following conditions holds.*

(a) $F(m-x)+F(m+x) \geq 1$ *for all* $x \in [0,1]$.

(b) $F^{-1}(t)+F^{-1}(1-t) \leq 2m$ *for all* $t \in (0,1)$.

(c) There exists $\xi \in (0,1)$ *such that* $f(m+x) \geq f(m-x)$ *for all* $x \in [0,\xi)$ *and* $f(m+x) \leq f(m-x)$ *for all* $x \in (\xi,1]$.

(d) $f(F^{-1}(t)) \leq f(F^{-1}(1-t))$ *for all* $t \in (0,1/2)$.

(e) For all $x_1 \neq x_2$ *such that* $f(x_1) = f(x_2) > 0$, $f'(x_1) \leq |f'(x_2)|$.

Proof. We will argue that each of (a)–(e) imply that the mean-median-mode inequality holds. Theorem 1 of van Zwet (1979) states that (a) is sufficient for the mean-median-mode inequality. van Zwet (1979) shows (b) is equivalent to (a). Corollaries 1 and 2 of van Zwet (1979) state that, respectively, (c) and (d) are sufficient for the mean-median-mode inequality. Finally, Timerding (1915) (cited in Runnenburg 1978) shows that (e) is sufficient for the mean-median-mode inequality to hold. Thus each of (a)–(e) imply that the mean-median-model inequality holds, implying $\mu \leq m$. Substituting $\mu \leq m$ into (36) gives the bound on ρ^*. *Q.E.D.*

To help understand the content of this fairly technical proposition, it is first useful to understand the relationship among conditions (a)–(e) and their relationship to the mean-median-mode inequality. van Zwet (1979) established the chain of implications:

$$\begin{array}{ccccccccc} & & & & (c) & & & & \\ & & & & \Downarrow & & & & \\ (e) & \Rightarrow & (d) & \Rightarrow & (b) & \Leftrightarrow & (a) & \Rightarrow & \text{mean-median-mode inequality} \end{array}$$

To see how the conditions can be used in applications, consider the simple case of triangle distributions shown in Figure 9. The density of a triangle distributions is linear to either side of its mode; the equation for

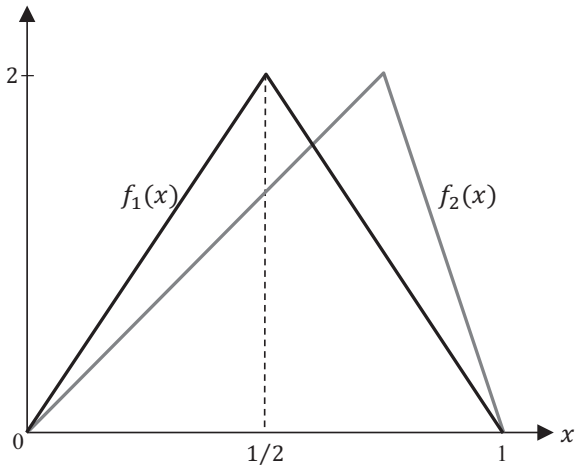


Figure 9: Density functions for triangle distributions with various modes.

such a distribution with support $[0,1]$ is

$$f(x) = \begin{cases} \dfrac{2x}{M} & x \leq M \\ \dfrac{2(1-x)}{1-M} & x > M. \end{cases} \tag{37}$$

Density $f_1(x)$ is associated with the symmetric triangular distribution with mode $M = 1/2$. The absolute values of the slope of the density to the left and right of the mode are equal, implying that condition (e) of Proposition 9 is satisfied (with equality). The proposition therefore implies $\rho^* \geq 1/2$ for this distribution. Indeed, we had already argued that Theorem 6 ensures $\rho^* \geq 1/2$ for symmetric distributions of rescaled values. Density $f_2(x)$ has a mode which is shifted to the right of $1/2$ and is not symmetric. Thus we cannot immediately conclude from Theorem 6 that $\rho^* \geq 1/2$. Still, we see that the density is steeper in absolute value to the right than the left of the mode, so it satisfies condition (e), implying $\rho^* \geq 1/2$ by Proposition 9.[14]

Dharmadhikari and Joag-dev (1988) provide a result that we can use to generalize Proposition 9 in that X is not even required to have a well-defined density. Before stating the proposition, some definitions are in order. The authors define a unimodal random variable X as having a distribution function $F(x)$ that is convex for $x < M$ and concave for $x > M$ (c.f. their Definition 1.1). Random variable X_1 weakly first order stochastic dominates X_2 if $F_1(x) \leq F_2(x)$ for all x in support of these distributions, where F_i is the distribution function associated with X_i. We have the following proposition.

Proposition 10. *Assume that X is a unimodal random variable for which* $\max(m-X,0)$ *weakly first order stochastic dominates* $\max(X-m,0)$. *Then* $\rho^* \geq 1/2$.

Proof. The statement of Theorem 1.14 of Dharmadhikari and Joag-dev (1988) relates to the reverse of the mean-median-mode inequality. One can repeat the proof, reversing the inequalities, to show that the

[14] One can calculate that $\rho^* \approx 0.54$ for the distribution associated with density $f_1(x)$ and $\rho^* \approx 0.57$ for that associated with $f_2(x)$.

mean-median-mode inequality holds—and thus $\mu \leq m$—for unimodal X for which $\max(m-X,0)$ first order stochastic dominates $\max(X-m,0)$. Substituting $\mu \leq m$ into (36) gives the bound on ρ^*. *Q.E.D.*

Although Proposition 10 generalizes the conditions under which $\rho^* \geq 1/2$ provided by Proposition 9, the latter conditions can be easier to verify in applications and thus Proposition 9 has independent merit.

Basu and DasGupta (1997) provide a different bound on the gap between the mean and median that can be used to bound ρ^*.

Proposition 11. *Assume that X is a unimodal random variable. Then*

$$\rho^* \geq \frac{1}{2}\left(1-\frac{\sigma\sqrt{0.6}}{\mu}\right). \tag{38}$$

Proof. Part (ii) of Corollary 4 of Basu and DasGupta (1997) states $|\mu-m|/\sigma \leq \sqrt{0.6}$. Rearranging, $m/\mu \geq 1-\sigma\sqrt{0.6}/\mu$. Substituting this inequality into (36) gives the bound on ρ^*. *Q.E.D.*

Tighter bounds can be derived for the beta distribution. We explained in Section 6.2 that the beta is useful in our setting because it has the correct support $[0,1]$ for rescaled consumer values. That section provided exact results for ρ^*, but the exact results required numerical optimization for each parameter configuration. Here we are interested in deriving bounds that can be calculated using a simple formula. One might imagine deriving such a formula by substituting analytical expressions for the mean and median of a beta into equation (36). The proof is not that simple because no analytical expression exists for the median of a beta. Instead we rely on Kerman's (2011) median approximation.

Proposition 12. *Suppose X has a beta distribution with parameters a and b such that $a,b>1$ (implying X is unimodal). Then*

$$\rho^* > \frac{1}{2.08}\left(\frac{a-1/3}{a+b-2/3}\right)\left(\frac{a+b}{a}\right). \tag{39}$$

Proof. Kerman (2011) offers the approximation formula for the beta median

$$m \approx \frac{a-1/2}{a+b-2/3}$$

showing it has a relative error of less than 4%. Hence

$$m \geq \frac{1}{1.04}\left(\frac{a-1/2}{a+b-2/3}\right).$$

Substituting this inequality for m in (36) as well as the beta mean $\mu = a/(a+b)$ gives (39). *Q.E.D.*

6.5. General Demand Curvature

The results of Anderson and Renault (2003) can be used to provide bounds on ρ^* that depend on a generalized notion of the curvature of demand, c-concavity or c-convexity, introduced by Caplin and Nalebuff

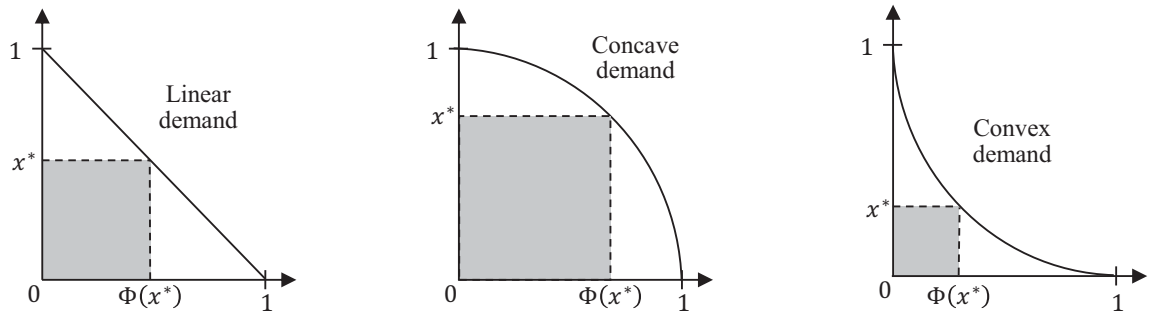


Figure 10: Surplus-extraction ratio depends on curvature of demand.

(1991).[15] Anderson and Renault (2003) show in their Proposition 1 that c-concavity of demand function Φ is equivalent to

$$\frac{\Phi''_X \Phi_X}{(\Phi'_X)^2} \leq 1-c \tag{40}$$

over the domain of Φ; c-convexity is equivalent to the reverse weak inequality holding. Observe that substituting $c = 1$ into (40) gives the definition of ordinary concavity (or convexity with the inequality reversed) and substituting $c = 0$ gives the definition of log-concavity (or log-convexity with the inequality reversed). The following result is a corollary of Anderson and Renault's (2003) results.

Proposition 13. *Suppose rescaled demand Φ in a market is twice continuously differentiable. If rescaled demand Φ in a market is c-concave for $c > -1$, then*

$$\rho^* \geq \begin{cases} \left(\frac{1}{1+c}\right)^{1/c} & c \neq 0 \\ 1/e & c = 0. \end{cases} \tag{41}$$

If Φ is c-convex for $c > -1$, then the reverse weak inequality holds.

Proof. Substituting $n = 1$, representing the monopoly market structure, into Proposition 5 of Anderson and Renault (2003), taking reciprocals, and noting the definition of $\rho^* = PS^*/TS^{**}$ gives equation (41). *Q.E.D.*

Bounds for concave or convex rescaled demand curves can be obtained by substituting $c = 1$ into (41), which yields $\rho^* \geq 1/2$ for concave demand and $\rho^* \leq 1/2$ for convex demand. Because a line is both concave and convex, we have that $\rho^* = 1/2$ for linear demand.

Figure 10 provides intuition for these results. The case of linear demand is shown in the first panel. Results from elementary geometry imply that the area of the largest rectangle that can be inscribed under a linear demand curve is half of the area under the curve, so by Lemma 2, which relates ρ^* to these areas, $\rho^* = 1/2$. The case of concave demand is shown in the second panel. As the figure indicates, the area of the largest rectangle that can be inscribed under this curve is at least half the area under this curve, so

[15] Caplin and Nalebuff (1991) call these curvature concepts ρ-concavity and ρ-convexity. We have relabeled them with c to avoid confusion with our notation for the surplus-extraction ratio, ρ^*.

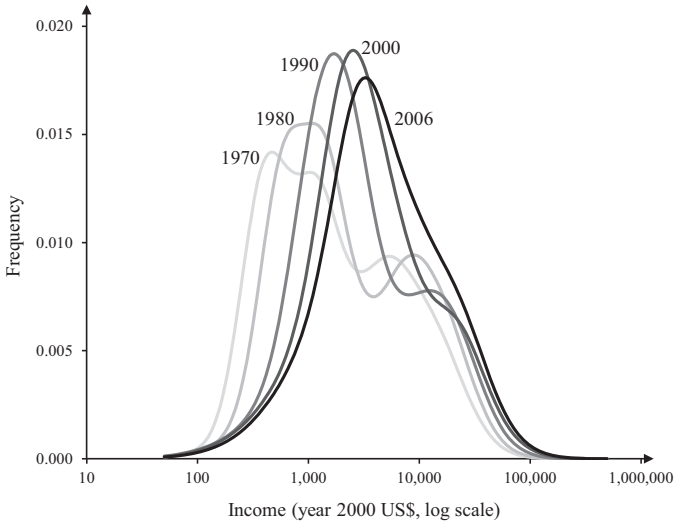


Figure 11: World income distribution from Pinkovskiy and Sala-i-Martin (2009).

$\rho^* \geq 1/2$. The case of convex demand is shown in the third panel. As the figure indicates, the area of the largest rectangle that can be inscribed under this curve is no more than half the area under this curve, so $\rho^* \leq 1/2$. That convex rescaled demand curves generate lower ρ^* than concave or linear is not surprising. They most resemble the STRZ demand achieving the lower bound on ρ^*.[16]

7. Calibrations Based on the World Distribution of Income

This section provides some calibrations of demand based on the world distribution of income. The calibrations will serve to show how to apply the STRZ decomposition works in practical settings and to show how tight the bound provided by the STRZ demand can be.

We use the world distribution of income from Pinkovskiy and Sala-i-Martin (2009). Figure 11 reproduces their Figure 21 in the form of a density function, with log scaling of income on the horizontal axis, for each decade since 1970, also including 2006. Focusing on the most recent, 2006, estimates and further assuming that consumers have unit income elasticity of demand for the product in question leads to the demand curve shown as the black curve in Figure 12. This has been rescaled so that the top income is given value 1 as the vertical intercept and so that the world population is scaled to have unit mass on the horizontal axis. For comparison, the STRZ demand with the same mean value is drawn as the lighter, grey curve. Producer surplus calculations will set the constant marginal cost to $c = 0$. This can either be viewed as a convenient value for the baseline or an approximation of true cost in certain markets such as software, digital media, or even small-molecule drugs.

[16]Because of its truncated endpoints, the STRZ demand, while convex over a range, is not globally convex. The truncations allow the STRZ demand to achieve a lower surplus-extraction ratio than a globally convex demand having the same μ. The fact from Proposition 13 that $\rho^* \leq 1/2$ for convex Φ seems inconsistent with the fact that $\underline{\rho}(1) = 1$. The inconsistency is resolved by noting that globally convex demand curves are unable to generate values of μ near 1.

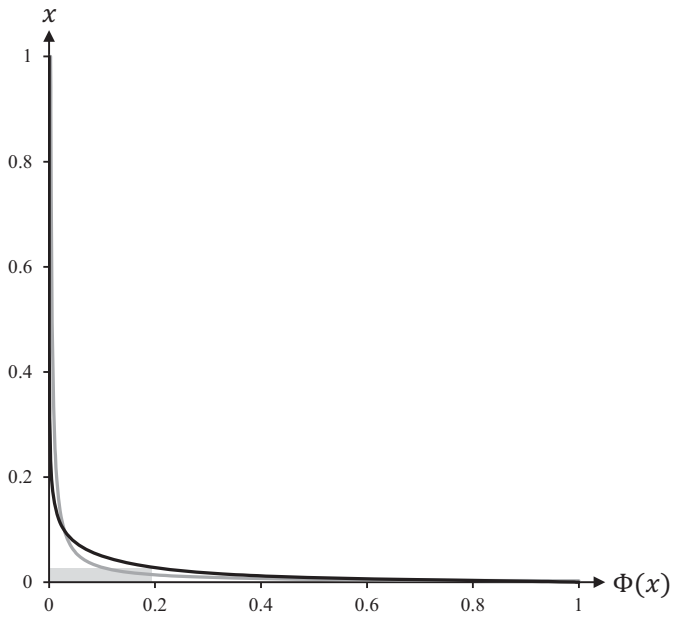


Figure 12: Calibrated demand based on 2006 distribution of world income. Black curve line is calibrated demand. For comparison, STRZ demand with same mean value is drawn as grey curve. Shaded rectangle is monopoly producer surplus for zero-cost product.

Visually, the calibrated demand curve is remarkably similar to its STRZ counterpart. Formally, one can compute that the index of Zipf-similarity, $Z = 0.83$. The blocky truncations which have been a prominent feature of STRZ demands drawn in previous figures are hardly noticeable in this calibration.

The mean rescaled consumer value is so low, $\mu = 0.019$, that the STRZ demand can only generate a surplus-extraction ratio of $\underline{\rho}(0.019) = 0.15$. The resulting surplus-extraction ratio for the calibrated demand curve is $\rho^* = 0.29$. That is, a monopolist would only be able to extract 29% of total surplus with a uniform price. Hence potential deadweight loss can be as high as $1-\rho^* = 71\%$ of total surplus. Producer surplus is shown as the area of the shaded rectangle. The monopolist sells to the richest 19% of the world population. Deadweight loss from pricing distortions, the area of the Harberger triangle, is 36% of total surplus, or 124% of producer surplus. The main message of the calibration is that one does not have to look far for highly Zipf-similar demands that generate huge distortions on both the investment and pricing margins.

Table 1 records these results for the 2006 calibration as well as earlier decades. Looking across columns, one can detect a slight sine pattern to the time series of each statistic, but the stronger message from the table is the stability of the results over time. This message can be visualized more clearly in Figure 13, which graphs the time series one of the more important of these statistics, the potential distortion $1-\rho^*$, as the black curve. For comparison, the baseline specification represented by the black curve is maintained across all of the panels. Each panel represents in effect a different comparative-statics exercise, varying one aspect of the calibration at a time. Variations over time in the black curve are quite small compared to some of the larger shifts in the curve arising from parameter changes in several of the panels.

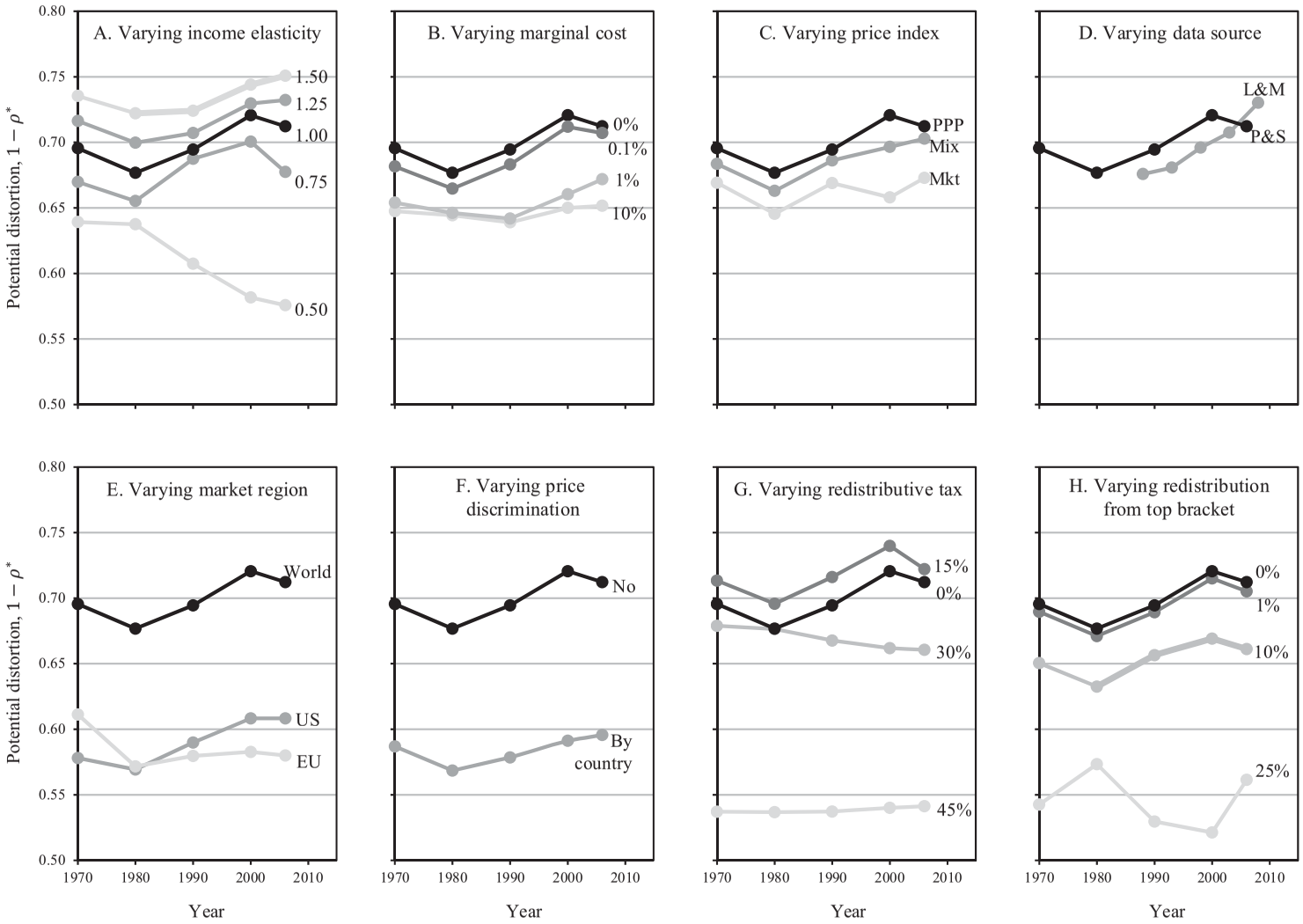


Figure 13: Baseline calibration compared to alternatives. Baseline calibration for potential distortion $1-\rho^*$, taken from third line of Table 1, graphed as the black curve in all panels, which share the same axes and scales. Panel A varies the income elasticity from the baseline value of 1.0 down to 0.5 and up to 1.5. Panel B varies the marginal cost of the good from the baseline value of 0% up to 10%, where the percentages are of the peak consumer value. Panel C compares the baseline calibration using purchasing-power-parity-adjusted incomes (PPP) to one using incomes measured at market exchange rates (the lightest grey curve labeled "mkt"). The medium-grey curve labeled "mix" assumes an income distribution involving a 50–50 mix of the two. Panel D compares the baseline calibration using Pinkovskiy and Sala-i-Martin's (2009) estimates of the income distribution to using Lakner and Milanovic's (2015) estimates. Panel E compares the baseline for the world market to the case in which good's market is just the United States or just the European Union. Panel F compares the baseline case of a uniform world price to the alternative of price discrimination across countries. Panel G varies the income tax from 0% up to 45%; tax proceeds are redistributed to consumers in equal lump sums. Panel H varies the size of the top bracket whose "excess" income is redistributed to consumers in equal lump sums.

Panel A varies consumers' income elasticity for the good in question in 0.25 increments from 0.50 to 1.50. The panel exhibits a robust monotonic relationship, with higher values of the income elasticity leading to higher potential distortions. In 2006, for example, the potential distortion is 58% when the income elasticity is 0.50, a substantial value but much less than the 75% potential distortion when the income elasticity is 1.50. Equation (27) can be used to decompose the increase in potential distortion moving the income elasticity from 0.50 to 1.50 into its constituents, showing it is due in roughly equal measure (54% versus 46% to be precise) to a greater Zipf-similarity and a smaller mean-to-peak ratio at the higher elasticity. Theory does not guarantee the monotonic relationship between elasticity and potential distortion observed in Panel A. Theory does guarantees that the distortion disappears in the limit as the income elasticity approaches zero: since consumers are homogeneous except for income differences in the calibration, setting the income elasticity to zero leads their values to be homogeneous, eliminating any distortion by Proposition 1.

Panel B examines the sensitivity of the results to changes in the good's marginal cost. The baseline calibration represented by the black curve involves costless production. Lighter grey curves represent calibrations for higher marginal costs. Marginal cost is measured as a percentage of peak consumer value; thus at marginal costs higher than 100%, the demand of even the highest value consumer ends up being choked off by a price that covers marginal cost. A marginal cost equal to 0.1% of peak value ends up being small in that it does not result in a marked change in potential distortion (or other variables not shown in the graph such as price or quantity) compared to the baseline calibration. Moving to yet higher marginal costs such as 1% or 10% of peak value causes more substantial declines in potential distortion. In 2006, for example, the potential distortion falls from 71% to 65% when marginal cost rises from 0 to 10%. Using equation (27), we can again decompose this change into its constituents, showing it is due now in precisely equal measure to a greater Zipf-similarity and a smaller mean-to-peak ratio of consumer values net of the higher marginal cost.

Before moving to more interesting counterfactuals, Panels C and D provide robustness checks, looking at alternative sources of data for the income distribution. Pinkovskiy and Sala-i-Martin (2009) adjust the

Table 1: Selected variables by decade from world income distribution calibrations

Variable	1970	1980	1990	2000	2006
Quantity q^*	0.17	0.19	0.15	0.14	0.19
Surplus-extraction ratio $\rho^* = PS^*/TS^{**}$	0.30	0.32	0.31	0.28	0.29
Potential distortion $1-\rho^*$	0.70	0.68	0.70	0.72	0.71
Zipf-similarity index Z	0.80	0.78	0.81	0.84	0.83
Relative Harberger deadweight loss $HDWL^*/TS^{**}$	0.33	0.32	0.37	0.41	0.36

incomes in their distributions by purchasing power parity (PPP), so that people with equal incomes in two different countries are roughly as well off. The baseline black curve in Panel C reflects this PPP adjustment. The PPP adjustment is natural when examining world poverty and inequality. Our goal is different: we are interested in the profit that a monopolist makes in selling to individuals in all these countries at a price in the monopolist's home currency. Thus, the market-exchange-rate adjusted incomes might be a better measure of people's ability to pay in this currency. Calibrations for the income distribution based on market-exchange-rate adjustment are given by the lightest grey curve. This adjustment leads to a modest reduction, less than four percentage points on average across years, in potential distortion. Perhaps the most realistic adjustment is the 50-50 mix of PPP and market-exchange-rate adjustment shown as the medium grey curve. We see this is quite close to the black curve reflecting the baseline PPP adjustment.

Panel F examines a different source for the income distribution, Lakner and Milanovic (2015). While different data sources may lead to different conclusions about poverty and other issues important to those authors, for our purposes, they result in fairly small differences in potential distortion, certainly small compared to the differences observed in some of the previous panels. The two curves would almost perfectly overlap if we shifted the Pinkovskiy and Sala-i-Martin (2009) curve ahead or the Lakner and Milanovic (2015) curve back by roughly seven years.

The remaining four panels move from robustness to more interesting counterfactual exercises. Panel E shows how the distortion changes if instead of the world market, the product is only sold in a single developed market such as the United States or the European Union. One might guess that surplus is more easily extracted in the single market, resulting in a smaller potential distortion (as a percentage of available social surplus). Panel E shows this is indeed the case. Potential distortion drops by roughly ten percentage points. Although the level shifts down for the individual markets, their time paths are remarkably similar to that for the world. Also remarkable is that the potential distortion is nearly identical whether the United States or the European Union is taken to be the single market. The decomposition formula (27) shows that the reduction in potential distortion moving from the world to a single developed market is mostly due to the fact that a demand curve based on the world income distribution is more Zipf-similar than one based on the U.S. or E.U. distribution. For example, over 70% of the decline in potential distortion moving from the world to the U.S. market in 2006 is due to a decline in Zipf-similarity.

Panel F examines shows how the potential distortion is affected by the ability to price discriminate. [17] The black curve is the benchmark case in which no price discrimination is possible; the monopolist has to sell on the world market at a single uniform price. The grey curve allows the monopolist to charge a different price in each country; i.e., the monopolist can discriminate across but not within countries. This form of price discrimination allows the monopolist to extract more surplus from the world market, reducing the potential distortion by roughly ten percentage points. Coincidentally, the decline in the potential distortion

[17] See Alessandria and Kaboski (2011) for a discussion of the extent of price discrimination across world markets in tradable goods.

is about equal to that from moving to a single developed market in the previous panel.

Panel G analyzes the role for a redistributive tax to reduce the potential distortion. In particular, the counterfactual involves the imposition of a proportional income tax at rate t, with tax revenues returned to consumers in equal lump sums. The idea is that this should compress extreme consumer values, making the demand curve less Zipf similar and the market more lucrative. Surprisingly, for small tax rates, we see the opposite effect, exacerbating the potential distortion. The exacerbation peaks at about the $t = 15\%$ rate shown, raising potential distortion from the baseline with no tax by an average of almost two percentage points. The firm ends up serving a similar group of relatively rich consumers but makes less money doing so because their after-tax incomes have fallen. Sufficiently high tax rates lead to a wholesale change in the monopolist's pricing strategy. For example, at the highest rate shown, 45%, the monopolist finds it profitable to serve almost all consumers since the poorest have substantial incomes after redistribution. Using this strategy, the producer surplus ratio, ρ^*, approaches the tax rate, $t = 45\%$, leaving the residual $1-\rho^*$, graphed as the lightest grey curve, in essence as a constant of height $1-45\% = 55\%$. We conclude that a redistributive tax can improve innovation incentives but can have the opposite effect so care must be exercised.

The last panel analyzes a different redistributive policy, focusing on the top 1% of consumers that received attention in recent discussions of income inequality. The policy takes the income these consumers have in excess of the marginal consumer just outside of the top category and redistributes that to all consumers in an equal lump sum. For example, in a population of 100 consumers, one with an income of \$100 and the rest with incomes of \$10, we take \$90 from the richest consumer and distribute 90 cents in a lump sum to all consumers, so that all 100 end up with \$10.90 as a result of the policy. As other alternatives, we pursue an analogous policy but expand the set of consumers considered to be in the top bracket, from whom income is redistributed, from 1% to 10% to 20%. Panel H presents the results. Unlike the tax policy, this more targeted policy reduces potential distortion in all cases examined. The policy redistributes income from the monopolist's inframarginal to its marginal purchasers, increasing producer surplus. By contrast, the 15% tax reduced the income of the monopolist's marginal purchasers, and so reduced producer surplus. The policy reduces potential distortion in rough proportion to the size of the top bracket. So, redistribution from the top 1% reduces the potential distortion by about one percentage point (1% of baseline distortion), redistribution from the top 10% reduces potential distortion by about five percentage points (7% of baseline distortion), and redistribution from the top 25% reduces potential distortion by about 15 percentage points (22% of baseline distortion).

8. Conclusion

As clarified by Makowski and Ostroy (1995, 2001), efficiency depends on agents' ability to capture the social benefits they create. We first show that the amount of surplus a monopolist cannot extract provides a

tight upper bound on deadweight loss from all sources (entry as well as pricing decisions).

We then show that this worst case is itself worst for the STRZ demand. This demand curve entails a Zipf distribution for consumers' valuations net of marginal cost, with a doubling of the value cutting the number of consumers with that value or higher in half. With this worst case in hand, we can relate the potential deadweight loss in a broad class of markets—those with finite demand intercepts, required for the needed rescaling—to the Zipf-similarity of the rescaled demand curve. Holding constant Zipf-similarity, the other factor limiting surplus extraction is the ratio of the mean to the peak of net consumer values. If the demand is sufficiently Zipf similar and the ratio of mean to peak net consumer value is sufficiently low, the ratio of producer to total surplus can be arbitrarily close to zero and distortions can come arbitrarily close to dissipating 100% of total surplus.

Our demand calibration based on the world distribution of income provides a measure of the practical relevance of the STRZ bound. The calibration is based on a number of assumptions (unit income elasticity, zero production cost), which may be good approximations for some products and not others but was a neutral starting point. The calibrated demand in Figure 12 looks remarkably similar to its STRZ lower bound. Both hug the axes so tightly that it would be hard for a monoplist to extract surplus for its product. According to our theory, 72% of total surplus could be dissipated due to the poor investment incentives demands with such shapes provide. As the income distribution evolved over time from 1970 to the present, calibrated potential deadweight loss was worst in 2000 and improved slightly since.

The Zipf perspective provides an argument for R&D distortions being greater for product than process innovations. The upper tail of the distribution of values may be Zipf-similar for a new product for which there are few good substitutes, but the upper tail of the distribution of values for a cost-reducing innovation is necessarily truncated at the price of the more expensive product it is replacing. The ratio of the mean-to-peak consumer value will thus tend to be lower with the product innovation, leading to a greater potential for deadweight loss.

To the extent that policymakers can identify which markets have Zipf-similar demands, they may wish to target R&D subsidies there. For example, the Zipf-similarity of the risk distribution (see Liljeros *et al.* 2001 for a characterization of the power-law distribution of number of sexual partners) may lead to a Zipf-similar demand for an HIV vaccine. This argument is consistent with claims made by industry observers (e.g., Thomas 2002) that an HIV vaccine would be far less luractive than the potential social benefit it would provide. Because the harm from HIV is so serious, total surplus and hence the absolute magnitude of the potential distortion from underinvestment, is likely high. This provides a rationale for programs such as the International AIDS Vaccine Initiative (IAVI) that support vaccine research. To take a very different example, the case for cities to subsidize sports stadia will be stronger if the distribution of valuation is Zipf-similar than if it is closer to being homogeneous.

Regarding policy toward price discrimination, our analysis shows that the potential deadweight loss of

banning price discrimination is greater the more Zipf-similar is demand. Conversely, the potential for price discrimination to increase producer surplus, and thus investment incentives, is greatest when the distribution of demand is Zipf-similar.

The conclusion has so far focused on deadweight loss at the investment/entry margin, rightly so because that is the margin that the theory shows can generate the greatest deadweight loss. However, the paper provides a suite of results taking investment/entry as given and just considering distortions at the pricing margin. The same STRZ demands that had the highest potential for deadweight loss at the investment/entry margin also has the highest potential for deadweight loss at the pricing margin. The policy implications of this set of theoretical results are quite interesting and deserve note. Facing a Zipf-similar distribution of net consumer values, the monopolist may be close to indifferent between widely disbursed prices that induce very different shares of the population to purchase the good and that realize very different proportions of potential consumer surplus. A tiny subsidy could be enough to drive the equilibrium toward that with high output and a small Harberger triangle. The outcome could also be achieved with an appropriate price ceiling. Unlike familiar cases in which a price cap might correct a pricing distortion but of course reduce producer surplus and impair investment/entry incentives, with a STRZ demand or demand close to it, a price cap would only have a small effect on investment/entry incentives if the firm were essentially indifferent between pre- and post-cap prices. Thus subsidies and price caps can potentially be very effective in markets with Zipf-similar demands.

Larger subsidies would be needed to completely eliminate distortions in other situations. Whether the policymaker wishes to eliminate pricing distortions when the demand curve is not Zipf-similar or the demand is Zipf-similar but the policymaker wishes to address investment/entry distortions in addition to price distortions, if taxes can be raised on a dollar-for-dollar basis with no additional social cost of public funds, the first best can be obtained by offering a subsidy of TS^{**}/q^{**} per unit sold at marginal cost. We showed that the gain from such a policy is tightly bounded above by, again, the unextracted surplus on the market, i.e., total minus producer surplus. In a range of cases—software, movies, music, some small molecule drugs—it is plausible that the policymaker knows this marginal cost and that it is close to zero. If the policymaker is constrained in the size of subsidies it can offer, although it may not attain the first best, it may still be able to ameliorate investment and pricing distortions with a modest subsidy, and the subsidy dollar can be leveraged most fully if the policymaker can target the subsidy to the lowest net valuation consumers.

Appendix A: Proofs Omitted from Text

Proof of Theorem 2: We will first analyze a ban on price discrimination. Let superscript d denote the equilibrium values of variables under a given form of price discrimination (as distinct from stars, denoting equilibrium values under linear pricing, and double stars, denoting socially optimal values). The distortion from banning price discrimination is

$$E^d W^d - E^* W^*, \tag{A1}$$

obviously maximized for the form of price discrimination generating the highest $E^d W^d$. First-degree price discrimination allows the firm to capture W^{**} as profit, inducing the firm to make the socially optimal entry decision E^{**}. Hence social welfare under first-degree price discrimination is $E^{**}W^{**}$. Substituting into (A1) yields $E^{**}W^{**} - E^*W^*$, which equals DWL^* by definition. Of course $E^d W^d$ cannot be higher than in the first best, so (A1) cannot be higher than DWL^*. Thus DWL^* is a tight upper bound on the distortion from banning price discrimination.

We next establish the bound on the welfare gain from the subsidy policy. Specifically, consider the policy of providing a per-unit subsidy of TS^{**}/q^{**} for all units the firm sells at marginal cost c. The firm can earn producer surplus TS^{**} from accepting the policy and charging c to all q^{**} who buy. There is no strategy that would allow the firm to earn more than TS^{**}, so it is indeed an equilibrium for it to accept the policy and charge c. To construct an upper bound, we will consider the best-case scenario of no social cost of public funds. Then producer surplus equals the transfer from the government and nets out of social surplus, leaving consumer surplus, which equals TS^{**} given marginal-cost pricing. Social welfare from the policy is thus W^{**}. The government can ensure the first best $E^{**}W^{**}$ by offering the policy if and only if $W^{**} > 0$. The gain from the policy is thus $E^{**}W^{**} - E^*W^* = DWL^*$. Of course welfare from any other policy cannot exceed $E^{**}W^{**}$, so we have constructed a tight upper bound on the gain from subsidy policies. *Q.E.D.*

Proof of Theorem 3: The analysis closely follows the proof of Theorem 1. Because the monopoly's entry decision depends on the same k as before, it continues to be characterized by the partition into the three subintervals shown in Figure 4. The only change is the computation of deadweight loss when k falls into the moderate subinterval (PS^*, TS^{**}). While deadweight loss still equals all of first-best welfare W^{**}, now $W^{**} = TS^{**} - (1-\beta)k$ because βk is not a social cost. Since W^{**} is decreasing in k, the supremum over the moderate subinterval of k is achieved at the lower boundary, $k = PS^*$, where it equals $TS^{**} - (1-\beta)PS^*$. This expression is decreasing in β, attaining its minimum value over feasible β for $\beta = 0$. This minimum value, $TS^{**} - PS^*$, was shown in the proof of Theorem 1 to exceed the potential deadweight loss in the other subintervals. Hence $TS^{**} - (1-\beta)PS^*$ is the bound on potential deadweight loss. To express as a proportion, divide through by TS^{**}, yielding equation (3). *Q.E.D.*

Proof of Theorem 4: Let $K_1 = \{k_1 \geq 0, k_i > TS^{**} | i = 2, \ldots, N\}$ be the set of fixed-cost vectors such that firm 1's can be any non-negative number but the rest of the potential entrants' are higher than total surplus. Then

$$\sup_{\{k_i \geq 0 | i = 1, \ldots, N\}} \left[\frac{DWL^*(C,N)}{TS^{**}} \right] \geq \sup_{K_1} \left[\frac{DWL^*(C,N)}{TS^{**}} \right] \tag{A2}$$

because $K_1 \subset \{k_i \geq 0 | i = 1, \ldots, N\}$. Consider possible entry equilibrium when fixed costs are in K_1. In that case, for all $i = 2, \ldots, N$, $k_i > TS^{**} \geq PS^* + CS^* \geq PS^* = \overline{PS}^*(C,1) \geq \overline{PS}^*(C,n)$. The first step follows from the restriction on fixed costs on the right-hand side of (A2), the second step from the relevant surplus definitions, the third step from $CS^* \geq 0$, the fourth step from assumption (4), and the last step from assumption (5). But $k_i > \overline{PS}^*(C,n)$ for all $n \in \{1, \ldots, N\}$ implies no firm $i = 2, \ldots, N$ enters by assumption (6), implying firm 1 is the only possible entrant, implying that equilibrium welfare equals monopoly welfare, W^*. For fixed costs in K_1, social welfare is negative if any firm $i = 2, \ldots, N$ enters, implying that a social planner

would have at most firm 1 enter, in turn implying that first-best welfare is the same as with a monopoly, W^{**}. Putting these results together, $DWL^*(C,N) = W^{**} - W^* = DWL^*$ for fixed costs in K_1. Thus

$$\sup_{K_1}\left[\frac{DWL^*(C,N)}{TS^{**}}\right] = \sup_{k_1\geq 0}\left(\frac{DWL^*}{TS^{**}}\right) = 1-\rho^*,$$

where the second equality follows from Theorem 1. *Q.E.D.*

Proof of Theorem 5: We will divide the search over values of k leading to the supremum on relative deadweight loss into two regions: $k > PS^*$ and $k < PS^*$. First, consider values of k satisfying $k > PS^*$. Then $n^* = 0$, implying $W^* = 0$, in turn implying $DWL(C,n^*) = W^{**} - W^* = W^{**} = TS^{**} - k$. Hence $\sup_{k>PS^*} DWL(C,n^*) = \sup_{k>PS^*}(TS^{**}-k) = TS^{**} - PS^*$. Dividing by TS^{**} yields

$$\sup_{k>PS^*}\left[\frac{DWL(C,n^*)}{TS^{**}}\right] = 1-\rho^*. \tag{A3}$$

Next, consider values of k satisfying $k \in (0, PS^*)$. We will show that k can be written

$$k = PS^*(C,n(k)) - \epsilon, \tag{A4}$$

for some

$$\epsilon \in \big[0, PS^*(C,n(k)) - PS^*(C,n(k)+1)\big), \tag{A5}$$

where $n(k) \in \mathbb{N}$ is the equilibrium number of entrants induced by k. To this end, note $PS^*(C,n) \leq PS^*(C,1) = PS^*$, where the first step follows from equation (9) and the second from equation (8). Thus $PS^*(C,n) \leq PS^*/n$, implying $\lim_{n\uparrow\infty} PS^*(C,n) \leq \lim_{n\uparrow\infty} PS^*/n = 0$. Equation (9) implies $PS^*(C,n)$ is strictly decreasing in n. Therefore, the set $\{PS(C,n)|n \in \mathbb{N}\}$ forms a partition of the interval $(0, PS^*)$. This means that for any $k \in (0, PS^*)$, we can find $n(k)$ such that $k \in \big(PS^*(C,n(k)+1), PS^*(C,n(k))\big]$. Equation (A4) follows.

Having established that key result, we proceed to characterize relative deadweight loss:

$$\begin{aligned} DWL(C,n(k)) &= W^{**} - W^* && \text{(A6)}\\ &= TS^{**} - k - [CS^*(C,n(k)) + n(k)PS^*(C,n(k)) - n(k)k] && \text{(A7)}\\ &= TS^{**} - CS^*(C,n(k)) - PS^*(C,n(k)) - [n(k)-1]\epsilon, && \text{(A8)} \end{aligned}$$

where (A8) follows by substituting from (A4). The supremum of (A6) over values of k in $(0, PS^*)$ is equivalent to the supremum of (A8) over values of $n(k)$ in $\mathbb{N}$ and over values of ϵ satisfying (A5). This supremum is approached for $n(k) = \hat{n}$ defined in the statement of the theorem and for $\epsilon \downarrow 0$. Substituting these values into (A8) and dividing by TS^{**} yields

$$\sup_{k\in(0,PS^*)}\left[\frac{DWL(C,n^*)}{TS^{**}}\right] = 1 - \frac{CS^*(C,\hat{n}) + PS^*(C,\hat{n})}{TS^{**}}. \tag{A9}$$

Equations (A3) and (A9) are the two candidates for the supremum over relative deadweight loss. Whichever is greater is the supremum. *Q.E.D.*

Proof of Lemma 1: To see the first equality in the statement of the lemma,

$$\mu = \int_0^1 x\,dF(x) = \left[1 - \int_0^1 F(x)dx\right] = \int_0^1 \bar{F}(x)dx = \int_0^1 \Phi(x)dx. \tag{A10}$$

The first equality follows from the definition of μ, the second from integration by parts, and the third from the definition of $\bar{F}(x)$. To see the last equality, recall $\bar{F}(x) = \Phi(x) - \Pr(X = x)$ as shown in the paragraph leading up to the lemma. Because $\Phi(x)$ is a rescaling of the left-continuous $Q(p)$, $\Phi(x)$ can only have a countable number of discontinuities. Hence $\bar{F}(x)$ only differs from $\Phi(x)$ for at most a countable set of x, implying their Riemann integrals are equal.

To see the second equality in the statement of the lemma,

$$TS^{**} = \int_c^{p^{\max}} Q(p)dp = \int_0^1 (p^{\max} - c)q^{**}\Phi(x)dx = (p^{\max} - c)q^{**}\mu. \tag{A11}$$

The first equality holds by definition. The second equality follows from substituting $p - c = (p^{\max} - c)x$ from (14) and substituting $Q(p) = q^{**}\Phi(x)$ from (15). The last equality follows from (A10). Rearranging (A11) gives the second equality in the statement of the lemma. *Q.E.D.*

Proof of Lemma 2: Starting with the definition of PS^*,

$$PS^* = \max_{p \in [c, p^{\max}]} \{(p - c)Q(p)\} = \max_{x \in [0,1]} \{(p^{\max} - c)q^{**}x\Phi(x)\} = (p^{\max} - c)q^{**}REC^*. \tag{A12}$$

The first equality follows from the definition of PS^*, the second from the rescaling embodied in (14) and (15), and the third from the definition of REC^*. Thus $\rho^* = PS^*/TS^{**} = REC^*/\mu$, where the first equality holds by definition of ρ^* and the second follows from substituting an expression for TS^{**} obtained by rearranging the statement of Lemma 1 and then simplifying.

We next show $\rho^* \in [0,1]$ under the maintained assumptions. The assumption that $p^0 > c$ implies $TS^{**} > 0$. The assumption that $p^{\max}$ is finite ensures REC^* and μ are well-defined. Hence, all of our expressions for ρ^* are well-defined. Now $PS^*, CS^*, HDWL^* \geq 0$ implies $0 \leq PS^* \leq PS^* + CS^* + HDWL^* = TS^* \leq TS^{**}$, ensuring $\rho^* \in [0,1]$. *Q.E.D.*

Proof of Proposition 2: We first argue that the implicit solution $A(\mu)$ of (17) is a well-defined function $A : (0,1) \to (0,1)$. We will initially work with its inverse, $\mu(A) = A(1 - \ln A)$, where $A \in (0,1)$. Now $\mu'(A) = -\ln A$, implying $\mu(A)$ is continuously differentiable and strictly increasing on $(0,1)$. Therefore, by the Inverse Function Theorem, its inverse, $A(\mu)$, exists. Further, $\lim_{A \downarrow 0} \mu(A) = 0$ by l'Hôpital's Rule, and $\mu(1) = 1$ by direct evaluation. Thus $\mu(A) \in (0,1)$ for all $A \in (0,1)$, implying $\mu(A)$ is a bijection on $(0,1)$, implying its inverse $A(\mu)$ is also a bijection on $(0,1)$. The fact that $\lim_{A \downarrow 0} \mu(A) = 0$ implies $\lim_{\mu \downarrow 0} A(\mu) = 0$, and the fact that $\mu(1) = 1$ implies $A(1) = 1$. By the Inverse Function Theorem, $A'(\mu) = 1/\mu'(A(\mu))$, which is positive for all $\mu \in (0,1)$ because $\mu'(A) > 0$ for all $A \in (0,1)$.

Consider a given market m in which rescaled demand is $\Phi(x)$ and the mean rescaled consumer value is at least $\mu \in (0,1)$. Let $REC^* = \max_{x \in [0,1]} x\Phi(x)$ denote producer surplus in market m. We will show $\underline{REC}(\mu) \leq REC^*$, with strict inequality unless $\Phi(x) = \underline{\Phi}(x, \mu)$ almost everywhere. If $\Phi(x) = \underline{\Phi}(x, \mu)$ almost everywhere, then $\underline{REC}(\mu) = REC^*$, and we are done. For the remainder of the proof, therefore, assume $\Phi(x) \neq \underline{\Phi}(x, \mu)$ for a positive measure of $x \in (0,1)$.

We have

$$\int_0^1 \underline{\Phi}(x, \mu)dx = \mu \leq \int_0^1 \Phi(x)dx,$$

where the first equality follows from Lemma 1 and the second from the fact that the mean rescaled value in market m is at least μ. Given $\int_0^1 \underline{\Phi}(x, \mu)dx$ is not greater than $\int_0^1 \Phi(x)dx$ and $\Phi(x) \neq \underline{\Phi}(x, \mu)$ for a positive measure of x, it must be that $\Phi(x) > \underline{\Phi}(x, \mu)$ for all x in some subset of positive measure $S \subseteq (0,1)$. For $x \in S$, $x\Phi(x) > x\underline{\Phi}(x, \mu) = \min\{A(\mu), x\}$ substituting from (16), implying either $x\Phi(x) > A(\mu)$ or $x\Phi(x) > x$. The latter inequality implies $\Phi(x) > 1$, a contradiction to $\Phi(x)$ being a proper rescaled demand curve with

range $[0,1]$. This proves that for all $x \in S$,

$$x\Phi(x) > A(\mu). \tag{A13}$$

It follows that, for $x \in S$, $REC^* \geq x\Phi(x) > A(\mu) = \underline{REC}(\mu)$. The first inequality holds because REC^* is the maximized value of $x\Phi(x)$ over $x \in [0,1]$. The next inequality follows from (A13) and the last equality from (18). *Q.E.D.*

Proof of Proposition 3: Since $\underline{\rho}(\mu)$, defined in equation (20), is a special case of a more general formula, equation (B5), derived in Appendix B, we will not present a separate derivation here. Instead, we will simply verify that $\underline{\rho}(\mu)$ satisfies (19). Rearranging equation (20) yields

$$W(-\mu/e) = \frac{-1}{\underline{\rho}(\mu)}. \tag{A14}$$

For now we are ignoring the specific branch of the Lambert W function and just using the function $W(\cdot)$ to represent a generic branch. As stated footnote 11, $W(z)$ is defined as the inverse relation of the function $z = We^W$. Inverting (A14) and multiplying through by -1, we obtain

$$\frac{\mu}{e} = \frac{1}{\underline{\rho}(\mu)} \exp\left(\frac{-1}{\underline{\rho}(\mu)}\right). \tag{A15}$$

Taking the natural log of (A15) and rearranging gives (19).

The Lambert W function is multivalued on the domain $[-1/e, 0]$ relevant to this proposition, having an upper branch, denoted $W_0(z)$, and a lower branch, denoted $W_{-1}(z)$, where $W_0(z) > -1$ for $z \in (-1/e, 0)$ and $W_1(z) < -1$ for $z \in (-1/e, 0)$. Beginning with the upper branch, the fact that $W_0(z) > -1$ implies $\underline{\rho}(\mu) > e/\mu > 1$, violating $\underline{\rho}(\mu) \leq 1$ by Lemma 2. Thus the only candidate is the lower branch. We will return to verify that $\underline{\rho}(\mu) \in [0,1]$ when the lower branch is used in equation (20), so the lower branch indeed constitutes a viable solution.

To find $\underline{\rho}'(\mu)$, differentiate both sides of (A14) with respect to μ, giving

$$\frac{-W'_{-1}(-\mu/e)}{e} = -\ln \underline{\rho}(\mu)\underline{\rho}'(\mu), \tag{A16}$$

or after rearranging,

$$\underline{\rho}'(\mu) = \frac{W'_{-1}(-\mu/e)}{e \ln \underline{\rho}(\mu)}. \tag{A17}$$

By implicit differentiation,

$$W'_{-1}(-\mu/e) = \frac{eW_{-1}(-\mu/e)}{\mu[1+W_{-1}(-\mu/e)]}. \tag{A18}$$

Substituting (A18) into (A17) and simplifying yields

$$\underline{\rho}'(\mu) = \frac{-W_{-1}(-\mu/e)}{\mu \ln \underline{\rho}(\mu)[1+W_{-1}(-\mu/e)]}. \tag{A19}$$

Now $W_{-1}(z) < -1$ on the lower branch for all $z \in (-1/e, 0)$. Further, $\underline{\rho}(\mu) \in (0,1)$ for $\mu \in (0,1)$, implying $\ln \underline{\rho}(\mu) < 0$. Thus equation (A19) is positive for all $\mu \in (0,1)$.

We next evaluate the limiting values of $\underline{\rho}(\mu)$. The standard result that $\lim_{z \downarrow 0} W_{-1}(z) = -\infty$ implies $\lim_{\mu \downarrow 0} \underline{\rho}(\mu) = -1/-\infty = 0$. The standard result that $W_{-1}(-1/e) = -1$ implies $\underline{\rho}(1) = -1/-1 = 1$.

The results in the previous two paragraphs imply that when the lower branch of the Lambert W function is used in the formula (20) for $\underline{\rho}(\mu)$, we have $\underline{\rho}(\mu) \in (0,1)$ for all $\mu \in (0,1)$. This verifies that the lower branch provides the viable solution for $\underline{\rho}(\mu)$.

This completes the proof. For reference, Table 2 computes $\underline{\rho}(\mu)$ over a grid of values of μ. *Q.E.D.*

Table 2: Lower Bound on Producer-Surplus Ratio, $\underline{\rho}(\mu)$, for a Grid of Means, μ

μ	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.000[a]	0.131	0.146	0.157	0.166	0.174	0.181	0.188	0.193	0.199
0.1	0.205	0.210	0.215	0.219	0.224	0.229	0.233	0.238	0.242	0.246
0.2	0.250	0.255	0.259	0.263	0.267	0.271	0.275	0.279	0.283	0.287
0.3	0.291	0.295	0.299	0.303	0.307	0.311	0.315	0.319	0.323	0.327
0.4	0.331	0.335	0.339	0.343	0.347	0.352	0.356	0.360	0.365	0.369
0.5	0.373	0.378	0.382	0.387	0.392	0.396	0.401	0.406	0.411	0.416
0.6	0.421	0.426	0.431	0.436	0.442	0.447	0.453	0.459	0.465	0.471
0.7	0.477	0.483	0.490	0.496	0.503	0.510	0.517	0.524	0.532	0.540
0.8	0.548	0.557	0.565	0.575	0.584	0.594	0.605	0.616	0.627	0.640
0.9	0.653	0.667	0.682	0.699	0.717	0.738	0.761	0.789	0.823	0.871
1.0	1.000									

Notes: Table entry is the value of $\underline{\rho}(\mu)$ corresponding to the value of μ equal to the sum of the row and column headings. Entries rounded to three decimals. [a]Result holds in limit.

Proof of Proposition 4: Suppose rescaled demand in market 1 is $\underline{\Phi}(x,\mu)$. One equilibrium in this market involves $x_1^* = 1$, i.e., the monopolist sells just to the mass of highest-value consumers at a price extracting all their consumer surplus. In this equilibrium,

$$\frac{HDWL_1^*}{TS_1^{**}} = 1 - \frac{CS_1^*}{TS_1^{**}} - \frac{PS_1^*}{TS_1^{**}} = 1 - \frac{PS_1^*}{TS_1^{**}} = 1 - \underline{\rho}(\mu), \tag{A20}$$

where the first equality follows from the definition of TS^{**}, the second from the fact that the equilibrium price extracts all consumer surplus, and the third from the definition of $\underline{\rho}(\mu)$ as the surplus-extraction ratio associated with $\underline{\Phi}(x,\mu)$. Consider market 2 with the same μ but a rescaled demand $\Phi_2(x)$ that is not almost everywhere identical to $\underline{\Phi}(x,\mu)$. Then

$$\frac{HDWL_2^*}{TS_2^{**}} = 1 - \frac{CS_2^*}{TS_2^{**}} - \frac{PS_2^*}{TS_2^{**}} \leq 1 - \frac{PS_2^*}{TS_2^{**}} = 1 - \rho_2^* < 1 - \underline{\rho}(\mu), \tag{A21}$$

where the first step follows from the definition of $HDWL^*$, the second step from $CS^*/TS^{**} \geq 0$, the third step from the definition $\rho = PS^*/TS^{**}$, and the last step from $\underline{\rho}(\mu) < \rho$ by Proposition 2.

Combining (A20) and (A21), Harberger deadweight loss as a proportion of total surplus is strictly lower in the second than the first market. *Q.E.D.*

Proof of Proposition 5: We have

$$\frac{HDWL^*}{TS^{**}} \leq 1 - \rho^* = Z[1 - \underline{\rho}(\mu)], \tag{A22}$$

where the inequality follows from (A21) and the equality from (22). The right-hand side of (A22) obviously approaches 0 as $Z \downarrow 0$. The result from Proposition 3 that $\lim_{\mu \downarrow 0} \underline{\rho}(\mu) = 0$ implies that the right-hand side of (A22) approaches 0 as $\mu \uparrow 1$. *Q.E.D.*

Proof of Proposition 6: The original demand curve $Q(p)$ is nonincreasing by assumption. Rescaled demand $\Phi(x)$ inherits this property, implying $\Phi(x^*) \leq \Phi(x^*(C,n))$ if the condition on prices in the statement of the proposition holds, i.e., if $x^*(C,n) \leq x^*$. Then

$$\begin{aligned} TS^* &= x^*\Phi(x^*) + \int_{x^*}^{1} \Phi(x)dx && \text{(A23)} \\ &= \int_0^1 \min\{\Phi(x^*), \Phi(x)\}dx && \text{(A24)} \\ &\leq \int_0^1 \min\{\Phi(x^*(C,n)), \Phi(x)\}dx && \text{(A25)} \\ &= TS^*(C,n). && \text{(A26)} \end{aligned}$$

Thus

$$\frac{HDWL^*(C,n)}{TS^{**}} = 1 - \frac{TS^*(C,n)}{TS^{**}} \leq 1 - \frac{TS^*}{TS^{**}} \leq Z[1 - \underline{\rho}(\mu)], \qquad \text{(A27)}$$

where the first step follows from surplus definitions, the next step from (A26), and the last step from (A22). *Q.E.D.*

Proof of Proposition 8: For the rescaled Pareto we have

$$\begin{aligned} REC^* &= \max_{x \in [0,1]} [x\Phi(x, \alpha, x_0)] && \text{(A28)} \\ &= \max_{x \in [0,1]} \left(\min\{x, x_0^\alpha x^{1-\alpha}\}\right) && \text{(A29)} \\ &= \max_{x \in [0,1]} \left(x_0^\alpha x^{1-\alpha}\right) && \text{(A30)} \\ &= \begin{cases} x_0^\alpha & \alpha < 1 \\ x_0 & \alpha \geq 1. \end{cases} && \text{(A31)} \end{aligned}$$

Equation (A28) is the definition of REC^*; (A29) holds by substituting from (34), and (A30) and (A31) follow from direct calculation.

We proceed by analyzing two cases: $\alpha < 1$ and $\alpha > 1$. First suppose $\alpha < 1$. Then

$$\begin{aligned} \rho^* &= \frac{REC^*}{\mu} && \text{(A32)} \\ &= \frac{1-\alpha}{1-\alpha x_0^{1-\alpha}}. && \text{(A33)} \end{aligned}$$

Equation (A32) follows from Lemma 2 and (A33) from substituting for μ from (35). Differentiating (A33),

$$\frac{\partial \rho^*}{\partial \alpha} = \frac{1 - \alpha \ln x_0 + \alpha^2 \ln x_0 - x^{\alpha-1}}{x^{\alpha-1}\left(\alpha x^{1-\alpha} - 1\right)^2}. \qquad \text{(A34)}$$

The denominator is obviously positive, so the sign of (A34) is determined by the sign of the numerator:

$$NUM(\alpha, x_0) = 1 - \alpha \ln x_0 + \alpha^2 \ln x_0 - x^{\alpha-1}. \tag{A35}$$

Now

$$\frac{\partial^2 NUM}{\partial \alpha^2} = 2 \ln x_0 - x^{\alpha-1} (\ln x_0)^2, \tag{A36}$$

which is obviously positive, implying $NUM(\alpha, x_0)$ is concave in α. Direct calculation shows $NUM(1, x_0) = 0$ and $\partial NUM(1, x_0)/\partial \alpha = 0$. Hence $NUM(\alpha, x_0) < 0$ for all $\alpha \in [0, 1)$. Thus ρ^* is decreasing in α for all $\alpha \in [0, 1)$.

Next suppose $\alpha > 1$. Then

$$\rho^* = \frac{1 - \alpha}{x_0^{\alpha-1} - \alpha}, \tag{A37}$$

substituting for REC^* from (A31) into (A32). Differentiating (A37),

$$\frac{\partial \rho^*}{\partial \alpha} = \frac{x_0 - x_0^{\alpha}[1 + (1-\alpha)\ln x_0]}{x_0 (\alpha - x^{\alpha-1})^2}. \tag{A38}$$

The denominator is obviously positive, so the sign of (A38) is determined by the sign of the numerator:

$$NUM(\alpha, x_0) = x_0 - x_0^{\alpha}[1 + (1-\alpha)\ln x_0] \tag{A39}$$

(we use *NUM* for the numerator of another expression to economize on notation). We have $\partial NUM(\alpha, x_0)/\partial \alpha = -(1-\alpha) x_0^{\alpha} (\ln x)^2 < 0$. Hence $NUM(\alpha, x_0)$ attains its lowest value in the limit $\alpha \uparrow \infty$. Taking limits,

$$\lim_{\alpha \uparrow \infty} NUM(\alpha, x_0) = x_0 + \ln x_0 \lim_{\alpha \uparrow \infty} (\alpha x_0^{\alpha}) = x_0, \tag{A40}$$

where l'Hôpital's Rule can be used to show $\lim_{\alpha \uparrow \infty} (\alpha x_0^{\alpha}) = 0$ since $x_0 \in (0, 1)$. This proves $NUM(\alpha, x_0) > 0$ for all $\alpha > 1$ and thus that ρ^* is increasing in α for all $\alpha > 1$.

In sum, we have proved that ρ^* is decreasing in α for $\alpha < 1$ and increasing in α for $\alpha > 1$. Taking limits $\alpha \uparrow 1$ in equations (A33) and (A37), one can show ρ^* is continuous in α at $\alpha = 1$. This proves that ρ^* is strictly quasiconvex in α, achieving a minimum $\rho^* = 1/(1 - \ln x_0)$ at $\alpha = 1$. *Q.E.D.*

Appendix B: Harberger Deadweight Loss under Cournot Competition

This appendix computes tight bounds on Harberger deadweight loss for a market with n homogeneous firms engaging in Cournot competition. For conciseness, we will suppress the argument indicating the competition model (earlier we used C for this purpose), adding n as an argument to earlier monopoly notation to indicate Cournot competition with that many firms. As discussed in the text, the demand curve maximizing relative Harberger deadweight loss has firms only serve the highest-demand consumers. For this outcome to be an equilibrium, each firm must weakly prefer the revenue obtained from serving a $1/n$ share of highest-demand consumers to any higher output. The worst case is generated by distributing the given mass μ under the demand curve such that each firm is indifferent among all these quantities. It can be shown that the following rescaled demand curve accomplishes this:

$$\underline{\Phi}(x,\mu,n) = \min\left\{A(\mu,n)\left(\frac{1}{x}+n-1\right),1\right\}, \tag{B1}$$

where the constant $A(\mu,n)$ is set to preserve the mean consumer value, i.e., $\mu = \int_0^1 \underline{\Phi}(x,\mu,n)dx$. Integrating equation (B1), one can show that $A(\mu,n)$ is the implicit solution for A in

$$\mu = \frac{A}{1-(n-1)A} - A\ln\left(\frac{A}{1-(n-1)A}\right) + A(n-1)\left[\frac{1-nA}{1-(n-1)A}\right]. \tag{B2}$$

Following equation (20), one can show that industry producer surplus equals $nA(\mu,n)$, so the industry producer surplus ratio equals

$$\underline{\rho}(\mu,n) = \frac{nA(\mu,n)}{\mu}. \tag{B3}$$

One can directly verify that $\underline{\Phi}(x,\mu,1) = \underline{\Phi}(x,\mu)$ and $A(\mu,1) = A(\mu)$, implying that the Cournot formulas (B1) and (B2) given here nest the monopoly formulas (16) and (17) given earlier. Worst-case Harberger deadweight loss in this Cournot market is $\underline{HDWL}^*(n) = 1-\underline{\rho}(\mu,n)$.

We will derive an analytical expression for $\underline{\rho}(\mu,n)$ in terms of the Lambert W function defined in footnote 11. Substituting

$$g(A,n) = \frac{A}{1-(n-1)A} \tag{B4}$$

into (B2) yields, after considerable rearranging,

$$\frac{-\mu}{g(A,n)} = \ln g(A,n) - 1 - (1-\mu)(n-1).$$

Exponentiating both sides and rearranging,

$$\frac{-\mu}{g(A,n)}\exp\left(\frac{-\mu}{g(A,n)}\right) = \frac{-\mu}{\exp((1-\mu)(n-1)+1)}.$$

The Lambert W function is the inverse of the left-hand-side function, implying

$$\frac{-\mu}{g(A,n)} = W\left(\frac{-\mu}{\exp((1-\mu)(n-1)+1)}\right).$$

Table 3: Worst-Case Harberger Deadweight Loss under Cournot Competition

	Number of Cournot competitors					
μ	$n=1$	$n=2$	$n=4$	$n=8$	$n=16$	$n=32$
0.2	0.750	0.617	0.470	0.332	0.218	0.136
0.4	0.669	0.532	0.394	0.274	0.180	0.113
0.6	0.579	0.447	0.325	0.225	0.149	0.094
0.8	0.452	0.338	0.243	0.168	0.112	0.072

Substituting for $g(A,n)$ from (B4) and then for $A(\mu)$ from (B3) yields

$$\underline{\rho}(\mu,n) = \frac{-n}{W_{-1}\left(\dfrac{-\mu}{\exp((1-\mu)(n-1)+1)}\right)-(n-1)\mu}. \tag{B5}$$

The formula uses the lower branch $W_{-1}(\cdot)$ of the Lambert W function, argued in the proof of Proposition 3 to be the relevant one in our context. Note that (20) is a special case of (B5) with $n=1$.

Table 3 applies the formula in (B5) to compute $1-\underline{\rho}(\mu,)$, the maximum possible Harberger deadweight loss in a Cournot model with n homogeneous competitors for various values of μ. Note the values in the $n=1$ column apply for the case of monopoly, for which the potential for Harberger deadweight loss is worse than any other market structure considered in Proposition 6.

Figure 14 graphs the demand curve from equation (B1) for various values of n. As n increases, the demand pivots clockwise through a point, but otherwise retains much of its Zipf shape. The shaded rectangles

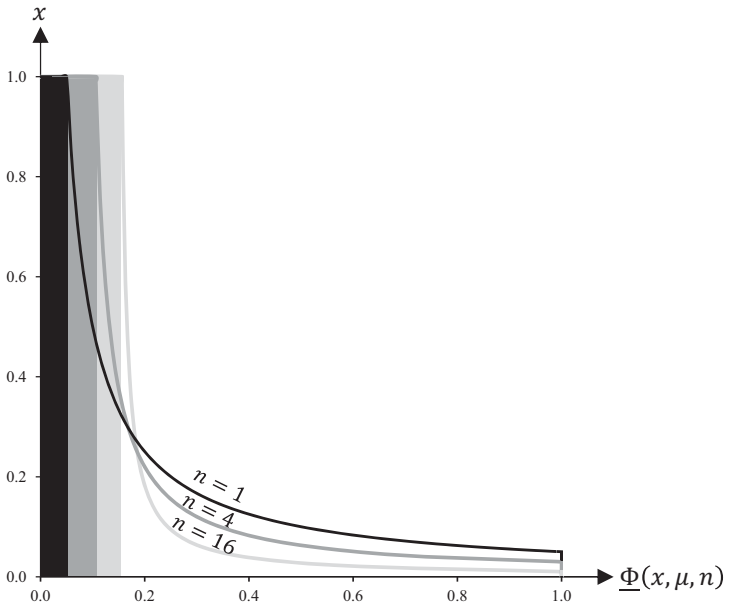


Figure 14: Maximizing Harberger deadweight loss under Cournot competition. Drawn for $\mu = 0.2$. Rectangles indicate industry producer surplus corresponding to demand curves of same shade.

indicate producer surplus added across the Cournot firms. The area of the black rectangle equals producer surplus earned by the one firm operating under the black demand curve, the area of black plus the dark grey rectangle equal producer surplus earned by the four firms operating under the dark grey demand curve, etc. By construction, there is no consumer surplus in equilibrium, so Harberger deadweight loss equals the area under demand outside of the relevant shaded rectangles. As the figure shows, Harberger deadweight loss shrinks as n increases from 1 to 4 to 16. More concretely, one can apply standard numerical methods to solve (B2) for specific values of μ and n and then substitute to find the maximum value of relative Harberger deadweight loss $1-\underline{\rho}(\mu,n)$. Table 3 undertakes this computation for a range of values of μ and n (including $n=1$, which covers the monopoly case with the highest potential for Harberger deadweight loss according to Proposition 4). From the table one can see $1-\underline{\rho}(0.2,1)=0.750$, $1-\underline{\rho}(0.2,4)=0.470$, and $1-\underline{\rho}(0.2,16)=0.218$. With as many as $n=16$ Cournot competitors, the potential for Harberger deadweight loss, even in the worst case, is limited.

References

Acemoglu, D. and J. Linn (2004). "Market Size in Innovation: Theory and Evidence from the Pharmaceutical Industry," *Quarterly Journal of Economics* 119: 1049–1090.

Alessandria, G. and J. P. Kaboski. (2011) "Pricing-to-Market and the Failure of Absolute PPP," *American Economic Journal: Macroeconomics* 3: 91–127.

Anderson, S. and R. Renault. (2003) "Efficiency and Surplus Bounds in Cournot Competition," *Journal of Economic Theory* 113: 253–264.

Basu, S. and A. DasGupta. (1997) "The Mean, Median, and Mode of Unimodal Distributions: A Characterization," *Theory of Probability and Its Applications* 41: 210–223.

Baumol, W. J., J. C. Panzar, and R. D. Willig. (1982) *Contestable Markets and the Theory of Industry Structure*. New York: Harcourt Brace Jovanovich.

Bergemann, D., B. Brooks, and S. Morris. (2014) "The Limits of Price Discrimination," Cowles Foundation working paper no. 1896RR.

Biehl, A. R. (2001) "Durable-Goods Monopoly with Stochastic Values," *Rand Journal of Economics* 32: 565–577.

Brooks, B. A. "Surveying and Selling: Belief and Surplus Extraction in Auctions," Princeton University working paper.

Budish, E., B. N. Roin, and H. Williams. (2013) "Do Fixed Patent Terms Distort Innovation? Evidence from Cancer Clinical Trials," National Bureau of Economic Research working paper no. 19430.

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

De Graba, P. (1995) "Buying Frenzies and Seller-Induced Excess Demand," *Rand Journal of Economics* 26: 331-342.

Dharmadhikari, S. and K. Joag-dev. (1988) *Unimodality, Convexity, and Applications*. Boston: Academic Press.

Dosi, G. (1988) "Sources, Procedures, and Microeconomic Effects of Innovation," *Journal of Economic Literature* 26: 1120Ű-1171.

Fabinger, M. and E. G. Weyl. (2014) "A Tractable Approach to Pass-Through Patterns with Applications to International Trade," SSRN working paper, available at http://ssrn.com/abstract=2194855.

Finkelstein, A. (2004). "Static and Dynamic Effect of Health Policy: Evidence from the Vaccine Industry," *Quarterly Journal of Economics* 119: 527–564.

Freeman, C. (1944) "The Economics of Technical Change," *Cambridge Journal of Economics* 18: 463Ű-514.

Furman, J. L. and S. Stern. (2011) "Climbing Atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research," *American Economic Review* 101: 1933-Ű1963.

Gabaix, X. (2009) "Power Laws in Economics and Finance," *Annual Review of Economics* 1: 255–293.

Garber, A. M., C. I. Jones, and P. Romer. (2006) "Insurance and Incentives for Medical Innovation," *Forum for Health Economics & Policy* 9: 1–27.

Harris, M. and A. Raviv. (1981) "A Theory of Monopoly Pricing Schemes with Demand Uncertainty," *American Economic Review* 71: 347–365.

Hartline, J. D. and T. Roughgarden. (2009) "Simple Versus Optimal Mechanisms," *Proceedings of the 10th ACM Conference on Electronic Commerce* 225–234.

Johnson, J. P. and D. P. Myatt. (2006) "On the Simple Economics of Advertising, Marketing, and Product Design," *American Economic Review* 96: 756–784.

Katz, M. L. (1987) "The Welfare Effects of Third-Degree Price Discrimination in Intermediate Good Markets," *American Economic Review* 77: 154–167.

Kerman, J. (2011) "A Closed-Form Approximation for the Median of the Beta Distribution," arXiv:1111.0433 [math.ST].

Kremer, M. and C. M. Snyder. (2003) "Why Are Drugs More Profitable Than Vaccines?" National Bureau of Economic Research working paper no. 9833.

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

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

Kremer, M. and C. M. Snyder. (2015) "Preventives Versus Treatments Redux: Tighter Bounds on Distortions in Innovation Incentives with an Application to the Global Demand for HIV Pharmaceuticals," *Review of Industrial Organization* 53: 235–273.

Liljeros, F., C. R. Edling, L. A. Nunes Amaral, H. E. Stanley, and Y. Åberg. (2001) "The Web of Human Sexual Contacts," *Nature* 411: 907–908.

Makowski, L. and J. M. Ostroy. (1995) "Appropriation and Efficiency: A Revision of the First Theorem of Welfare Economics," *American Economic Review* 85: 808–827.

Makowski, L. and J. M. Ostroy. (2001) "Perfect Competition and the Creativity of the Market," *Journal of Economic Literature* 39: 479–535.

Malueg, D. A. (1993) "Bounding the Welfare Effects of Third-Degree Price Discrimination," *American Economic Review* 83: 1011–1021.

Maleug, D. A. and C. M. Snyder (2006) "Bounding the Relative Profitability of Price Discrimination," *International Journal of Industrial Organization* 24: 995-1011.

Neeman, Z. (2003) "The Effectiveness of English Auctions," *Games and Economic Behavior* 43: 214–238.

Newell, R., A. Jaffee, and R. N. Stavins. (1999) "The Induced Innovation Hypothesis and Energy-Saving Technological Change," *Quarterly Journal of Economics* 114: 907–940.

Pinkovskiy, M. and X. Sala-i-Martin. (2009) "Parametric Estimations of the World Distribution of Income," National Bureau of Economic Research working paper no. 15433.

Romer, P. (1994) "New Goods, Old Theory, and the Welfare Costs of Trade Restrictions," *Journal of Development Economics* 43: 5–38.

Runnenburg, J. T. (1978) "Mean, Median, Mode," *Statistica Neerlandica* 32: 73–79.

Sala-i-Martin, X. (2006) "The World Distribution of Income: Falling Poverty and . . . Convergence, Period," *Quarterly Journal of Economics* 121: 351–397.

Thomas, P. (2002) "The Economics of Vaccines," *Harvard Medical International (HMI) World*. September/October.

Timerding, H. E. (1915) *Die Analyse des Zufalls*. Braunschweig (in Leiden University Library).

van Zwet, W. R. (1979) "Mean, Median, Mode II," *Statistica Neerlandica* 33: 1–5.

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

Weyl, G. and J. Tirole. (2012) "Market Power Screens Willingness-to-Pay," *Quarterly Journal of Economics* 127: 1971–2003.