Quotas in General Equilibrium

David R. Baqaee UCLA

Kunal Sangani Northwestern*

November 4, 2025

Abstract

We study economies with quotas and other quantity-based distortions. Unlike economies with wedge-based distortions, economies with quotas are constrained efficient, not subject to the theory of second best, and satisfy macro-envelope conditions. Thus, the aggregate effects of changes in technologies, distortions, and policy can be summarized by a small set of sufficient statistics. We provide a nonparametric and nonlinear characterization of how quota and productivity changes affect aggregate output, and we derive the welfare costs of misallocation from an inverse demand system that maps quota prices to quota levels. We illustrate the framework by quantifying the effects of reforms in several settings: raising the cap on H-1B visas, relaxing single-family zoning in U.S. cities, eliminating New York City's taxi medallions, phasing out U.S. quotas on Chinese textiles and apparel, and removing capital controls in Argentina. Across these applications, our method flexibly measures the costs of quota distortions and the gains from reform.

*Emails: baqaee@econ.ucla.edu, sangani@northwestern.edu. We thank Ariel Burstein, Pablo Fajgelbaum, Chang He, Oleg Itskhoki, Pete Klenow, Hannes Malmberg, Alireza Tahbaz-Salehi, as well as seminar participants at Boston University, Stanford, and UCLA for valuable comments. We are also grateful to Judith Dean, Amit Khandelwal, and Peter Schott for sharing data on quota prices for Chinese clothing and textile exports.

1 Introduction

Quotas—and quantity-based distortions more broadly—are pervasive across a wide range of markets. Policies such as import quotas, visa caps, zoning restrictions, emissions limits, and local content requirements directly restrict quantities of activities or inputs, without regard to prices. Missing markets likewise constrain quantities regardless of shadow prices: the absence of credit markets limits trade across time, and the absence of insurance markets limits transactions across states of nature.

A classic approach to analyzing such distortions is to recast them as implicit taxes (or wedges). Reforms can then be evaluated by mapping changes in the underlying distortions to changes in effective tax rates. But constructing this mapping from quotas to effective tax rates typically requires detailed knowledge of the economy's structure, including input-output linkages, elasticities of substitution in production and consumption, and wedges elsewhere in the economy.

We analyze quota distortions in economies with general production functions, input-output linkages, and any number of factors and goods. We show that quotas, like wedges, can be used to implement any distorted allocation of resources in an economy. Unlike economies with wedge distortions, however, economies with quotas are constrained efficient: allocations maximize output subject to the quotas. Consequently, these economies satisfy macro-envelope conditions and are not subject to the theory of second best, which greatly simplifies comparative statics.

We use this property of quota economies to develop three sets of results. First, we provide first-order approximations for the effect of quota changes and productivity shocks on output. Second, we characterize nonlinearities in the effects of quota changes on output. Finally, we derive expressions for the misallocation cost of quota distortions. We discuss each of these results in turn.

To a first order, the effect of changing a quota on output is summarized by the rents of producers who own the rights to operate under the quota. Intuitively, because the economy is constrained efficient, the rents earned by quota holders precisely reflect the value of permitting a marginal increase in the restricted activity. If a quota does not bind, quota holders earn zero rents, and adjusting the quota has no first-order effect on output.[1]

[1] This is because the marginal value of any resource is equated across uses when the quota is non-binding, so even if the quota diverts resources, there is no first-order effect on output. In Online Appendix E, we extend our framework to allow for rent-seeking, in which agents waste productive resources competing for rents, à la Bhagwati (1965) or Krueger (1974) (see Section 8 for a summary). In this case, even starting at the efficient point, a just-binding quota can reduce output since it diverts resources towards competition for rents, which has zero marginal value. Rent-seeking is not unique to quotas and can occur with taxes as well (see, e.g., Liu 2019).

However, when a quota binds, quota holders earn positive rents, and loosening the quota raises aggregate output by an elasticity equal to rents divided by GDP.

Likewise, the effect of a productivity shock on output is proportional to the affected producer's initial sales less the rents of quota holders. When rents are zero, the effect of a productivity shock is given by the sales of the affected producer, as in Hulten's (1978) theorem. When rents are positive, the effect of productivity shocks is dampened relative to Hulten's theorem because the resources saved from an increase in productivity are diverted to lower marginal value users.

These comparative statics rely on only a few sufficient statistics: quota rents and sales. The parsimony comes from constrained efficiency. Relaxing a binding quota (holding all other quotas fixed) reallocates resources from unconstrained uses toward the constrained producer. Alleviating this constraint thus always increases output, and the gain in the marginal revenue product from redirecting resources to the constrained user is precisely reflected by the quota's rents—i.e., the constrained producer's profit margin. By contrast, in economies with tax- or wedge-like distortions, cutting one tax can reallocate resources throughout the economy, potentially away from producers whose marginal value is even higher than that of the targeted firm. Analyzing wedges therefore typically requires rich information about the economy's structure, which is not needed when distortions take the form of quotas.

Since the effects of marginal quota changes depend on a quota's rents, the nonlinear effects of large reforms depend on how rents change in response to quota changes. Thus, computing these nonlinear effects requires knowing the *quota demand system*, which describes rents (or, equivalently, quota prices) as a function of each quota's levels. If total rents earned by quota holders remain roughly constant as a quota is tightened or relaxed, then the log-linear approximation provides a good estimate for the effects of a large reform. But if rents change with the quota level, then the effects of a large reform can diverge sharply from the first-order approximation.

There are two ways to obtain the quota demand system: (i) specify a structural, microfounded model or (ii) estimate it directly from the data. The first approach builds up the quota demand system from the economy's input-output structure and microeconomic elasticities of substitution. Given a fully-specified model of the economy, we provide a method for computing the nonlinear effects of large reforms that accommodates the possibility that other quotas in the economy switch from slack to binding or vice versa in response to the reform. The second approach recovers the quota demand system by specifying a reduced-form model of the quota demand system and estimating its parameters using variation in quotas across markets and over time (subject to the usual qualifiers about identifying elasticities of demand).

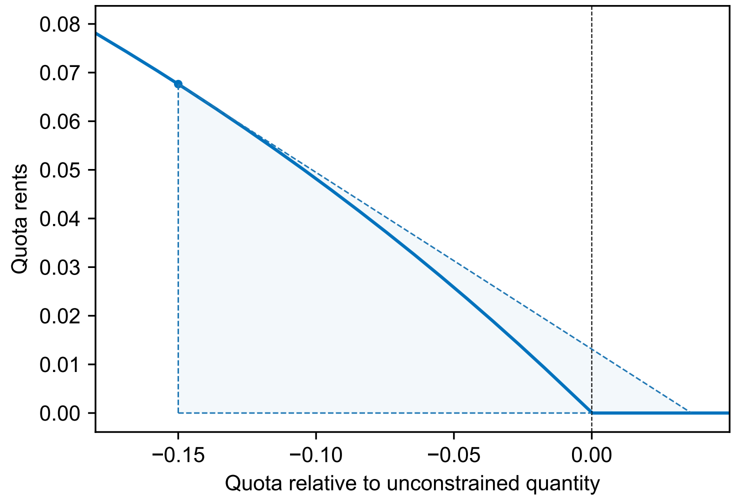
Even without fully specifying the economic environment, the nonlinear effects of a large quota change can be approximated to a second order using the local response of rents to quota changes. Whether rents rise or fall as a quota is tightened determines whether output is locally log concave or log convex with respect to quota changes. Interestingly, while output is always log concave with respect to quota changes near the efficient point, it can become log convex when the equilibrium is sufficiently far from efficiency. That is, in economies with preexisting distortions, nonlinearities can amplify the benefits of large liberalizations relative to small ones and mitigate the losses from further distortions. These cases occur when total rents generated by a quota rise, rather than fall, as the quota is relaxed—i.e., when the economy is on the wrong side of the "Laffer curve" for revenues generated by a quota. We show this also implies that when a quota is set to maximize the real rents it generates (taking other quotas as given), output is always log concave in quota changes.

Our characterization of effects of quota changes extends to dynamic economies, where quotas may vary across time periods or states. By treating goods as indexed by both commodity and date, à la Arrow–Debreu, we can apply our results to analyze the effects of a sequence of quota changes. In an explicitly dynamic setting, the first-order effect of a change to a future quota depends on the present discounted value of rents for that dated good as a share of wealth (i.e., the net present value of income), and nonlinear effects depend on how those rents change in response to a reform. When perpetual rights to produce under a quota are traded on asset markets, we show that the effect of an anticipated path of quota changes on output can be estimated to a second-order using the asset price and its change on the announcement date.

Our final set of results characterizes the deadweight loss caused by quota distortions. The key insight is the following. If we know how rents respond to quota changes on the margin, then we can estimate how much a quota must be relaxed to reach the efficient frontier. Figure 1 illustrates this graphically in a stylized example. The figure shows quota rents as a function of the quota level. Starting with an existing quota distortion, we can extrapolate linearly to estimate the change in the quota needed to reach the point where rents are zero. This is precisely the level at which the distortion ceases to bind. Thus, to a second order, the gains from removing the quota are approximated by the area of the shaded triangle in Figure 1. This approximation requires only two statistics: rents earned at the distorted point and their response to local variation in the quota level.

This idea extends to economies with multiple quotas, where the costs of misallocation can be estimated using the vector of initial quota rents and the matrix of cross-price

Figure 1: Estimating the distance to the frontier due to a quota distortion.



elasticities of the quota demand system. This elasticity matrix allows us to infer how much any individual quota needs to be relaxed to cease binding, as well as interactions between quotas, which depend on how rents earned by holders of one quota change when another quota is relaxed or tightened. Once this matrix is estimated, it can be used to calculate the gains from removing a single quota, any subset of quotas, or eliminating quotas altogether to achieve the first-best allocation.

We demonstrate the applicability of our framework using several empirical examples. Specifically, we examine:

1. the gains from raising the cap on H-1B visas;

2. the gains from relaxing zoning restrictions on single-family housing;

3. the output cost of the restriction on taxicab medallions in New York City, and the extent to which the entry of ride-sharing companies has relaxed it;

4. how the gains from phasing out a subset of U.S. quotas on Chinese textiles and apparel imports compare with removing all quotas; and

5. the cost of Argentina's restrictions on capital outflows.

Each policy directly regulates quantities. Our framework delivers (approximate) answers in each case while imposing minimal structure on the rest of the economy, using sufficient statistics from quota rental markets, natural experiments, and microdata.

The outline of the paper is as follows. Section 2 sets up the framework and shows that any feasible distorted allocation can be decentralized with quotas. Section 3 characterizes the first-order effects of quota and productivity changes on output. Section 4 characterizes nonlinear effects of quota changes, and Section 5 applies these results to explicitly dynamic economies. Section 6 presents results on the distance to the efficient frontier. We illustrate our results in several empirical examples in Section 7. Section 8 describes extensions of our baseline framework, including (i) hybrid economies that feature both wedge and quota distortions, (ii) the effects of quota changes in economies with externalities, and (iii) the effects of quotas with rent-seeking. Section 9 concludes.

Related literature. This paper is related to a large literature on the costs of misallocation. The classic approach, dating back to Harberger (1954), models misallocation using wedges. The wedge approach has been successfully applied across a range of domains, such as growth accounting (Basu and Fernald 2002), analyzing the drivers of business cycles (Chari et al. 2007), explaining cross-country income differences (Restuccia and Rogerson 2008; Hsieh and Klenow 2009), productivity measurement (Petrin and Levinsohn 2012), calculating social losses from financial frictions and market power (Itskhoki and Moll 2019; Bigio and La'O 2020; Peters 2020; Edmond et al. 2023), estimating the benefits of reform and liberalization (De Loecker et al. 2016; Bau and Matray 2023), and analyzing monetary non-neutrality (La'O and Tahbaz-Salehi 2022; Rubbo 2023; Baqaee et al. 2024).

Baqaee and Farhi (2020) provide a general characterization of comparative statics for economies with wedge distortions. Our paper provides an analogous characterization for economies where distortions take the form of quotas rather than wedges. Since economies with quotas are constrained efficient, we are further able to characterize nonlinearities away from the first best in terms of price elasticities and expenditure shares. In contrast, in economies with wedge distortions, the first-order response of output to shocks away from the first best already depends on price elasticities (Baqaee and Farhi 2020), and so nonlinearities generally depend on super-elasticities and other higher-order derivatives, which are much more challenging to deal with.

This paper is also related to a literature that studies how microeconomic shocks affect aggregate efficiency, dating back to Domar (1961) and Hulten (1978). Carvalho and Tahbaz-Salehi (2019) and Baqaee and Rubbo (2023) provide recent surveys. One branch of this literature focuses on how micro shocks affect aggregate output in efficient economies—e.g., Foerster et al. (2011), Gabaix (2011), Acemoglu et al. (2012), Atalay (2017), and Baqaee and Farhi (2019)—while the other emphasizes the role of inefficiencies—e.g., Baqaee (2018), Grassi (2017), Liu (2019), Reischer (2019), and Buera and Trachter (2024). Our paper is at the intersection of these two branches, since the economies we study feature distortions but are constrained efficient.

Finally, we relate to studies that examine the costs of specific quantity-based constraints using quantitative models. For example, Feenstra (1988) estimates the cost of import quotas on Japanese automobiles, and Feenstra (1992) surveys evidence on losses from quotas and other protectionist trade measures across a wider array of imported goods. Khandelwal et al. (2013) estimate the costs of quotas on Chinese textile and clothing imports. Other studies estimate the costs of misallocation induced by constraints on housing supply (see e.g., Glaeser and Gyourko 2018; Hsieh and Moretti 2019). We illustrate our sufficient statistics methodology using some of these examples.[2,3]

2 Framework

In this section, we set up our framework, define an equilibrium with quotas, and show that any feasible allocation can be implemented using quotas.

2.1 Setup

Output is the maximizer of a constant-returns aggregator of final demand for goods $1, ..., N$,

$$Y = \max_{\{c_1,...,c_N\}} \mathcal{D}(c_1, ..., c_N),$$

subject to the budget constraint,

$$\sum_{i}^{N} p_i c_i = \sum_{f=1}^{F} w_f L_f + \sum_{i=1}^{N} \Pi_i,$$

where c_i is the representative household's final demand for good i, p_i is its price, w_f is the wage of factor f, L_f is the supply of factor f, and Π_i is the profit earned by producers of good i. In principle, i can index goods, as well as states of nature and time. We require that all final demands c_i are non-negative and assume that $\mathcal{D}$ is weakly increasing in each argument. We take nominal output as the numeraire throughout the paper, i.e.,

[2] Falvey (1979), Anderson (1985), and Boorstein and Feenstra (1991) emphasize that industry-level quotas can distort the consumption choices of households across varieties within an industry by causing relative prices to change. For example, higher quality varieties, which have higher prices, experience a smaller proportional increase in their price relative to lower quality, lower price varieties, when industry output is subject to a quota. This type of misallocation arises endogenously in our framework and is captured by our formulas for the aggregate cost of quotas.

[3] Our paper is not closely related to the public finance literature that studies whether policymakers should use quotas or taxes to achieve policy objectives, like raising revenues, under uncertainty (see, for example, Weitzman 1974 or Dasgupta and Stiglitz 1977).

$\sum_i^N p_i c_i = 1$.

Producers of i take prices as given and produce using a neoclassical production function,

$$A_i F_i(x_{i1}, ..., x_{iN}, L_{i1}, ..., L_{iF}),$$

where x_{ij} is the quantity of good j used in the production of good i, L_{if} is the quantity of factor f used by i, and A_i is a Hicks-neutral productivity shifter. We assume that F_i has constant returns to scale and is weakly increasing in each argument, and we require that all inputs x_{ij} and L_{if} are non-negative. These assumptions are less restrictive than they may appear: decreasing returns to scale can be captured by adding quasi-fixed factors, and input-augmenting productivity shifters can be captured by adding intermediaries that sell good j to producers of i.

A *quota* restricts the output of good i at a quantity y_i^*,

$$y_i = \min\{y_i^*, A_i F_i(x_{i1}, ..., x_{iN}, L_{i1}, ..., L_{iF})\}.$$

That is, the total quantity of good i available to consumers and other producers cannot exceed y_i^*. We do not need to specify the assignment of the quota across producers of i because of constant returns.[4] While we model quotas as restrictions on output, our framework can also accommodate input quotas: a quota on the use of input j by producers of good i can be represented as an output quota on an intermediary that supplies j to i's producers.

Gross profits for producers of good i are total revenues less costs of intermediate inputs and factors,

$$\Pi_i = p_i y_i - \sum_{j=1}^{N} p_j x_{ij} - \sum_{f=1}^{F} w_f L_{if}.$$

As anticipated by the representative household's budget constraint, gross profits of all producers are rebated to households lump sum. Since each production function F_i has constant returns to scale, gross profits in the absence of quotas are zero, but may be strictly positive when quotas are binding. We refer to these profits as *rents*.

Note that, in our baseline environment, the economy is efficient absent quota distortions. In Section 8, we extend our results to economies that feature wedge distortions in addition to quota distortions, and to environments with preexisting inefficiencies due to externalities.[5]

[4] That is, we could assume that each good is produced by a representative firm that takes its output price and all input prices as given. For this reason, we refer to quotas on goods and firms interchangeably below.

[5] In both cases, the effects of quotas on output depend on the results in the main text, plus additional

Resource constraints for each good $1 \le i \le N$ and each factor $1 \le f \le F$ are

$$c_i + \sum_{j=1}^{N} x_{ji} \le y_i \qquad \text{and} \qquad \sum_{i=1}^{N} L_{if} \le L_f.$$

We denote the Domar weight of good i—i.e., the sales of i as a share of income—by $\lambda_i = p_i y_i$, and the Domar weight of factor f by $\Lambda_f = w_f L_f$. (Recall that total income is the numeraire.)

Definition 1 (Equilibrium with quotas)**.** Given quotas y_i^*, productivities A_i, production functions F_i, and factor supplies L_f, an *equilibrium with quotas* is a set of prices p_i, factor wages w_f, outputs y_i, final demands c_i, and intermediate and factor input choices x_{ij} and L_{if} such that: final demand maximizes the final demand aggregator subject to the budget constraint; each producer maximizes profits taking prices as given; $y_i \le y_i^*$ for each good with a quota; and resource constraints for all goods and factors are satisfied.

2.2 Implementing an Allocation Using Quotas

We end this section by making a simple observation: any feasible allocation—i.e., any allocation of goods and factors that obeys production technologies and resource constraints—can be implemented as the decentralized equilibrium of an economy with quotas.

Definition 2. An allocation $\{y_i, c_i, x_{i1}, ...x_{iN}, L_{i1}, ..., L_{iF}\}_{1\le i\le N}$ is *feasible* if c_i, x_{ij}, and L_{if} are all non-negative, $y_i = A_i F_i(x_{i1}, ..., x_{iN}, L_{i1}, ...L_{iF})$ for all i, and the resource constraints hold: $c_i + \sum_{j=1}^{N} x_{ji} \le y_i$ for all i and $\sum_{i=1}^{N} L_{if} \le L_f$ for all f.

Proposition 1 (Implementation via implicit quotas)**.** *Suppose an allocation $\mathcal{X}$ is feasible. Then there exists an economy with quotas in which the allocation of the decentralized equilibrium is $\mathcal{X}$. Moreover, given these quotas, the allocation $\mathcal{X}$ is efficient.*

By introducing additional producers and using quotas, one can guarantee that the competitive equilibrium yields any desired feasible allocation. First, to ensure that the use of good j in the production of i is equal to x_{ij}, one can create a new producer k such that i's use of j flows through k. Then, introducing a quota on the output of good k at $y_k^* = x_{ij}$ guarantees that the use of good j by i is at most x_{ij}. Further quotas on every other use of good j, combined with the fact that the final demand aggregator is increasing in

terms. For economies with both quotas and wedges, the additional terms depend on how quantities for producers with wedges respond to quota changes. In the case of externalities, the additional terms depend on the willingness-to-pay to reduce the regulated activity.

all goods, can also guarantee that the use of good j by i is at least x_{ij}. Thus, given these quotas, the decentralized equilibrium with competitive firms yields exactly the desired allocation.[6]

Since the allocation can be decentralized as the equilibrium of the competitive economy, the first welfare theorem also implies that the allocation is constrained efficient. That is, the allocation $\mathcal{X}$ maximizes output among the set of allocations in the production possibilities frontier of the economy with quotas.

The following stylized example of a small open economy shows how quotas can implement any feasible allocation. We return to this example to illustrate several results throughout the paper.[7]

Example 1 (Small Open Economy)**.** Consider a small open economy in which labor (L) is the sole domestic factor and is used to produce y_d units of a domestic good (d). Import–export firms trade the domestic good for a foreign good (f) and sell the foreign good to households. The quantity of the foreign good imported, y_f, is given by $y_f = A_f x_{fd}$, where x_{fd} is the amount of domestic good used for trade and $1/A_f$ are the number of units of domestic good that must be exchanged to import one unit of the foreign good. (A_f can reflect exogenous terms-of-trade as well as iceberg trade costs.) We impose that trade is balanced. Household welfare is given by the constant elasticity of substitution (CES) aggregate,

$$Y = \left(\omega c_d^{\frac{\theta-1}{\theta}} + (1-\omega) c_f^{\frac{\theta-1}{\theta}}\right)^{\frac{\theta}{\theta-1}},$$

where c_d and c_f are household consumption of the domestic and foreign goods, θ controls the Armington trade elasticity, and ω is a taste shifter that determines the degree of home bias.

An allocation of resources in this economy can be expressed as a tuple, (y_d, c_f, c_d). The set of feasible allocations in this economy is $\{(y_d, c_f, c_d) \in \mathbb{R}^3_+ \,|\, c_f/A_f + c_d \leq y_d \leq L\}$. Since this set is three dimensional, we can implement any allocation in this set by introducing three quotas: a quota on labor use, which controls y_d, a quota on imports of the foreign good by import–export firms, which controls c_f, and a quota on the consumption of the domestic good, which controls c_d.

[6] This result generalizes the classic trade result on the equivalence of tariffs and quotas (see e.g., Bhagwati 1965), since any allocation that can be obtained as the decentralized equilibrium of an economy with wedges can also be implemented using quotas. In fact, quotas can even implement some feasible allocations (e.g., a desired input mix under Leontief production) that cannot be implemented with any wedges.

[7] While our general model is of a closed economy, rather than an open economy, this stylized model of a small open economy is a special case of our framework since balanced trade with exogenous terms-of-trade is equivalent to having a linear technology that converts domestic goods into foreign goods with some exogenous rate of transformation.

3 First-Order Effects

How do changes in distortions and technologies affect output in an economy with quotas? In this section, we characterize the response of output to quota changes and productivity shocks up to a first-order approximation.

3.1 First-Order Effects of Quota and Productivity Changes

Proposition 2 describes the change in output resulting from changes to quotas and producer productivities.

Proposition 2 (First-order effects with quotas)**.** *To a first order, the change in output resulting from changes in quotas $d\log y_i^*$ and productivities $d\log A_i$ is*

$$d\log Y = \sum_{i^*} \Pi_i d\log y_i^* + \sum_i (\lambda_i - \Pi_i)\, d\log A_i.$$

If all quotas are non-binding, then $d\log Y = \sum_i \lambda_i d\log A_i$.

The elasticity of aggregate output to a change in quota i is the rents earned by that quota Π_i. Positive rents indicate that the quota is a binding constraint on production. Thus, when rents are positive, relaxing the quota constraint increases the production of the good and total output. Note that calculating the effect of relaxing a quota does not require specifying where in the economy the additional resources used in the production of i will come from. Because the economy is constrained efficient, producer i's rents exactly reflect the value of assigning it more resources relative to unconstrained producers.

Likewise, the elasticity of output with respect to i's productivity is i's sales minus rents. If rents are positive, then a binding quota prevents producers from expanding their output when their productivity rises. Rather than increasing their output, the increase in productivity thus frees up some of the resources that were required to produce the quota amount. Constrained efficiency implies that the value of those freed-up resources is proportional to their Domar weight, i.e., the costs of constrained producers.

In an economy without any binding quota distortions, all rents are zero. In this case, Proposition 2 shows that the comparative statics converge to familiar results for efficient economies. Specifically, the introduction of marginal distortions has no first-order effect on output, since efficiency equates the marginal benefit of inputs across all uses, and the elasticity of output to productivity shocks is exactly equal to the sales shares of affected producers, as in Hulten's (1978) theorem.

We illustrate these results in the small open economy from Example 1.

Example 2 (Small Open Economy with Import Quota)**.** Consider the small open economy from Example 1, and suppose the only binding quota is the import quota y_f^*. We apply Proposition 2 to see how changes in the import quota and iceberg trade cost affect output.

The effect of a change in the import quota by $d \log y_f^*$ is

$$\frac{d \log Y}{d \log y_f^*} = \Pi_f. \tag{1}$$

That is, the welfare gains from increasing the import quota are proportional to the rents earned by import–export firms, Π_f. These rents are equal to import–export firms' sales, λ_f, minus the costs of imports, $p_d x_{fd}$.[8]

Likewise, the gains from reducing trade costs—i.e., increasing the productivity A_f of import–export firms—are given by

$$\frac{d \log Y}{d \log A_f} = \lambda_f - \Pi_f. \tag{2}$$

Due to the import quota, a reduction in trade costs does not actually increase household consumption of imported goods. But it does reduce the amount of domestic good that is required for exchange with the foreign good. As a result, the reduction in trade costs increases welfare by increasing the quantity of domestic good that remains for consumption by households. Thus, the output gains from a reduction in trade costs are also equal to the household expenditure share on the domestic good $(1 - \lambda_f)$, multiplied by the ratio of the amount of the domestic good used for trade to the amount consumed, i.e., $\lambda_f - \Pi_f = (1 - \lambda_f)(y_d - c_d)/c_d$.

3.2 Comparison to Economies with Wedge Distortions

It is useful to contrast the effect of shocks in an economy with quotas to the effect of shocks in an economy with wedge distortions.

Definition 3 (Equilibrium with wedges)**.** Given wedges τ_i, productivities A_i, production functions F_i, and factor supplies L_f, an *equilibrium with wedges* is a set of prices p_i, factor wages w_f, outputs y_i, final demands c_i, and intermediate and factor input choices x_{ij} and L_{if} such that: final demand maximizes the final demand aggregator subject to the budget constraint; each producer minimizes costs taking prices as given; the price of good i equals

[8] In terms of primitives, when y_f^* is binding, $y_d = L$, $c_d = L - A_f y_f^*$, $\lambda_f = 1/(1 + \frac{\omega}{1-\omega}(L/y_f^* - 1/A_f)^{\frac{\theta-1}{\theta}})$, and $\Pi_f = \lambda_f - p_d x_{fd} = \lambda_f - (1 - \lambda_f)(y_d/c_d - 1)$.

its marginal cost times the exogenous wedge τ_i; wedge revenues $\Pi_i = (1 - 1/\tau_i)\, p_i y_i$ are rebated to the representative household; and resource constraints for all goods and factors are satisfied.[9]

Proposition 3 characterizes the effects of wedge changes and productivity shocks on output in an economy with wedge distortions, summarizing results developed by Petrin and Levinsohn (2012) and Baqaee and Farhi (2020).

Proposition 3 (First-order effects with wedges: Petrin and Levinsohn 2012)**.** *In an economy with wedge distortions, the effect of wedge changes $d \log \tau_i$ and productivity shocks $d \log A_i$ on output is*

$$d \log Y = \sum_i \sum_j \Pi_i \left[\frac{\partial \log y_i}{\partial \log \tau_j} d \log \tau_j + \frac{\partial \log y_i}{\partial \log A_j} d \log A_j \right] + \sum_i (\lambda_i - \Pi_i)\, d \log A_i, \qquad (3)$$

where $\partial \log y_i / \partial \log \tau_j$ and $\partial \log y_i / \partial \log A_j$ are general-equilibrium elasticities of y_i with respect to changes in wedge τ_j and productivity A_j, respectively. If $\tau_i = 1$ for all i, then $d \log Y = \sum_i \lambda_i d \log A_i$.

As in an economy with quotas, if profits for all producers are initially zero, $\Pi_i = 0$, marginal wedge distortions have no first-order effect on output, and the effect of productivity shocks is given by Hulten's theorem. However, if there are existing distortions, the effect of wedge shocks and productivity shocks on output depends on how the quantities of all producers with non-zero profits respond to the shocks.[10] Computing these responses generally requires information about elasticities of substitution in production and consumption, input-output linkages, and so on. Moreover, in economies with multiple wedge distortions, there is no guarantee that removing the wedge on one producer will improve efficiency and output, due to the theory of second best (Lipsey and Lancaster 1956).

The usefulness of Proposition 2 over Proposition 3 depends on the extent to which quotas can be treated as primitives. If the mapping from primitive shocks to changes in quotas is itself complicated, then Proposition 2 is less useful. For example, if the primitive economy features taxes, and we represent that allocation using quotas instead,

[9] The assumption that wedges are applied to output prices is without loss of generality, since user-good-specific wedges can be modeled by introducing intermediaries with wedges.

[10] Note that, even if an economy with wedges and an economy with quotas share the same physical allocation of resources, quota rents and wedge revenues across the two economies may differ. This means that applying Proposition 2 to an economy with wedges, and treating wedge revenues as quota rents, can lead to inaccurate results. In Appendix C.1, we provide a set of restrictions on wedges such that quota rents and wedge revenues coincide.

then all quotas may need to adjust endogenously in response to changes in one tax rate. In this case, calculating the endogenous changes in quotas ultimately requires the same information about the structure of the economy that is required to calculate the effects of changes in wedges (i.e., information about the input-output structure, elasticities of substitution, returns to scale, etc.). However, in cases where the primitive distortions are quantity-based, then computing the effects of quota changes using the equivalent wedge representation in (3) requires information about the economy that, given Proposition 2, is unnecessary.

Example 3 (Small Open Economy with Import Tariff)**.** Suppose the allocation in our small open economy from Examples 1–2 is implemented with an import tariff rather than an import quota. Following Proposition 3, the effect of a change in the import tariff $d\log\tau_f$ on welfare is

$$\frac{d\log Y}{d\log\tau_f} = \Pi_f \frac{d\log y_f^*}{d\log\tau_f} = -\theta\Pi_f\frac{c_d}{y_d}.$$

Increases in the import tariff reduce welfare. The effect is stronger when the economy is more open, as measured by the ratio of the domestic good used for domestic consumption c_d/y_d, and when the trade elasticity θ is high, because a higher trade elasticity leads to a greater reduction in imports. Note that calculating the effect of changes to the import quota in (1) required only the rents earned by importers. In comparison, calculating the effect of tariff changes requires additional structural parameters: the trade elasticity θ and information about the economy's structure (in this case, the share of domestic good used for consumption, c_d/y_d).

Following Proposition 3, the effect of a decline in trade costs (i.e., $d\log A_f > 0$) is:

$$\frac{d\log Y}{d\log A_f} = \lambda_f - \Pi_f + \Pi_f\frac{d\log y_f^*}{d\log A_f} = \lambda_f - \Pi_f + \frac{\Pi_f}{1-\Pi_f}\left[\left(\lambda_f - \Pi_f\right) + \theta\left(1-\lambda_f\right)\right].$$

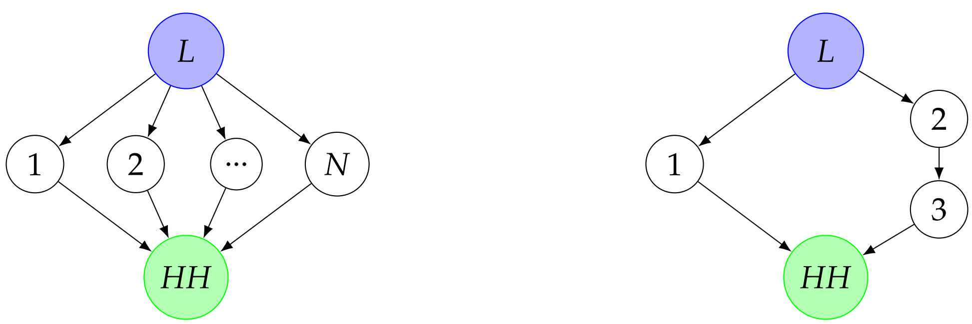
Computing the effect of the decline in trade costs in the tariff economy again requires knowing the trade elasticity θ, which is not necessary to compute the effect of the decline in trade costs in the quota economy.

Note also that the effect of a decline in trade costs in the economy with quotas generally *differs* from the effect of an identical decline in the economy with an import quota in (2). The two expressions only coincide when the level of imports is undistorted or the economy is in autarky (in both of these cases, $\Pi_f = 0$). The difference arises because tariffs allow the quantity of imports to adjust when trade costs change, whereas the binding import quota fixes import levels. In other words, despite the two economies sharing the same initial allocation of resources, the effect of changes in trade costs across the two economies generally differs depending on whether the primitive distortion takes the form of a quota or tax.

The trade cost shock in the previous example highlights that whether distortions take the form of quotas or wedges matters for the reallocations that take place in response to shocks. In economies with quotas, when a quota is relaxed, resources are always reallocated to a constrained producer from unconstrained uses. In economies with wedge distortions, reducing the wedge on a producer can reallocate resources throughout the economy, even from parts of the economy that are more constrained than the producer whose wedge is reduced. This difference is at the root of why economies with quotas are constrained efficient, but not economies with wedges.

We illustrate why the reallocations triggered by relaxing a quota avoid the usual challenges of the Theory of Second Best in the two following examples.

Figure 2: Illustrative examples.



(a) Horizontal economy.

(b) Interdependent producers.

Example 4 (Reallocations Under Quotas vs. Wedges)**.** Consider the horizontal economy illustrated in Figure 2a. Firms $1, \ldots, N$ use labor to produce varieties. A representative household has CES preferences over these varieties with an elasticity of substitution θ. We compare how relaxing a distortion on firm 1 affects output when distortions are implemented with wedges versus with quotas.

If distortions are implemented with wedges, we can apply Proposition 3 to calculate the effect of a change in the wedge on firm 1 that increases 1's output by $d \log y_1$:

$$\frac{d \log Y}{d \log \tau_1} = \Pi_1 \frac{d \log y_1}{d \log \tau_1} - \frac{l_1}{1 - l_1} \left(\sum_{i \neq 1} \Pi_i \right) \frac{d \log y_1}{d \log \tau_1}, \qquad \text{(Wedge economy)}$$

where $l_1 = L_1 / L$ is the share of labor used by firm 1. When the wedge on firm 1 is relaxed, the resources gained by firm 1 come proportionally from the cross-section of other firms. Even when $\Pi_1 > 0$, firm 1 may be less constrained than the average firm in the horizontal economy, and the overall output effect can be negative. In other words, the presence of multiple distortions means that reducing one distortion may actually exacerbate other distortions and reduce, rather than improve, efficiency.

If the allocation were instead implemented with quotas, Proposition 2 describes the effect of relaxing the constraint on firm 1:

$$\frac{d\log Y}{d\log y_1^*} = \Pi_1. \qquad \text{(Quota economy)}$$

In the economy with quotas, the output of any other firms with binding quotas is unchanged to a first-order, and so the resources reallocated to firm 1 as the quota is relaxed come only from initially unconstrained firms. These unconstrained firms are precisely those where the marginal benefit of resources is initially lowest, and so the reallocation of resources from them toward firm 1 always weakly improves output and strictly improves output if the quota on firm 1 is binding (i.e., $\Pi_1 > 0$).

Example 5 (Interdependent Producers)**.** Next, consider the economy in Figure 2b: firm 1 produces a consumption good using labor, firm 2 produces an intermediate that is used by firm 3 to produce a consumption good, and households have CES preferences over the consumption goods produced by firms 1 and 3 with an elasticity of substitution θ.

Suppose first that an allocation of resources in this economy is implemented with wedges τ_1, τ_2, and τ_3. Applying Proposition 3, the effect of reducing the wedge τ_2 is

$$\frac{d\log Y}{d\log \tau_2} = \sum_i \Pi_i \frac{\partial \log y_i}{\partial \log \tau_2} d\log \tau_2 = \theta \left[\Pi_1 - (\Pi_1 + \Pi_2 + \Pi_3)\, l_1\right], \qquad \text{(Wedge economy)}$$

where $l_1 = L_1/L$ is the share of labor used by firm 1. Notice that this effect can be positive or negative depending on firms' initial profits. That is, removing the distortion does not unambiguously increase output. Moreover, comparing τ_2 to τ_1 alone is not sufficient to identify whether removing the wedge on firm 2 increases output, because of the interdependence between firm 2 and firm 3. The importance of these interdependencies for evaluating policies was emphasized by McKenzie (1951).

If the same allocation were instead implemented with quotas, the effect of relaxing the quota on firm 2 is instead

$$\frac{d\log Y}{d\log y_2^*} = \Pi_2. \qquad \text{(Quota economy)}$$

In the economy with quotas, the rents of firm 2 reflect the difference in marginal product between firm 2 and firm 1. Thus, relaxing the quantity constraint on firm 2 always weakly increases output and strictly increases output when these rents are non-zero.

4 Nonlinearities

While the previous section analyzed marginal quota changes, evaluating major reforms requires understanding the economy's nonlinear response to large shocks. This section characterizes nonlinearities in the effects of quota changes on output.

4.1 Nonlinear Effects of Quotas

Since the first-order effect of a marginal quota change is always given by Proposition 2, the total effect of moving from quota vector $\boldsymbol{y}_0^*$ to $\boldsymbol{y}_1^*$ can be obtained by integrating marginal effects.

Proposition 4 (Nonlinear effects of quotas)**.** *The effect of changing a vector of quotas from $\boldsymbol{y}_0^*$ to $\boldsymbol{y}_1^*$ on output is*

$$\Delta \log Y = \int_0^1 \sum_i \Pi_i(\boldsymbol{y}^*(s)) \frac{d \log \boldsymbol{y}_i^*}{ds} ds, \tag{4}$$

where $\boldsymbol{\Pi}(\boldsymbol{y}^*)$ is the vector of rents given quotas $\boldsymbol{y}^*$ and $\boldsymbol{y}^*(s)$ is any smooth function of $s \in [0,1]$ with end points $\boldsymbol{y}^*(0) = \boldsymbol{y}_0^*$ and $\boldsymbol{y}^*(1) = \boldsymbol{y}_1^*$.

The difference between the nonlinear effects of quota changes in Proposition 4 and the first-order approximation in Proposition 2 depends on how rents $\boldsymbol{\Pi}(\boldsymbol{y}^*)$ evolve as the reform becomes larger. Thus, the first-order approximation in Proposition 2 provides an accurate estimate of the effects of a large reform if rents are unchanged by the reform. But if rents change as quotas are relaxed or tightened, then the effects of a large reform can differ substantially from the first-order approximation. We refer to $\boldsymbol{\Pi}(\boldsymbol{y}^*)$ as the *quota demand system*, because it describes rents (and thus quota prices) as a function of all quota levels.

In the rest of this section, we use Proposition 4 to study the nonlinear effect of changes in quotas. First, we use (4) to show that the elasticity of rents to quota changes can be used to estimate the effects of a reform to a second order. Second, we show how, under some assumptions, the full nonlinear effects of a change in quotas can be computed using equation (4).

4.2 Second-Order Approximation

Proposition 5 characterizes the effect of quota changes on output to a second order by twice differentiating (4). This provides an analogue to Baqaee and Farhi (2019) for economies with quota distortions.

Proposition 5 (Nonlinear effects of quotas: Second-order approximation)**.** *The effect of a vector of quota changes $\Delta \log \mathbf{y}^*$ on output to a second order is*

$$\Delta \log Y \approx \mathbf{\Pi}' \Delta \log \mathbf{y}^* + \frac{1}{2} (\Delta \log \mathbf{y}^*)' H (\Delta \log \mathbf{y}^*),$$

where H is a symmetric matrix with $H_{ij} = \partial \Pi_i / \partial \log y_j^*$ equal to the semi-elasticity of rents Π_i to the quota on producer j.[11] *When there is a change to only a single quota y_i^*, the effect on output to a second order is*

$$\Delta \log Y \approx \Pi_i \Delta \log y_i^* + \frac{1}{2} \frac{d\Pi_i}{d \log y_i^*} (\Delta \log y_i^*)^2.$$

If tightening a quota leads to an increase in its rents, then rising rents amplify the output losses that result from a large reduction in the quota level. Conversely, if rents fall as a quota is tightened, then nonlinearities partially mitigate the output losses from a large shock. Note that, since output is maximized when all quotas are slack, output is always concave with respect to quota changes around efficiency.

When multiple quotas change, the nonlinear interactions between quotas are captured by the off-diagonal elements of H. If H_{ij} is negative, then relaxing the quota on j lowers rents earned by i. In this case, relaxing the quota on i and j simultaneously increases output by less than the sum of the effect of relaxing the quota on i or j one at a time.

To obtain the matrix H, one can either estimate it directly using data on rents and exogenous variation in $\boldsymbol{y}$ or build it up by explicitly aggregating from microeconomic expenditure shares and elasticities of substitution. The latter procedure is described in Appendix Proposition B1, which shows how to compute H building up from the input-output matrix and microeconomic elasticities of substitution. Computing H, and thus the second-order effects of a quota reform, is relatively tractable because H depends only on expenditure shares and elasticities of substitution. In contrast, the second-order effects of wedge changes in economies with wedges require substantially more information. Since the first-order effects of wedges depend on the entire input-output table, all elasticities of substitution, and all initial distortions, the second-order effects of wedges require all

[11] H is symmetric because $\Pi_i = d \log Y / d \log y_i^*$, so H is the Hessian of log output with respect to quotas. Ultimately, this symmetry is a consequence of the fact that final demand maximizes a homothetic aggregator. If final demand does not maximize a homothetic aggregator, then H needs to be adjusted to account for income (and income distribution) effects (see Baqaee and Burstein 2023 for a discussion along these lines).

this information as well as higher-order derivatives of preferences and technologies (i.e., super-elasticities).

A consequence of Proposition 5 is that, if a quota y_i^* is chosen to maximize the real rents it generates, then the effect of a quota change on output to a second order depends on rents alone.

Proposition 6 (Second-order approximation with a monopolist)**.** *Suppose y_i^* is chosen to maximize real rents (rents divided by the consumer price index), taking all other producers' production technologies and quotas as given. Then, the effect of changes in the y_i^* to a second order are*

$$\Delta \log Y \approx \Pi_i \Delta \log y_i^* - \frac{1}{2}\Pi_i^2 (\Delta \log y_i^*)^2.$$

Note that at the output quantity that maximizes real rents, the derivative of *nominal* rents with respect to output is negative ($d\Pi / d \log y_i^* < 0$), because the monopolist internalizes the effect of restricting its supply on aggregate output. Thus, at the profit-maximizing quantity, output is log concave with respect to changes in quantity. Nonlinearities thereby amplify the losses from further output cuts by the monopolist and moderate the gains from increases in production. The larger the monopolist's rents, the faster the gains from increasing the quota peter out relative to the first-order approximation.

We illustrate Propositions 5 and 6 in two examples that consider the nonlinear effects of a single quota and interactions between multiple quotas.

Example 6 (Nonlinearities in a Small Open Economy)**.** Consider the small open economy from Example 1, and suppose there is an import quota y_f^*. We consider how nonlinearities shape the welfare effects of a change in the import quota. Applying Proposition 5, we find that the response of welfare to a change in the import quota is

$$\Delta \log Y \approx \Pi_f \Delta \log y_f^* + \frac{1}{2}\left[\frac{\Pi_f}{1-\Pi_f} - \frac{1}{\theta}\frac{\lambda_f}{1-\lambda_f}\right]\left(1-\Pi_f\right)^2 \left(\Delta \log y_f^*\right)^2. \tag{5}$$

The first term in (5) reflects the first-order effect of quota changes from Proposition 2. The second term in (5) reflects nonlinearities, which depend on how rents earned by import–export firms respond to quota changes. We compute $d\Pi_f / d \log y_f^*$ for this economy using Appendix Proposition B1, which expresses these semi-elasticities in terms of microeconomic elasticities and expenditure shares.

Note that the sign of this second-order term depends on the initial level of rents, Π_f, as well as the Armington elasticity θ and the initial foreign good share λ_f. Close to efficiency, this term is negative because $\Pi_f \approx 0$, meaning that output is log concave in quota changes: nonlinearities exacerbate the negative effects of tightening the quota and dampen the positive effects of relaxing it. However, away from the efficient point, the second-order term may be positive if $\theta > 1$. A positive second-order term implies that an increase in the import quota increases rents. When this is the case, nonlinearities amplify the benefits of positive shocks and mitigate further losses from negative shocks.

Figure 3 illustrates these results in a numerical example with $\theta > 1$. The left panel shows that rents Π_f follow a hump-shaped "Laffer curve" in the import quota y_f^*. Starting at the point where the quota is just binding (i.e., $d \log y_f^* = 0$), tightening the quota increases rents. But when the quota is sufficiently tight, further tightening it in fact causes rents to fall. This non-monotonic path of rents means that output, shown in the right panel of Figure 3, switches from log concave in the region near the efficient point to log convex in the quota at points sufficiently far from the efficient frontier. Intuitively, if the quota is sufficiently tight, importing has effectively been outlawed, importers earn little rents, and log changes in the quota have little effect. As predicted by Proposition 6, the quota that maximizes real rents (indicated by the dashed red lines in Example 6) is in the region where welfare is log concave.[12]

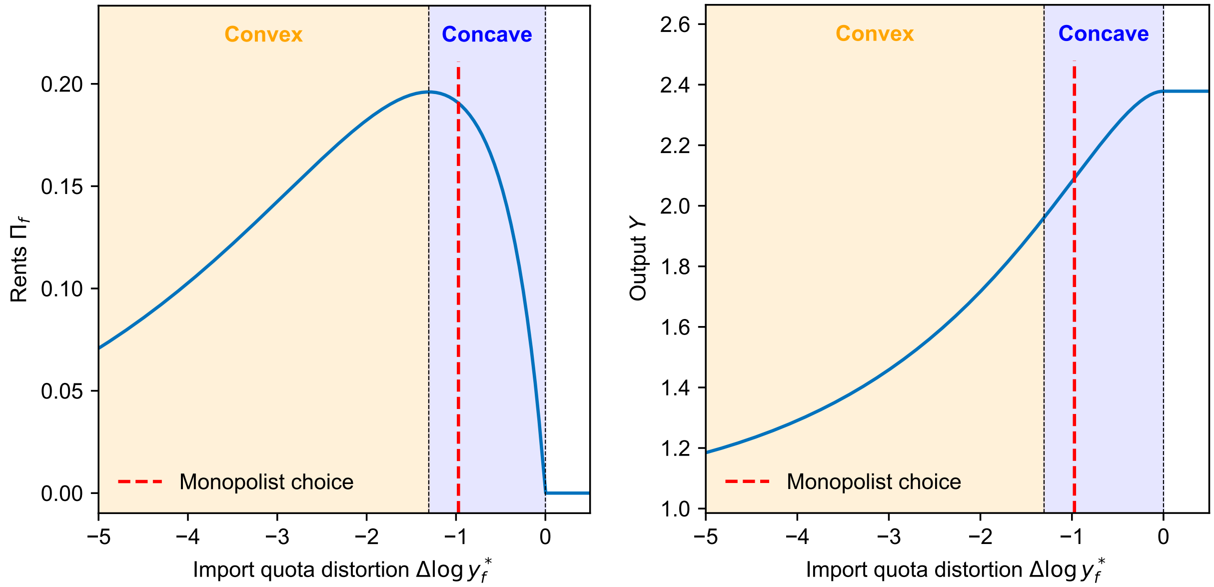
Example 7 (Horizontal Economy with Multiple Quotas)**.** In an economy with multiple quotas, changes in quotas can interact with one another. Consider the horizontal economy from Figure 2a, and suppose there are binding quotas on two firms, y_1^* and y_2^*. Following Proposition 5, the effect on aggregate output is given by

$$\Delta \log Y \approx \underbrace{\Pi_1 \Delta \log y_1^* + \Pi_2 \Delta \log y_2^*}_{\text{First order}} + \underbrace{(1/2)\left(H_{11}\left(\Delta \log y_1^*\right)^2 + H_{22}\left(\Delta \log y_2^*\right)^2 + 2H_{12}\left(\Delta \log y_1^*\right)\left(\Delta \log y_2^*\right)\right)}_{\text{Second order}}.$$

How the interaction between the two quota changes affects output depends on the sign of H_{12}. When H_{12} is positive, relaxing the quota on one firm increases the rents that accrue to the second quota. Thus, relaxing both quotas together amplifies efficiency gains relative to loosening each quota independently. Conversely, when H_{12} is negative, relaxing one quota makes the second quota less binding, and hence reduces the incremental gains that would be achieved from also relaxing the second quota.

[12]Curiously, if the economy is in the convex region, sufficiently far from the efficient point, then random variation in quotas can actually be welfare improving due to convexity. This relates to the debate between Oi (1961) and Samuelson (1972) about the desirability of policy-induced price instability. Samuelson (1972) showed that in efficient equilibria, policy-induced price instability harms welfare. This example shows that this result may not hold once the economy is sufficiently far from the efficient point.

Figure 3: Nonlinearities in a small open economy.



(a) Rents (Π_f).

(b) Output (Y).

Note: Simulation of the small open economy from Example 1 with $A_f = 1$, $\omega = 0.5$, and $\theta = 1.8$. The x-axis shows the log difference between the import quota and the undistorted level of imports. The thick dashed line is the import level chosen by a monopolist import–export firm to maximize real rents.

Using Appendix Proposition B1, we can write H in terms of the firms' sales shares and the household's elasticity of substitution. Using this, we find that H_{12} is positive if

$$\theta < 1 - \frac{(\lambda_1 - \Pi_1)(\lambda_2 - \Pi_2)}{(1 - \lambda_1 - \lambda_2)\Pi_1 \Pi_2}.$$

Two insights emerge. First, when the economy is efficient and $\Pi_1 = \Pi_2 = 0$, H_{12} is always negative, and thus the gains from relaxing both quotas around the efficient point are always lower than the sum of the gains from relaxing each quota individually regardless of the elasticity of substitution. The intuition is that, when both quotas are just binding, tightening the quota on firm 1 pushes more resources to firm 2 and thus makes the existing quota on firm 2 more restrictive. But the effects of positive rents at both firms can be undone by relaxing the quota solely on firm 1—thus the incremental gains from relaxing both quotas is less than the gains from relaxing each quota individually.

Second, when $\Pi_1, \Pi_2 > 0$, a necessary condition for H_{12} to be positive in this economy is that the firms' outputs are complements ($\theta < 1$). Intuitively, when outputs are complements, an increase in the supply of output by firm 1 increases the marginal value of outputs from firm 2. This force amplifies the gains from relaxing the quotas on both firms together compared to relaxing each individually. When θ is sufficiently low and firms have sufficiently high initial rents, this force can lead the net effect of relaxing both quotas together to be greater than each alone.

4.3 Nonlinearities Beyond Second Order

To estimate the effects of large reforms beyond a second-order approximation, Proposition 4 requires not only the elasticities of rents to marginal quota changes, but the quota demand system $\mathbf{\Pi}(\boldsymbol{y}^*(s))$ over the entire region $s \in [0, 1]$. We discuss how one can obtain $\mathbf{\Pi}(\boldsymbol{y}^*)$ by either (i) estimating it directly or (ii) specifying a structural model of the economy.

Estimating $\mathbf{\Pi}(\boldsymbol{y}^*)$ directly. One way to recover $\mathbf{\Pi}(\boldsymbol{y}^*)$ is to specify an empirical model of the quota demand system and estimate it using variation in quotas and rents across markets and over time. This approach parallels demand estimation methods standard in the industrial organization literature and the estimation of general-equilibrium factor demand systems, as in Adao et al. (2017). With the estimated demand system in hand, the nonlinear effects of any counterfactual change in quotas can be computed using Proposition 4. We provide an example of this approach in Section 7.

For ex-post, rather than counterfactual, questions, if we observe changes to a quota and its rents over time, and if we assume rents did not change in response to other shocks (i.e. other shocks are absent or cross elasticities with respect to other shocks are zero), then Proposition 4 can be used directly to estimate the nonlinear effects on output. Specifically, given time-series data on rents and the quota level, we can approximate the integral in (4) using

$$\Delta \log Y \approx \sum_{s=0}^{T-1} \left(\Pi_i(s) + \frac{1}{2} \Delta \Pi_i(s) \right) \Delta \log y_i^*(s), \tag{6}$$

where $s = 0, ..., T$ indexes observations, $\Delta \Pi_i(s) = \Pi_i(s+1) - \Pi_i(s)$, and $\Delta \log y_i^*(s) = \log[y_i^*(s+1)/y_i^*(s)]$. In effect, Proposition 4 integrates the area under the rent curve for a quota, and (6) approximates this area using a Riemann sum.

Estimating $\mathbf{\Pi}(\boldsymbol{y}^*)$ using a structural model. An alternative approach to obtaining $\mathbf{\Pi}(\boldsymbol{y}^*)$ is to use a fully-specified structural model of the economy. In Online Appendix B.3, we show how to numerically evaluate the integral in (4) given a counterfactual shock of interest $\Delta \log \boldsymbol{y}^*$, the initial input-output table, and elasticities of substitution.

One challenge with this approach is that quotas are only occasionally binding and may endogenously transition between slack and binding as we integrate. Rents have a kink at the point where a quota just binds: the right-derivative is zero and the left-derivative is strictly negative. Thus, when a quota is just-binding, it is crucial to evaluate derivatives from the appropriate direction to capture how further quota changes will affect rents.

Our method in Online Appendix B.3 overcomes this challenge by integrating the effects of quota changes along the reform path, updating rents using the structural model, and checking at each step whether quotas switch from slack to binding or vice versa. By keeping track of which quotas are slack and binding, we ensure that we evaluate the derivatives of quota rents from the appropriate direction. In the following example, we illustrate how this procedure can be used to compute the impact of a large reform in an economy with multiple quotas that transition from slack to binding along the reform path.

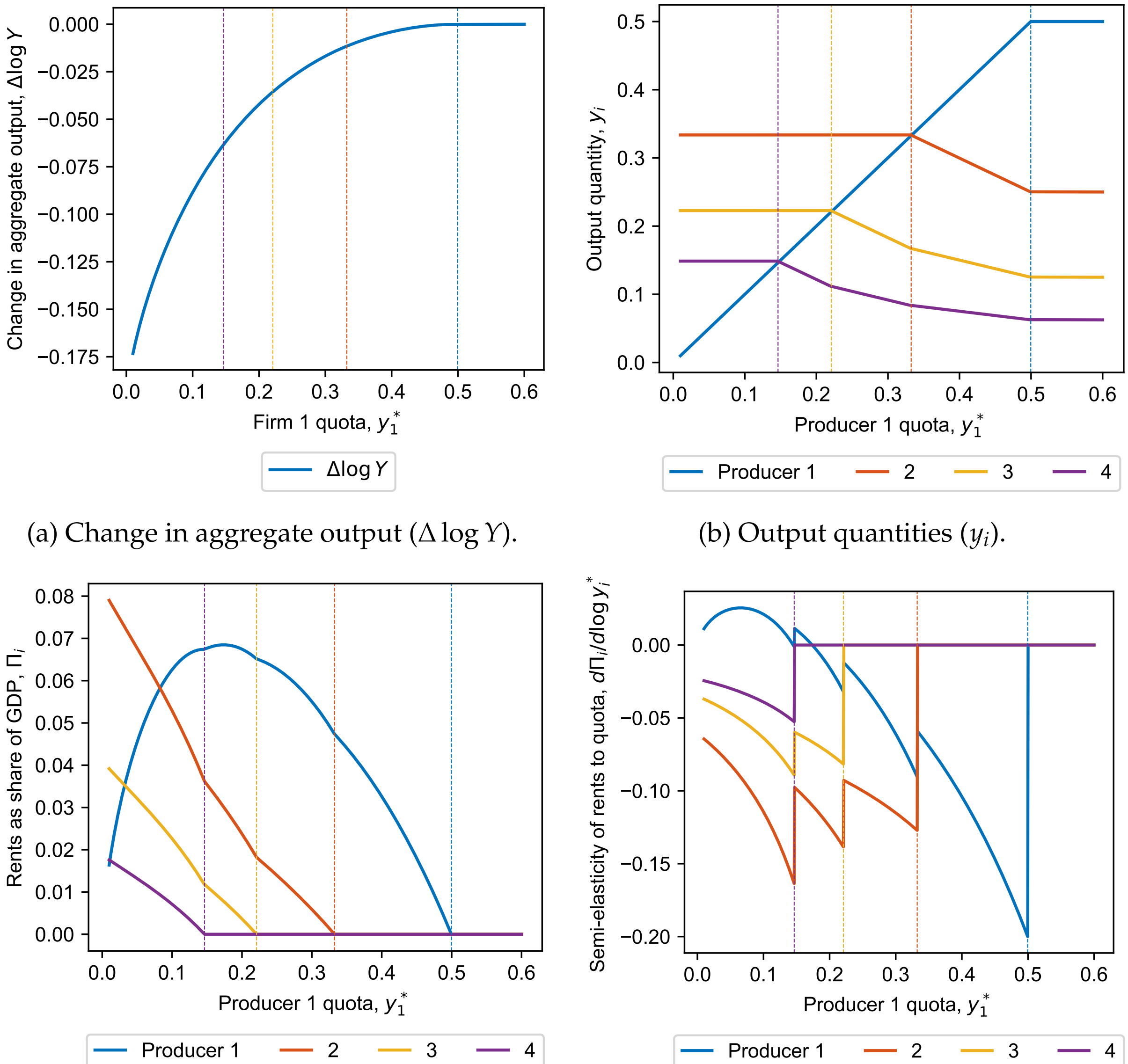
Example 8 (Horizontal economy with multiple quotas)**.** Consider the horizontal economy from Figure 2a. Suppose rents for all producers are initially zero, prices are initially $p_i = 1$ for $i \in \{1, ..., N\}$, and initially,

$$\begin{aligned} y_i &= \frac{1}{2^i} && \text{for } \; i \in \{1, ..., N-1\}, && \text{and} && y_N = \frac{1}{2^{N-1}}, \\ y_i^* &= \frac{1}{(0.75)^{i-1}} y_i && \text{for } \; i \in \{1, ..., N-1\}, && \text{and} && y_N^* > 1. \end{aligned}$$

In words, the undistorted sizes of producers follow a geometric series. The quota for the first producer is just-binding when $y_1 = 0.5$, the quota for producers $2, ..., N-1$ are initially slack, and the quota for producer N never binds.

Figure 4 shows the effect on output from varying the quota on producer 1 between $y_1^* = 0.6$ to $y_1^* = 0.01$. Reading panel (a) from right to left, output is initially unchanged as the quota is varied from 0.5 to 0.6, since all quotas are slack over this range. However, the quota on the first producer becomes binding when $y_1^* = 0.5$, at which point further tightening the quota leads to output losses. As shown in panel (b), when y_1^* is reduced below 0.5, the output quantities of the other producers rise. At the point where each producer's output quantity meets its quota, the quota on that producer becomes binding, and rents increase sharply from zero, as shown in panel (c). Thus, each time a producer's quota transitions from slack to binding, the semi-elasticity of all producers' rents to their quantities change discontinuously (panel (d)). Detecting when each quota binds along the reform path allows us to calculate the effect of the large change in producer 1's quota.

Figure 4: Effects of a large quota change in an economy with multiple quotas.



(a) Change in aggregate output ($\Delta \log Y$).

(b) Output quantities (y_i).

(c) Rents (Π_i).

(d) Rent semi-elasticities ($d\Pi_i/d \log y_i^*$).

Note: Effects of changing the quota on firm 1 from $y_1^* = 0.6$ to $y_1^* = 0.01$ in a horizontal economy with $N = 5$ firms. The vertical dotted lines indicate where the algorithm detects a previously slack quota start to bind.

5 Effects of Quotas in a Dynamic Economy

In this section, we discuss the effects of quota changes in dynamic economies. To apply our results to such economies, we index goods by both commodity and date. In this representation, a quota on good i at date t is a quota on the dated good (i, t). The relevant rents, denoted by $\Pi_i(t)$, are therefore the present discounted value of rents earned by the quota on i at date t, expressed as a share of the net present value of income across all dates (i.e., total wealth).

With this interpretation, our results above apply directly: the effect of changes to the quota on good i in each period, $\{\Delta \log y_i^*(t)\}_{t=0}^{\infty}$, depends on the rents $\Pi_i(t)$ associated with each dated good and how those rents respond to the liberalization. (Note that $\Delta \log y_i^*(t)$ denotes the change in $y_i^*(t)$, the quota level at date t for good i, across counterfactuals. This is not to be confused with differences in the quota level across different points in time.)

If there is an asset market where perpetual rights to produce under the quota can be traded, we can go further. The following proposition shows how to estimate the effect of an announced path of quota changes in a dynamic economy using price of the asset and its reaction to the shock on impact.

Proposition 7 (Effect of a path of quota changes)**.** *Suppose there is an initial steady-state quota $\bar{y}_i$. At time zero, a path of future quotas $\{y_i^*(t)\}_{t=0}^{\infty}$ is announced that differs from the steady state by $\Delta y_i^*(t) = y_i^*(t) - \bar{y}_i = h\epsilon(t)$, where h is scalar controlling the size of the shock. Assume that interest rates follow an exogenous path $r(t)$ and that rents earned by the quota at each time t are only a function of the contemporaneous quota level $y_i^*(t)$.*[13] *Then, the effect of the path of quota changes on output to a second order in h is*

$$\Delta \log Y \approx P_i \frac{\int_0^{\infty} e^{-\int_0^t r(s)ds} \Delta y_i^*(t) dt}{\int_0^{\infty} e^{-\int_0^t r(s)ds} dt} + \frac{1}{2} \Delta P_i \frac{\int_0^{\infty} e^{-\int_0^t r(s)ds} \left(\Delta y_i^*(t)\right)^2 dt}{\int_0^{\infty} e^{-\int_0^t r(s)ds} y_i^*(t) dt},$$

where P_i is the price of a perpetual license to produce under the quota as a share of wealth and ΔP_i is the change in the asset price on impact.[14]

The first term of Proposition 7 reflects the first-order effects of the path of quota changes on output. Intuitively, the effect of the quota change in each period t on contemporaneous income is given by rents earned per unit of the quota times the change in the quota level; the effect on the present discounted value of real income is thus summarized by the asset price of a quota license, which reflects the stream of future rents, and the discount factor-weighted average of future quota changes. Since the first-order effect depends on the asset price of a quota license, the second-order effect depends on how this asset price changes in response to the liberalization, as indicated by asset price change ΔP_i in the second term.

A special case of Proposition 7 is when the liberalization entails a one-time, permanent change in a quota level from the initial steady state $\bar{y}_i$ to a new steady-state level $\bar{y}_i + \Delta y_i^*$,

[13] This condition would be satisfied if, for example, goods cannot be stored across periods.

[14] Continuous time is not crucial for these results. The effect of a path of quota changes in a discrete-time economy can likewise be calculated by replacing the integrals in Proposition 7 with sums.

which we analyze in Corollary 1.

Corollary 1 (One-time, permanent quota changes)**.** *Suppose a one-time, persistent change in the quota y_i^* from initial steady-state level $\bar{y}_i$ to $\bar{y}_i + \Delta y_i^*$ is announced at time zero. Then, the change in output to a second order is*

$$\Delta \log Y \approx P_i \Delta y_i^* + \frac{1}{2} \Delta P_i \Delta y_i^* \approx V_i \Delta \log y_i^* + \frac{1}{2} \Delta V_i \Delta \log y_i^*,$$

where $V_i = P_i y_i^*$ is the market value of permits as a share of wealth.

Corollary 1 shows that the effects of moving from one permanent quota level to another are given by the same expression that we derived in Proposition 5. This also means that the effect of a sequence of unanticipated, permanent quota changes $1, ..., T$ on output are given by cumulating the expression in Corollary 1:

$$\Delta \log Y \approx \sum_{s=0}^{T-1} \left(V_i(s) + \frac{1}{2} \Delta V_i(s) \right) \Delta \log y_i^*(s),$$

where $\Delta V_i(s) = V_i(s+1) - V_i(s)$ is the change in the market value of permits as a share of wealth in reaction to each unanticipated, permanent quota change. This result extends the static case in Equation (6) to characterize the effect of a sequence of unanticipated reforms on output in a dynamic setting. We provide examples of how to use these formulas in Section 7.

6 Distance to the Frontier

In this section, we characterize the misallocation costs of quotas—that is, the output loss relative to the efficient frontier where quota distortions are removed. We provide three nonparametric expressions for the distance to the frontier. These expressions can be used to analyze the effect of relaxing a single quota or relaxing multiple quotas at once.

Proposition 8 (Distance to the frontier)**.** *Let $\mathbf{y}^*$ be a vector of quotas and $\mathbf{y}^{\mathbf{eff}}$ be the vector of output quantities that would result if quotas on producers $i \in \mathcal{I}^*$ were relaxed to the point of being non-binding. Let $\mathbf{\Pi}(\mathbf{y}^*)$ be the vector of producers' rents given quotas $\mathbf{y}^*$, and define the vector of quantity distortions $\Delta \log \mathbf{y}^* = \log \mathbf{y}^* - \log \mathbf{y}^{\mathbf{eff}}$. Let H be the symmetric matrix with $H_{ij} = \partial \Pi_i / \partial \log y_j^*$ equal to the semi-elasticity of rents Π_i to changes in the quota on producer j.*

The output gains from relaxing all quotas $i \in \mathcal{I}$ up to a second order in the quantity distortions

$\Delta \log \mathbf{y}^*$ is

$$\Delta \log Y \approx -\frac{1}{2}\mathbf{\Pi}' \Delta \log \mathbf{y}^*; \tag{7}$$

or

$$\Delta \log Y \approx -\frac{1}{2}\left(\Delta \log \mathbf{y}^*\right)' H \left(\Delta \log \mathbf{y}^*\right); \tag{8}$$

or,

$$\Delta \log Y \approx -\frac{1}{2}\mathbf{\Pi}' H^{-1} \mathbf{\Pi}. \tag{9}$$

For (8) and (9), the matrix H can be evaluated either at the equilibrium with quotas or as a directional derivative from the binding side at the equilibrium where all quotas in $\mathcal{I}^*$ are relaxed.

Equation (7) expresses the distance to the frontier in terms of rents and the size of quantity distortions. When distortions are small, the effect of removing distortions to a second order can be calculated by averaging the first-order effect of changing quotas at the distorted equilibrium, given by Proposition 2, and the first-order effect of changing quotas at the efficient point, which is zero by the envelope theorem.

Alternatively, rents close to the efficient point can also be estimated by specializing the nonlinear effects from Proposition 5 to an economy that is initially efficient. Since rents at the efficient point are zero, the first-order term disappears, and we are left with (8). The matrix H, which captures the response of rents on each quota to changes in other quotas, describes the misallocation cost of a vector of quantity distortions. We note that, for the second-order approximation in (8), these semi-elasticities can be calculated at either the efficient point or at the observed inefficient allocation.

Both expressions in (7) and (8) require knowing the size of quantity distortions $\Delta \log \mathbf{y}^*$, or equivalently, the output quantities that would prevail if there were no quotas. For cases where it is difficult to ascertain the size of quantity distortions, Equation (9) provides a formula for the efficiency gains from removing quotas in terms of observed rents and the inverse of the semi-elasticities matrix H. The intuition for (9) comes from the fact that rents of unconstrained firms are zero. Thus, we can express the efficiency gains from removing quotas in terms of their initial rents and the rate at which rents change as the quotas are relaxed (described by H).

The expressions in Proposition 8 can be used to estimate the efficiency gains from relaxing all or any subset of quotas. To build intuition, Corollary 2 specializes the expressions from Proposition 8 to the case of removing a single quota.

Corollary 2 (Efficiency gains from removing a single quota)**.** *Let Π_i be the rents of producer i, and let $\Delta \log y_i^* = \log y_i^* - \log y_i^{eff}$ be the log-difference between the quota on i and the level of i's output that would obtain without a quota, holding quotas on all other producers fixed. The* *efficiency gains from removing the quota on producer i up to the second order in* $\Delta \log y_i^*$ *can be estimated using any of the three following expressions:*

$$\Delta \log Y \approx -\frac{1}{2}\Pi_i \Delta \log y_i^*. \tag{Option 1}$$

$$\Delta \log Y \approx -\frac{1}{2}\frac{\partial \Pi_i}{\partial \log y_i^*}(\Delta \log y_i^*)^2. \tag{Option 2}$$

$$\Delta \log Y \approx \frac{1}{2}\Pi_i \left[-\frac{d \log \Pi_i}{d \log y_i^*}\right]^{-1} \tag{Option 3}$$

The expressions labeled Options 1–3 in Corollary 2 correspond to the equations (7)–(9) in Proposition 8. The final expression, labeled Option 3, rewrites the efficiency gains from removing a quota in terms of the elasticity of rents with respect to the quota (rather than the semi-elasticity). The efficiency gain is inversely related to the elasticity of rents with respect to the quota because, fixing the level of initial rents, if rents fall quickly as the quota is relaxed, a small change in the quota level is required to take the economy to the unconstrained point. Conversely, if rents fall slowly as the quota is relaxed, the distance to the unconstrained point is large, since it will take a large change in the quota level to restore rents to zero.

The elasticity $d \log \Pi_i / d \log y_i^*$ can also be useful to differentiate empirically between situations where the quota on a producer is close to or far from its unconstrained level of production. If the quota is close to the unconstrained level, the elasticity $d \log \Pi_i / d \log y_i^*$ must be negative, since rents must fall to zero as the level of the quota rises to the point where it is no longer binding. Hence, if the elasticity $d \log \Pi_i / d \log y_i^*$ at an initial equilibrium is positive—i.e., an increase in the quota raises rents—then the economy must be far from the efficient frontier. In this case, the assumption that the quantity distortion is small is violated, and the expressions in Corollary 2 cease to be a reasonable approximation for the efficiency gains. We provide an empirical example of this in Section 7.

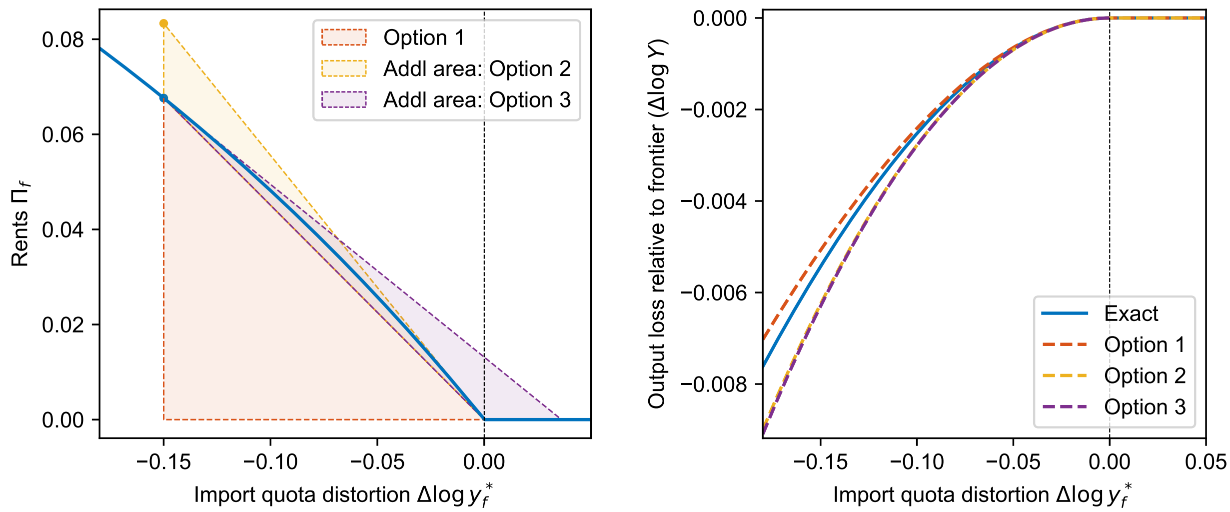
Example 9 (Distance to Frontier in Small Open Economy)**.** We illustrate the welfare gains from removing the import quota y_f^* in the small open economy from Example 1. We apply each of our three expressions for the distance to the frontier in turn.

First, Equation (7) shows that we can estimate the distance to the frontier using the rents of constrained import–export firms and the size of the distortion,

$$\Delta \log Y \approx -\frac{1}{2}\Pi_f \Delta \log y_f^*. \tag{Option 1}$$

Figure 5 illustrates. For a given quantity distortion $\Delta \log y_f^*$, the estimated distance to

Figure 5: Distance to the frontier in a small open economy.



(a) Import–export firms' rents.

(b) Distance to the frontier.

Note: The solid line in panel (a) shows rents as a function of the log difference between the unconstrained level of imports and the quota y_f^*. The three shaded triangles are approximations for the distance to the frontier. Panel (b) shows output as a function of the import quota distortion (solid line), alongside the three second-order approximations (dashed). This example uses $A_f = 1$, $\omega = 0.5$, and $\theta = 1.8$.

the frontier is given by multiplying the quantity distortion by the resulting rents Π_f and dividing by two. This formula approximates the area under the rent function and thus the output gains from moving to the efficient frontier.

Second, Equation (8) replaces the level of rents, Π_f, with the semi-elasticity of rents with respect to the quota times the size of the distortion,

$$\Delta \log Y \approx -\frac{1}{2}\frac{d\Pi_f}{d\log y_f^*}(\Delta \log y_f^*)^2 = \frac{1}{2\theta}\frac{\lambda_f}{1-\lambda_f}(\Delta \log y_f^*)^2. \qquad \text{(Option 2)}$$

The second equality expresses the semi-elasticity of rents with respect to the import quota in terms of households' expenditures on the foreign good, λ_f, and the trade elasticity, θ (using the explicit characterization in Online Appendix B). In Figure 5, this approximation for the distance to the frontier corresponds to estimating rents by extrapolating out from the efficient point where $\Delta \log y_f^* = 0$, and then multiplying those estimated rents by the size of the distortion $\Delta \log y_f^*$ and one-half.

Third, Equation (9) estimates the size of the distortion, $\Delta \log y_f^*$, using the elasticity of import–export firms' rents to import quota changes around the initial, distorted allocation,

$$\Delta \log Y \approx \frac{1}{2}\Pi_f \left[-\frac{d \log \Pi_f}{d \log y_f^*}\right]^{-1} = \frac{\theta}{2}\frac{\Pi_f^2}{1-\Pi_f}\frac{1-\lambda_f}{\lambda_f(1-\Pi_f)-\theta\Pi_f(1-\lambda_f)}. \qquad \text{(Option 3)}$$

The second equality expresses the elasticity of rents to quota changes in terms of the trade elasticity θ, sales of the foreign good λ_f, and rents Π_f. For small Π_f, holding fixed rents Π_f and sales λ_f, the distance to the frontier is increasing in the trade elasticity θ. This is because a higher elasticity implies that a greater change in the import quota level is required to achieve the undistorted allocation (i.e., rents are relatively unresponsive to quota changes).

Figure 5 provides a graphic illustration of Option 3. Starting with a given distortion $\Delta \log y_f^*$, this approximation uses the level of rents Π_f and estimates the size of the distortion $\Delta \log y_f^*$ by extrapolating rents from the inefficient point. As shown in the right panel of Figure 5, this expression, as well as the two alternatives, closely approximates the true distance to the frontier even for substantial import quota distortions.

Example 10 (Horizontal Economy with Multiple Quotas)**.** To illustrate nonlinear interactions between quotas, consider the horizontal economy with quotas y_1^* and y_2^* from Example 7. Applying Proposition 8 shows the efficiency gains from relaxing both quotas y_1^* and y_2^* are

$$\Delta \log Y \approx -\frac{1}{2}\left(\Pi_1^2 H_{11}^{-1} + \Pi_1^2 H_{22}^{-1}\right) - \Pi_1\Pi_2 H_{12}^{-1}.$$

The final term, $-\Pi_1\Pi_2 H_{12}^{-1}$, describes the additional efficiency gain that results from relaxing both quotas together compared to the sum of the efficiency gains realized from relaxing each quota individually. If H_{12}^{-1} is positive, the gains from relaxing one quota partially offset the gains from relaxing the other. On the other hand, if H_{12}^{-1} is negative, relaxing each quota amplifies the additional efficiency gains associated with the other.

Since the matrix H is negative definite at the efficient point, the sign of H_{12}^{-1} near efficiency is given by the sign of $-H_{12} = -\partial \Pi_1 / \partial \log y_2^*$. Since rents at the efficient point are zero, it must always be the case that $H_{12} = \partial \Pi_1 / \partial \log y_2^* \leq 0$. Thus, around the efficient point, relaxing the quota on firm 2 always weakly decreases the rents of firm 1, H_{12}^{-1} is weakly positive, and the gains from relaxing the two quotas must always (weakly) offset each other.

7 Empirical Applications

We demonstrate how to apply our results in several empirical examples. The first two empirical examples, which consider the cap on H-1B visas and zoning restrictions on single-family housing, illustrate how to apply our results on the first-order effects of quota changes from Section 3. The following three examples, on taxicab medallions in New York City, U.S. quotas on Chinese textile and clothing exports, and Argentina's capital controls, each illustrate various results on nonlinearities and the distance to the frontier from Sections 4–6.

7.1 H-1B Visa Quota

The H-1B visa allows U.S. firms to employ high-skill foreign workers. Since the mid-2000s, the total number of visas issued has been capped at 85,000, with 20,000 of the slots reserved for immigrants holding a master's or higher degree from a U.S. university. We can use our results to estimate the efficiency gains that would result from relaxing the cap on H-1B visa quotas.

Our measure of the rents that accrue to winners of the H-1B visa lottery comes from Clemens (2013), who compares earnings of winners and losers of the 2007 H-1B lottery within a pool of Indian software workers employed at the same firm. In 2007, the U.S. government received more applications than needed to fill the H-1B quota within the first two days of the application window and chose which H-1B visa applications to process by random lottery. Earnings for workers whose applications were processed—those who won the lottery—were \$12,641 higher two years after the lottery than their colleagues who lost the lottery (\$18,823 in 2025 dollars). This difference in earnings is an intention-to-treat estimate: it captures the rise in earnings for workers allocated the right to immigrate under the H-1B program, regardless of whether the worker ultimately decided to immigrate.

If we assume that software workers are paid their marginal product, then the first-order efficiency gains from expanding the H-1B cap can be computed from this statistic alone. (If workers are not paid their marginal product and some of the rents from H-1B visa awards accrue to employers, rather than employees, then these estimates are a lower bound for the gains from increasing the cap.) We apply Proposition 2 to get

$$d\log Y = \Pi_i d\log y_i^* \approx \frac{\Pi_i}{y_i^*} dy_i^*.$$

That is, the efficiency gain in dollars from increasing the H-1B cap by one slot is equal to the per-person rents of visa holders today. This means that for example, doubling the number of available visas in 2007 would have increased world output by \$1.60B in 2025 dollars.

This figure reflects efficiency gains in *world* output from increasing the number of H-1B visas. It does not include reallocations in output from the rest of the world to the U.S., e.g. from moving workers to the U.S. from other countries. Assuming all other distortions take the form of quotas and are held fixed, the additional increase in U.S. output—and commensurate reduction in output in the rest of the world—from moving workers to the U.S. is equal to the workers' earnings minus the rents they receive, $85{,}000 \times \$46{,}450 \approx \$3.95$B in 2025 dollars (using earnings of lottery losers from Clemens 2013).

7.2 Zoning Restrictions on Single-Family Housing

Next, consider the potential efficiency gains from relaxing zoning restrictions on single-family housing across U.S. cities. To estimate the rents that accrue to zoning restrictions, we use data on "zoning taxes" for 24 metropolitan statistical areas (MSAs) from Gyourko and Krimmel (2021). They measure these zoning taxes by comparing land prices for vacant parcels purchased to build new single-family housing units—which include the rights to supply single-family housing—with land prices for nearby parcels that have existing single-family homes. This comparison isolates the value of permits to build a new single-family housing unit from the value of the land itself.

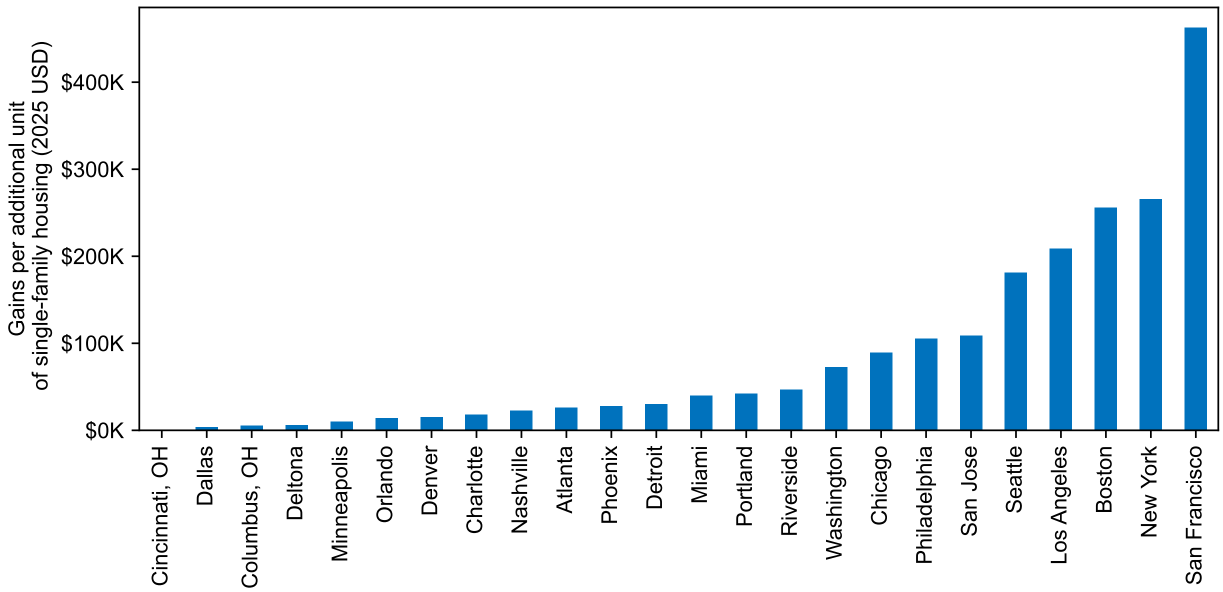
Figure 6 shows the estimated gains associated with relaxing zoning restrictions to increase the supply of single-family housing in each MSA.[15] Supplying an additional unit of single-family housing is associated with efficiency gains of over \$450,000 in San Francisco, and over \$200,000 in other coastal cities like New York, Boston, and Los Angeles.

Policymakers often state housing policies in terms of the number of permits they plan to make available, as these permits directly control the supply of housing in zoning-constrained cities.[16] Modeling zoning restrictions as quantity distortions allows one to map these proposals to expand the supply of housing permits directly into efficiency gains. Moreover, modeling zoning restrictions as quotas has the advantage of requiring less information than modeling them as wedge distortions. Using the wedge approach, we would need to estimate the reduction in zoning wedges necessary to achieve a target increase in housing, which depends on underlying elasticities of supply and demand for

[15] Gyourko and Krimmel (2021) observe several vacant parcel sales in each MSA. To estimate zoning rents per unit of single-family housing in each MSA, we use the median of estimated zoning taxes per quarter acre in each MSA and divide this estimate by the median acreage of single-family homes in the MSA.

[16] For example, California state mandates require that San Francisco approve the creation of 82,000 new housing units by 2031. See `https://www.sfchronicle.com/projects/2023/san-francisco-housing/`.

Figure 6: Gains from expanding the supply of single-family housing across U.S. cities.



Note: The figure shows efficiency gains from increasing the supply of single-family housing by one unit. The estimates apply Proposition 2, using data on zoning taxes from Gyourko and Krimmel (2021).

housing across U.S. cities. In contrast, Proposition 2 allows us to directly use proposed quantity changes without having to map quantities to wedges and back.

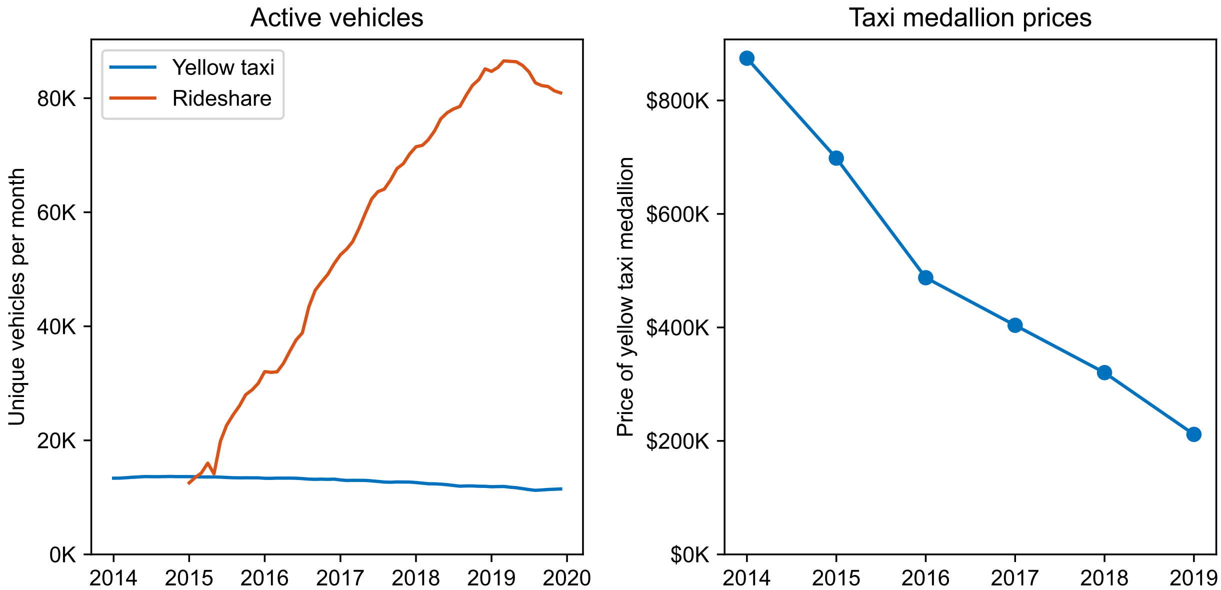
7.3 Taxicab Medallions in New York City

Our next empirical example studies the efficiency costs of the taxicab medallion system in New York City. Taxicab medallions are required to operate a taxi; the city of New York created the taxicab medallion system in 1937 to restrict the total supply of taxicabs. We exploit the growth of rideshare apps such as Uber and Lyft in New York to estimate the efficiency gains from relaxing these restrictions on the supply of taxis.

The first panel of Figure 7 shows how the number of taxi and rideshare vehicles in New York from 2014 to 2019. The number of unique taxis active each month has stayed around 13,000, just under the total 13,587 taxi medallions available from the New York Taxi and Limousine Commission. The number of rideshare vehicles, on the other hand, grew nearly sevenfold from about 12,500 in January 2015 to over 85,000 by mid-2019. During this time, the transaction prices of taxi medallions also fell dramatically, from nearly $1 million dollars at its peak in 2014 to $200,000 in 2019.[17]

[17]Similar trends unfolded in other U.S. cities when rideshare apps entered the market. For example, medallion prices in both Boston and Chicago dropped 30–40 percent from 2015 to 2016. See `https:`

Figure 7: Changes in New York taxi market from 2014–2019.



Note: Monthly unique vehicles are from aggregated reports from the NYC Taxi and Limousine Commission. Taxi medallion prices are annual averages of prices for medallion transfers, from the NYC Taxi and Limousine Commission.

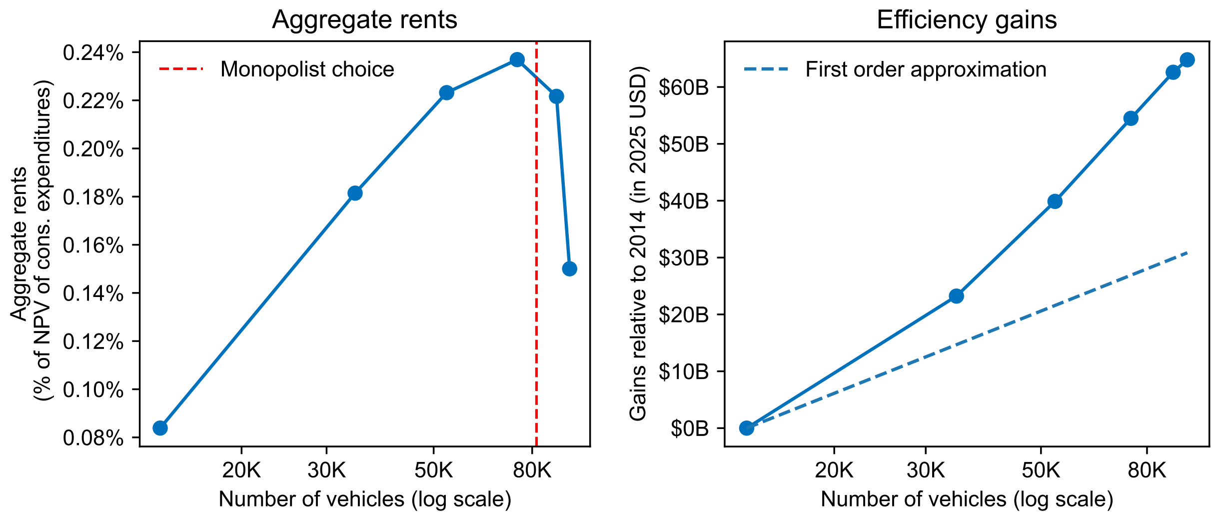
We use how taxi medallion prices fall with the entry of rideshare vehicles to estimate the output gains from relaxing the quota on the number of taxis in New York. For this exercise, we make two assumptions. First, we assume that ride-sharing services are a perfect substitute to taxis, and hence the introduction of ride-sharing services is equivalent to relaxing the quota on the number of vehicles in the market. Second, we assume that taxi medallion prices reflect the discounted value of all future rents accruing to owners of vehicles that are approved to provide rides in New York.[18]

The left panel of Figure 8 shows the aggregate value of rents accruing to taxis and rideshare vehicles as the number of vehicles increased from 2014 to 2019. The number of vehicles in the market was initially so low that initial increases in the number of vehicles in fact increased the aggregate rents earned by these vehicles. Since initially $d\Pi_i/d\log y_i^* > 0$, the market was in the region where output is log convex with respect to quota changes (as seen from Proposition 5). Moreover, the fact that aggregate rents rose

//www.foxnews.com/opinion/are-taxi-medallions-too-big-to-fail-too.

[18]We find similar results if we instead calculate taxicab drivers' excess profits using the change in taxis' revenues as Uber and Lyft entered the market. From 2014 to 2019, revenues per taxi fell by about $40,000 annually, while the change in taxi medallion prices over this period corresponds to a decline in annual rents per taxi of about $37,000. The advantage of using medallion prices is that they isolate changes in rents expected to accrue to medallion owners from other changes in costs that affect revenues and profits.

Figure 8: Rents and efficiency gains in the New York taxi market from 2014–2019.



Note: Aggregate rents are medallion transaction prices times the number of active vehicles. Rents shown as a share of the NPV of consumer expenditures, calculated using BLS Consumer Expenditure Surveys Northeast MSA statistics with a 4% discount rate. The dashed red line in the left panel marks the number of vehicles estimated to maximize real rents. Efficiency gains in the right panel are calculated by cumulating (10) and are expressed in 2025 USD.

as the quota was relaxed means that the initial number of medallions was below the level that a monopolist would have chosen. Using Proposition 6, we estimate that a monopolist would maximize real rents with around 82,000 vehicles, six times higher than the initial number of medallions.[19]

Under the assumption that changes in the number of vehicles reflect unanticipated, permanent shocks each year, we can approximate the efficiency gains from relaxing the medallion quota to a second order in each year using Corollary 1,

$$\Delta \log Y_t \approx \left(V_{it} + \frac{1}{2}\Delta V_{it}\right)\Delta \log y^*_{it}. \tag{10}$$

As shown in Figure 8, these gains are largest in 2014 and 2015 as ride-share vehicles first enter the market, and by 2019 cumulate to nearly $65B in efficiency gains.[20] The first column of Table 1 shows that these gains translate into $7,237 per household in the New York City metro area, or 2.6% of the present value of current and future household

[19]To estimate the rent-maximizing number of vehicles, we fit a third-degree polynomial to rents as a function of number of vehicles and find the level of vehicles at which $d\Pi_i/d\log y^*_i = -\Pi_i$.

[20]Medallion prices were rising up until 2014, when Uber and Lyft entered the market, suggesting that market participants were not anticipating their entry prior to 2014.

Table 1: Estimated efficiency gains from relaxing capacity constraint on New York taxis.

	Change from 2014–2019 Unanticipated	Change from 2014–2019 Anticipated	Distance to frontier
Output gains	\$64.8B	\$81.2B	\$2.3B
Gains per New York MSA household	\$7,237	\$9,066	\$253
% of NPV of transportation expenditures	2.58%	3.23%	0.09%

Note: New York MSA consumer units and transportation expenditures are from the BLS Consumer Expenditure Surveys 2013–2014 northeast MSA statistics. The net present value of transportation expenditures is calculated using annual transportation expenditures in 2013–2014 and a 4% discount rate. All dollar values are in 2025 USD, converted from 2014 USD using the GDP implicit price deflator from the BEA.

transportation expenditures.

Note that because rents initially rose with the quota level, nonlinearities amplify the efficiency gains from liberalization relative to what we would estimate using a first-order approximation. Indeed, using the initial quota rents in 2014 and the log change in vehicles from 2014 to 2019 to calculate a first-order approximation yields estimated efficiency gains of \$31B, or less than half of the efficiency gains that we estimate when we account for nonlinearities in (10).

If we instead assume that the liberalization of the taxicab market was anticipated starting with the entry of Uber and Lyft in 2014, the efficiency gains are given by Proposition 7. Using the change in medallion prices from 2014 to 2015 as an estimate of the asset price reaction to the announced path of quota changes, we estimate efficiency gains of \$81.2B, or \$9,066 per household in the New York City metro area.[21]

Of course, even in 2019, the market is not efficient, since the supply of vehicles is determined by the number of medallions and by imperfectly-competitive ride-share companies. We use Equation (9) from Proposition 8 to estimate the distance to the frontier in 2019, using the level of aggregate rents in 2019 and the elasticity of aggregate rents to changes in the number of vehicles from 2018 to 2019.[22] The final column of Table 1 shows that the remaining distance to the frontier is small compared to the efficiency gains achieved from 2014 to 2019. In particular, increasing the number of vehicles to the efficient level would only add a further \$253 in gains per household in the New York MSA.

[21] Assuming a 4% discount rate, our estimates correspond to an annual gain of \$2.6–3.2B 2025 USD. These gains are a similar order of magnitude to estimates from Cohen et al. (2016), who use estimates of consumer price elasticity along the demand curve for Uber rides to calculate an annual consumer surplus of \$2.88B in 2015 USD—or \$3.8B in 2025 USD—from Uber across New York, Los Angeles, Chicago, and San Francisco.

[22] The rent share fell 39 log points from 2018 to 2019, as the number of vehicles increased by 6.3 log points. Dividing one by the other gives us an elasticity of -6.3.

7.4 U.S. Quotas on Chinese Textile & Clothing Exports

We now illustrate how nonlinear interactions between multiple quotas affects efficiency gains. We use the phase-out of textile and clothing quotas under the World Trade Organization (WTO) Agreement on Textile and Clothing (ATC). From 1975 to 1994, the Multi-Fiber Agreement (MFA) imposed quotas on exports of textiles and clothing from developing countries to the U.S. and European Union. These quotas were particularly binding on China—whose textile and clothing exports to the U.S. rose dramatically when these quotas were relaxed (Dean 1990). As part of the WTO's Uruguay Round, the Agreement on Textile and Clothing (ATC) introduced a plan for phasing out these quotas over the period from 1995 to 2005.

The removal of quotas on textile and clothing goods in phases over this period allows us to study the interactions between sets of quotas. We focus in particular on quotas on China, and on the interaction between the quotas that were lifted as part of Phase III of the ATC in 2002 and Phase IV of the ATC in 2005.[23] Goods with quotas lifted in Phase III included knit fabrics, gloves, dressing gowns, brassieres, and textile luggage products; while a broader set of quotas on silk, wool, and cotton textiles, carpets, and most apparel categories were not lifted until Phase IV in 2005.

To study the interaction between the quotas relaxed in Phase III and Phase IV, we estimate the matrix of semi-elasticities of rents to quota changes, H. We proceed in two steps. First, we estimate the size of the initial quota distortions on each set of goods using the response of exports to the phase-out of each set of quotas. Second, we combine these estimates with data on initial quota rents to back out the semi-elasticities in H.

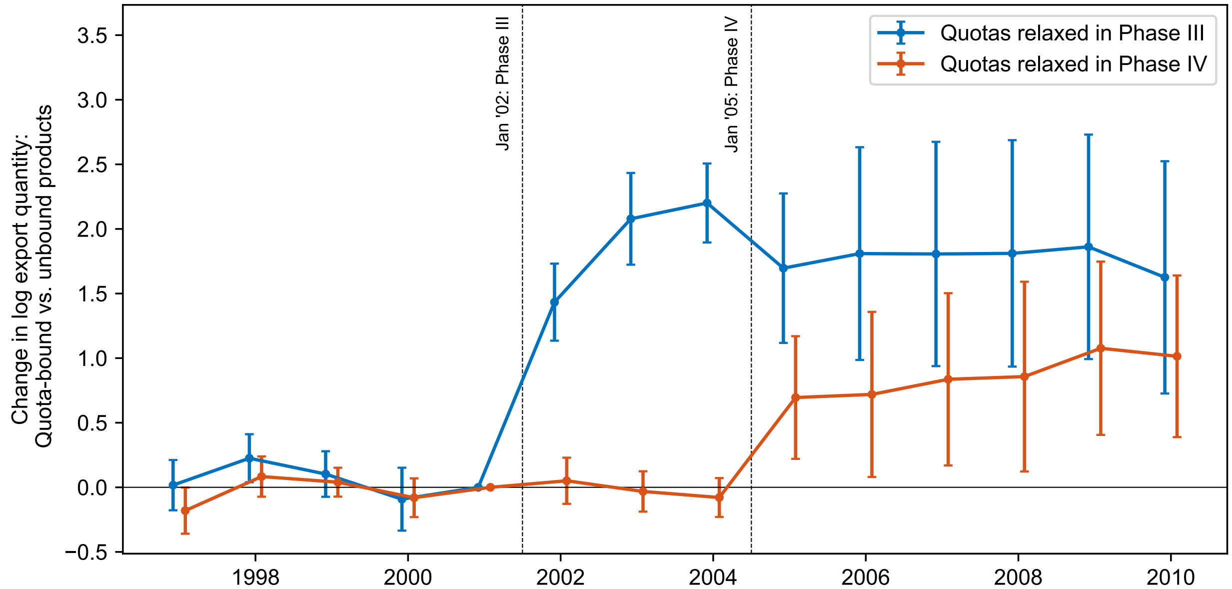
We first estimate the effects of the Phase III and Phase IV quota removals on exports using the specification,

$$\begin{aligned}\log y_{ict} &= \beta_t^{\text{Phase III}}\Big(\text{Binding}_c \times 1\{c \text{ quota relaxed in Phase III}\} \times 1\{\text{year} = t\}\Big) \\ &+ \beta_t^{\text{Phase IV}}\Big(\text{Binding}_c \times 1\{c \text{ quota relaxed in Phase IV}\} \times 1\{\text{year} = t\}\Big) + \alpha_t + \delta_i + \varepsilon_{ict}, \quad (11)\end{aligned}$$

where y_{ict} is the quantity of exports of HS-10 code i in category c from China to the U.S. in year t, Binding_c indicates whether the quota on category c was initially binding, and α_t and δ_i are year and HS-10 code fixed effects. For each group of goods, the β_t coefficients

[23] Although the ATC officially required quotas to be removed in four phases from 1995 to 2005, the structure of the agreement allowed the U.S. (and the E.U.) to defer the removal of most binding quotas until the final two phases of the agreement. During Phase I (1995) and Phase II (1998), the U.S. strategically liberalized non-binding quotas or low-restriction categories; the real impact of the ATC materialized in Phase III (2002) and Phase IV (2005), when the U.S. began lifting quotas that had been actively constraining trade (Chiron 2004).

Figure 9: Differential changes in export quantity for products with initially binding quotas.



Note: The blue and red lines plot estimates for $\beta_t^{\text{Phase III}}$ and $\beta_t^{\text{Phase IV}}$, respectively, from specification (11). The sample includes 14,975 observations across 1,931 HS-10 codes. Standard errors are two-way clustered by category and year. Error bars indicate 95 percent confidence intervals.

estimate the change in export quantities for goods with initially binding quotas relative to other goods also included in the ATC whose quotas were non-binding. Our identifying assumption is that other factors that affect export quantities for products with initially binding quotas relative to other clothing and textile products with non-binding quotas are uncorrelated with the timing of the MFA phase-out.

We estimate (11) using data on Chinese exports of all clothing and textile goods to the U.S. at the HS-10 level from the Office of Textiles and Apparel (OTEXA) and data on quota fillrates from the U.S. MFA/ATC database created by Brambilla et al. (2010). Following Brambilla et al. (2010), we define a quota as binding if the fill rate (i.e., realized exports as a percent of the quota allowance) exceeds 90 percent.

Figure 9 plots the estimated coefficients for $\beta_t^{\text{Phase III}}$ and $\beta_t^{\text{Phase IV}}$ from specification (11). Phase III of the ATC in 2002 led to a large increase in exports for products whose quotas expired in 2002. Exports for HS-10 codes in the Phase III group with initially binding quotas rose by more than 180 log points from 2002–2004 relative to products with non-binding quotas. The final Phase IV of the ATC in 2005 led to a small decline in exports for Phase III group products relative to 2002–2004, and an 80 log point rise in exports for

HS-10 codes in the Phase IV group.[24]

We combine these estimates with data on quota annual license prices to estimate the matrix of semi-elasticities of rents to quota changes.[25] We measure the initial aggregate rents of quota holders for Phase III and Phase IV products by multiplying quota license prices in 2001 by the quantity of exports in those product categories in 2001. Assuming that rents for Phase III and Phase IV group products go to zero when quotas are relaxed, we then solve for the matrix H by solving the following system of equations:

$$\begin{aligned}
\Pi_{\text{Phase III}} &= \beta^{\text{Phase III}}_{\text{2003-2004}} H_{11}, \\
\Pi_{\text{Phase III}} &= \beta^{\text{Phase III}}_{\text{2006-2007}} H_{11} + \beta^{\text{Phase IV}}_{\text{2006-2007}} H_{12}, \\
\Pi_{\text{Phase IV}} &= \beta^{\text{Phase III}}_{\text{2006-2007}} H_{12} + \beta^{\text{Phase IV}}_{\text{2006-2007}} H_{22}.
\end{aligned}$$

where $\beta^{x}_{\text{2003-2004}}$ and $\beta^{x}_{\text{2006-2007}}$ are two-year averages of the effect of the quota phase-out on export quantities for goods in group x. First, the increase in export quantities for Phase III products after 2002 identifies the semi-elasticity of rents for Phase III products to their quotas, H_{11}, holding the Phase IV quotas fixed.[26] Second, the change in exports of Phase III products after 2005 allows us to estimate the cross-product elasticity H_{12}. Finally, since the symmetry of H guarantees $H_{12} = H_{21}$, we can estimate the semi-elasticity of rents for Phase IV products to their quotas, H_{22}. Solving this system of equation yields

$$\mathbf{\Pi} = \begin{bmatrix} \Pi_{\text{Phase III}} \\ \Pi_{\text{Phase IV}} \end{bmatrix} = \begin{bmatrix} \$520\text{M} \\ \$1583\text{M} \end{bmatrix}, \qquad \frac{d \log \mathbf{\Pi}}{d \log \mathbf{y}^*} = \begin{bmatrix} -0.472 & -0.200 \\ -0.066 & -1.149 \end{bmatrix}$$

Note that the off-diagonal entries H_{12} and H_{21} are negative. The negative cross-term is identified by the decline in export quantities for Phase III products when Phase IV quotas were lifted in 2005. As discussed in Example 10, $H_{12} < 0$ implies $H^{-1}_{12} > 0$ and that the gains from relaxing both the Phase III and Phase IV quotas together are smaller than the sum of the gains from relaxing each subset of quotas individually.

The magnitude of this interaction is quantified in Table 2, which estimates the efficiency

[24] U.S. textile and clothing industry groups lobbied for new quotas on a subset of categories after 2005, though the new quotas were in most cases substantially higher than the expiring ATC quotas. We find similar results if we exclude products that had new quotas imposed after 2005 from our estimation.

[25] We are grateful to Amit Khandelwal and Judith Dean for sharing data on these quota prices, which were originally scraped from chinaquota.com. Chinese firms were required to buy these licenses each year to export to countries under the MFA.

[26] While the phases of the ATC technically required changes in Phase IV products' quota levels even before the quotas were completely relaxed in 2005, we assume that Phase IV quotas were held fixed since our estimates of $\beta^{\text{Phase IV}}_t$ for $t \in \{2002, 2003, 2004\}$ are not significantly different from zero. Instead specifying $\Pi_{\text{Phase III}} = \beta^{\text{Phase III}}_{\text{2003-2004}} H_{11} + \beta^{\text{Phase IV}}_{\text{2003-2004}} H_{12}$ does not meaningfully alter our results.

Table 2: Gains from relaxing textile/clothing quotas on Chinese exports to the U.S.

Intervention	Annual efficiency gains (millions of 2025 USD)
(A) Relaxing Phase III quotas only	\$565
(B) Relaxing Phase IV quotas only	\$706
(C) Relaxing both Phase III and IV quotas	\$1,075
Difference: C − (A + B)	\$196

gains from either relaxing the quotas either individually or jointly using Equation (9) from Proposition 8. Starting from the quota levels in 2001, we estimate that relaxing either the Phase III or Phase IV quotas alone would have increased efficiency by \$565 and \$706 million, respectively. Relaxing all quotas together raises efficiency by \$1,075 million—about \$196 million less than the sum of the gains from relaxing each set of quotas in isolation.

A key advantage of estimating the matrix H is that it allows us to evaluate the effects of unobserved, counterfactual reforms without fully specifying a structural model. For instance, we can estimate the efficiency gains from the removal of Phase IV quotas while holding the Phase III quotas in place. More broadly, working with quotas allows us to quantify the costs of protectionist policy and gains from free trade without specifying a large-scale general equilibrium model or relying on assumptions about trade elasticities and other structural parameters.

7.5 Argentina's Capital Controls

In our final example, we use Proposition 8 to estimate the distance to the frontier in the context of restrictions on capital outflows imposed by Argentina. On September 1, 2019, the Argentine government reimposed capital controls following a four-year period with no restrictions on capital flows. The restrictions initially limited U.S. dollar purchases by individuals to \$10,000 per month and imposed tighter controls on corporate access to foreign exchange. Following this imposition of capital controls, capital outflows fell from an average of \$7.2B per month in the free market period to under \$1.5B.

We use two approaches to estimate the efficiency losses due to these quotas on capital outflows. The first approach applies Option 1 from Corollary 2, which expresses the distance to the frontier in terms of the rents accruing to quota holders and the size of the distortion. In the context of Argentina, transactions that are permitted under the capital outflow restrictions typically exchange Argentine pesos for dollars at the official exchange rate, which grants pesos a substantial premium relative to other market exchange rates.[27] Assuming that currency exchange in the black market is unconstrained, we can measure the rents of quota holders permitted to make transactions at the official rate using the gap between the official and black market exchange rates, $\Pi_i = (\log e/\bar{e})\, y_i^*$, where e and $\bar{e}$ are the black market and official Argentina peso–USD exchange rates and y_i^* is the allowed quantity of capital outflows. Thus, Option 1 becomes

$$\Delta \log Y \approx -\frac{1}{2}\Pi_i \Delta \log y_i^* \approx -\frac{1}{2}(\log e/\bar{e})\, \Delta y_i^*.$$

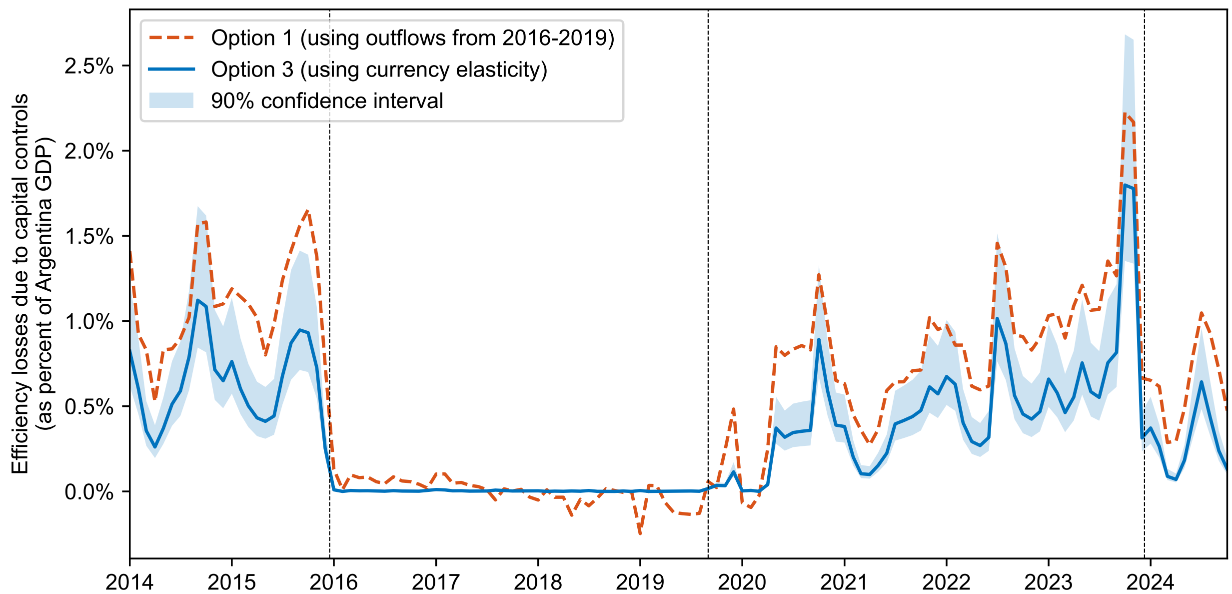
The dashed line in Figure 10 plots the distance to the frontier estimated using Option 1. We use the most popular black market exchange rate, known as the "Dólar Blue," to measure the rents earned by quota holders with the license to exchange pesos at the official exchange rate. We measure the size of the distortion Δy_i^* as the difference between the (restricted) level of capital outflows and the average level of outflows during the period without capital controls from January 2016 to August 2019. Since the reinstatement of capital controls in September 2019, the estimated efficiency losses due to capital controls average 0.8 percent of Argentina's GDP and reach a high of 2.2 percent of GDP just before the devaluation of the peso in late 2023.[28]

A disadvantage of this first approach is the strict assumption that the efficient level of capital outflows during the period with capital controls is equal to the observed level of outflows during the period without controls. Our second approach instead uses Option 3 to back out the size of the distortion using the level of rents and the responsiveness of rents to outflows. For these restrictions on capital outflows, we need to know the responsiveness of rents to outflows—that is, how additional outflows would change the official exchange

[27] Under Argentina's capital controls, there are multiple regulated channels for converting pesos to U.S. dollars, some of which involve exchanges at different rates than the official rate. For example, the *contado con liquidación* (CCL) and *dólar MEP* channels, which involve buying and selling securities to obtain dollars, trade at an exchange rate above the official rate but below black-market rates, and the *dólar soja* grants higher-than-official exchange rates to soybean exporters. The Argentine central bank's (BRCA) monthly reports aggregate all regulated transactions using the official exchange rate, so we use the official rate for our calculations. For the period from December 23, 2019 to December 22, 2024, we adjust the official exchange rate to account for the fact that transactions at the official rate were subject to the Impuesto PAIS tax. Not adjusting for this tax would lead us to overestimate rents from exchanging at the official rate and overestimate the distance to the frontier.

[28] During the period where foreign exchange transactions were subject to both quantity constraints and additional taxes, our estimates reflect the efficiency gains that would be realized if the combination of quota and wedge distortions were relaxed to obtain the undistorted level of outflows. Provided our assumption that the rest of the economy is constrained efficient, our expressions for the distance to the efficient frontier apply whether the distortion in capital outflows is the result of a quota, a wedge, or both, since eliminating these distortions entirely leads to the same, efficient allocation regardless of the form of the distortion.

Figure 10: Estimated efficiency losses due to Argentina's capital controls.



Note: The three vertical lines correspond to the end of capital controls on Dec. 17, 2015, the reinstatement of capital controls on Sep. 1, 2019, and the devaluation of the peso on Dec. 10, 2023. The wedge between market and official Argentine exchange rates is calculated using the Dólar Blue and official exchange rates from Refinitiv. Option 1 calculates the size of the distortion as the difference in monthly capital outflows relative to the average from Jan 2016 to Sep 2019, using data from the Central Bank of Argentina (BCRA). Option 3 applies a currency elasticity of $\theta = 1.692$ (se: 0.358) from Adler et al. (2019).

rate and thus shrink the gap between the black market and official exchange rates.

A common statistic used to summarize the responsiveness of exchange rates to outflows is the depreciation in nominal exchange rates caused by purchases of foreign currency equal to one percent of GDP (Blanchard et al. 2015; Adler et al. 2019). Denoting the *currency elasticity* of nominal exchange rates to outflows as a share of GDP by θ, we can express the distance to the efficient level of capital outflows as

$$\Delta y_i^* = \frac{1}{\theta}\,\text{GDP}\,(\log e/\bar{e})\,.$$

Lower values of θ imply a greater size of distortion, since more capital outflows would be required to close the gap between the black market exchange rate e and the official exchange rate $\bar{e}$.

Combining this expression with the previous, we can express the efficiency losses due to capital controls as a share of Argentina's GDP in terms of the currency elasticity θ and the gap between market and official exchange rates,

$$\frac{\Delta Y}{\mathrm{GDP}} \approx -\frac{1}{2}\frac{1}{\theta}\left(\log e/\bar{e}\right)^2.$$

The distance to the frontier is greater when the current elasticity θ is low. The distance to the frontier also scales quadratically in the gap between black market and official exchange rates, because a higher gap implies both higher rents per dollar of capital flow and implies a greater quantity distortion relative to the frontier.

The solid line in Figure 10 plots the distance to the efficient frontier over time using this second approach, applying estimates of the currency elasticity $\theta = 1.692$ from Adler et al. (2019).[29] The efficiency losses due to capital controls estimated using this approach are broadly similar to the estimates of the distance to the frontier from Option 1. The estimates also indicate that changes since late 2023 have substantially lowered the distance to the frontier. A sharp devaluation of the peso on December 13, 2023 lowered the efficiency losses to below 0.5 percent of GDP. Growing confidence in the peso in late 2024 also narrowed the gap between the black market and official exchange rates, despite the fact that permitted capital outflows have remained low, bringing the distance to the frontier to under 0.2 percent of GDP in October 2024.

8 Extensions

This section describes extensions of our framework developed in the Online Appendix.

Hybrid economies. In the main text, we focus on economies that exhibit either quota distortions or wedge distortions, but not both. In "hybrid economies" that feature both quota and wedge distortions, the effect of quota changes depends on both quota rents, as in Proposition 2, and the endogenous response of quantities for producers with wedge distortions, as in Proposition 3.

Proposition 9 (First-order effects in hybrid economies)**.** *Consider an economy in which some producers have output quotas and others have output wedges. For any producer i with an output*

[29] Adler et al. (2019) estimate that outflows equal to one percent of GDP lead to 1.5–2.0 percent depreciation in nominal exchange rates. For Argentina, these estimates imply that \$1B of outflows in 2023 results in a depreciation in the Argentine peso by 0.26%. These estimates align with other available estimates: for example, using exogenous global capital flow shocks, Blanchard et al. (2015) estimate that outflows equal to one percent of GDP lead to a 1.5% depreciation in nominal exchange rates. Estimates of the impact of order flows on currency markets are also quantitatively similar. For example, Evans and Lyons (2002) find that \$1B of net purchases in 1996 leads to an 0.54% appreciation (or, converting to 2023 dollars, \$1B in 2023 USD outflows leads to a currency depreciation of 0.30%).

quota, the effect of a change in i's quota on output is

$$\frac{d \log Y}{d \log y_i^*} = \Pi_i + \sum_{j \in \mathcal{W}} \Pi_j \frac{d \log y_j}{d \log y_i^*},$$

where $\mathcal{W}$ is the set of producers with output wedges.

When the set of producers with output wedges is empty, when all producers with wedges have zero profits (i.e., the rest of the economy is constrained efficient), or when the quantities of producers with wedges are unresponsive to quota changes (e.g., because the resources used by producers with quota and wedge distortions are non-overlapping), the effects of a quota change coincide with our results in Proposition 2. If these conditions are not satisfied, then the effect of quota changes depends on how quantities for producers with output wedges respond to the quota change.

In Appendix C, we further extend Proposition 9 to characterize nonlinearities and the effects of removing a quota altogether in hybrid economies. Unlike economies with quotas alone, hybrid economies are not generally constrained efficient, and thus removing a quota distortion does not necessarily increase output.

Ex ante results for CES economies. Our results characterize the distance to the frontier and the nonlinear effects of quotas in terms of the quota demand system. In many of the examples in the main text, we estimate or calibrate this quota demand system directly using variation in quotas. In Appendix B, we also provide a characterization of the quota demand system building up from microeconomic information: the input-output structure of the economy and microeconomic elasticities of substitution. For these results, we focus on general input-output economies in which all producers have constant elasticity of substitution (CES) production technologies. We also provide an algorithm for computing the effects of large quota changes using a structural model.

Externalities. Thus far, we have assumed that quotas (or wedges) are the only source of distortion in the economy. Appendix D considers the welfare effects of quota changes in environments where quotas are used to address preexisting externalities. If the quota is initially chosen to maximize welfare, then by construction the first-order effect of marginal quota changes on welfare is zero. In this case, the rents earned by quota holders exactly measure the marginal external cost of the regulated activity. The effects of large quota changes are described by Proposition 5, plus two additional terms that reflect how the willingness-to-pay to avoid the externality varies with real output and with the scale of the harmful activity.

Rent-seeking. In Appendix E, we extend the framework to allow for rent-seeking, in which agents destroy resources to acquire permits. When there is free entry into rent-seeking, so that resources destroyed have exactly the same value as the rents they generate, as in Krueger (1974), the comparative statics of output with respect to quota changes include an additional term that depends on how the quota change affects labor income relative to rents. This effect can result in first-order losses associated with quota changes even starting at the efficient allocation.

9 Conclusion

This paper analyzes economies with quotas and other quantity-based distortions. These economies are constrained efficient, allowing us to develop nonparametric results for the effects of microeconomic shocks and the misallocation costs of quotas, relying only on a small set of sufficient statistics. Table 3 summarizes our results on the effects of quota changes and the sufficient statistics required for each type of counterfactual exercise.

Table 3: Summary of analytic results and sufficient statistics.

Counterfactual	Order of approximation	Sufficient statistics
Marginal quota changes	First-order	Rents $\mathbf{\Pi}$
Large quota changes	Second-order	Rents $\mathbf{\Pi}$, elasticity matrix, H
Large quota changes	Exact	Quota demand system $\mathbf{\Pi}(\boldsymbol{y}^*)$
Announced path of quota changes	Second-order	Quota asset price P, change on announcement, ΔP
Distance to frontier	Second-order	Rents $\mathbf{\Pi}$, elasticity matrix H

Our results can be used to evaluate policy experiments and to characterize the social costs of quota distortions in many settings. The empirical examples we develop—H-1B visas, zoning restrictions, taxicab medallions, import quotas, and capital controls—illustrate how one can measure the sufficient statistics necessary to apply our results.

References

Acemoglu, D., V. M. Carvalho, A. Ozdaglar, and A. Tahbaz-Salehi (2012). The network origins of aggregate fluctuations. *Econometrica 80*(5), 1977–2016.

Adao, R., A. Costinot, and D. Donaldson (2017). Nonparametric counterfactual predictions in neoclassical models of international trade. *American Economic Review 107*(3), 633–689.

Adler, G., N. Lisack, and R. C. Mano (2019). Unveiling the effects of foreign exchange intervention: A panel approach. *Emerging Markets Review* (100620).

Anderson, J. E. (1985). The relative inefficiency of quotas: The cheese case. *American Economic Review 75*(1), 178–190.

Atalay, E. (2017). How important are sectoral shocks? *American Economic Journal: Macroeconomics 9*(4), 254–280.

Baqaee, D. R. (2018). Cascading failures in production networks. *Econometrica 86*(5), 1819–1838.

Baqaee, D. R. and A. Burstein (2023). Welfare and output with income effects and taste shocks. *The Quarterly Journal of Economics 138*(2), 769–834.

Baqaee, D. R. and E. Farhi (2019). The macroeconomic impact of microeconomic shocks: Beyond Hulten's theorem. *Econometrica 87*(4), 1155–1203.

Baqaee, D. R. and E. Farhi (2020). Productivity and misallocation in general equilibrium. *The Quarterly Journal of Economics 135*(1), 105–163.

Baqaee, D. R., E. Farhi, and K. Sangani (2024). The supply-side effects of monetary policy. *Journal of Political Economy 132*(4), 1065–1112.

Baqaee, D. R. and E. Rubbo (2023). Micro propagation and macro aggregation. *Annual Review of Economics 15*(1), 91–123.

Basu, S. and J. G. Fernald (2002). Aggregate productivity and aggregate technology. *European Economic Review 46*(6), 963–991.

Bau, N. and A. Matray (2023). Misallocation and capital market integration: Evidence from India. *Econometrica 91*(1), 67–106.

Bhagwati, J. (1965). *Trade, growth and the balance of payments*, Chapter On the equivalence between tariffs and quotas, pp. 53–67. Rand McNally.

Bigio, S. and J. La'O (2020). Distortions in production networks. *The Quarterly Journal of Economics 135*(4), 2187–2253.

Blanchard, O., G. Adler, and I. de Carvalho Filho (2015). Can foreign exchange intervention stem exchange rate pressures from global capital flow shocks. IMF Working Paper.

Boorstein, R. and R. C. Feenstra (1991). *International Trade and Trade Policy*, Chapter Quality upgrading and its welfare cost in U.S. steel imports, 1969–74, pp. 167–186. The MIT

Press.

Brambilla, I., A. K. Khandelwal, and P. K. Schott (2010). *China's Growing Role in World Trade*, Chapter China's Experience under the Multi-Fiber Arrangement (MFA) and the Agreement on Textiles and Clothing (ATC), pp. 345–387. University of Chicago Press.

Buera, F. J. and N. Trachter (2024). Sectoral development multipliers. Technical Report 32230, National Bureau of Economic Research.

Carvalho, V. M. and A. Tahbaz-Salehi (2019). Production networks: A primer. *Annual Review of Economics 11*(1), 635–663.

Chari, V. V., P. J. Kehoe, and E. R. McGrattan (2007). Business cycle accounting. *Econometrica 75*(3), 781–836.

Chiron, C. (2004). Influences of quotas, tariffs, and bilateral trade agreement on post 2005 apparel trade. Technical report, Harvard Center for Textile and Apparel Research.

Clemens, M. A. (2013). Why do programmers earn more in Houston than Hyderabad? Evidence from randomized processing of US visas. *American Economic Review 103*(3), 198–202.

Cohen, P., R. Hahn, J. Hall, S. Levitt, and R. Metcalfe (2016). Using big data to estimate consumer surplus: The case of Uber. Technical Report 22627, National Bureau of Economic Research.

Dasgupta, P. and J. E. Stiglitz (1977). Tariffs vs. quotas as revenue raising devices under uncertainty. *American Economic Review 67*(5), 975–981.

De Loecker, J., P. Goldberg, A. K. Khandelwal, and N. Pavcnik (2016). Prices, markups, and trade reform. *Econometrica 84*(2), 445–510.

Dean, J. M. (1990). The effects of the US MFA on small exporters. *The Review of Economics and Statistics 72*(1), 63–69.

Domar, E. D. (1961). On the measurement of technological change. *The Economic Journal 71*(284), 709–729.

Edmond, C., V. Midrigan, and D. Y. Xu (2023). How costly are markups? *Journal of Political Economy 131*(7), 1619–1675.

Evans, M. D. D. and R. K. Lyons (2002). Order flow and exchange rate dynamics. *Journal of Political Economy 110*(1), 170–180.

Falvey, R. E. (1979). The composition of trade within import-restricted product categories. *Journal of Political Economy 87*(5), 1105–1114.

Feenstra, R. C. (1988). Quality change under trade restraints in Japanese autos. *The Quarterly Journal of Economics 103*(1), 131–146.

Feenstra, R. C. (1992). How costly is protectionism? *Journal of Economic Perspectives 6*(3), 159–178.

Foerster, A. T., P.-D. Sarte, and M. W. Watson (2011). Sectoral versus aggregate shocks: A structural factor analysis of industrial production. *Journal of Political Economy 119*(1), 1–38.

Gabaix, X. (2011). The granular origins of aggregate fluctuations. *Econometrica 79*(3), 733–772.

Glaeser, E. and J. Gyourko (2018). The economic implications of housing supply. *Journal of Economic Perspectives 32*(1), 3–30.

Grassi, B. (2017). IO in I-O: Size, industrial organization, and the input-output network make a firm structurally important. Working paper.

Gyourko, J. and J. Krimmel (2021). The impact of local residential land use restrictions on land values across and within single family housing markets. *Journal of Urban Economics 126*, 103374.

Harberger, A. C. (1954). Monopoly and resource allocation. *American Economic Review 44*(2), 77–87.

Hsieh, C.-T. and P. J. Klenow (2009). Misallocation and manufacturing TFP in China and India. *The Quarterly Journal of Economics 124*(4), 1403–1448.

Hsieh, C.-T. and E. Moretti (2019). Housing constraints and spatial misallocation. *American Economic Journal: Macroeconomics 11*(2), 1–39.

Hulten, C. R. (1978). Growth accounting with intermediate inputs. *The Review of Economic Studies 45*(3), 511–518.

Itskhoki, O. and B. Moll (2019). Optimal development policies with financial frictions. *Econometrica 87*(1), 139–173.

Khandelwal, A. K., P. K. Schott, and S.-J. Wei (2013). Trade liberalization and embedded institutional reform: Evidence from Chinese exporters. *American Economic Review 103*(6), 2169–2195.

Krueger, A. O. (1974). The political economy of the rent-seeking society. *American Economic Review 64*(3), 291–303.

La'O, J. and A. Tahbaz-Salehi (2022). Optimal monetary policy in production networks. *Econometrica 90*(3), 1295–1336.

Lipsey, R. G. and K. Lancaster (1956). The general theory of second best. *The Review of Economic Studies 24*(1), 11–32.

Liu, E. (2019). Industrial policies in production networks. *The Quarterly Journal of Economics 134*(4), 1883–1948.

McKenzie, L. W. (1951). Ideal output and the interdependence of firms. *The Economic Journal 61*(244), 795–803.

Oi, W. Y. (1961). The desirability of price instability under perfect competition. *Economet-*

rica, 58–64.

Peters, M. (2020). Heterogeneous markups, growth, and endogenous misallocation. *Econometrica 88*(5), 2037–2073.

Petrin, A. and J. Levinsohn (2012). Measuring aggregate productivity growth using plant-level data. *The RAND Journal of Economics 43*(4), 705–725.

Reischer, M. (2019). Finance-thy-neighbor: Trade credit origins of aggregate fluctuations. Working paper.

Restuccia, D. and R. Rogerson (2008). Policy distortions and aggregate productivity with heterogeneous establishments. *Review of Economic Dynamics 11*(4), 707–720.

Rubbo, E. (2023). Networks, Phillips curves and monetary policy. *Econometrica 91*(4), 1417–1455.

Samuelson, P. A. (1972). The consumer does benefit from feasible price stability. *The Quarterly Journal of Economics 86*(3), 476–493.

Weitzman, M. L. (1974). Prices vs. quantities. *The Review of Economic Studies 41*(4), 477–491.

Online Appendix

(Not for publication)

A Proofs

Proof of Proposition 1. Consider a feasible allocation $\mathcal{X} = \{y_i, c_i, x_{i1}, ...x_{iN}, L_{i1}, ..., L_{iF}\}_{1\leq i\leq N}$. For ease of notation, denote the representative household by the index zero, so that $c_i = x_{0i}$. We implement the allocation $\mathcal{X}$ by introducing $N \times (N + F + 1)$ additional nodes with quotas. Each node is placed between user $i \in \{0, ..., N\}$ and resource $j \in \{1, ..., N, 1, ..., F\}$ with a quota of x_{ij}. These quotas ensure that the use of resource j by i is at most x_{ij}. Since producers' production functions F_i and the demand aggregator $\mathcal{D}$ are each weakly increasing in all arguments, the use of resource j by i is also at least x_{ij}. Thus, these quotas ensure that the decentralized equilibrium allocation exactly coincides with $\mathcal{X}$.

Since the allocation is implemented as a competitive equilibrium in the economy with quotas, the first welfare theorem implies that the allocation is efficient (subject to the quota constraints). ∎

Proof of Proposition 2. For an economy with quotas with N producers and F factors, we construct an *isomorphic economy* with a set of producers $\{1, ..., N, 1^q, ..., N^q\}$ and factors $\{1, ..., F, 1^*, ..., N^*\}$. The production functions of producers $1, ..., N$ and the supply of each factor $1, ..., F$ are as in the economy with quotas. For the additional factors $1^*, ..., N^*$ and additional producers $1^q, ..., N^q$, the supply of factor i^* is $L_{i^*} = y_i^*$, and the production function of producer i^q is

$$y_{i^q} = \min\{y_i, L_{i^*}\}.$$

Let $\widehat{\lambda}_i$ and $\widehat{\Lambda}_f$ denote the Domar weights of producers and factors in the isomorphic economy, and let λ_i, Λ_f, and Π_i denote the Domar weights and rents in the economy with quotas. It is straightforward to verify that $\widehat{\Lambda}_{i^*} = \Pi_i$, $\widehat{\lambda}_i = \lambda_i - \Pi_i$, and $\widehat{\lambda}_{i^q} = \lambda_i$.

Applying Hulten's theorem, the response of output to factor supply and productivity changes in the isomorphic economy is

$$d\log Y = \sum_i \widehat{\Lambda}_{i^*} d\log L_{i^*} + \sum_i \widehat{\lambda}_i d\log A_i.$$

Thus, in the economy with quotas,

$$d\log Y = \sum_i \Pi_i d\log y_i^* + \sum_i (\lambda_i - \Pi_i) d\log A_i.$$

∎

Proof of Proposition 3. The resource constraint for each good i, $y_i = c_i + \sum_j x_{ji}$, implies that

$$d \log y_i = \frac{c_i}{y_i} d \log c_i + \sum_j \frac{x_{ji}}{y_i} d \log x_{ji}.$$

Cost minimization implies that for any producer j,

$$d \log y_j = d \log A_j + \sum_i \frac{p_i x_{ji}}{p_j y_j / \tau_j} d \log x_{ji} + \sum_f \frac{w_f L_{jf}}{p_j y_j / \tau_j} d \log L_{jf}.$$

We use these equations in lines 2 and 3 of the following:

$$\begin{aligned}
d \log Y &= \sum_i p_i c_i d \log c_i \\
&= \sum_i p_i y_i d \log y_i - \sum_i \sum_j p_i x_{ij} d \log x_{ji} \\
&= \sum_i p_i y_i d \log y_i - \sum_j \left(p_j y_j / \tau_j\right) \left(d \log y_j - d \log A_j\right) + \sum_j \sum_f w_f L_{jf} d \log L_{jf} \\
&= \sum_i \lambda_i \left(1 - 1/\tau_i\right) d \log y_i + \sum_i \lambda_i \left(1 - \left(1 - 1/\tau_i\right)\right) \left(d \log A_i\right) + \sum_f \Lambda_f d \log L_f.
\end{aligned}$$

Given exogenous shocks $d \log \tau_j$ and $d \log A_j$, to a first order,

$$d \log y_i = \sum_j \frac{\partial \log y_i}{\partial \log \tau_j} d \log \tau_j + \frac{\partial \log y_i}{\partial \log A_j} d \log A_j.$$

Substituting this and $\Pi_i = \lambda_i (1 - 1/\tau_i)$ above completes the proof. ■

Proof of Proposition 4. Follows from substituting Proposition 2 into

$$\Delta \log Y = \int_0^1 \frac{d \log Y}{ds} ds.$$

■

Proof of Proposition 5. From Proposition 2,

$$d \log Y = \sum_i \Pi_i d \log y_i^*.$$

Thus,

$$d^2 \log Y = \sum_i \left[\sum_j \frac{d\Pi_i}{d \log y_j^*} d \log y_j^* \right] d \log y_i^*.$$

Writing this expression in matrix form completes the proof. ■

Proof of Proposition 6. The quantity y_i is chosen to maximize real rents, taking all other quotas as given,

$$y_i = \arg\max_y \frac{\Pi_i(y)}{P(y)} = \arg\max_y \Pi_i(y) Y(y).$$

From the first order condition and Proposition 2,

$$\frac{d \log \Pi_i}{d \log y_i} = -\frac{d \log Y}{d \log y_i} = -\Pi_i.$$

Thus,

$$d^2 \log Y = \frac{d\Pi_i}{d \log y_i} (d \log y_i)^2 = \Pi_i \frac{d \log \Pi_i}{d \log y_i} (d \log y_i)^2 = -\Pi_i^2 (d \log y_i)^2.$$

■

Proof of Proposition 7. Let $y_i^*(t; h) = \bar{y}_i + h\epsilon(t)$ denote the quota at t given a shock of size h, where $\bar{y}_i$ is the steady-state quota level before the shock announcement. The effect of the announced quota path on output is given by integrating over the marginal effects of incremental changes to the quotas, given by Proposition 2,

$$\Delta \log Y = \int_0^h \int_0^\infty \Pi_{it} \left(y_i^*(t; h') \right) \frac{d \log y_i^*}{dh'} dt dh'. \tag{12}$$

Rents Π_{it} are a function of $y_i^*(t; h)$ given our assumption that rents earned by the quota at time t are only a function of the contemporaneous quota level. Note that Π_{it} is the present discounted value of rents earned by quota i in period t as a share of total wealth. We can write

$$\Pi_{it} \left(y_i^*(t; h') \right) = e^{-\int_0^t r(s) ds} y_i^* R_i \left(y_i^*(t; h') \right),$$

where $r(t)$ is the (exogenous) interest rate at time t and $R_i(y_i^*(t; h))$ is the rents earned per quota unit in time t dollars as a share of wealth. Use this to rewrite (12) as

$$\Delta \log Y = \int_0^h \int_0^\infty e^{-\int_0^t r(s) ds} \underbrace{R_i \left(y^*(t; h') \right)}_{\substack{\text{Quota rents} \\ \text{(per unit)}}} \underbrace{\frac{dy^*(t; h')}{dh'}}_{\substack{\text{Change in} \\ \text{quota}}} dt dh'. \tag{13}$$

We will take a second-order approximation of (13) around $h = 0$.

$$\frac{d\left[\Delta \log Y\right]}{dh} = \int_0^\infty e^{-\int_0^t r(s)ds} R_i\left(y_i^*(t;h)\right)\epsilon\left(t\right)dt,$$

$$\frac{d^2\left[\Delta \log Y\right]}{dh^2} = \int_0^\infty e^{-\int_0^t r(s)ds} \frac{dR_i\left(y_i^*(t;h)\right)}{dy_i^*}\left[\epsilon\left(t\right)\right]^2 dt.$$

Thus, to a second order in h,

$$\Delta \log Y = \int_0^\infty e^{-\int_0^t r(s)ds} R_i\left(y_i^*(t;0)\right)\left[h\epsilon\left(t\right)\right]dt + \frac{1}{2}\int_0^\infty e^{-\int_0^t r(s)ds} \frac{dR_i\left(y_i^*(t;0)\right)}{dy_i^*}\left[h\epsilon\left(t\right)\right]^2 dt + O(h^3). \quad (14)$$

The price of a perpetual license to produce under the quota is given by the present discounted value of rents earned by a unit of the quota in all future periods,

$$P_i\left(0;h\right) = \int_0^\infty e^{-\int_0^t r(s)ds} R_i\left(y_i^*\left(t;h\right)\right)dt.$$

Use this expression to write the asset price given $h = 0$ and the change in the asset price given an announced $h > 0$:

$$P_i = P_i\left(0;0\right) = \int_0^\infty e^{-\int_0^t r(s)ds} R_i\left(y_i^*\left(t;0\right)\right)dt,$$

$$\Delta P_i = P_i\left(0;h\right) - P_i\left(0;0\right) = \int_0^\infty e^{-\int_0^t r(s)ds} \frac{dR_i\left(y_i^*\left(t;0\right)\right)}{dy_i^*}\left[h\epsilon\left(t\right)\right]dt + O\left(h^2\right).$$

Substituting these expressions into (14) concludes the proof. ■

Proof of Corollary 1. Starting from Proposition 7, and substituting in $h\epsilon\left(t\right) = \Delta y_i^*$ yields

$$\Delta \log Y = P_i \Delta y_i^* + \frac{1}{2}\Delta P_i \Delta y_i^* + O(h^3).$$

We can use $\log y_i^*$ and $V_i = P_i y_i^*$ to rewrite

$$\Delta \log y_i^* = \frac{1}{y_i^*}\Delta y_i^* - \frac{1}{2}\frac{1}{y_i^{*2}}\left(\Delta y_i^*\right)^2 + O\left(h^3\right),$$

$$\Delta V_i = P_i \Delta y_i^* + \Delta P_i y_i^* + \frac{dP_i}{dy_i^*}\left(\Delta y_i^*\right)^2 + O\left(h^3\right).$$

We use these expressions to show that $\left(V_i + \frac{1}{2}\Delta V_i\right) \Delta \log y_i$ is equal to $P_i \Delta y_i^* + \frac{1}{2}\Delta P_i \Delta y_i^*$ to a second-order in h:

$$
\begin{aligned}
&\left(V_i + \frac{1}{2}\Delta V_i\right) \Delta \log y_i + O\left(h^3\right) \\
&\quad = \left(P_i y_i^* + \frac{1}{2}\left(P_i \Delta y_i^* + \Delta P_i y_i^* + \frac{dP_i}{dy_i^*}\left(\Delta y_i^*\right)^2\right)\right)\left(\frac{1}{y_i^*}\Delta y_i^* - \frac{1}{2}\frac{1}{y_i^{*2}}\left(\Delta y_i^*\right)^2\right) + O\left(h^3\right) \\
&\quad = P_i \Delta y_i^* - \frac{1}{2} P_i \frac{1}{y_i^*}\left(\Delta y_i^*\right)^2 + \frac{1}{2} P_i \frac{1}{y_i^*}\left(\Delta y_i^*\right)^2 + \frac{1}{2}\Delta P_i \Delta y_i^* + O\left(h^3\right) \\
&\quad = P_i \Delta y_i^* + \frac{1}{2}\Delta P_i \Delta y_i^* + O\left(h^3\right).
\end{aligned}
$$

■

Proof of Proposition 8. Starting at the efficient point where distortions are just-binding, rents $\mathbf{\Pi} = 0$. To a second order, the change in output from distortions $\Delta \log \mathbf{y}^* = \log \mathbf{y}^* - \log \mathbf{y}^{\mathbf{eff}}$ starting from this point is given by Proposition 5,

$$\log Y - \log Y^{\text{eff}} \approx \frac{1}{2} \left(\Delta \log \mathbf{y}^*\right)' H \left(\Delta \log \mathbf{y}^*\right).$$

Multiplying by negative one yields the expression for the distance to the frontier, $\Delta \log Y = \log Y^{\text{eff}} - \log Y$, given in Equation (8).

Starting at the point where rents are zero, to a first order,

$$\Pi_i \approx \sum_j \frac{d\Pi_i}{d \log y_j^*} \Delta \log y_j^* \qquad \Rightarrow \qquad \mathbf{\Pi} \approx H \Delta \log \mathbf{y}^*.$$

Substituting into Equation (8) yields Equation (7). Finally, substituting

$$\Delta \log \mathbf{y}^* \approx H^{-1} \mathbf{\Pi}$$

into Equation (7) yields Equation (9). ■

Proof of Proposition 9. Follows immediately from Proposition 3. ■

B Structural Model

In the main text, we consider examples where we can estimate the quota demand system using *ex post* variation in quotas. In this appendix, we provide *ex ante* results that characterize the quota demand system and the effects of large quota changes on output in terms of the input-output structure of the economy. These results exploit an isomorphism between economies with quotas and efficient economies.

For the results in this appendix, we focus on economies in which all producers have constant elasticity of substitution (CES) production technologies. The results can be generalized to non-CES production functions following the approach in Baqaee and Farhi (2019). So, given an economy with quotas that features N producers and F factors, we assume that each producer has a CES production function given by

$$y_i = A_i \left(\sum_{j=1}^{N} \omega_{ij} x_{ij}^{\frac{\theta_i - 1}{\theta_i}} + \sum_{f=1}^{F} \omega_{if} L_{if}^{\frac{\theta_i - 1}{\theta_i}} \right)^{\frac{\theta_i}{\theta_i - 1}},$$

where y_i is the output of producer i, x_{ij} is i's use of intermediate inputs from producer j, L_{if} is i's use of factor f, ω_{ij} and ω_{if} are positive constants, and θ_i is the elasticity of substitution in production across i's inputs. We further assume, without loss of generality, that household consumption is equal to the output of the first producer, so that $Y = y_1$.

B.1 Isomorphic Economy: Definition and Notation

Given an economy with quotas with N producers and F factors, we define an *isomorphic economy* with a set of producers $\{1, ..., N, 1^q, ..., N^q\}$ and set of factors $\{1, ..., F, 1^*, ..., N^*\}$. In words, the isomorphic economy includes N additional producers (which we denote with superscripts q) and N additional factors (which we denote with asterisks). Let $\mathcal{N}$ denote the original set of producers $\{1, ..., N\}$, let $\mathcal{N}^q$ denote the set of additional, fictitious producers in the isomorphic economy $\{1^q, ..., N^q\}$, let $\widehat{\mathcal{N}} = \mathcal{N} \cup \mathcal{N}^q$ denote the full set of producers in the isomorphic economy, and let $\widehat{\mathcal{F}} = \{1, ..., F, 1^*, ..., N^*\}$ denote the set of factors in the isomorphic economy. In the text that follows, we will use hats to denote other variables in the isomorphic economy.

The *input-output* matrix $\widehat{\Omega}$ of the isomorphic economy is defined as follows. For producers $i \in \mathcal{N}$,

$$\widehat{\Omega}_{ij} = 0 \text{ for } j \in \mathcal{N}, \qquad \widehat{\Omega}_{ij^q} = \frac{p_j x_{ij}}{\lambda_i - \Pi_i} \text{ for } j \in \mathcal{N}, \qquad \widehat{\Omega}_{if} = \frac{w_f L_{if}}{\lambda_i - \Pi_i} \text{ for } j \in \mathcal{F}.$$

That is, each element of $\widehat{\Omega}$ is the total expenses of producer i on good j, as a share of the total costs (sales minus profits) of producer i. Note that all intermediate inputs used by firm i are purchased from the fictitious producers j^q rather than directly from producer j.

For each fictitious producer $i^q \in \mathcal{N}^q$,

$$\widehat{\Omega}_{i^q i} = \frac{\lambda_i - \Pi_i}{\lambda_i}, \qquad \widehat{\Omega}_{i^q i^*} = \frac{\Pi_i}{\lambda_i}, \qquad \widehat{\Omega}_{i^q j} = 0 \text{ for all } j \notin \{i, i^*\}.$$

Finally, $\widehat{\Omega}_{fj} = 0$ for all $f \in \widehat{\mathcal{F}}$ and for all j.

For producers $i \in \mathcal{N}$, the elasticity of substitution across inputs in the isomorphic economy is equal to that in the original economy with quotas, i.e., $\widehat{\theta}_i = \theta_i$. For the fictitious producers $\{1^q, ..., N^q\}$, $\widehat{\theta}_{i^q} = -1$. That is, the fictitious producer i^q has a Leontief production function in the output of producer i and the fictitious factor i^*. Thus, output of producer i^q is

$$y_{i^q} = \min\{y_i, y_i^*\}.$$

Denote the *Leontief inverse* of the isomorphic economy by $\widehat{\Psi} = (I - \widehat{\Omega})^{-1}$. The first row of $\widehat{\Psi}$ describes the sales of each producer as a fraction of nominal GDP, i.e. $\widehat{\lambda} = \widehat{\Psi}^{(1)}$, in the isomorphic economy.

When comparing the economy with quotas and the isomorphic economy, note that for $i = 1, ..., N$, $\widehat{\lambda}_{i^q} = \lambda_i$ and $\widehat{\lambda}_i = \lambda_i - \Pi_i$. For the fictitious factors $1^*, ..., N^*$, factor income shares in the isomorphic economy are equal to quota rents, $\widehat{\Lambda}_{i^*} = \Pi_i$. For the remaining factors $1, ..., F$, factor income shares in both economies coincide, $\widehat{\Lambda}_f = \Lambda_f$.

B.2 Analytic Results for Elasticities of Rents to Quota Levels

With these definitions for the isomorphic economy, we can apply results from efficient economies to characterize how rents in the economy with quotas respond to shocks. Following Baqaee and Farhi (2019), we define the *input-output covariance operator*

$$Cov_{\widehat{\Omega}^{(j)}}\left(\widehat{\Psi}_{(f)}, \widehat{\Psi}_{(g)}\right) = \sum_{k \in \widehat{\mathcal{N}} \cup \widehat{\mathcal{F}}} \widehat{\Omega}_{jk} \widehat{\Psi}_{kf} \widehat{\Psi}_{kg} - \left(\sum_{k \in \widehat{\mathcal{N}} \cup \widehat{\mathcal{F}}} \widehat{\Omega}_{jk} \widehat{\Psi}_{kf}\right)\left(\sum_{k \in \widehat{\mathcal{N}} \cup \widehat{\mathcal{F}}} \widehat{\Omega}_{jk} \widehat{\Psi}_{kg}\right).$$

Proposition B1 applies Proposition 9 from Baqaee and Farhi (2019) to characterize the local derivative of the quota demand system, H, using the input-output structure of the isomorphic economy.

Proposition B1 (Quota demand system)**.** *Define the $|\mathcal{F}| \times |\mathcal{F}|$ matrix $\widehat{H}$, with the i'th row of*

$\widehat{H}$ *equal to,*

$$\widehat{H}_{(i)} = (I - \Gamma)^{-1} \delta^i,$$

where the matrix Γ *and vector* δ^i *are*

$$\Gamma_{fg} = -\frac{1}{\widehat{\Lambda}_g} \left(\sum_j \widehat{\lambda}_j \left(\widehat{\theta}_j - 1 \right) Cov_{\widehat{\Omega}^{(j)}} \left(\widehat{\Psi}_{(f)}, \widehat{\Psi}_{(g)} \right) \right) \qquad \text{if } \widehat{\Lambda}_g > 0 \ \text{ else } 0,$$
$$\delta^i_f = \sum_j \widehat{\lambda}_j \left(\widehat{\theta}_j - 1 \right) Cov_{\widehat{\Omega}^{(j)}} \left(\widehat{\Psi}_{(f)}, \widehat{\Psi}_{(i)} \right).$$

H is the submatrix formed by the last N rows and N columns of $\widehat{H}$.

The system of equations in Proposition B1 describes how income shares for each factor in the isomorphic economy respond to a change in the supply of factor i. Since rents in the economy with quotas correspond to income shares for fictitious factors $1^*, ..., N^*$, the entries of H are given by the lower right $N \times N$ submatrix of $\widehat{H}$.

When the economy with quotas has a single factor and a single quota, we can solve for the semi-elasticity of rents to the quota in closed form, shown in Corollary B1.

Corollary B1 (Quota demand system with a single factor and single quota)**.** *Suppose there is a single factor and a single quota y^*_i. In response to a change in quota y^*_i, the response of rents Π_i is*

$$\frac{d\Pi_i}{d \log y^*_i} = \frac{\alpha \Pi_i (1 - \Pi_i)}{\alpha + \Pi_i (1 - \Pi_i)}.$$

where

$$\alpha = \sum_j \widehat{\lambda}_j \left(\widehat{\theta}_j - 1 \right) Var_{\widehat{\Omega}^{(j)}} \left(\widehat{\Psi}_{(i^*)} \right).$$

B.3 Iterative Algorithm for Effects of Large Quota Changes

We can solve for the effects of large quota changes on output by chaining together the effects of infinitesimal quota changes. When doing so, it is important to keep track of quotas that transition from slack to binding, since the semi-elasticity of rents with respect to the quota changes discontinuously at the point where a quota becomes binding. Algorithm B1 presents an iterative method for computing the effects of large quota changes.

Algorithm B1 (Iterative algorithm for large quota changes)**.** *For a vector of quota changes $\Delta \log y^*$, the algorithm proceeds by discretizing the total change into a sequence of small increments $\{d \log y^*\}$ and cumulating the first-order effect of each incremental change.*

1. *Initialize the input-output matrix for the isomorphic economy $\widehat{\Omega}$, and initialize $\Delta \log Y = 0$. Let y denote the vector of initial quantities for any producers with (slack or binding) quotas. Define a small constant ϵ (e.g., $\epsilon = 1 \times 10^{-8}$).*

2. *For each increment change to the vector of quotas $d \log y^*$, repeat the following steps (a)–(h):*

   (a) Using $\widehat{\Omega}$, construct $\widehat{\Psi} = (I - \widehat{\Omega})^{-1}$ and $\widehat{\lambda} = \widehat{\Psi}^{(1)}$.

   (b) Calculate the change in output using Proposition 2:

   $$d \log Y = \sum_i \widehat{\Lambda}_{i^*} \, d \log y_i^*.$$

   Update $\Delta \log Y^{+1} = \Delta \log Y + d \log Y$.

   (c) Compute $\widehat{H}$ using Proposition B1.

   (d) Compute the change in the price vector and the input-output matrix,

   $$d \log \widehat{p} = -\sum_i \left(\widehat{\Psi}_{(i^*)} - \sum_f \frac{\widehat{\Psi}_{(f)}}{\widehat{\Lambda}_f} \widehat{H}_{fi^*} \right) d \log y_i^*.$$
   $$d\widehat{\Omega}_{ji} = \left(\theta_j - 1\right) Cov_{\widehat{\Omega}^{(j)}} \left(-d \log \widehat{p}, I_{(i)}\right).$$

   (e) Define the updated input-output matrix $\widehat{\Omega}^{+1} = \widehat{\Omega} + d\widehat{\Omega}$. For each i, check if $\widehat{\Omega}^{+1}_{i^q i^*} < 0$. If so, set $\widehat{\Omega}^{+1}_{i^q i^*} = 0$ and $\widehat{\Omega}^{+1}_{i^q i} = 1$.

   (f) Construct $\widehat{\Psi}^{+1}$ and $\widehat{\lambda}^{+1}$ using $\widehat{\Omega}^{+1}$ as in step (a). For all producers with (slack or binding) quotas, construct $y_i^{+1} = \exp\left(\log y_i + \log \widehat{\lambda}_i^{+1} - \log \widehat{\lambda}_i - d \log \widehat{p}_i\right)$.

   (g) For each i where $\widehat{\Lambda}_{i^*} = 0$, check if $y_i^{+1} \geq y_i^*$. If so, set $\widehat{\Omega}^{+1}_{i^q i^*} = \epsilon$ and $\widehat{\Omega}^{+1}_{i^q i} = 1 - \epsilon$.

   (h) Set $\widehat{\Omega} = \widehat{\Omega}^{+1}$, $y = y^{+1}$, and $\Delta \log Y = \Delta \log Y^{+1}$.

The crucial steps in the algorithm are steps 2(e) and 2(g). Step 2(e) checks whether the price of any quota turns negative. If so, the input-output matrix for the economy is updated to reflect that the quota is slack. Step 2(g) checks if any previously slack quota becomes binding by checking whether the output quantity for any producer exceeds its quota. In that case, the algorithm assigns the just-binding quota a small positive weight, ϵ. This ensures that, in the next iteration, directional derivatives of factor income shares with respect to the quota are taken from the binding side rather than the slack side.

Example 8 in the main text illustrates how we can use Algorithm B1 to evaluate the effects of large quota reforms in economies with multiple quotas.

C Additional Results for Wedge and Hybrid Economies

This appendix derives additional results comparing wedge and hybrid economies to economies with quotas. Section C.1 takes up the issue that economies with quotas and wedges with identical allocations of resources may have differerent prices and presents sufficient conditions such that prices, sales, and quota rents / wedge revenues coincide. In Section C.2, we extend Proposition 9 to characterize nonlinear effects of quotas and the gains (or losses) from removing a quota in a hybrid economies that feature both quota and wedge distortions.

C.1 Mapping Wedges to Quotas

A challenge when comparing economies with quotas and wedges is that two economies that share the same physical allocation of resources, when implemented via implicit quotas or wedges, may have different prices, sales shares, and profits. This challenge stems from the fact that it is often possible to implement a given allocation of resources with many different sets of wedges. To take an example, consider a horizontal economy in which firms use labor to produce differentiated varieties, which are then consumed by a representative household. In this economy, doubling all firms' markups increases firms' prices and profits and reduces labor's share of income without affecting the allocation of resources.

We can eliminate this indeterminacy by imposing restrictions on wedges. Proposition C1 presents restrictions that ensure that if the allocation of resources in a wedge economy coincides with a quota economy, then the observable prices, sales, and profits also coincide.

Proposition C1 (Matching observables in wedge and quota economies)**.** *Consider an economy with quotas in which all producer prices p_i and factor wages w_f are strictly positive. Consider a second economy in which the same allocation of resources is implemented with wedges, τ. If*

(i) $\tau_i \geq 1$ *for all i, and*

(ii) for each good or factor i, either the good is directly consumed by the household $c_i > 0$ or there exists some producer j such that $\partial F_j / \partial x_{ji} > 0$ and $\tau_j = 1$,

then prices and sales are identical across the two economies.

Proof. Since the first welfare theorem holds, the equilibrium in the economy with quotas maximizes the consumption aggregator subject to the feasibility constraints, quotas, and factor supplies,

$$Y = \max \mathcal{D}(c_1, ..., c_N) + \sum_i \psi_i \left(F_i\left(x_{i1}, ..., x_{iN}, L_{i1}, ..., L_{iF}\right) - y_i\right)$$
$$+ \sum_i \phi_{i^*}\left(y_{i^*} - y_i\right) + \sum_i \rho_i \left(y_i - \sum_j x_{ji} - c_i\right) + \sum_f \rho_f \left(L_f - \sum_j L_{jf}\right). \quad (15)$$

where ψ_i, ϕ_{i^*}, ρ_i, and ρ_f are Lagrange multipliers. The assumption that prices p_i and wages w_f are strictly positive in the economy with quotas implies that $\rho_i, \psi_i > 0$ for all i and $\rho_f > 0$ for all f.

For any good i, since ρ_i is the Lagrange multiplier on its resource constraint and ϕ_{i^*} is the Lagrange multiplier on its quota constraint, we can solve for the wedge between the price of good i and its marginal cost,

$$\tau_i = \frac{\rho_i}{\rho_i - \phi_{i^*}}.$$

We now show that the vector of these wedges τ must satisfy the conditions in the proposition: for each i, either (1) i is directly consumed by the representative household, or (2) for all users j where $\partial F_j / \partial x_{ji} > 0$, there is at least one producer j such that $\phi_{j^*} = 0$ and $\tau_j = 1$.

We prove by contradiction. Suppose there is a good i that is not consumed by the household, where $\phi_{j^*} > 0$ for all j where $\partial F_j / \partial x_{ji} > 0$. Since $\rho_i > 0$, we must have

$$\sum_j x_{ji} = y_i.$$

Moreover, since $\rho_j > 0$ and $\phi_{j^*} > 0$ for each j where $\partial F_j / \partial x_{ji} > 0$, we must have

$$\begin{aligned} y_j &= F_j\left(x_{j1}, ..., x_{ji}, ..., x_{jN}, L_{j1}, ..., L_{jF}\right). \\ y_j &= y_{j^*}. \end{aligned}$$

From (15), the change in output from an exogenous increase in y_i is equal to $\rho_i > 0$. Note that y_i is not consumed directly. Moreover, for all producers j where $\partial F_j / \partial x_{ji} > 0$, we have that $y_j = y_j^*$. Thus, the exogenous increase in y_i has no effect on $c_1, ..., c_N$ and has no effect on output, in contradiction with the value of an exogenous increase in y_i being strictly positive. ■

The first condition that $\tau_i \geq 1$ for all producers ensures that profits in the wedge economy are weakly positive. This is necessary to match observables across the wedge and quota economies, because quota rents must be weakly positive (they are strictly positive when quotas are binding or else zero).

The second condition requires that one user of each factor or good in the economy (which may be the representative household) has a wedge $\tau_i = 1$. In an economy with quotas, if all users of a good have binding quotas, the price of that good must be equal to zero. The assumption that all prices and wages in the economy with quotas are strictly positive thus implies that at least one user of each factor or good must be unconstrained.

Together, the first and second conditions also ensure that the wedges that map to a given quota allocation are unique. Since all producers must have weakly positive profits, and at least one producers' profits must be exactly zero among users of each good, one cannot scale up wedges across firms while continuing to satisfy these requirements. Thus, the conditions in Proposition C1 identify the unique vector of wedges that generate the same allocation and prices as a given set of quotas. Note that when these conditions are satisfied, wedge revenues for each producer in the wedge economy exactly equal the rents earned by the quota on the corresponding producer's output in the quota economy.

Example 11 (Small Open Economy)**.** Consider the small open economy from Example 1, and suppose the only binding quota is the quota on imports y_f^*. Suppose we have an identical economy (the "tariff economy") where, instead of an import quota, there is an import tariff τ_f and a tax on consumption of the domestic good τ_d. Given total production of the domestic good y_d and the domestic-good consumption tax τ_d, the tariff τ_f that implements the same import quantity y_f^* as the economy with quotas is

$$\tau_f = \frac{1-\omega}{\omega}\tau_d A_f \left(\frac{y_d - y_{f^*}/A_f}{y_{f^*}}\right)^{\frac{1}{\theta}}. \tag{16}$$

Notice that the import tariff τ_f and the tax on domestic good consumption τ_d can be scaled by an arbitrary factor without altering the import quantity.

Setting the tax on the domestic good $\tau_d = 1$ leads prices, sales, and profits to coincide across the tariff economy and the quota economy. For example, in the quota economy, the quota holders earn rents Π_f. It is straightforward to verify that in the tariff economy with $\tau_d = 1$, the same Π_f is generated as tariff revenue instead.

C.2 Nonlinearities in Hybrid Economies

In the main text, Proposition 9 characterizes the first-order effect of a quota change on output in an economy that features both quota and wedge distortions. In this appendix, we consider the nonlinear effects of quota changes in such "hybrid" economies. Proposition C2 characterizes the effect of a quota change on output in a hybrid economy to a second order.

Proposition C2 (Nonlinearities: Hybrid economies)**.** *Consider an economy that features both quotas and wedges. Let $\mathcal{Q}$ denote the set of producers with output quotas and $\mathcal{W}$ the set of producers with output wedges. The effect of a change in the quota on producer $i \in \mathcal{Q}$, $\Delta \log y_i$, on output to a second order is*

$$\Delta \log Y \approx \Pi_i \Delta \log y_i^* + \sum_{j \in \mathcal{W}} \Pi_j \Delta \log y_j + \frac{1}{2}\left(\Delta\Pi_i \Delta \log y_i + \sum_{j \in \mathcal{W}} \Delta\Pi_j \Delta \log y_j \right),$$

where $\Delta \log y_j$ is the change in producer j's output induced by the quota change, $\Delta\Pi_i$ is the change in rents for producer i, and $\Delta\Pi_j$ is the change in the wedge revenues for producer j.

Proof. Proposition 9 shows that

$$d \log Y = \Pi_i d \log y_i^* + \sum_{j \in \mathcal{W}} \Pi_j \frac{d \log y_j}{d \log y_i^*} d \log y_i^*.$$

Taking the derivative,

$$d^2 \log Y = \frac{d\Pi_i}{d \log y_i^*} d \log y_i^* + \sum_{j \in \mathcal{W}} \frac{d\Pi_j}{d \log y_i^*} \frac{d \log y_j}{d \log y_i^*} \left(d \log y_i^* \right)^2 + \sum_{j \in \mathcal{W}} \Pi_j \frac{d^2 \log y_j}{d \log y_i^{*2}} \left(d \log y_i^* \right)^2.$$

Thus,

$$\begin{aligned}
\Delta \log Y &\approx d \log Y + \frac{1}{2} d^2 \log Y \\
&= \left(\Pi_i + \sum_{j \in \mathcal{W}} \Pi_j \frac{d \log y_j}{d \log y_i^*} \right) \Delta \log y_i^* \\
&\quad + \frac{1}{2} \left(\frac{d\Pi_i}{d \log y_i^*} + \sum_{j \in \mathcal{W}} \frac{d\Pi_j}{d \log y_i^*} \frac{d \log y_j}{d \log y_i^*} + \Pi_j \frac{d^2 \log y_j}{d \log y_i^{*2}} \right) (\Delta \log y_i^*)^2.
\end{aligned}$$

Substituting $\Delta \log y_j \approx d \log y_j + \frac{1}{2} d^2 \log y_j$ and $\Delta\Pi_j \approx d\Pi_j + \frac{1}{2} d^2 \Pi_j$ and keeping only first- and second-order terms completes the proof. ■

Intuitively, since the first-order effects of quota changes on output in a hybrid economy depend on wedge revenues and quantity changes for all producers with wedges, the nonlinear effects of quota changes also depend on changes in wedge revenues for all producers for wedges. It is also worth emphasizing that for the second-order expansion in Proposition C2, one must know how the quantity for each producer j with a wedge distortion responds *to a second order* to the change in i's quota. If one observes a quota change in a hybrid economy, however, one can directly measure quantity changes for other producers $\Delta \log y_j$ in the data to calculate the change in output implied by Proposition C2.

A useful case of Proposition C2 is when the quota on producer i is removed entirely. In this case, the effect on output can be simplified to the expression in Corollary C1.

Corollary C1 (Effect of removing a quota: Hybrid economies)**.** *The effect of removing the quota on producer i on output in an economy that features both quotas and wedges is, to a second order,*

$$\Delta \log Y \approx \frac{1}{2}\Pi_i \Delta \log y_i^* + \sum_{j \in \mathcal{W}} \left(\Pi_j + \frac{1}{2}\Delta\Pi_j\right) \Delta \log y_j,$$

where Π_j *is the initial wedge revenues for producer* j *and* $\Delta\Pi_j$ *is the change in wedge revenues induced by removing the quota on i.*

Note that, unlike in an economy with quotas, the effect of removing a quota distortion in a hybrid economy is not guaranteed to improve efficiency and output. This is because the hybrid economy is not generally constrained efficient, and so the gains from removing the quota on i can be offset by reallocations across producers with wedges that exacerbate existing distortions.

D Optimal Quotas with Externalities

Quotas are sometimes used to correct for negative externalities. In this appendix, we characterize the effect of quota changes on welfare in the presence of an externality. When the initial quota level is chosen optimally, quota rents exactly reflect the marginal willingness to pay to limit the constrained activity. Starting at this point, the effect of large quota changes depends on the nonlinear effects that we characterize in the main text as well as additional terms that depend on how the willingness to pay changes with real output and with the level of the constrained activity.

Suppose that welfare is given by

$$\mathcal{U}(Y, y_i),$$

where Y is real output, y_i is the level of output for the activity with an externality, and $\mathcal{U}$ is assumed to be strictly increasing in the first argument and decreasing in the second. For exposition, we will refer to y_i as the level of pollution; of course, y_i could refer to any activity that directly affects social welfare beyond its contribution to real output.

For any given combination (Y, y_i), define W as the level of real output that would yield the same welfare if pollution were fixed at the level y_i^0:

$$\mathcal{U}(W, y_i^0) = \mathcal{U}(Y, y_i).$$

Note that W is a money-metric utility, since it expresses the utility of (Y, y_i) in units of real output holding the level of pollution fixed at y_i^0. Since $\mathcal{U}$ is strictly increasing in its first argument, we can directly express the money-metric utility W as a function of real output Y and the level of pollution y_i, where this function is parameterized by the benchmark pollution level y_i^0:

$$W = \mathcal{W}(Y, y_i; y_i^0).$$

Proposition D1 characterizes the marginal external cost of pollution—i.e., the direct effect of an increase in pollution on money-metric utility—assuming that the initial quota is chosen optimally to maximize welfare.

Proposition D1 (Marginal external cost)**.** *Let y_i^0 denote the welfare-maximizing level of pollution. Starting with a quota on pollution at $y_i^* = y_i^0$, the marginal external cost of a proportional increase in pollution is equal to the quota rents,*

$$\frac{\partial \log \mathcal{W}(Y, y_i; y_i^*)}{\partial \log y_i} = -\Pi_i.$$

Proof. Given that y_i^0 maximizes welfare, the first-order effect of changes in y_i on welfare starting at $y_i = y_i^0$ is zero:

$$\frac{d\log W}{d\log y_i} = \frac{\partial \log \mathcal{W}(Y, y_i; y_i^0)}{\partial \log Y}\frac{d\log Y}{d\log y_i^*} + \frac{\partial \log \mathcal{W}(Y, y_i; y_i^0)}{\partial \log y_i} = 0. \tag{17}$$

Given our definition of W,

$$\mathcal{W}(Y, y_i; y_i^0)\Big|_{y_i=y_i^0} = Y, \qquad \text{and} \qquad \frac{\partial \log \mathcal{W}(Y, y_i; y_i^0)}{\partial \log Y}\Bigg|_{y_i=y_i^0} = 1.$$

Substituting this in and applying Proposition 2 yields the result. ■

When the quota on pollution maximizes welfare, Proposition D1 shows that the rents earned by quota holders reflect the direct welfare cost of a marginal increase in pollution. The intuition is analogous to that behind a Pigouvian tax: just as the optimal tax equals the marginal external cost of pollution, the effective tax rate induced by the optimal quota reflects the direct effect of pollution on welfare. As a result, the direct welfare cost of a proportional increase in pollution is equal to the effective tax rate induced by the optimal quota times the quota level, or the total rents earned by the quota.

If a quota is chosen optimally, then the Envelope Theorem implies that marginal changes in the quota have no first-order effect on welfare. In Proposition D2, we characterize the effect of large quota deviations from the optimal quota level on welfare to a second order.

Proposition D2 (Welfare Effects of Quota Changes)**.** *Let y_i^0 denote the welfare-maximizing level of pollution. Starting with a quota on pollution at $y_i^* = y_i^0$, the effect of a quota change $\Delta \log y_i^*$ on welfare is*

$$\Delta \log W \approx \left[\frac{1}{2}\frac{d\Pi_i}{d\log y_i^*} + \Pi_i \frac{\partial^2 \log \mathcal{W}\left(Y, y_i; y_i^*\right)}{\partial \log Y \partial \log y_i} + \frac{1}{2}\frac{\partial^2 \log \mathcal{W}\left(Y, y_i; y_i^*\right)}{\partial \log y_i^2}\right]\left(\Delta \log y_i^*\right)^2 + h.o.t.$$

where h.o.t are terms of order $(\Delta \log y_i^*)^3$.

Proof. Taking the derivative of (17) with respect to $\log y_i$, we have

$$\frac{d^2 \log W}{d\log y_i^2} = \left[\frac{\partial^2 \log \mathcal{W}\left(Y, y_i; y_i^0\right)}{\partial \log Y^2}\frac{d\log Y}{d\log y_i^*} + 2\frac{\partial^2 \log \mathcal{W}\left(Y, y_i; y_i^0\right)}{\partial \log Y \partial \log y_i}\right]\frac{d\log Y}{d\log y_i^*}$$

$$+\frac{\partial \log \mathcal{W}\left(Y, y_i; y_i^0\right)}{\partial \log Y}\frac{d^2 \log Y}{\left(d \log y_i^*\right)^2}+\frac{\partial^2 \log \mathcal{W}\left(Y, y_i; y_i^0\right)}{\partial \log y_i^2}$$

Given our definition of W, evaluated at $y_i = y_i^0$, we have:

$$\left.\frac{\partial \log \mathcal{W}(Y, y_i; y_i^0)}{\partial \log Y}\right|_{y_i=y_i^0} = 1, \qquad \text{and} \qquad \left.\frac{\partial^2 \log \mathcal{W}(Y, y_i; y_i^0)}{\partial \log Y^2}\right|_{y_i=y_i^0} = 0.$$

Further substituting in our results from Proposition 2 and Proposition 5, we get

$$\frac{d^2 \log W}{d \log y_i^2} = 2\frac{\partial^2 \log \mathcal{W}\left(Y, y_i; y_i^0\right)}{\partial \log Y \partial \log y_i}\Pi_i + \frac{d\Pi_i}{d \log y_i^*} + \frac{\partial^2 \log \mathcal{W}\left(Y, y_i; y_i^0\right)}{\partial \log y_i^2}. \tag{18}$$

Using

$$\Delta \log W \approx \frac{d \log W}{d \log y_i}(\Delta \log y_i^*) + \frac{1}{2}\frac{d^2 \log W}{d \log y_i^2}(\Delta \log y_i^*)^2 + h.o.t.,$$

substituting in (17) and (18) concludes the proof. ■

The welfare effects of the quota change in Proposition D2 includes three terms. The first term, which depends on how quota rents change with the quota level, is familiar from Proposition 5 in the main text and captures nonlinearities in the effect of quota changes on real output. The second term depends on how the marginal external cost of pollution changes with real output. For example, if the marginal external cost of pollution increases as real output rises (i.e., $\partial \log \mathcal{W}/\partial \log y_i$ becomes more negative), then deviations in the quota from the optimal quota level are more costly. The third term depends on how the marginal external cost of pollution changes with the pollution level. For example, if money-metric utility is concave with respect to pollution, then deviations from the optimal quota level are more costly. These second and third terms can be measured using surveys or other instruments to capture how the willingness-to-pay to reduce pollution changes with the level of real output and with the level of pollution.

E Rent-Seeking

In this section, we extend our baseline framework to allow for rent-seeking, in which productive resources are wasted in acquiring quota permits. We characterize the effect of quota changes on output with rent-seeking and illustrate our results in a small open economy.

E.1 Setup with Rent-Seeking

For each quota y_i^*, we assume that the government sells permits to engage in the production of good i. The government sets the price of permits at h_{i^*}. Revenues from permit sales are rebated to households lump sum.

There is a unit mass of households, and each household is endowed with one unit of labor that can be devoted to production work or rent-seeking. Hence, the unit mass of available labor is split into labor used for production, L, and rentier labor, $R = 1 - L$. Rentier households expend their labor acquiring quota permits, rather than engaging in production work, and earn rents from licensing these permits to producers.

For each quota i^*, free entry determines the mass of rentier households. Thus, the earnings from becoming a permit owner for activity i^* are equal to wages from production work:

$$\underbrace{\frac{\Pi_i}{R_{i^*}}}_{\substack{\text{Profits}\\ \text{per owner}}} - \underbrace{\frac{h_{i^*} y_i}{R_{i^*}}}_{\substack{\text{Permit costs}\\ \text{per owner}}} = w_L, \tag{19}$$

where R_{i^*} is the mass of rentier households for activity i^*, and w_L is the wage for production labor. Thus, the shares of labor devoted to rent-seeking and production labor are,

$$R = \sum_{i \in \mathcal{I}^*} R_{i^*} = \sum_{i \in \mathcal{I}^*} \max\left\{0, \frac{\Pi_i - h_{i^*} y_i^*}{w_L}\right\}, \qquad \text{and} \qquad L = 1 - R.$$

We denote the total profits of permit owners for sector i in excess of government permit costs by $\Pi_i^{\text{excess}} = \Pi_i - h_{i^*} y_i^*$.

Given quotas y_i^* and permit prices h_{i^*}, an equilibrium is a set of prices p_i, factor wages w_f, outputs y_i, final demands c_i, intermediate and factor input choices x_{ij} and L_{if}, and labor allocations L and R_{i^*} such that: (1) as before, final demand maximizes the final demand aggregator subject to the budget constraint; each sector minimizes costs; and resource constraints for all goods and factors are satisfied; additionally, (2) free entry for rentier labor in each constrained sector holds; and (3) the sum of production labor and the mass of rentier households is equal to the total mass of households.

E.2 First-Order Effects of Quota Changes with Rent-Seeking

We present results on the first-order effects of quota changes on output in economies with rent-seeking. To begin, we first characterize how the share of rentier labor depends on the quota permit prices.

Lemma 1 (Permit prices and rentiers)**.** *The share of rentier households for quota y_i^* depends on whether the permit price $h_{i^*} \lesseqgtr \Pi_i / y_i^*$.*

1. *If the permit is **correctly priced** ($h_{i^*} = \Pi_i / y_i^*$), then $R_{i^*} = 0$.*
2. *If the permit is **under-priced** ($h_{i^*} < \Pi_i / y_i^*$), then the share of households that are rentiers for i^* is*
$$R_{i^*} = \frac{\Pi_i^{excess}}{w_L L + \sum_{k \in \mathcal{I}^*} \Pi_k^{excess}}.$$
3. *If the permit is **over-priced** ($h_{i^*} > \Pi_i / y_i^*$), output of sector i in equilibrium is $y_i < y_i^*$, and the equilibrium is equivalent to implementing a correctly priced, stricter quota y_i.*

Whether a positive share of households become rentiers for a quota y_i^* depends on whether permit prices are set above or below a threshold, Π_i / y_i^*. Intuitively, when $h_{i^*} y_i^* = \Pi_i$, rents earned by permit owners are exactly offset by the costs of obtaining a permit. Hence, households are indifferent between owning a permit and not, and there is no loss in the supply of production labor.[30]

If the permit price is below this threshold, the share of households that become rentiers is proportional to the profits made by sector i in excess of permit costs. The higher these excess profits, the more households must become rentiers to equate rents per owner with production work wages. Relative to when permits are correctly priced, output is lower when permits are under-priced due to the loss in production labor.

Finally, when the permit price is above the threshold, the profits from engaging in the constrained activity are lower than the costs of obtaining a permit to do so. Hence, the level of the activity must drop to some level $y_i < y_i^*$ that equates profits with permit costs. If the permit price set by the government is high enough, there may be no level of the

[30]Since this price equates the rents earned from the permit with its cost, this is also the price that would obtain if the government auctioned off the permit. Note that the government may also be able to achieve the same result of no loss in production labor by using a different mechanism to allocate permits, such as assigning permits by random lottery or exogenously to some subset of households.

activity y_i at which profits and permit costs are equated, in which case the permit cost is equivalent to shutting down the market for i.

Since an over-priced permit can always be re-expressed as a correctly priced permit at a different quota level, we assume without loss in the following results that all permits are under-priced or correctly priced. With these results in place, we characterize the first-order effect of changes in quotas and permit costs on output in Proposition E3.

Proposition E3 (First-order effects with rent-seeking)**.** *Suppose all permits are under-priced or correctly priced. The change in output resulting from changes in quotas y_i^* and permit costs h_{i^*} is*

$$d \log Y = \sum_{i^*} \Pi_i d \log y_i^* + \Lambda_L d \log L,$$

where the change in production labor $d \log L$ is

$$d \log L = R d \log \Lambda_L - \sum_{i^*} R_{i^*} d \log \Pi_i^{excess}. \tag{20}$$

and where $d \log \Lambda_L$ and $d \log \Pi_i^{excess}$ are changes in production labor income and in excess profits.

The effect of a change in a quota on output consists of a direct effect and an indirect effect. The direct effect of the change in the quota on output is $\Pi_i d \log y_i^*$ and is exactly equal to the effect of the quota change in an economy without rent-seeking (Proposition 2). The indirect effect of the quota on output depends on how the quota affects the supply of production labor, which in turn depends on changes in the share of income going to production labor versus excess profits. If excess profits increase relative to labor income, then the profitability of being a rentier is increasing relative to production labor, and more households to opt out of production work. Conversely, if labor income rises relative to excess profits, the supply of production labor increases. In both cases, the change in the quota thus has an additional effect on output by changing the supply of production labor.

Unlike quotas, changes in permit costs $d \log h_{i^*}$ do not directly affect output (provided that permits are not over-priced). However, changes in permit costs can affect output indirectly by changing excess profits, and thus influencing the supply of production labor. In particular, an increase in permit costs decreases the excess profits available to rentiers, and hence increases the labor available for production work.

We focus on two special cases of Proposition 2, where permits are always correctly priced or always free. These two limiting cases reflect the extremes where changes in profits are completely dissipated by entry of rentier households or are cleared by changes in permit prices. Corollary E1 takes the case where all permits are correctly priced, and Corollary E2 takes the case where all permits are free.

Corollary E1 (Comparative statics with correctly priced permits). *Suppose permits are always correctly priced. Then, quota changes do not affect production labor, and the effects of quota changes on output are given by Proposition 2.*

Corollary E2 (Comparative statics with free permits). *Suppose all permits are free ($h_{i^*} = 0$ for all i^*). Then, the changes in output resulting from changes in quotas y_i^* are*

$$d \log Y = \sum_{i^*} \Pi_i d \log y_i^* + \left(R d\Lambda_L - \sum_{i^*} L d\Pi_i \right).$$

If labor is the only factor, then $d \log Y = \sum_{i^*} \Pi_i d \log y_i^* - \sum_{i^*} d\Pi_i$. If profits for all sectors are initially zero, then $d \log Y = -\sum_{i^*} d\Pi_i$.

When permits are correctly priced, permit costs exactly offset profits, and so households allocate all available labor to production work. This means that there are no indirect effects of quota changes on production work. Thus, the effect of a quota change on output is limited to the direct effects characterized in Proposition 2.

In contrast, when permits are free, changes in quotas lead to changes in profits, which lead to entry or exit of households into rent-seeking. Thus, in addition to their direct effect on output, quota changes indirectly affect output by changing the supply of production labor. These indirect effects are non-zero even when quota profits are initially zero. Corollary E2 shows that when quotas are just-binding, tightening a quota has a first-order, negative effect on output.

Example 12 (Small Open Economy). Consider the small open economy from Example 1. We compare the effect of changes in the import quota y_f^* on welfare when permits are correctly priced (i.e., there is no rent-seeking) or free.

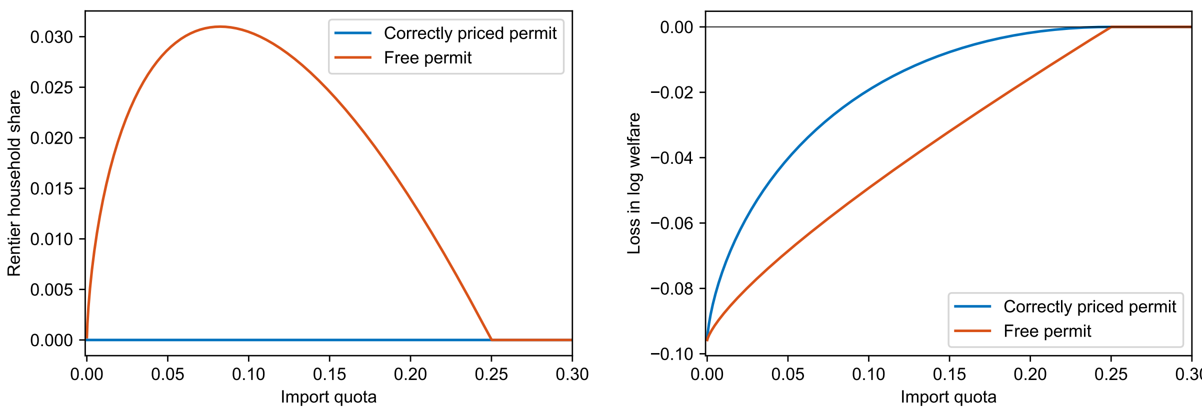
Applying Corollary E1 and Corollary E2 yields:

$$\frac{d \log Y}{d \log y_f^*} = \Pi_f, \qquad \text{(Without rent-seeking)}$$

$$\frac{d \log Y}{d \log y_f^*} = \Pi_f + \underbrace{\frac{\lambda_f - \Pi_f \left(\lambda_f + \theta \left(1 - \lambda_f \right) \right)}{\lambda_f + \theta \left(1 - \lambda_f \right)}}_{\text{Effect of change in production labor}}. \qquad \text{(With rent-seeking)}$$

When import permits are correctly priced, the elasticity of welfare to the import quota is equal to the quota profits (i.e., the government revenues from selling permits). When import permits are instead free, a change in the import quota also affects welfare by changing the supply of production labor. This change in the supply of production labor

Figure 11: Effect of import quota on share of rentier households and welfare.



(a) Rentier households.

(b) Welfare.

depends in turn on how the excess rents earned by permit owners change with the quota. Given a foreign expenditure share λ_f, welfare is less elastic to changes in the quota when the Armington elasticity θ is high. Intuitively, the ability for households to substitute from the foreign good to the domestic good restricts the ability of import–export firms to make large profits and thus limits the extent to which households forego production work to become rentiers.

Figure 11 illustrates the effects of the import quota on the share of rentier households and welfare. We choose an Armington elasticity of $\theta = 4$ and $A_f = 1$, and we choose ω so that the unconstrained expenditure share on imports is 0.25. When permits are correctly priced, all labor is used for production work regardless of the level of the import quota. Moreover, starting at the point where the import quota is just-binding, marginal changes in the quota have no first-order effect on welfare.

In contrast, when permits are free, starting at the point where the import quota is just-binding, a small reduction in the import quota leads some households to reallocate their labor toward rent-seeking, resulting in a loss in production labor and a first-order decline in welfare. As the import quota is reduced further, welfare declines and the share of rentier households initially grows. However, at some point the import quota becomes so tight that total profits of import–export firms falls (even though profits per unit of the foreign imported good rises). In the limit with autarky, import–export firms have no profits, and hence the level of welfare is the same regardless of how permits are priced.