Rationing Under Sticky Prices

Tom D. Holden, Deutsche Bundesbank* 11/12/2024

Abstract: Following the Covid pandemic, the Suez Canal blockage and the Ukrainian war, many goods experienced stockouts and delivery delays. But if prices are flexible, then production cost increases pass through to prices, and all goods remain available. Only if prices are sticky might firms ration demand through stockouts or delivery delays, to avoid selling goods at a price below marginal cost. However, the standard assumption in solving sticky price models is that firms sell the entire quantity demanded at their price. This paper investigates the consequences of allowing firms to ration under sticky prices, in a continuous time model with idiosyncratic demand shocks and endogenous price rigidity. Rationing helps the model match empirical results from both micro & macro data. It produces a convex, backward bending Phillips curve, yet lower monetary non-neutrality and significantly higher optimal inflation.

Keywords: rationing, Phillips curve, inflation, New Keynesian.

JEL codes: E31, E52, D45

PRELIMINARY AND INCOMPLETE

Latest version and slides available at https://www.tholden.org/.

* Address: Mainzer Landstraße 46, 60325, Frankfurt am Main, Germany.
E-mail: thomas.holden@gmail.com. Website: https://www.tholden.org/.
The views expressed in this paper are those of the author and do not represent the views of the Deutsche Bundesbank, the Eurosystem or its staff.
The author would like to thank Klaus Adam, Justin Bloesch, Corina Boar, Richard Dennis, Keshav Dogra, Mike Golosov, Marcus Hagedorn, Cosmin Ilut, Peter Karadi, Campbell Leith, Vivien Lewis, Jochen Mankart, Giovanni Nicolò, Frank Schorfheide, Luminita Stevens, Harald Uhlig, Henning Weber, Nils Wehrhöfer and Elisabeth Wieland for helpful comments.

1 Introduction

Economies worldwide ground to a halt under supply constraints in the early 2020s. Covid restrictions prevented many people from working. The Suez Canal was blocked by the ship Ever Given, preventing goods from reaching Europe from Asia. The Russian invasion of Ukraine led to the end of Russia's gas exports to Europe. These supply constraints were accompanied by high inflation, stockouts in some consumer goods (Cavallo & Kryvtsov 2023) and delivery delays for goods such as cars.[1]

Stockouts and delivery delays are both forms of rationing, as they are both ultimately a choice of the supplier. While supply disruptions increase marginal costs, still marginal costs remain finite. If a firm desperately wanted a production input while the Suez Canal was blocked, they could have put it in an airplane instead. If car manufacturers really wanted microchips delivered in 2022 rather than 2023, they could have offered semiconductor manufacturers high enough prices to get them to switch from producing chips for GPUs and mobile phones. Instead, they sold consumers the substitute good "car-in-2023" instead of the good "car-in-2022" they were ideally looking for.

Firms had another choice though. They could have raised prices. If prices had risen with the increase in marginal costs, then all goods would have remained available. Consumers who were prepared to pay could still have obtained the goods they wanted. Thus, sticky prices are essential for supply disruptions to lead to stockouts or other forms of rationing.

Rationing is also common in normal times. Over 10% of all goods are out of stock in normal times, according to the evidence of Cavallo & Kryvtsov (2023). This paper builds a dynamic model of rationing under sticky prices to understand the implications of rationing for monetary policy and the broader macroeconomy.

Unfortunately, prior dynamic models of sticky prices have all been solved

[1] See e.g. https://www.thedrive.com/news/new-cars-piling-up-at-german-port-will-mean-longer-wait-for-us-buyers, https://www.cnbc.com/2021/05/07/chip-shortage-is-starting-to-have-major-real-world-consequences.html, or https://www.thisismoney.co.uk/money/cars/article-11831443/How-long-wait-new-car-delivered-revealed.html.

under the simplifying assumption that firms satisfy all demand at their posted price, even if that results in them selling at a price below marginal cost. This is true for both Calvo, Rotemberg and menu-cost approaches to modelling price rigidities. While tractable, this seems deeply implausible.

If a firm cannot adjust their nominal price, then their real price will be declining over time. A lower real price implies higher demand for their good, and so higher sales. With short-run decreasing returns to scale, higher sales in turn means higher real marginal costs. So, the firm's real price is declining, while their real marginal cost is increasing. If the price remains fixed, eventually the firm's marginal cost will equal or exceed its price. No firm would want to continue to sell their good in this state. Instead, they would ration demand, only selling up to the quantity at which price equals marginal cost.

Does rationing really matter in practice? I will present new retail scanner data evidence that supports the ubiquity of rationing, but a simple back of the envelope calculation is also instructive. Perhaps one reason the prior literature has been happy to rule out rationing is that they have had a misleading calculation in mind: *"Mark-ups are 10%, inflation is 2%, prices are updated at least once per year, real prices will not hit marginal cost."* But this is not the right calculation when firms face short-run decreasing returns to scale. The estimates of Abraham et al. (2024) using data from Belgian firms imply that around $\frac{1}{3}$ of all labour and intermediate inputs are fixed at annual frequency, implying a total share of fixed inputs in production, α, of around $\frac{5}{9}$.[2] Thus, firm marginal costs are roughly proportional to $y^{\frac{\alpha}{1-\alpha}} = y^{\frac{5}{4}}$, where y is their output. Meanwhile, firms face demand proportional to $\left(\frac{p}{P}\right)^{-\epsilon}$, where p is their nominal price, P is the price level, and $\epsilon \approx 10$ in standard calibrations. So, if the price level increases by 2% (over a year, say), but the firm's nominal price stays fixed, then

[2] From Table 3, column (3) or (4) of Abraham et al. (2024), we see that we cannot reject that the share of all capital inputs that are fixed is 100% at a 1% (or lower) significance level, and we cannot reject that the shares of all labour or intermediate inputs includes that are fixed are both 33% at a 10% (or lower) significance level. Ignoring intermediates, with a capital share of $\frac{1}{3}$, this gives a total fixed share in production of $1 \times \frac{1}{3} + \frac{1}{3} \times \frac{2}{3} = \frac{5}{9}$. Boehm, Flaaen & Pandalai-Nayar (2019) find that intermediates are perfect complements to other inputs, so given their fixed share $\frac{1}{3}$ is less than $\frac{5}{9}$, we are justified in taking $\frac{5}{9}$ as the overall fixed share.

firm sales increase by $2\% \times 10 = 20\%$, which means marginal costs increase by $\frac{5}{4} \times 20\% = 25\%$. A 25% rise in marginal costs is more than enough to erode standard calibrations of firm level mark-ups. Thus, we should expect firms with one year old prices to be rationing.

This simple calculation is likely to understate firms' incentives to ration. Firstly, firms face high frequency fluctuations in demand. At times of high demand, marginal costs will be high, making rationing more tempting. Secondly, inflation can be much higher than 2%. It was 9% over the period from June 2021 to June 2022 in the U.S..[3] With 9% inflation and a fixed nominal price, it would take less than a quarter for marginal costs to have risen by 25%. Thirdly, demand is also growing over time due to aggregate income growth. Even holding wages fixed, 2% demand growth implies 2.5% increase in marginal costs. Finally, marginal costs are increasing over time due to irregular replacement of broken machines, and imperfect maintenance. Firms face non-convex adjustment costs in new investment (Cooper & Haltiwanger 2006; Khan & Thomas 2008) and maintenance rates are below depreciation rates (Kabir, Tan & Vardishvili 2024).[4] Thus, in between installations of new machines, capital stocks will be declining and marginal costs will be increasing.[5]

A natural question is why firms with price near marginal cost do not just

[3] https://fred.stlouisfed.org/series/CPIAUCSL.

[4] Kabir, Tan & Vardishvili (2024) find that annual maintenance expenditure is around 6.2% of the value of the capital stock, while their (caveated) estimate of annual depreciation is around 9.4% of the value of the capital stock.

[5] How much on average capital stocks are decreasing over the life of a price will depend on just how often firms make significant capital investments, and how correlated these times are with price change times. It seems natural to suppose that any firm going to the significant trouble of installing new machines would also take the much smaller step of updating its price at the same time. Using data extracted from Figure 1 of Cooper & Haltiwanger (2006) reveals that in any year, around 57% of all firms do not invest enough to cover depreciation (6.9% in their data) plus 2% growth, and 49% of all firms do not invest enough to cover just depreciation. This suggests that firms increase their capital stock less often than they update prices. (The price adjustment estimates of Blanco et al. (2024b) imply around 24% of firm prices last for at least a year.) This is consistent with net investments being accompanied by price changes.
Adam & Weber (2019) stress declining firm marginal costs over the firm life cycle. This is not inconsistent with rising marginal costs over the life of a price if productivity improvements (perhaps brought about by the installation of new machines) are accompanied by price changes.

update their price to restore their mark-up.[6] In a Golosov Lucas (2007) menu cost economy, with constant returns and no micro or macro uncertainty, it is clear that paying the menu cost is always optimal for a firm with price equal to marginal cost. Then, waiting t weeks to change prices is dominated by changing prices now but setting a higher price such that in t weeks your price is what you would have set had you waited.

However, any micro or macro uncertainty can destroy this result. Once there is uncertainty, it can be optimal to tolerate rationing or a price below marginal cost in order to avoid repeated price changes. For example, suppose your price is currently too low, but you expect aggregate or idiosyncratic productivity to improve soon (perhaps due to mean reversion), at which time your current price will be comfortably above marginal cost. Modern menu cost models rely on random menu costs (Dotsey, King & Wolman 1999) and free price change opportunities (Nakamura & Steinsson 2010) to match the micro data, so in these models there is an even greater incentive to temporarily tolerate rationing or a price below marginal cost. Maybe now the menu cost is high, but next period it could be much lower. At the risk of oversimplifying, modern menu cost models work hard to look more like a Calvo model, and in a Calvo model, many firms get stuck with price below marginal cost. For example, price change hazard functions appear flat (Klenow & Kryvtsov 2008; Nakamura & Steinsson 2008; Klenow & Malin 2010), so old prices (with a higher probability of being lower than marginal cost) are no more likely to be adjusted than new prices.

This also means that models that allow for rationing will be consistent with much lower price adjustment frictions than models that do not. In a model without rationing, firms risk substantial losses if they do not adjust their price. To match the data in which they do not adjust their price despite this, the price adjustment frictions must be large. In a model with rationing though, the firm can always guarantee weekly positive profits no matter how old its price is, thus smaller adjustment frictions are needed to match the observed low frequency of price adjustment. If your prior is that adjustment frictions, like menu costs, are small, then you should place greater posterior weight on models with

[6] A version of this point was made in Barro (1977).

rationing, such as the one I present in this paper in Section 3.

My basic model is in continuous time, with Calvo-type price rigidity.[7] In my preferred variant of the model, firms are owned by conglomerates, who can choose the arrival rate of price adjustment opportunities for the firms they manage, following Blanco et al. (2024b). This provides aggregate state dependence, while matching the flat adjustment hazard functions found by Klenow & Kryvtsov (2008), Nakamura & Steinsson (2008) and Klenow & Malin (2010)

At all points in time, firms can freely choose their sales. Optimally, they will meet demand if they can do so with price above marginal cost, otherwise they will just produce up to the point at which price equals marginal cost, rationing demand. To smooth out the kink introduced by this decision, I assume that firms face demand shocks that are independent both across firms and over time. With a carefully chosen density, the model then admits aggregation with a finite dimensional state vector, permitting analytic results and easy simulation. Whereas the standard model without rationing is unstable at high inflation levels, the model with rationing is robustly stable, with reasonable behaviour even under extreme shocks.

I show that the model generates a convex, backward bending Phillips curve, in line with the evidence surveyed in the next section. The convexity emerges from the fact that high demand leads to high rationing. For the same reason, allowing rationing reduces the overall degree of monetary non-neutrality. The model also generates a robustly positive output maximizing inflation level, around 1%, far higher than the near 0% level in the absence of rationing. This result is driven by the low efficiency costs of inflation under rationing. Intuitively, rationing prevents firms with old, highly distorted prices from selling huge quantities, reducing overall misallocation. I provide further intuition for the relatively high productivity of economies with rationing in Section 4 when presenting the results.

The model also matches a range of further empirical evidence presented in

[7] Early continuous time New Keynesian models were developed by Posch, Rubio-Ramírez & Fernández-Villaverde (2011), (2018).

the next section, despite only introducing one new parameter in the basic variant. This evidence includes new evidence from supermarket scanner data on sales over the life of a price, which supports the ubiquity of rationing. Section 5 of the paper considers various extensions of the base model, both to demonstrate robustness of the key conclusions, and to build a quantitative model with which to examine the broader empirical implications of rationing, particularly following supply shocks.

Prior literature. Important early work examining rationing with sticky prices includes Barro & Grossman (1971), Drèze (1975) and Svensson (1984). Barro & Grossman look at outcomes in a one period model when both aggregate output and aggregate labour may be rationed. Drèze examines at equilibrium existence with the possibility of rationing in an Arrow-Debreu setup with price inequality constraints. Svensson looks at rationing in a dynamic monetary model with a single good. A little more recently, Corsetti & Pesenti (2005) worked in a proto-New Keynesian framework with prices set one period in advance, and were careful to restrict their model's shocks to ensure the absence of rationing.

I am aware of three papers that look at rationing in a modern (New Keynesian) setting. Huo & Ríos-Rull (2020) and Gerke et al. (2023) look at the rationing of labour supply that comes from sticky wages, but omit rationing on the price side. These papers both have infinite dimensional state vectors, which makes it challenging to understand all the details of their mechanics. Hahn (2022) looks at rationing under price rigidity in the steady state of a New Keynesian model with Calvo price frictions. While he is able to derive some interesting comparative statics results, his approach is not tractable for looking at dynamics, so he provides no dynamic results. Without idiosyncratic shocks, he also cannot hope to produce an empirically reasonable path of output over the life of a price, even in steady state, as we will see in Subsection 2.1.

Another relevant strand of the literature looks at stockouts in models of inventories. Contributions include Alessandria, Kaboski & Midrigan (2010), Kryvtsov & Midrigan (2013) and Bils (2016). They demonstrate the importance of inventory dynamics for a variety of macro questions. However, in all of these

papers, firms always meet demand if they have stock available, even if the marginal value of that stock to the firm is greater than the price at which they can sell the good. Thus, in these models too, firms would like to ration in some circumstances. For the sake of tractability, my model will not feature inventories, but combining inventories and rationing is a promising avenue for future research.

2 Empirical evidence for rationing

The previous arguments suggest rationing should be widespread. In line with this, Cavallo & Kryvtsov (2023) found that around 11% of all goods in their data were out of stock in 2019, using daily web-scraped data from 70 large retailers in the U.S., Canada, China, France, Germany, Japan and Spain.[8] This may understate the true prevalence of rationing, since retailers can encourage consumers to substitute away from particular goods by, for example, lowering their ranking in search results, worsening their position on physical shelves, or by reducing advertising. Encouraging such substitution helps reduce stockouts, which may provide a reputational benefit for the store. "Shrinkflation" may also mask rationing. If I want 400 grams of cereal, but it is now in sold in 375-gram boxes, I am unlikely to buy two boxes.

Unsurprisingly, Cavallo & Kryvtsov (2023) found that stockouts increased massively during the Covid pandemic. More interestingly though, they found that in 2022 (January to August), still 23% of goods were out of stock.[9] By 2022 many of the direct effects of Covid had subsided, but inflation was picking up worldwide. Thus, in line with the story of the model I will present, it appears that high inflation leads to large amounts of rationing.

I will shortly present further micro-evidence on the prevalence of rationing. In particular, using retail scanner data, I show that quantities sold are concave in the age of a price. Thus, goods with young prices experience relatively high output growth, while goods with older prices experience relatively low output growth. This fits with quantities lying on the demand curve for young prices, with inflation driving real price declines and hence sales increases, and

[8] The number 11% was extracted from Figure 2 of Cavallo & Kryvtsov (2023).

[9] The number 23% was extracted from Figure 2 of Cavallo & Kryvtsov (2023).

quantities lying on the supply curve for older prices, with increasing marginal costs driving ever tighter rationing.

Rationing will also help to explain two important sets of macro facts. Firstly, rationing will help to explain the observed convexity of the Phillips curve. For pre-Covid evidence on this, see, for example, Kumar & Orrenius (2016), Babb & Detmeister (2017) or Forbes, Gagnon & Collins (2022). The fact that the inflation of 2022 was not accompanied by huge output booms provides "natural experiment" evidence in further support of such convexity. Under rationing, such convexity emerges naturally. When demand is already high, further demand increases just lead to increased rationing, rather than increased output. As firms with sufficiently high prices will not ration, increases in rationing tilt the welfare relevant price index towards such highly priced firms, increasing the aggregate price level.

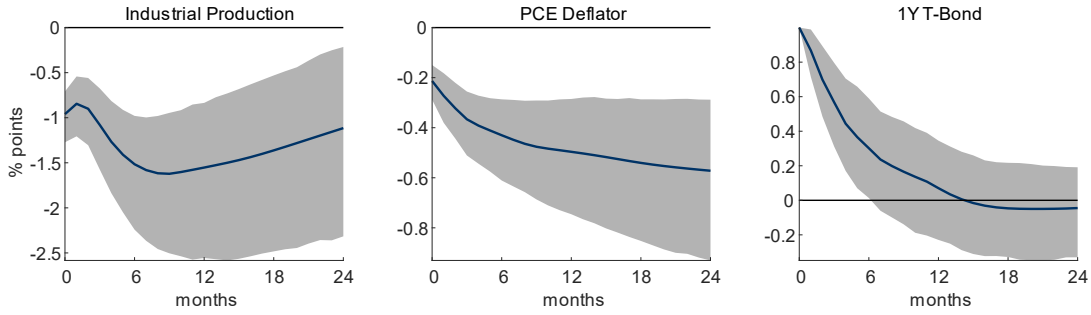


Figure 1: Impulse response to a monetary policy shock. Informationally robust specification from Figure 3 of Miranda-Agrippino & Ricco (2021), but estimated using PCEPI in place of CPI.
95% credible bands highlighted.

Secondly, recent estimates of the response to monetary shocks from Miranda-Agrippino & Ricco (2021) and Bauer & Swanson (2023) suggest that monetary shocks cause an immediate jump in both the price level and output. For reference, Figure 1 plots the impulse response to a monetary policy shock following the informationally robust specification from Figure 3 of Miranda-Agrippino & Ricco (2021), but estimated using PCEPI in place of CPI. Calvo or Rotemberg type models of price rigidity can never generate a jump in the price level following a shock, as the price level is a state variable in these models. By contrast, in a model with rationing the price level is no longer a state variable,

since jumps in levels of rationing cause jumps in the aggregate price index. Jumps in rationing cause jumps in the aggregate price level, since they lead the weight placed on goods with relatively higher prices to jump up, due to rationing of lower price goods.

One challenge to this explanation is that Miranda-Agrippino & Ricco (2021) and Bauer & Swanson (2023) measure the price level with the CPI index, which in the absence of missing data would use essentially fixed weights.[10] In continuous time, a fixed weight index can only jump if a positive measure of firms adjust their price, which never happens in Calvo type models. However, in practice, the CPI data collectors have to deal with many missing prices, for which they then use imputation based on price growth of other items. Stockouts (from rationing) are a major source of missing prices. Thus, the CPI imputation procedure ascribes average price changes from non-rationed goods to rationed goods. Since rationed goods are less likely to have changed price, this produces greater aggregate inflation than the true fixed weight index after an inflationary shock, bringing the CPI index closer to the welfare price index.

While menu cost models can also potentially generate a jump in prices after a shock without rationing, cleanly identified monetary policy shocks are small, and so are unlikely to lead to large amounts of price resetting. For example, Blanco et al. (2024a) calibrate a menu cost model to match both micro price data and the aggregate response of the price change frequency to inflation, and find that 1% increases in the money supply are mostly absorbed by output, not prices, in the short run.

Let me end this section by stressing that this paper is not about a fundamentally different model of price rigidity. Rather, it is about relaxing a simplifying assumption previously used in solving such models. As such, whatever evidence supports your favourite sticky price model will probably also support the same model extended to allow for rationing.

2.1 Evidence from scanner data

I will now present new evidence from micro scanner data to support

[10] Pre-2023 weights were updated biennially, since 2023 they are updated annually.

rationing being widespread. By looking directly at quantities sold, I can measure not only stockouts, but also less direct forms of rationing, such as changes in product placement. I use data from a former chain of Chicago supermarkets called "Dominick's Finer Foods", made freely available by the Kilts Center for Marketing at Chicago Booth.[11] The data covers the period 1989 to 1994, during which time annual PCEPI inflation was between around 2% and around 5%.[12] While newer data is always preferable, supermarket practices have not changed so dramatically in the last thirty years, and the use of open data ensures replicability.

The data records the prices and quantities sold of products from 29 broad categories,[13] from 93 stores, over 399 weeks. The 29 broad categories are further refined into 92 narrower categories.[14] Where possible, I use the item code information provided by the supermarket to match goods which are newer versions of former products. I treat goods at different stores as being distinct. For each good, at each store, I drop the following observations:

- Those with price equal to the first price observed for the good. (We do not observe the start of the first price spell, so we cannot construct price age for those observations.)
- Those with price equal to the final price observed for the good. (Maybe the good disappeared due to changing tastes, in which case the concavity in sales over the span of the final price could reflect demand, not supply.)
- Those with price less than the cumulative maximum price for the good at

[11] https://www.chicagobooth.edu/research/kilts/research-data/dominicks.

[12] https://fred.stlouisfed.org/series/PCEPI.

[13] Analgesics, Bath Soap, Bathroom Tissues, Beer, Bottled Juices, Canned Soup, Canned Tuna, Cereals, Cheeses, Cigarettes, Cookies, Crackers, Dish Detergent, Fabric Softeners, Front-end-candies, Frozen Dinners, Frozen Entrees, Frozen Juices, Grooming Products, Laundry Detergents, Oatmeal, Paper Towels, Refrigerated Juices, Shampoos, Snack Crackers, Soaps, Soft Drinks, Toothbrushes, Toothpastes.

[14] The split into narrower categories was unavailable for "Refrigerated Juices", so I allocated goods in this category into the following eleven narrower categories based on their description field: Orange Juice, Orange Drinks, Apple Juice and Cider, Cranberry Juices and Cranberry Juice Blends, Other Fruit/Vegetable Juices, Fruit Punch and Mixed Fruit Drinks, Lemonade, Iced Tea, Dairy-based Drinks and Shakes, Puddings, Colored Easter Eggs. The CSV file giving the allocation of items to categories is contained in the replication materials for this paper.

that store. (This ensures we are only looking at sales after a price rise, not a price cut. It would be unsurprising if sales initially increased after a price cut. We want to pick up the increase in sales after a price rise coming from inflation eroding real prices. This filter also takes out sales during which demand may be distorted by different advertising levels.)

- Those occurring at the same time as a change in price, or the week after a missing observation (which could have hidden a change in price). (Keeping observations the period of a price change could be a source of endogeneity, due to the same demand shock influencing both quantities sold and the decision to change prices.)
- Those with a price age greater than four years. (There are relatively few prices that ever last so long. Including them would reduce estimation reliability due to the use of average output over the life of a price in my regression specification.)

I estimate the following linear model for quantity sold as a function of the age of the price:

$$\frac{y_{i,j,t} - y_{i,j,t-1}}{\bar{y}_{i,j}} = \beta_{A(i,j,t)} + \gamma_{i,t} + \sigma^{(1)}_{i,A(i,j,t)}\sigma^{(2)}_{i,j}\sigma^{(3)}_{i,t}\varepsilon_{i,j,t}.$$

Here, i indexes narrow category, store pairs (92 narrow categories × 93 stores = 8,556 narrow category, store pairs). j indexes product, price pairs, of which there are 947,660 (the same product receives a different j in two periods if its price differs). t indexes time in weeks. $A(i,j,t)$ is the age in weeks of the j^{th} product-price from category-store i at t,[15] and $y_{i,j,t}$ is the number of units sold of this item, that week. $\bar{y}_{i,j}$ is the average of $y_{i,j,t}$ over the life of the price.

The left-hand side of this specification gives a measure of sales growth that is robust to the presence of zeros in $y_{i,j,t}$. Working in differences, not levels, ensures consistency even when products experience $I(1)$ demand shocks, due to entry or exit of substitute products, for example. On the right-hand side, $\beta_{A(i,j,t)}$ gives age fixed effects, our prime variable of interest. $\gamma_{i,t}$ gives category-

[15] For goods without missing observations, new prices start with age one (assuming that the price change occurred at the end of the previous week), so the first observed age will be two, as one week is dropped due to the price change. For goods with some missing observations, we renormalize ages so that the first included observation is age two.

store-time fixed effects to mop up changes in demand for specific category types in specific locations at specific times (think of the demand for candy around Halloween, concentrated in family neighbourhoods). I model heteroskedasticity in the residual by category-store combined (separately) with age, product-price and time. This substantially improves the efficiency of my estimates.

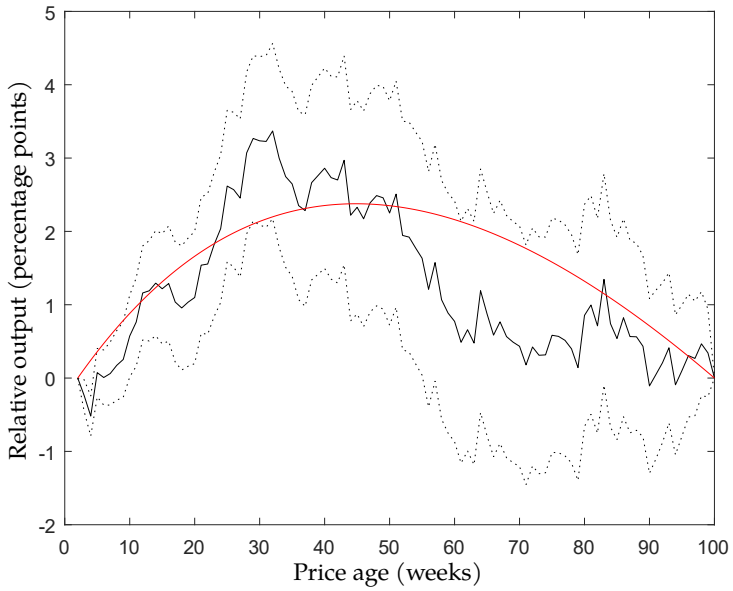


Figure 2: Average output over the life of a price ($100 \sum_{a=3}^{\text{AGE}} \beta_a$).
The effect is identified up to a linear trend, so I normalize to zero at ages 2 and 100.
The black solid line gives the estimates. The dashed lines give 99% confidence bands.
The red line gives the prediction of the model from Section 3.[16]

After differencing, I am left with 21,474,126 observations. Estimating the model on these observations by feasible generalized least squares gives the estimates summarized in Figure 2. This figure plots $100 \sum_{a=3}^{\text{AGE}} \beta_a$ as a function of AGE in the black solid line. I.e., it plots the average level of sales over the life of a price. Due to the category-store-time fixed effects, this is only identified up to a linear trend, so the plot is normalized so that the impact is zero for age 2 and age 100. The dashed lines give 99% confidence bands, constructed with three-way clustered standard errors (Cameron, Gelbach & Miller 2011), with

[16] In the notation of equation (3) from Section 3 this is $100 \log y_{\tau,t}$, detrended to be 0 at 2 and 100 weeks.

groups indexed by category-store combined (separately) with age, product-price and time, so with indices $(i, A(i,j,t))$, (i,t) and (i,j). The first grouping allows for heterogeneity in the effects of age across categories and stores. The second allows for time-varying correlation between the residuals of all products in a category and store. The third allows for arbitrary correlation across time for the residuals from any particular product and price.

We see that relative to the normalization, sales grow for around 30 weeks, before starting to decline. This is consistent with firms rationing demand for products with old prices. While a good's nominal price is fixed, its real price is declining, leading to higher sales. But with decreasing returns, higher sales mean higher marginal costs. Eventually, marginal costs are higher than prices, so the firm rations demand. Under rationing, sales are a decreasing function of real price (due to decreasing returns again), so sales are then declining in price age. The result is that sales are a concave function of price age, as we see here. Without any rationing, log-sales would be linear in firm age (as long as demand is roughly isoelastic), so after normalizing we would not find any statistically significant difference from zero.[17]

The red line in Figure 2 plots the prediction of the basic model I will present in Section 3. This is not a calibration target of the model, so it is reassuring how well the model performs. You might be surprised by the small size of the predicted effect of price age on sales, though. After all, if the price elasticity of demand is -10, then with 2% inflation, over 30 weeks sales should have increased by over 11% without rationing. If firms started rationing from week 30 on, then that would still imply a normalized peak impact of over 7.5%.[18] However, my model is one in which firms face idiosyncratic demand shocks that are independent across time. These shocks mean that for any price age, a firm's expected sales is a mix of their sales when their demand shock is high, so they ration, and their sales when their demand shock is low, so they meet demand. This reduces the sensitivity of average sales to price age, matching the

[17] Standard calibrations of Kimball (1995) demand can generate concavity in sales over the life of a price, without any rationing.

[18] Something like this would be true in the Hahn (2022) model, for example.

data. A model without idiosyncratic demand shocks would predict implausibly high average sales growth for young prices.

3 The basic model

I will now present my basic model of rationing under sticky prices. Throughout the paper, I stick to the convention that upper case letters denote aggregate variables, while lowercase Latin letters denote firm specific variables. The model is in continuous time, with time measured in years throughout. Letters without time subscripts denote steady-state values. For simplicity, there is no aggregate uncertainty: I will only look at the impact of prior probability zero "MIT" shocks.

3.1 Firms and aggregators

The model will feature a continuum of firms of measure one. Firms are only able to adjust their price when they are hit by a shock from a non-homogenous Poisson process. In particular, price change opportunities arrive at time t with rate $\lambda_t > 0$, where $\int_{-\infty}^{t} \lambda_v \, \mathrm{d}v = \infty$ for all t. As a result, the time t density of firms that last adjusted their price at time τ is given by $\lambda_\tau e^{-\int_\tau^t \lambda_v \, \mathrm{d}v}$. Note that, as required $\int_{-\infty}^{t} \lambda_\tau e^{-\int_\tau^t \lambda_v \, \mathrm{d}v} \, d\tau = \int_{-\infty}^{t} \frac{\mathrm{d}}{\mathrm{d}\tau} e^{-\int_\tau^t \lambda_v \, \mathrm{d}v} \, d\tau = 1 - e^{-\int_{-\infty}^t \lambda_v \, \mathrm{d}v} = 1$. I index firms with the time at which they last updated their price τ, so this density will appear frequently.

Firms will face demand shocks that are independent both across firms, and across time t. This means that over even an arbitrarily small interval of time, a firm will face all possible values of the demand shock. I write $y_{\zeta,\tau,t}$ for the output of a firm at time t, that last updated their price at time τ, that is hit by a demand shock of level $\zeta \in [0,1]$. Demand shocks ζ will be drawn from a Beta$(\theta, 1)$ distribution, where $\theta > 0$, meaning they have probability density function $g(\zeta) = \theta \zeta^{\theta-1}$. This implies the mean of the demand shock is $\frac{\theta}{\theta+1}$ and the variance of the demand shock is $\frac{\theta}{(\theta+1)^2(\theta+2)} \approx \frac{1}{\theta^2}$. Demand shocks are essential for tractability as they smooth out the kink introduced by the rationing decision. This particular distribution for the demand shocks is needed for the model to have a finite dimensional state. θ is the only non-standard parameter in the entire model. I will calibrate it to match the evidence from Cavallo &

Kryvtsov (2023) that around 11% of all goods are rationed in normal times.

The aggregate good Y_t is produced by a competitive industry of "aggregators" with access to the technology:

$$Y_t = D^{-\frac{\epsilon}{\epsilon-1}} \left[\int_{-\infty}^{t} \lambda_\tau e^{-\int_\tau^t \lambda_v \, \mathrm{d}v} \int_0^1 \zeta \, y_{\zeta,\tau,t}^{\frac{\epsilon-1}{\epsilon}} \, g(\zeta) \, \mathrm{d}\zeta \, \mathrm{d}\tau \right]^{\frac{\epsilon}{\epsilon-1}}. \tag{1}$$

Here, $\epsilon > 1$ is the elasticity of substitution across varieties, and $D = \frac{\theta}{\theta+1}$ is a scale factor chosen to ensure that if $y_{\zeta,\tau,t}$ is one for all ζ and τ, then $Y_t = 1$. This aggregator is essentially the standard Dixit-Stiglitz one. The only changes are the weighting by the density of firms that last updated at time τ, and the inner integral over the possible draws of the demand shock. Demand is higher for varieties receiving a higher draw of ζ. To understand the inner integral, you should think of there being a positive measure of firms that last updated their price at time τ. Of these infinitely many firms, a density $g(\zeta)$ will receive demand shock ζ at time t.

Like normal, aggregators choose their input quantities to maximize their profits:

$$P_t Y_t - \int_{-\infty}^{t} \lambda_\tau e^{-\int_\tau^t \lambda_v \, \mathrm{d}v} p_\tau \int_0^1 y_{\zeta,\tau,t} \, g(\zeta) \, \mathrm{d}\zeta \, \mathrm{d}\tau,$$

where P_t is the aggregate price, and p_τ is the price of all varieties that last updated their price at time τ.[19] In doing so, they face the supply constraints $y_{\zeta,\tau,t} \leq \bar{y}_{\zeta,\tau,t}$ for all ζ, τ and t. The sales limits $\bar{y}_{\zeta,\tau,t}$ will be chosen by firms. The first order conditions of this problem imply that firms face the demand constraint:

$$y_{\zeta,\tau,t} \leq \left(\frac{D}{\zeta} \frac{p_\tau}{P_t} \right)^{-\epsilon} Y_t. \tag{2}$$

Demand places an upper bound on firm sales, not a lower bound.

Unlike in a standard model, aggregators will make profits when rationing is allowed. The presence of sales limits mean that the aggregators face decreasing returns to scale, and so positive aggregator profits are consistent with perfect competition. Another way to see this is to note that the true price index would integrate over a sum of the actual price of goods, and the Lagrange

[19] We are assuming here that all firms updating their price at the same time will choose the same price. This will be true in equilibrium.

multipliers on the sales limits, but aggregators do not "pay" the Lagrange multipliers, resulting in profit.

Firms produce output using the decreasing returns to scale production function:

$$y_{\zeta,\tau,t} = v_{\zeta,\tau,t}^{1-\alpha}, \text{ where } v_{\zeta,\tau,t} = A_t l_{\zeta,\tau,t}.$$

Here, $v_{\zeta,\tau,t}$ is their effective labour input, $l_{\zeta,\tau,t}$ is their actual labour input, $A_t > 0$ is aggregate productivity and $\alpha \in (0,1)$ is the fixed share in production. The use of letter v for the effective labour input anticipates the extended model in which $v_{\zeta,\tau,t}$ will be a bundle of variable inputs. Labour will be supplied at the aggregate wage W_t. For convenience, we define the wage of effective labour by $\widehat{W}_t := \frac{W_t}{A_t}$.

Firms' flow of real production profits is given by:

$$o_{\zeta,\tau,t} = \frac{p_\tau}{P_t} y_{\zeta,\tau,t} - \widehat{W}_t v_{\zeta,\tau,t} = \frac{p_\tau}{P_t} v_{\zeta,\tau,t}^{1-\alpha} - \widehat{W}_t v_{\zeta,\tau,t}.$$

I assume firms can choose how much to produce at all points in time, after learning their demand shock. Thus, $v_{\zeta,\tau,t}$ (or $l_{\zeta,\tau,t}$) is a choice variable for the firm. Note that no matter the price p_τ, $o_{\zeta,\tau,t} = 0$ if $v_{\zeta,\tau,t} = 0$, but:

$$\frac{\mathrm{d}\tilde{o}_{\zeta,\tau,t}}{\mathrm{d}v_{\zeta,\tau,t}} = (1-\alpha)\frac{p_\tau}{P_t} v_{\zeta,\tau,t}^{-\alpha} - \widehat{W}_t \to \infty$$

as $v_{\zeta,\tau,t} \to \infty$. Thus, the firm can always ensure positive production profits by choosing a small enough $v_{\zeta,\tau,t}$. A small enough $v_{\zeta,\tau,t}$ will also satisfy the firm's demand constraint, (2), and hence the firm will always make strictly positive profits, and will always choose $v_{\zeta,\tau,t} > 0$ so $y_{\zeta,\tau,t} > 0$.

Firms choose $v_{\zeta,\tau,t}$ to maximize $o_{\zeta,\tau,t}$ subject to the demand constraint, (2). In Appendix A I show that this leads them to choose:

$$v_{\zeta,\tau,t} = \min\left\{\left[\left(\frac{D}{\zeta}\frac{p_\tau}{P_t}\right)^{-\epsilon} Y_t\right]^{\frac{1}{1-\alpha}}, \left(\frac{p_\tau}{P_t}\frac{1-\alpha}{\widehat{W}_t}\right)^{\frac{1}{\alpha}}\right\},$$

so:

$$y_{\zeta,\tau,t} = \min\left\{\left(\frac{D}{\zeta}\frac{p_\tau}{P_t}\right)^{-\epsilon} Y_t, \left(\frac{p_\tau}{P_t}\frac{1-\alpha}{\widehat{W}_t}\right)^{\frac{1-\alpha}{\alpha}}\right\}.$$

In both of these expressions, the first term in the curly brackets gives the outcome without rationing, in which firms meet demand. In this case, price is above marginal cost, and sales are decreasing in the good's real price. The

second term in the curly brackets in these expressions gives the outcome with rationing. In this case, price equals marginal cost, and sales are increasing in the good's real price. Note that the firm can calculate their maximum output in advance of the realisation of the shock. Thus, rationing does not require the firm to possess implausible amounts of information.

High values of ζ mean higher demand, and so make rationing more likely. To be specific, define:

$$\bar{\zeta}_{\tau,t} := D\left(\frac{p_\tau}{P_t}\right)^{1+\frac{1-\alpha}{\epsilon\alpha}}\left(\frac{1-\alpha}{\widehat{W}_t}\right)^{\frac{1-\alpha}{\epsilon\alpha}} Y_t^{-\frac{1}{\epsilon}},$$

then the firm will always ration if $\zeta > \bar{\zeta}_{\tau,t}$, and the firm will never ration if $\zeta < \bar{\zeta}_{\tau,t}$. High values of $\bar{\zeta}_{\tau,t}$ mean that rationing only takes place with extreme draws of the demand shock, whereas low values of $\bar{\zeta}_{\tau,t}$ mean rationing is likely. Increases in aggregate demand Y_t reduce $\bar{\zeta}_{\tau,t}$, increasing the chance of rationing. Likewise, when effective wages $\widehat{W}_t$ are high, so marginal costs are high, rationing is likely. Finally, note that having a high real price makes rationing less likely.

In the limit as $\lambda_t \to \infty$ for all t, the model tends to one with quasi-flexible prices. In this limit, firms continuously adjust their prices, but still set prices at t before the realisation of their time t demand shock. I show in Appendix A that firms still ration with positive probability in this limit (i.e., $\bar{\zeta}_{\tau,t} \leq 1$), as long as $\theta \leq \frac{\alpha\epsilon-1}{1-\alpha}\epsilon$, which will hold in any reasonable calibration. Thus, we should also expect $\bar{\zeta}_{\tau,t} \leq 1$ when $\lambda_t < \infty$ and prices are sticky, meaning there is rationing for at least some values of the demand shock. In all numerical exercises I will check that $\bar{\zeta}_{\tau,t} \leq 1$ for all τ and t.

Returning to the general case with $\lambda_t < \infty$, and assuming that $\bar{\zeta}_{\tau,t} \leq 1$, a firm's expected output before the demand shock is realized is:[20]

$$\begin{aligned} y_{\tau,t} &:= \int_0^1 y_{\zeta,\tau,t}\, g(\zeta)\, \mathrm{d}\zeta \\ &= \left(\frac{1-\alpha}{\widehat{W}_t}\frac{p_\tau}{P_t}\right)^{\frac{1-\alpha}{\alpha}} - \frac{\epsilon}{\theta+\epsilon} D^\theta Y_t^{-\frac{\theta}{\epsilon}}\left(\frac{1-\alpha}{\widehat{W}_t}\right)^{\frac{\theta+\epsilon 1-\alpha}{\epsilon \alpha}}\left(\frac{p_\tau}{P_t}\right)^{\theta+\frac{\theta+\epsilon 1-\alpha}{\epsilon \alpha}}. \end{aligned} \tag{3}$$

This has a part that is increasing in the good's real price and a part that is

[20] See Appendix A for derivations of this and subsequent results.

decreasing. The combination of the two gives log-concavity in $\frac{p_\tau}{P_t}$, generating the concave log-sales over the life of a price that we already plotted in the red line of Figure 2.[21]

Again, assuming that $\bar{\zeta}_{\tau,t} \leq 1$,[22] a firm's expected profits before the realization of the demand shock is given by:

$$\begin{aligned} o_{\tau,t} &:= \int_0^1 o_{\zeta,\tau,t}\, g(\zeta)\, \mathrm{d}\zeta \\ &= \alpha \left(\frac{1-\alpha}{\widehat{W}_t}\right)^{\frac{1-\alpha}{\alpha}} \left(\frac{p_\tau}{P_t}\right)^{\frac{1}{\alpha}} \\ &\quad - \frac{\epsilon}{\theta+\epsilon} \frac{\epsilon\alpha}{(1-\alpha)\theta+\epsilon} D^\theta \left(\frac{1-\alpha}{\widehat{W}_t}\right)^{\frac{\theta+\epsilon\frac{1-\alpha}{\alpha}}{\epsilon}} Y_t^{-\frac{\theta}{\epsilon}} \left(\frac{p_\tau}{P_t}\right)^{\theta+\frac{1}{\alpha}+\frac{\theta\frac{1-\alpha}{\alpha}}{\epsilon}}. \end{aligned} \tag{4}$$

This is also log-concave in $\frac{p_\tau}{P_t}$.[23]

3.2 State dynamics and the short-run Phillips curve

The basic model will have three state variables, though calculating excess demand will require a fourth. However, all these state variables will take the same form:

$$X_{j,t} := \int_{-\infty}^t \lambda_\tau e^{-\int_\tau^t \lambda_v \,\mathrm{d}v} p_\tau^{\chi_{j,1}} \,\mathrm{d}\tau,$$

where $j \in \mathbb{N}$ and $\chi_{j,1}$ is a constant to be defined.[24] This implies that:

$$\dot{X}_{j,t} = \lambda_t \left(p_t^{\chi_{j,1}} - X_{j,t}\right),$$

where, as usual, dots above variables denote time derivatives.

Total demand for the variable production input, effective labour, is given by:

$$V_t := \int_{-\infty}^t \lambda_\tau e^{-\int_\tau^t \lambda_v \,\mathrm{d}v} \int_0^1 v_{\zeta,\tau,t}\, g(\zeta)\,\mathrm{d}\zeta\,\mathrm{d}\tau.$$

Assuming $\bar{\zeta}_{\tau,t} \leq 1$ for all τ and t, I show in Appendix A that:

$$V_t = -\frac{\epsilon}{(1-\alpha)\theta+\epsilon} D^\theta \left(\frac{1-\alpha}{\widehat{W}_t}\right)^{\frac{1}{\alpha}+\frac{\theta\frac{1-\alpha}{\alpha}}{\epsilon}} Y_t^{-\frac{\theta}{\epsilon}} P_t^{-\chi_{1,1}} X_{1,t} + \left(\frac{1-\alpha}{\widehat{W}_t}\right)^{\frac{1}{\alpha}} P_t^{-\chi_{2,1}} X_{2,t}, \tag{5}$$

where $\chi_{1,1} := \theta + \frac{1}{\alpha} + \frac{\theta}{\epsilon}\frac{1-\alpha}{\alpha}$ and $\chi_{2,1} := \frac{1}{\alpha}$. Labour market clearing implies $V_t = A_t L_t$, where L_t is the household's labour supply.

---

[21] To see log-concavity in price, write this expression as $Ax^a - Bx^{a+b}$, where $x = \frac{p_\tau}{P_t}$, $A, a, B, b > 0$ and $A - Bx^b > 0$ (as $\bar{\zeta}_{\tau,t} \leq 1$). Then the second derivative of its logarithm is $-\frac{a}{x^2} - x^{-2}\left(A - Bx^b\right)^{-2} Bbx^b\left[(b-1)A + Bx^b\right]$. As long as $\theta > 1$, $b > 1$, so this is negative.

[22] I cover the $\bar{\zeta}_{\tau,t} > 1$ case in Appendix A.

[23] By an identical argument to that of Footnote 21.

[24] The subscript ", 1" anticipates the fact that other powers will enter this integral in the extended model.

Next, evaluating the integrals in the definition of the aggregator Y_t, equation (1), implies:

$$1 = -\frac{\theta}{\theta+1}\frac{\epsilon-1}{\theta+\epsilon}D^{\theta}\left(\frac{1-\alpha}{\widehat{W}_t}\right)^{\frac{\theta+\epsilon}{\epsilon}\frac{1-\alpha}{\alpha}} Y_t^{-\frac{\theta+\epsilon}{\epsilon}} P_t^{-\chi_{1,1}} X_{1,t}$$
$$+\frac{\theta}{\theta+1}D^{-1}\left(\frac{1-\alpha}{\widehat{W}_t}\right)^{\frac{\epsilon-1}{\epsilon}\frac{1-\alpha}{\alpha}} Y_t^{-\frac{\epsilon-1}{\epsilon}} P_t^{-\chi_{3,1}} X_{3,t}, \tag{6}$$

where $\chi_{3,1} := \frac{\epsilon-1}{\epsilon}\frac{1-\alpha}{\alpha}$, and where I again assume $\bar{\zeta}_{\tau,t} \leq 1$ for all τ and t.[25]

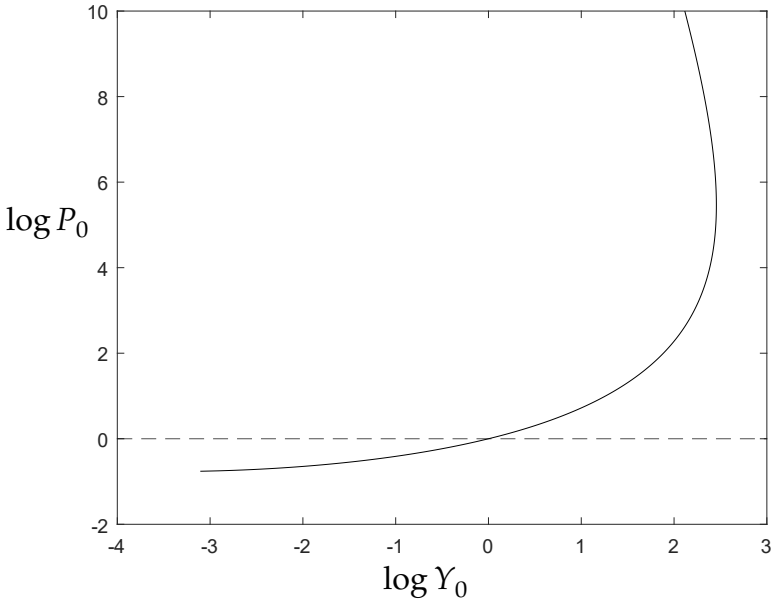


Figure 3: The model's short-run Phillips curve (solid line), and the short-run Phillips curve without rationing (dashed line). Percent deviation from steady state.

If $P_t = \exp(\pi t)$ for $t < 0$, with all state variables at steady state at time 0, how does Y_0 vary with a jump in P_0, assuming inflation continues at π after time 0?

Holding fixed the values of the three states, $X_{1,t}$, $X_{2,t}$ and $X_{3,t}$, equations (5) and (6) can be combined with labour market clearing ($V_t = A_t L_t$) and the household's labour first order condition (to be given) to produce four equations in five unknowns (V_t, L_t, $\widehat{W}_t$, Y_t and P_t). Plotting the set of points satisfying these equations in (Y_t, P_t)-space gives the model's short-run Phillips curve. I do this in Figure 3, under the model's baseline calibration which I will describe shortly. This figure answers the following question. Suppose that for all $t < 0$, $P_t = \exp(\pi t)$, meaning inflation was constant at π, and suppose all state

[25] Again, proven in Appendix A.

variables were at steady state at time 0. Then, suppose that at time 0, an unexpected monetary expansion/contraction caused the price level to jump to P_0 from 1, where it would have been had no shock arrived. How does Y_0 vary with P_0, assuming that $P_t = P_0 \exp(\pi t)$ for $t \geq 0$?

We see that the model's short-run Phillips curve is convex and backwards-bending. Expansionary monetary policy can produce a jump in prices by generating a jump in rationing, which tilts the weights of the welfare relevant price index away from goods with old (low) prices which are likely to ration. Large enough monetary expansions generate so much rationing that output falls. This result is completely independent of price setting, as it is an impact result, before prices have adjusted.

The slope of this Phillips curve around the point (0,0) is 0.53 in my calibration. This is not a calibration target, yet it almost exactly matches the slope of the short-run Phillips curve derived from Figure 1, 0.52. This estimate was produced by taking the ratio of the initial jump in prices following a monetary shock to the initial jump in industrial production, multiplied by the ratio of the monthly standard deviation of Brave-Butters-Kelley monthly Real Gross Domestic Product growth (Brave, Cole & Kelley 2019; Brave, Butters & Kelley 2019) to the monthly standard deviation of Industrial Production.[26] Thus, my model appears to match well the short-run Phillips curve we see in the data.

The convexity and backward-bending of the short-run Phillips curve can be seen analytically from equation (6) in the special case in which wages are fixed in the short-run. With wages fixed, totally differentiating equation (6) implies that:

$$\frac{\mathrm{d}\log P_t}{\mathrm{d}\log Y_t} = \frac{1}{\epsilon}\frac{\epsilon - 1 - (\theta+1)\mathcal{A}_t}{(\chi_{1,1} - \chi_{3,1})\mathcal{A}_t - \chi_{3,1}}, \tag{7}$$

where:

$$\mathcal{A}_t := \frac{\theta}{\theta+1}\frac{\epsilon-1}{\theta+\epsilon} D^{\theta} \left(\frac{1-\alpha}{\widehat{W}_t}\right)^{\frac{\theta+\epsilon 1-\alpha}{\epsilon \quad \alpha}} Y_t^{-\frac{\theta+\epsilon}{\epsilon}} P_t^{-\chi_{1,1}} X_{1,t}.$$

[26] Over the same sample as used by Miranda-Agrippino & Ricco (2021), from January 1979 to December 2014, with both series converted to continuously compounded growth rates.

When $\mathcal{A}_t$ is very small, corresponding to a large monetary expansion, the numerator of (7) is positive while the denominator is negative, implying $\frac{\mathrm{d}\log P_t}{\mathrm{d}\log Y_t} < 0$, meaning the Phillips curve is backward bending. For larger $\mathcal{A}_t$, corresponding to a monetary contraction or a smaller expansion, the denominator of (7) is positive, so then $\frac{\mathrm{d}\log P_t}{\mathrm{d}\log Y_t}$ is decreasing in $\mathcal{A}_t$. Thus, larger monetary expansions mean lower $\mathcal{A}_t$ and higher $\frac{\mathrm{d}\log P_t}{\mathrm{d}\log Y_t}$, i.e. a steeper Phillips curve.

3.3 Price setting

Just as all of the model's state variables take a similar form, so to do all of the forward-looking expressions that appear in the first order condition for firms' optimal price. In particular, they all take the form:

$$z_{j,\tau} := D^{\omega_{j,6}} \int_\tau^\infty e^{-\int_\tau^t (\lambda_v + r_v)\,\mathrm{d}v} \widehat{W}_t^{\omega_{j,2}} Y_t^{\omega_{j,3}} P_t^{\omega_{j,4}}\,\mathrm{d}t,$$

for $j \in \mathbb{N}$, and constants $\omega_{j,2}$, $\omega_{j,3}$, $\omega_{j,4}$ and $\omega_{j,6}$ to be defined. (The eccentric numbering here will seem more reasonable once I present the extended model in Section 5!) Here, r_t is the real interest rate at t. Differentiating the definition of $z_{j,t}$ implies it satisfies the following differential equation:

$$\dot{z}_{j,\tau} = -\widehat{W}_\tau^{\omega_{j,2}} Y_\tau^{\omega_{j,3}} P_\tau^{\omega_{j,4}} D^{\omega_{j,6}} + (\lambda_\tau + r_\tau) z_{j,\tau}.$$

Firms updating their price at a time τ choose p_τ to maximize their value over the life of the price:

$$\begin{aligned} o_\tau &:= \int_\tau^\infty e^{-\int_\tau^t (\lambda_v + r_v)\,\mathrm{d}v} o_{\tau,t}\,\mathrm{d}t \\ &= -\frac{\epsilon}{\theta+\epsilon}\frac{\alpha\epsilon}{(1-\alpha)\theta+\epsilon}(1-\alpha)^{-\omega_{1,2}} p_\tau^{-\omega_{1,4}} z_{1,\tau} + \alpha(1-\alpha)^{-\omega_{2,2}} p_\tau^{-\omega_{2,4}} z_{2,\tau}, \end{aligned}$$

where $\omega_{1,2} := -\frac{\theta+\epsilon}{\epsilon}\frac{1-\alpha}{\alpha}$, $\omega_{1,3} := -\frac{\theta}{\epsilon}$, $\omega_{1,4} := -\chi_{1,1} = -\left(\theta + \frac{1}{\alpha} + \frac{\theta}{\epsilon}\frac{1-\alpha}{\alpha}\right)$, $\omega_{1,6} := \theta$, $\omega_{2,2} := -\frac{1-\alpha}{\alpha}$, $\omega_{2,3} := 0$, $\omega_{2,4} := -\frac{1}{\alpha}$, $\omega_{2,6} := 0$.[27] Thus, firms optimally set p_τ such that:

$$\epsilon\left[\frac{\epsilon}{\theta(1-\alpha)+\epsilon} - \frac{\epsilon-1}{\theta+\epsilon}\right](1-\alpha)^{\frac{\theta 1-\alpha}{\epsilon\ \alpha}} p_\tau^{\theta+\frac{\theta 1-\alpha}{\epsilon\ \alpha}} z_{1,\tau} = z_{2,\tau}.$$

In the quasi-flexible price limit with $\lambda_\tau \to \infty$, this implies they would set the price p_τ^{QF} with:

[27] See Appendix A for this derivation and those of the rest of the results in this Subsection. I continue to assume $\bar{\zeta}_{\tau,t} \leq 1$ for all τ and t throughout.

$$\frac{p_\tau^{\text{QF}}}{P_\tau} = \left[\left[\frac{\epsilon^2}{\theta(1-\alpha)+\epsilon} - \epsilon \frac{\epsilon-1}{\theta+\epsilon} \right]^{-\frac{\alpha\epsilon}{\theta}} D^{-\alpha\epsilon} Y_\tau^\alpha \left(\frac{\widehat{W}_\tau}{1-\alpha} \right)^{1-\alpha} \right]^{\frac{1}{1+(\epsilon-1)\alpha}}.$$

For comparison, without rationing in the quasi-flexible price limit, firms would set the price p_τ^{QFNR}:[28]

$$\frac{p_\tau^{\text{QFNR}}}{P_\tau} = \left[\left[(1-\alpha) \left(\frac{\epsilon}{\epsilon-1} \right) \frac{\theta+\epsilon}{\theta(1-\alpha)+\epsilon} \right]^{1-\alpha} D^{-\alpha\epsilon} Y_t^\alpha \left(\frac{\widehat{W}_\tau}{1-\alpha} \right)^{1-\alpha} \right]^{\frac{1}{1+(\epsilon-1)\alpha}}$$

The two expressions agree when $\alpha = \frac{\theta+\epsilon}{\theta+\epsilon^2}$. At this point, the derivatives of the ratio $\frac{p_\tau^{\text{QFNR}}}{p_\tau^{\text{QF}}}$ with respect to α, ϵ or θ are all zero, and the second derivatives of the ratio with respect to those variables are all positive. Thus, at least locally around $\alpha = \frac{\theta+\epsilon}{\theta+\epsilon^2}$, with quasi-flexible prices, firms set higher prices if they cannot ration than if rationing is allowed. This is intuitive. If rationing is not allowed, firms worry about making large losses if demand is very high. To protect against this, they set a higher price.

3.4 Price adjustment rate choice

I will present results for two variants of this basic model. In one, λ_t will be exogenously fixed at λ. In the second, λ_t will be endogenized, broadly following Blanco et al. (2024b). This is important as I wish to analyse the effects of changing steady-state inflation, and it is not plausible to assume that λ_t remains fixed as the long-run inflation rate increases. Higher trend inflation should mean more frequent price adjustment.

To endogenize λ_t, I assume that all firms are owned by conglomerates, with each conglomerate owning countably many firms (still a measure zero subset of the set of all firms). Each conglomerate will choose the rate of price adjustment λ_t for the firms it owns, to maximize average firm value over its firms minus a price adjustment cost of $\frac{1}{2}\kappa\lambda_t^2$ labour units. This cost function has the reasonable property that if there is no price adjustment ($\lambda_t = 0$) then there are no costs, unlike the adjustment function chosen by Blanco et al. (2024b). However, as a one parameter adjustment cost function I will not be able to calibrate it to hit the observed variability of λ_t. (Blanco et al. (2024b) use a two-

[28] See Appendix B for the model without rationing.

parameter function.) It will turn out though that with this function, calibrating to match the observed average λ_t will also get close to matching the observed variability of λ_t.

The derivation of the conglomerate's first order condition is a little involved, so I confine it to Appendix A, but the first order condition itself is quite simple. Define the total flow of profits at t by:

$$\begin{aligned} O_t &:= \int_{-\infty}^{t} \lambda_\tau e^{-\int_\tau^t \lambda_v \,\mathrm{d}v} o_{\tau,t} \,\mathrm{d}\tau \\ &= -\frac{\epsilon}{\theta+\epsilon}\frac{\alpha\epsilon}{(1-\alpha)\theta+\epsilon} D^\theta \left(\frac{1-\alpha}{\widehat{W}_t}\right)^{\frac{\theta+\epsilon}{\epsilon}\frac{1-\alpha}{\alpha}} Y_t^{-\frac{\theta}{\epsilon}} P_t^{-\chi_{1,1}} X_{1,t} \\ &\quad + \alpha \left(\frac{1-\alpha}{\widehat{W}_t}\right)^{\frac{1-\alpha}{\alpha}} P_t^{-\chi_{2,1}} X_{2,t}. \end{aligned} \tag{8}$$

using equation (4), and define the total value of all firms at time s over the lives of their current prices by:

$$Q_s^* := \int_{-\infty}^{s} \lambda_\tau e^{-\int_\tau^s \lambda_v \,\mathrm{d}v} \int_s^\infty e^{-\int_s^t (\lambda_v + r_v)\,\mathrm{d}v} o_{\tau,t} \,\mathrm{d}t \,\mathrm{d}\tau.$$

Then:

$$\dot{Q}_t^* = \lambda_t o_t - O_t + r_t Q_t^*,$$

and the conglomerate's first order condition implies:

$$\kappa \lambda_t W_t = o_t - Q_t^*.$$

This is easy to understand. The right-hand side is the benefit of increasing the price adjustment rate. Firms that update their price will have value o_t (over the life of their new price), while those that do not update their price on average have value Q_t^* (over the lives of their current prices). The left-hand side is the marginal cost of increasing the price adjustment rate.

3.5 Households and monetary policy

In period t the representative household maximizes:

$$\int_\tau^\infty e^{-\int_\tau^t \rho_v \,\mathrm{d}v} \left[\log Y_t - \Psi_t \frac{1}{1+\nu}\left(L_t + \frac{1}{2}\kappa\lambda_t^2\right)^{1+\nu}\right] \mathrm{d}t,$$

where $\nu > 0$ and for all t, $\Psi_t > 0$ and $\rho_t > 0$, with $\int_t^\infty \rho_v \,\mathrm{d}v = \infty$. Note that we have defined L_t so that it just includes production labour, not labour used in price adjustment. In the simpler specification with exogenous λ_t we set $\kappa = 0$ so households only get disutility from productive labour supply.

The household faces the budget constraint:

$$Y_t + \frac{\dot{B}_t^{(i)}}{P_t} + \dot{B}_t^{(r)} = W_t L_t + i_t \frac{B_t^{(i)}}{P_t} + r_t B_t^{(r)} + \mathrm{T}_t,$$

where $B_t^{(i)}$ are their holdings of nominal bonds, which return i_t, $B_t^{(r)}$ are their holdings of real bonds, which return r_t, and where T_t contains all profits from owning firms and aggregators. The household's first order conditions then imply:

$$\Psi_t \left(L_t + \frac{1}{2} \kappa \lambda_t^2 \right)^{\nu} = \frac{W_t}{Y_t}, \qquad r_t = \rho_t + \frac{\dot{Y}_t}{Y_t}, \qquad i_t = r_t + \pi_t,$$

where $\pi_t = \frac{\dot{P}_t}{P_t}$.

I assume that the central bank sets the nominal interest rate according to the "real rate rule" of Holden (2024), so in particular:

$$i_t = r_t + \pi_t^* + \phi(\pi_t - \pi_t^*),$$

where $\phi > 1$ and where π_t^* is an exogenous inflation target. Combining this equation with the Fisher equation derived above implies $\pi_t = \pi_t^*$ for all t. Hence, inflation will be effectively exogenous. This is helpful as we are interested in the relationship between output and inflation. The clearest way to study this relationship is to make one of the two exogenous. Making output exogenous risks multiplicity due to the backward bending Phillips curve, so it is more sensible to make inflation exogenous, as here. I can still study monetary policy shocks in this environment, as the central bank can undertake expansionary policy by increasing π_t^*, and contractionary by decreasing it.

3.6 Other aggregates

Since aggregators will make profits when rationing is allowed, it is useful to define the real value of goods sold at t, "RGDP_t". This will be less than Y_t as RGDP_t does not include aggregator profits. Using the definition of average output from equation (3), this is defined by:[29]

$$\begin{aligned} \mathrm{RGDP}_t &:= \int_{-\infty}^{t} \lambda_\tau e^{-\int_\tau^t \lambda_\upsilon \,\mathrm{d}\upsilon} \frac{p_\tau}{P_t} y_{\tau,t} \,\mathrm{d}\tau \\ &= -\frac{\epsilon}{\theta+\epsilon} D^\theta \left(\frac{1-\alpha}{\widehat{W}_t} \right)^{\frac{\theta+\epsilon}{\epsilon}\frac{1-\alpha}{\alpha}} Y_t^{-\frac{\theta}{\epsilon}} P_t^{-\chi_{1,1}} \mathrm{X}_{1,t} + \left(\frac{1-\alpha}{\widehat{W}_t} \right)^{\frac{1-\alpha}{\alpha}} P_t^{-\chi_{2,1}} \mathrm{X}_{2,t}. \end{aligned}$$

Thus, by equations (5) and (8):

[29] As ever, see Appendix A for this derivation and those of the rest of the results in this Subsection. I continue to assume $\bar{\zeta}_{\tau,t} \leq 1$ for all τ and t throughout.

$$\widehat{W}_t V_t + O_t = -\left[\frac{(1-\alpha)\epsilon}{(1-\alpha)\theta+\epsilon} + \frac{\epsilon}{\theta+\epsilon}\frac{\alpha\epsilon}{(1-\alpha)\theta+\epsilon}\right] D^{\theta}\left(\frac{1-\alpha}{\widehat{W}_t}\right)^{\frac{\theta+\epsilon 1-\alpha}{\epsilon \quad \alpha}} Y_t^{-\frac{\theta}{\epsilon}} P_t^{-\chi_{1,1}} X_{1,t} + [(1-\alpha)+\alpha]\left(\frac{1-\alpha}{\widehat{W}_t}\right)^{\frac{1-\alpha}{\alpha}} P_t^{-\chi_{2,1}} X_{2,t} = \text{RGDP}_t.$$

So, as expected, labour income plus total firm profits equals the real value of goods sold.

It is also helpful to define a measure of excess demand. To do this, I first define Y_t^* to the be the value Y_t would take in a counter-factual economy in which all firms meet demand rather than rationing, but holding the demand curves fixed at their curves in the "actual" economy. I.e.:

$$Y_t^* := D^{-\frac{\epsilon}{\epsilon-1}}\left[\int_{-\infty}^{t} \lambda_\tau e^{-\int_\tau^t \lambda_v \,\mathrm{d}v} \int_0^1 \zeta\left[\left(\frac{D}{\zeta}\frac{p_\tau}{P_t}\right)^{-\epsilon} Y_t\right]^{\frac{\epsilon-1}{\epsilon}} g(\zeta)\,\mathrm{d}\zeta\,\mathrm{d}\tau\right]^{\frac{\epsilon}{\epsilon-1}}.$$

Then my measure of excess demand is:

$$\frac{Y_t^*}{Y_t} = \left[\frac{\theta}{\theta+\epsilon} D^{-\epsilon} P_t^{-\chi_{0,1}} X_{0,t}\right]^{\frac{\epsilon}{\epsilon-1}},$$

where $\chi_{0,1} := -(\epsilon-1)$. In a version of this model without any rationing (developed in Appendix B), $X_{0,t}$ also appears.

I will call the quantity $1-\frac{Y_t}{Y_t^*}$ the excess demand share. It is the share of demand in excess of what is supplied in the total demand. It provides a simple measure of the probability that a random good is out of stock. If this quantity is zero, there is no rationing, and if it is one then no goods are sold at all.

We also need a measure of aggregate productivity. First, imagine that a constrained social planner wants to maximize aggregate output at t by choosing $v_{\zeta,\tau,t}$ for all $\zeta \in [0,1]$ and $\tau \leq t$ subject to a fixed total effective labour supply, V_t. Then, I show in Appendix A that their choices imply total output Y_t of:

$$Y_t^{\text{SP}} := \left[\frac{\theta+1}{\theta+\frac{\epsilon}{1+\alpha(\epsilon-1)}}\right]^{\frac{1+\alpha(\epsilon-1)}{\epsilon-1}} \left(\frac{\theta+1}{\theta}V_t\right)^{1-\alpha}.$$

Given this, the natural measure of the economy's productivity is $\frac{Y_t}{Y_t^{\text{SP}}}$.

Finally, we need to define an analogue to the economy's CPI index, as CPI's short-run fixed weights mean it can differ quite substantially from the welfare

relevant price index, which PCEPI should track better. The complication is that the CPI data collectors impute missing prices (such as those of rationed goods) by assuming they had the average price growth of observed prices. Since rationed goods are less likely to have had a price change, this tends to push CPI growth up, away from a true equal weight index. A natural way to define a true equal weight index would be to define P_t^{EQUAL} by:

$$\log P_t^{\text{EQUAL}} = \int_{-\infty}^{t} \lambda_\tau e^{-\int_\tau^t \lambda_v \, dv} \log p_\tau \, d\tau.$$

If prices were differentiable, then the growth in P_t^{EQUAL} would be the average price growth of individual goods. With our Calvo setting, we instead have:

$$\frac{\mathrm{d} \log P_t^{\text{EQUAL}}}{\mathrm{d}t} = \lambda_t (\log p_t - \log P_t^{\text{EQUAL}}).$$

To capture at least some of the effects of the imputation procedure in CPI though, I assume the CPI price level P_t^{CPI} instead satisfies:

$$\frac{\mathrm{d} \log P_t^{\text{CPI}}}{\mathrm{d}t} = \lambda_t (\log p_t - \log P_t^{\text{CPI}}) + \left(1 - \frac{Y_t}{Y_t^*}\right) \frac{\mathrm{d} \log P_t^{\text{CPI}}}{\mathrm{d}t}.$$

This adjustment is intended to capture the fact that roughly $1 - \frac{Y_t}{Y_t^*}$ of all goods would be found to be out of stock by the CPI data collectors, and thus ascribed price growth equal to the overall price growth. This implies:

$$\frac{\mathrm{d} \log P_t^{\text{CPI}}}{\mathrm{d}t} = \frac{Y_t^*}{Y_t} \lambda_t \log \left(\frac{p_t}{P_t^{\text{CPI}}} \right).$$

3.7 Detrended variables and stability

For the sake of simulation, it is helpful to define detrended versions of the model's variables which should be stationary. The differential equations followed by these detrended variables will also inform us about the model's stability.

For the state variables, I define $\hat{X}_{j,t} := \frac{X_{j,t}}{P_t^{\chi_{j,1}}}$ for $j \in \mathbb{N}$, and I define $\hat{p}_t := \frac{p_t}{P_t}$. Then:

$$\dot{\hat{X}}_{j,t} = \lambda_t \hat{p}_t^{\chi_{j,1}} - (\lambda_t + \chi_{j,1} \pi_t) \hat{X}_{j,t}.$$

Given a path of $\hat{p}_t$, this differential equation is stable if and only if $\lambda_t + \chi_{j,1} \pi_t > 0$, in which case when $\hat{X}_{j,t}$ is high, it will be pushed back towards trend. Recall that with rationing, my model has the state variables $X_{1,t}$, $X_{2,t}$ and $X_{3,t}$, with $\chi_{1,1} = \theta + \frac{1}{\alpha} + \frac{\theta}{\epsilon} \frac{1-\alpha}{\alpha} > 0$, $\chi_{2,1} = \frac{1}{\alpha} > 0$ and $\chi_{3,1} = \frac{\epsilon - 1}{\epsilon} \frac{1-\alpha}{\alpha} > 0$. Thus, as long as

inflation does not go too negative, all three state variables will be stable. By contrast, the state variable of the model without rationing is $X_{5,t}$ with $\chi_{5,1} := -\frac{\epsilon}{1-\alpha} < 0$ (see Appendix B). Since this is negative, if inflation gets too high then the state variable can explode to infinity, with output collapsing to zero. Marsal, Rabitsch & Kaszab (2023) and Holden, Marsal & Rabitsch (2024) show that this instability is a major problem for empirically plausible calibrations. It is not even clear that a valid global solution exists to the basic New Keynesian model. Luckily, all of these problems go away when rationing is allowed.

For the forward-looking variables, I define $\hat{z}_{j,t} := \frac{z_{j,t}}{P_t^{\omega_{j,4}}}$ for $j \in \mathbb{N}$, so:

$$\dot{\hat{z}}_{j,t} = -\widehat{W}_t^{\omega_{j,2}} Y_t^{\omega_{j,3}} + \left(\lambda_t + r_t - \omega_{j,4}\pi_t\right)\hat{z}_{j,t}.$$

Remembering that this equation is solved backwards in time, given the paths of other variables, "stability" requires $\omega_{j,4} < 0$. The forward-looking variables with rationing were $z_{1,t}$ and $z_{2,t}$, with $\omega_{1,4} = -\left(\frac{1}{\alpha} + \theta + \frac{\theta}{\epsilon}\frac{1-\alpha}{\alpha}\right) < 0$ and $\omega_{2,4} = -\frac{1}{\alpha} < 0$, so both variables are well behaved. Again, without rationing, neither of the two forward looking variables have this "stability" property.

3.8 Parameterization and calibration

I will show results for four variants of the base model. These differ by whether λ_t is endogenous or exogenous, and by whether rationing is allowed or not (see Appendix B for the model without rationing). I set most parameters to standard values across all four variants. I set $\rho := 2\%$ and $\pi := \pi^* := 2\%$, unless otherwise stated. Following Smets & Wouters (2007), I set $\epsilon := 10$ and $\nu := 2$. Following Blanco et al. (2024b) I set λ so that the fraction of firms that adjust their price over a quarter is 0.297 in steady state. This means $\lambda := -4\log(1 - 0.297)$. In the model variants with endogenous λ_t, I set κ so that conglomerates endogenously choose this level of λ_t in steady state (with $\pi = 2\%$). At this level of κ, with rationing allowed, 0.6% of all labour is used for price adjustment, in line with Blanco et al. (2024b). By contrast, in the variant of the model without rationing, but with endogenous λ_t, 4.8% of all labour is used for price adjustment. This illustrates the degree to which rationing reduces the price adjustment frictions needed to match the data.

I set $\alpha := \frac{5}{9}$ following the argument of the introduction and the evidence of Abraham et al. (2024). Note that $\frac{5}{9}$ was the fixed share found by Abraham et al.

(2024) at annual frequency. At higher frequencies, perhaps even higher calibrations of α would be justified. Choosing $\frac{5}{9}$ is thus relatively conservative.

I normalize units by setting $A := 1$ and in the variants of the model with exogenous λ_t, I also normalize $\Psi := 1$. In the variant of the model with endogenous λ_t, I adjust Ψ so that the variants with endogenous λ_t have the same amount of production labour in steady state as the variants with exogenous λ_t.

This leaves θ, which is the one new parameter of the model. I set $\theta := 31$, as at this level the model implies that in steady state the excess demand share $1 - \frac{Y}{Y^*} = 11\%$, matching the 11% stockouts found in 2019 by Cavallo & Kryvtsov (2023). (Recall that $1 - \frac{Y}{Y^*}$ gives a simple measure of the probability that a random good is out of stock.) Setting $\theta = 31$ implies the mean of ζ is 0.97 and its standard deviation is 0.03. This does not seem like an implausibly high level for an idiosyncratic demand shock.

4 Results

I will first present comparative static results varying the steady-state inflation rate. I will then provide some further intuition for why rationing emerges as so beneficial from these comparative statics. I then present dynamic results, in the form of impulse responses to monetary policy shocks.

Before showing the comparative statics results though, let me remind you of two important results we have already seen. In Figure 2, I showed that the model can match the concavity of firm sales over the life of a price that we see in scanner data. Then, in Subsection 3.2, I showed that the model can match the slope of the short-run Phillips curve estimated from the jumps in industrial production and the price level in Figure 1. Thus, my one new parameter θ is already enough to match three facts simultaneously, including the 11% stock-out share in normal times from Cavallo & Kryvtsov (2023) that I used as a calibration target.

4.1 Comparative statics

This Subsection will present quite a number of graphs. In all the following plots, black solid lines are from the model with rationing but with exogenous

λ_t. Black dashed lines are from the model without rationing and still with exogenous λ_t. (See Appendix B for the model without rationing.) Blue solid lines are from the model with rationing and with endogenous λ_t. Finally, blue dashed lines are from the model without rationing but with endogenous λ_t.

A first question to answer is how rationing varies as the rate of inflation varies. Figure 6 answers this question using three different measures of rationing. The left panel looks at the average probability of experiencing at least some rationing across firms. I.e. it plots the steady-state value of:

$$\int_{-\infty}^{t} \lambda_\tau e^{-\int_\tau^t \lambda_v \, \mathrm{d}v} \Pr\left(\zeta \geq \bar{\zeta}_{\tau,t}\right) \mathrm{d}\tau,$$

as a function of inflation. This measure of rationing does not distinguish between firms that ration demand by a lot, and firms that barely ration at all. By this measure of rationing, firms ration around half the time.

The middle panel plots my preferred measure of how much firms ration, the excess demand share, which measures how much less people get than they wanted. To recap, the excess demand share is given by $1 - \frac{Y}{Y^*}$.

The rightmost panel plots the profits that are being made by aggregators, as a share of total output. If there was no rationing, this profit share would be zero. With rationing, aggregators make profits as they effectively face decreasing returns. If they wish to double output, they cannot just double inputs, due to firm rationing. Hence, aggregator profits also measure the amount of rationing.

All three measures are increasing in the steady-state inflation rate. This is driven by the fact that when inflation is high, mark-ups are eroded quickly, leading to greater rationing.

Figure 7 examines this mechanism in more detail. The leftmost plot shows the probability of being rationed (at least a little) at a firm that has just updated their price. This is actually decreasing in the trend inflation rate. This is because when inflation is high, firms resetting their price choose a high initial mark-up, to protect themselves against future mark-up erosion. The other two panels plot how the probability of being rationed at a firm varies over the life of a price, as a function of the long-run inflation level. The dark blue lines show that if the steady-state inflation rate is 0%, then the probability of being rationed is constant over the life of a price. However, if the steady-state inflation rate is 4%,

then the probability of being rationed increases sharply over the life of a price, as shown by the dark red lines. It turns out that the fact that rationing is increasing in the trend inflation level for old prices dominates, producing the results we already saw in Figure 6.

Figure 8 shows how output and welfare change with the long-run level of inflation. With rationing and exogenous λ_t, the output maximizing level of inflation is around 1%, and welfare is increasing in inflation over all of the range considered here. (For high enough inflation, firms with new prices no longer ration with positive probability, and welfare starts to decline.) Without rationing, output and welfare are lower at all levels of inflation. The gap is particularly stark at high levels of inflation. Inflation is bounded above in the model without rationing, and output falls to zero as inflation approaches this level. Allowing conglomerates to choose price adjustment rates improves outcomes for low levels of inflation but renders positive inflation unambiguously worse than zero inflation. At higher levels of trend inflation, a substantial amount of labour goes to price adjustment, which is costly.

Figure 9 looks in more detail at how rationing might be improving welfare relative to economies without rationing. It shows that when rationing is allowed, productivity (measured by $\log \frac{Y}{Y^{SP}}$) is increasing in inflation. Higher level of inflation are actually reducing misallocation across firms!

Figure 9 also shows that while aggregate mark-ups (measured by $\frac{(1-\alpha)Y}{WL}$) are increasing in the trend inflation rate when rationing is allowed, and higher than without rationing, still firms' excess profit shares ($\frac{O}{Y} - \alpha$) are decreasing in inflation, and lower than without rationing. This is explained by the increase in aggregator profits as inflation increases under rationing. Aggregator profits equal output minus labour income, minus firm profits. So, aggregator profits appear as an increase in the inverse labour share measure of aggregate mark-ups. However, since they reflect the effective decreasing returns faced by aggregators, they do not reflect a monopoly distortion. Firm excess profits do measure a monopoly distortion, but they are decreasing in inflation, as inflation erodes mark-ups. Excess firm profits are lower when rationing is allowed than when it is not allowed, since in the absence of rationing firms set a high initial

mark-up to guard against future negative profits.

4.2 Why might rationing be desirable?

The results of the previous subsection imply welfare is higher in economies with rationing than in those without rationing.[30] This may be surprising. Is rationing not a bad thing?

Rationing's welfare benefits are primarily a consequence of the fact that in standard models, the firms with the most distorted prices are selling a lot, since the most distorted prices are very old and hence very low. High production by these firms with old prices pushes up marginal costs for all firms, in turn reducing output for firms with relatively undistorted prices. Thus, without rationing, demand is shifted from firms with undistorted prices to firms with distorted prices. By contrast, if rationing is allowed, then these firms with old, highly distorted, prices will limit sales through rationing. With relatively low production of goods with old prices, there will be less pressure on marginal costs for firms with new prices, so those firms will produce more. Demand is shifted from firms with distorted prices to firms with undistorted ones, at least relative to the no rationing benchmark.

We also saw that excess firm profits were lower with rationing allowed, as firms do not need to set high initial mark-ups to guard against future losses. Thus, with rationing, prices are closer to the efficient level, resulting in lower aggregate distortion. The lower excess firm profits with rationing is all the more surprising since, without rationing, firms with old prices are making losses, not profits. For average firm profits to be higher without rationing, firms with new prices must be setting very high mark-ups when they cannot ration.

Furthermore, note that if a firm could adjust their price after observing their demand shock, they would choose a price that is increasing in ζ. Thus, fully flexible prices lead to reduced sales when ζ is high compared to the sticky or quasi-flexible benchmarks without rationing. Rationing also limits sales when ζ is high, so it is intuitive that increased rationing can bring the economy closer to the fully flexible benchmark.

[30] See also the results and discussion in Hahn (2022), who also examined static outcomes under rationing with sticky prices, but without idiosyncratic demand shocks.

At the micro level (looking at demand and supply of a single good), with arbitrary demand and cost curves and a fixed price, it is ambiguous whether welfare is higher with rationing or with production of the full quantity demanded. But, in reality, we expect the demand curve ($\frac{p}{P} \propto y^{-\frac{1}{\epsilon}} \approx y^{-\frac{1}{10}}$) to be flatter than the marginal cost curve ($\text{MC} \propto y^{\frac{\alpha}{1-\alpha}} \approx y^{\frac{5}{4}}$). In this case, we can see graphically that welfare should be higher when rationing is allowed than when firms are forced to satisfy demand, as shown in Figure 4. While the graphical argument of Figure 4 strictly only applies with linear marginal costs and linear demand, this result is more general. In Appendix C.1 I show that microeconomic welfare is higher with rationing with general isoelastic demand and marginal costs.

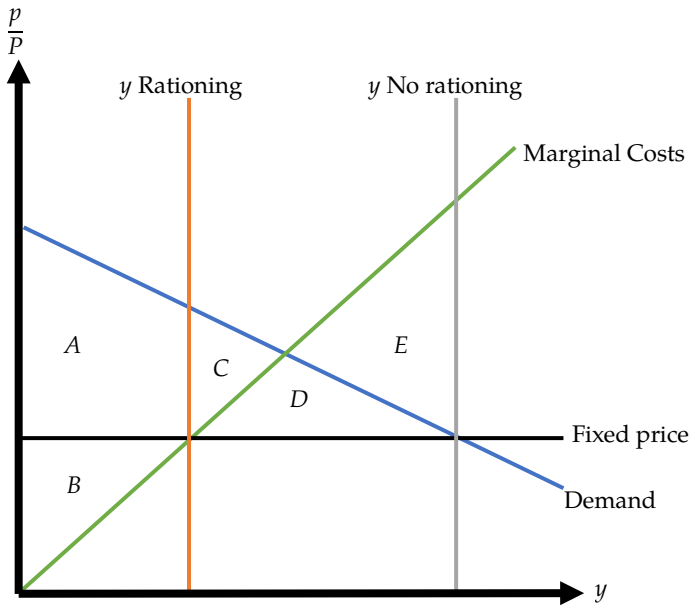


Figure 4: The microeconomics of rationing.
With rationing allowed, production is given by the orange line, and welfare is $A + B$.
Without rationing, production is given by the grey line, and welfare is $A + B + C - E$.
With demand flatter than marginal costs, $E > C$, and so welfare is higher with rationing.

The benefits of rationing are even clearer if supply constraints really do mean that marginal costs go to infinity at some finite output level $\bar{y}$ for some product, as depicted in Figure 5. Then, if the quantity demanded at the current price is greater than $\bar{y}$, there is no way the micro market can clear without

rationing, holding macro quantities fixed. Instead, as the firm increases production to try to satisfy demand, more and more of the economy's resources are devoted to this one micro market. This decreases aggregate production, pushing down demand for all products, including the current one, until demand for it is below $\bar{y}$. Thus, without rationing, macro quantities may have to move to clear a micro market, producing substantial distortions. With rationing, the micro equilibrium is given by the point at which price equals marginal cost, as standard.

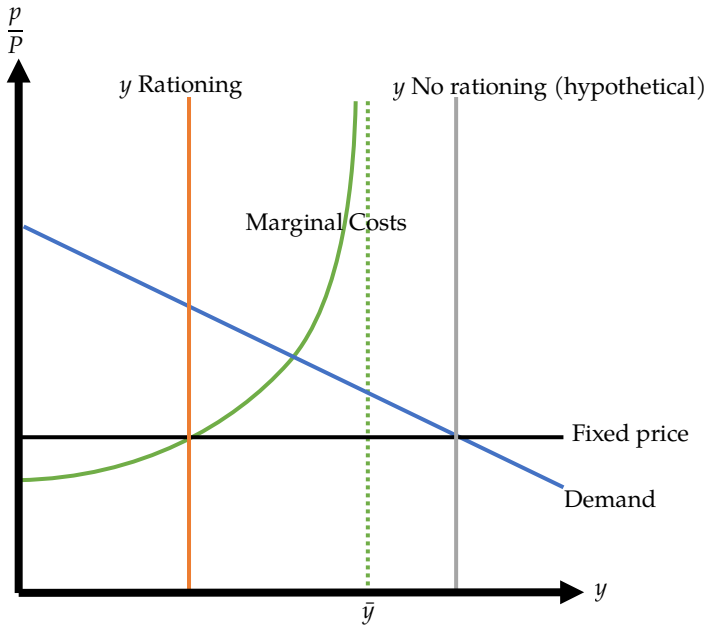
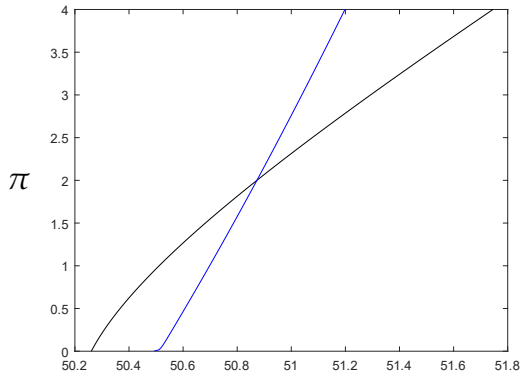
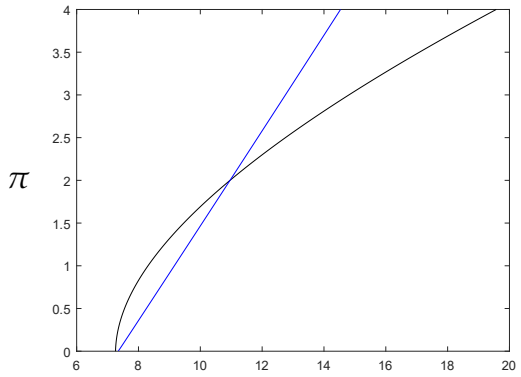


Figure 5: The microeconomics of rationing with supply constraints.
With rationing allowed, production is given by the orange line.
Without rationing, production should be given by the grey line, but it is impossible to ever produce this much, as maximum output is $\bar{y}$, the dashed green line.

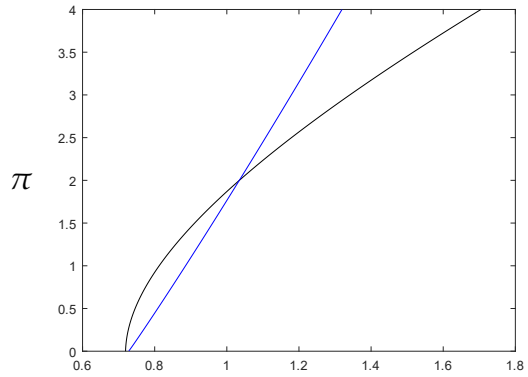
4.3 Results figures



Average probability across goods that at least some demand is rationed (percent).



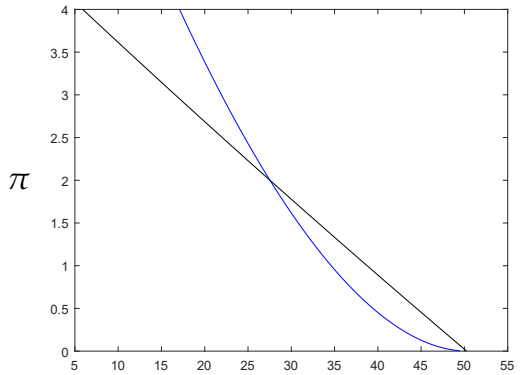
Excess demand share: $1 - \frac{Y}{Y^*}$ (percent).



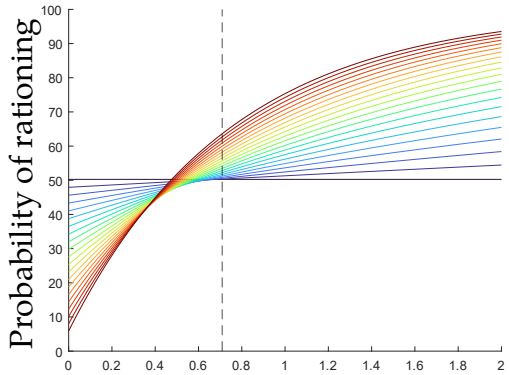
Aggregator profit share of output: $1 - \frac{\text{RGDP}}{Y}$ (percent).

Figure 6: Measures of rationing as a function of inflation (percent).

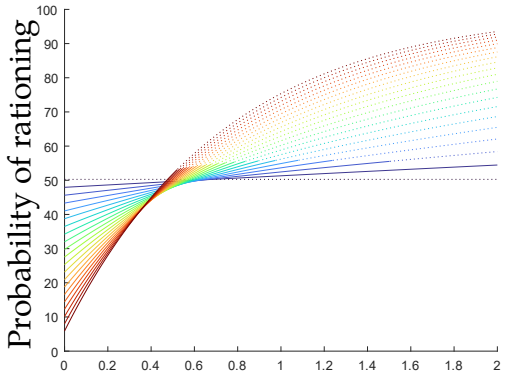
Black solid lines are the model with rationing, with exogenous λ_t. Blue solid lines are the model with rationing, with endogenous λ_t.



Probability of rationing with new prices (percent) as a function of inflation (percent).
Black solid lines are the model with rationing, with exogenous λ_t. Blue solid lines are the model with rationing, with endogenous λ_t.

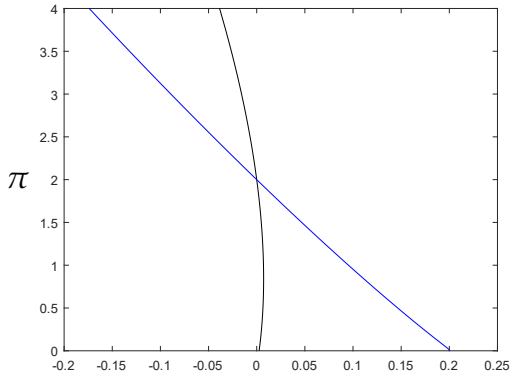


Time since price adjustment (years).
Model with exogenous λ_t.
Dashed line marks the mean life of a price.
Dark blue corresponds to 0% inflation.
Dark red corresponds to 4% inflation.

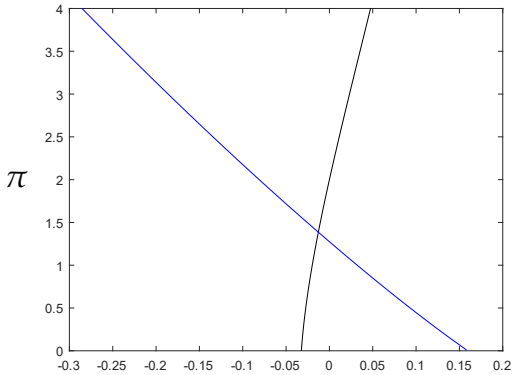


Time since price adjustment (years).
Model with endogenous λ_t.
Lines become dashed after the mean life of a price.
Dark blue corresponds to 0% inflation.
Dark red corresponds to 4% inflation.

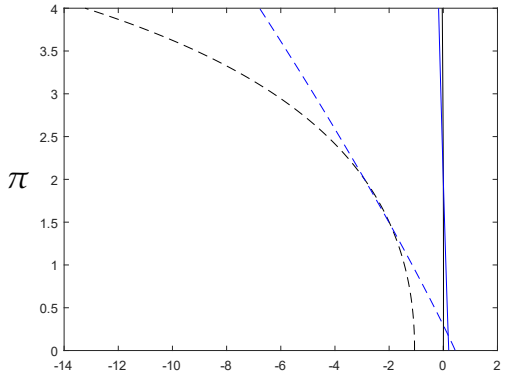
Figure 7: Which firms ration?



Relative output: $\log Y$ (percent).



Welfare: $100\left(\log Y - \Psi \frac{1}{1+\nu}\left(L + \frac{1}{2}\kappa\lambda^2\right)^{1+\nu}\right)$.



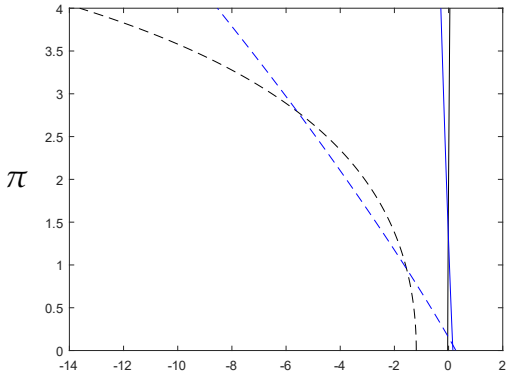
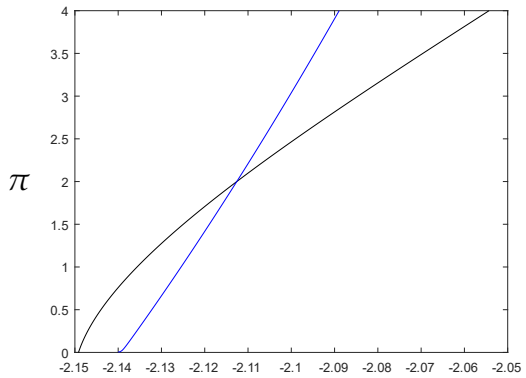
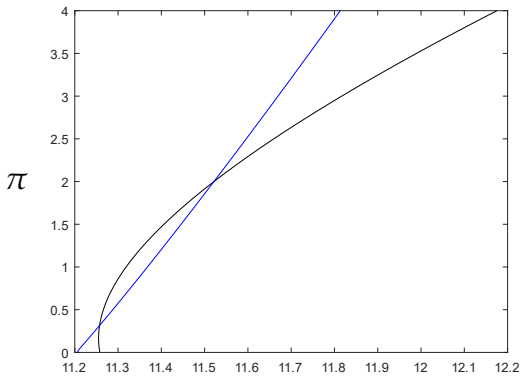


Figure 8: Output and welfare as a function of inflation (percent).

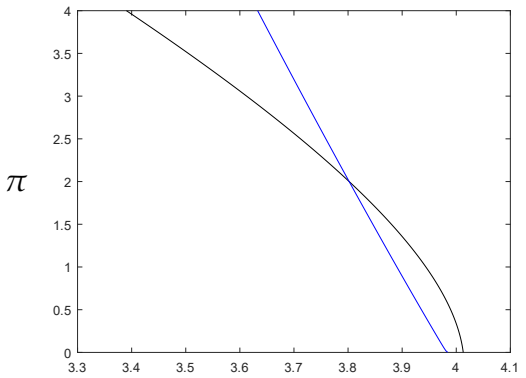
Black solid lines are the model with rationing, with exogenous λ_t. Black dashed lines are the model without rationing, with exogenous λ_t. Blue solid lines are the model with rationing, with endogenous λ_t. Blue dashed lines are the model without rationing, with endogenous λ_t. Top plots only include the results with rationing. Bottom plots include all four.



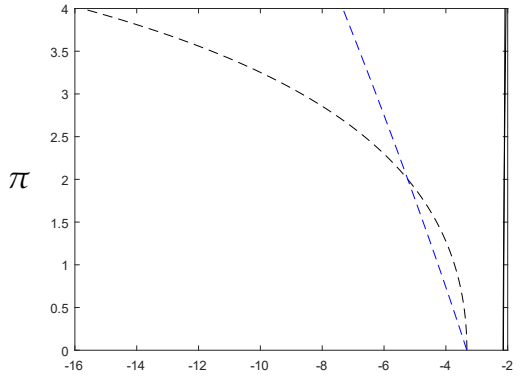
TFP loss from efficient benchmark: $\log \frac{Y}{Y^{SP}}$ (percent).

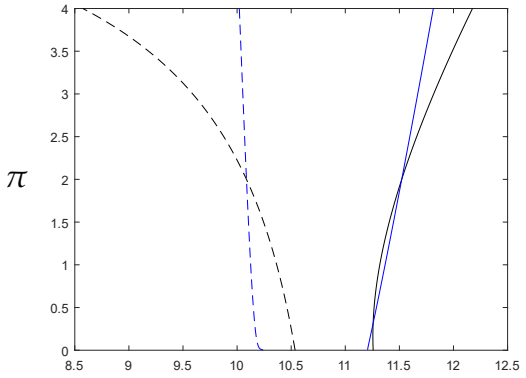


Aggregate mark-ups: $\log \frac{(1-\alpha)Y}{WL}$ (percent).



Excess firm profit share of output: $\frac{O}{Y} - \alpha$ (percent).





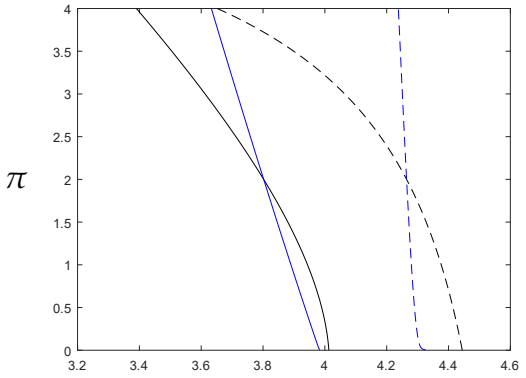
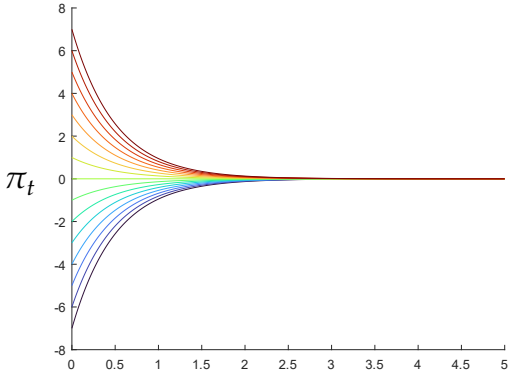
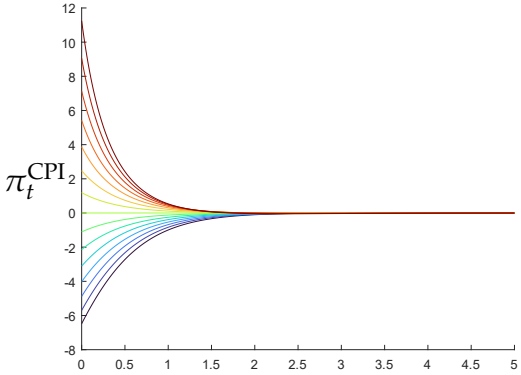


Figure 9: Measures of rationing as a function of inflation (percent).

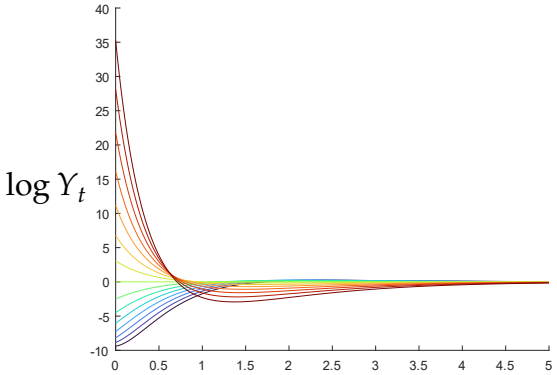
Black solid lines are the model with rationing, with exogenous λ_t. Black dashed lines are the model without rationing, with exogenous λ_t.
Blue solid lines are the model with rationing, with endogenous λ_t. Blue dashed lines are the model without rationing, with endogenous λ_t.
Top plots only include the results with rationing. Bottom plots include all four.



Driving inflation shocks (percent) as a function of time (years).

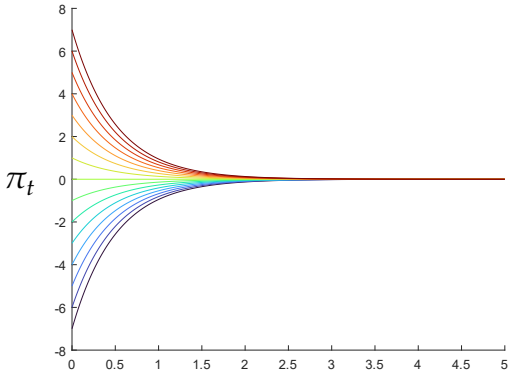


CPI inflation (percent) as a function of time (years).

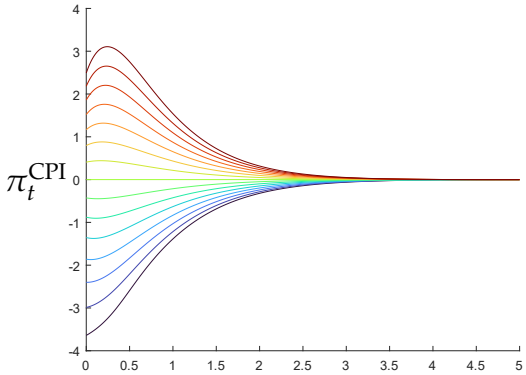


Output (percent) as a function of time (years).

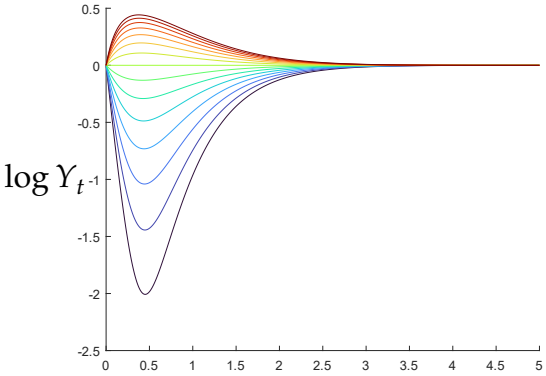
Figure 10: Impulse responses to monetary shocks without rationing, with exogenous λ_t.
Colours are consistent across subplots. All responses are relative to steady state.



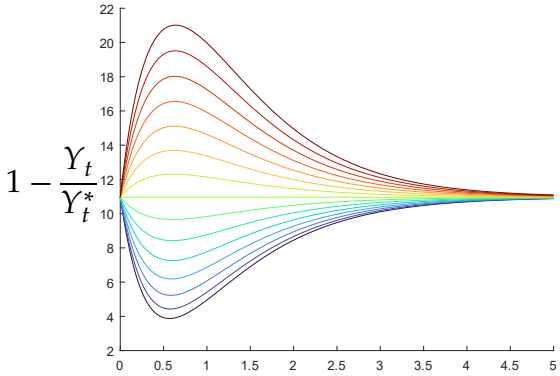
Driving inflation shocks (percent) as a function of time (years).



CPI inflation (percent) as a function of time (years).



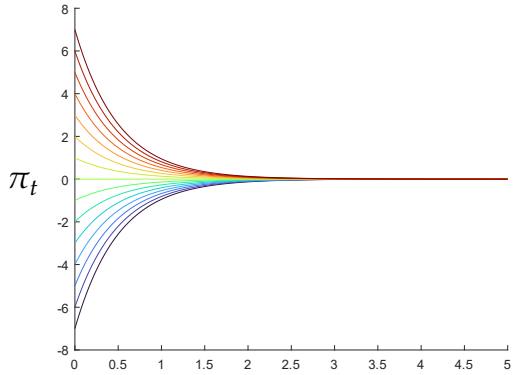
Output (percent) as a function of time (years).



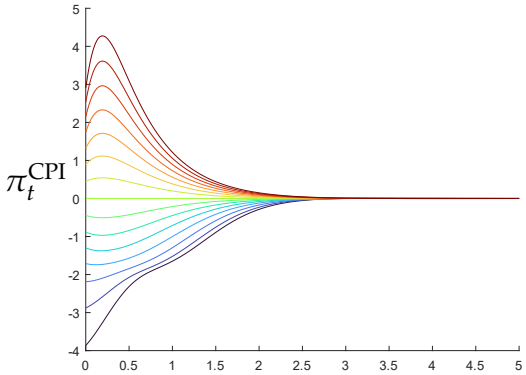
Excess demand share (percent) as a function of time (years).

Figure 11: Impulse responses to monetary shocks with rationing, with exogenous λ_t.

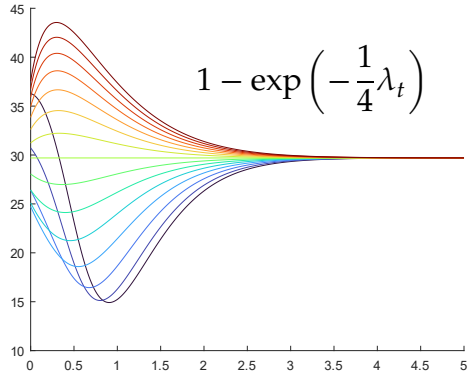
Colours are consistent across subplots. Inflation and output responses are relative to steady state.



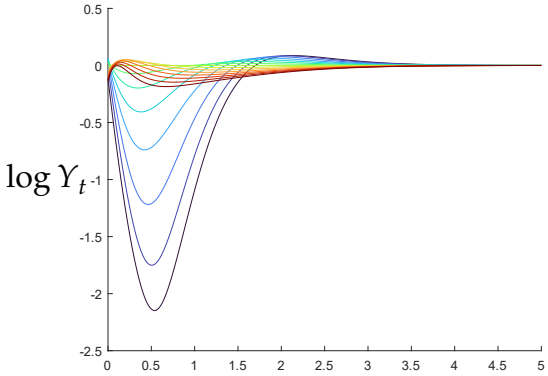
Driving inflation shocks (percent) as a function of time (years).



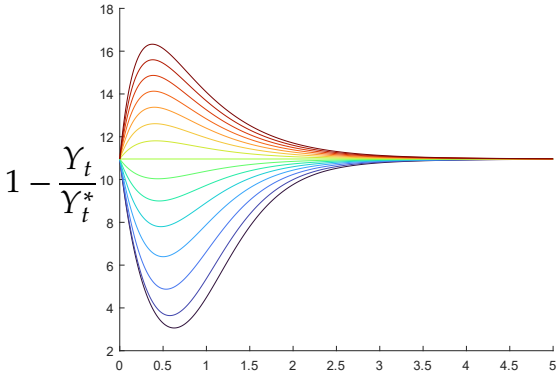
CPI inflation (percent) as a function of time (years).



Quarterly price adjustment probability (percent) as a function of time (years).



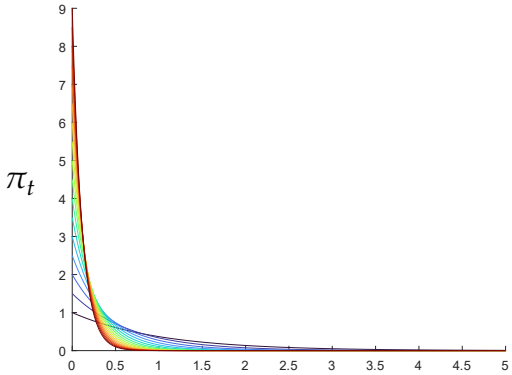
Output response (percent) as a function of time (years).



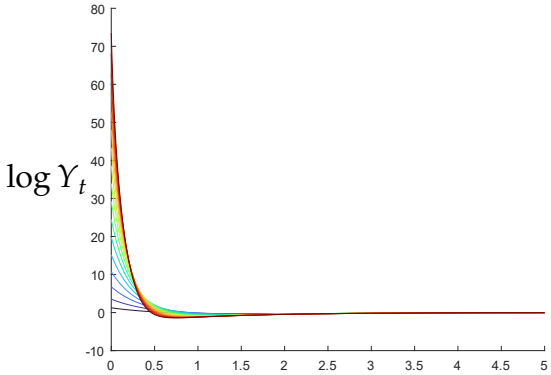
Excess demand share (percent) response as a function of time (years).

Figure 12: Impulse responses to monetary shocks with rationing, with endogenous λ_t.

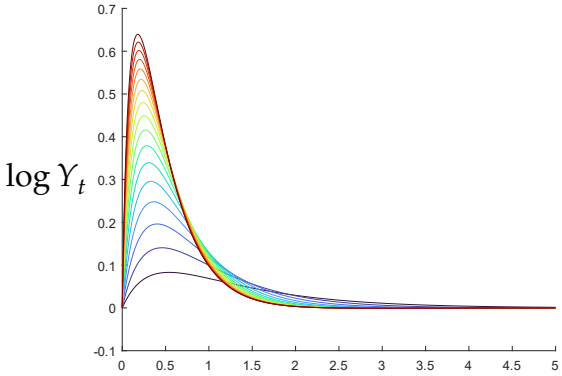
Colours are consistent across subplots. Inflation and output responses are relative to steady state.



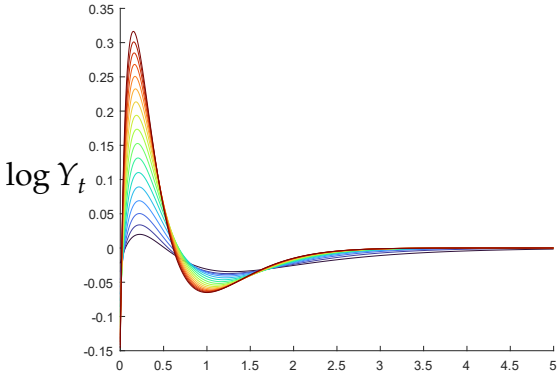
Driving inflation shocks (percent) as a function of time (years).



Output (percent) as a function of time (years), without rationing, with exogenous λ_t.



Output (percent) as a function of time (years), with rationing, with exogenous λ_t.



Output (percent) as a function of time (years), with rationing, with endogenous λ_t.

Figure 13: Impulse responses to monetary shocks with varying persistence.

Colours are consistent across subplots. Responses are relative to steady state. All shocks increase P_∞ by 1%.

4.4 Dynamics

I now examine the behaviour of the model in response to unexpected monetary policy shocks which vary π_t^* and hence π_t. I assume these shocks have prior probability zero, "MIT shock" style, and I assume the economy begins in steady state. In this simple model, other potential shocks are of limited interest due to divine coincidence.[31] (In the extended model of Section 5, I look at supply shocks, modelled as shocks to the intermediate input share in production.)

In all but the final exercise I consider, the driving shocks will take the form:

$$\pi_t^* = 2\% + \text{shock} \times \exp(-2t),$$

for $t \geq 0$, where "shock" varies from +7% to −7%. In the plots, the +7% shock will be in dark red, while the −7% shock will be in dark blue, with intermediate shocks in intermediary colours of the rainbow. A +7% inflation shock is intended to capture something like the 2022 inflation, where CPI inflation at least hit 9%.

Figure 10 plots the impulse responses to these shocks in an economy without rationing and with exogenous λ_t. (See Appendix B for details of the model without rationing.) The +7% shock to inflation increases output by about 35%, whereas the −7% shock decreases output by less than 10%. Thus, we see that positive inflation shocks are amplified, while negative inflation shocks are dampened. This is counter factual. It corresponds to a concave Phillips curve, not a convex one as in the data.

This concavity comes from the fact that without rationing, the definition of aggregate output implies that firms that can update their price must choose a $\hat{p}_0 = \frac{p_0}{P_0}$ that is log-convex in π_0,[32] as the fact consumers substitute away from expensive goods means that new prices have to increase more than proportionally to hit a given level of aggregate inflation. The only way this can

[31] Shocks to productivity, A_t, and to the disutility of labour supply, Ψ_t, have identical effects to their effects under quasi-flexible prices if monetary policy holds inflation constant. Shocks to ρ_t have very slightly different impacts to their effects under quasi-flexible prices, due to the presence of trend inflation.

[32] Without rationing, we have that $\log \hat{p}_t = -\frac{1}{\epsilon-1} \log \left[\hat{X}_{0,t} \left[1 - (\epsilon - 1) \frac{\pi_t}{\lambda_t} \right] \right]$, where $\hat{X}_{0,t}$ is actually constant in equilibrium.

be an optimal choice for firms is if Y_0 is also log-convex in π_0,[33] else they would be better off reducing their prices a little when π_0 is high to capture more sales.

Figure 10 also plots my approximation to CPI inflation for the model without rationing. The response of CPI inflation is also convex in true inflation, however it is not convex enough for the output-CPI Phillips curve of the model to end up convex.

Figure 11 performs the same exercise in the model with rationing, still with exogenous λ_t. The first thing to note is that now the impact on output is heavily muted. With rationing, monetary non-neutralities are smaller. This is because changes in rationing give the model another margin of adjustment, rather than forcing aggregate output to respond. Positive shocks lead to increases in rationing, dampening the output impact. Negative shocks reduce rationing, cushioning the output impact. This is clear from the bottom right subplot, which shows the impact on the excess demand share. Note that the excess demand share hits about 21% in response to the +7% shock, close to the 23% stockout level Cavallo & Kryvtsov (2023) found for 2022. (A fourth moment approximately hit with my one new parameter, θ!)

The dampening is particularly powerful for positive shocks, which produce big increases in firm marginal costs, and hence large amounts of rationing. As a result, now positive inflation shocks have a smaller impact on output than negative inflation shocks. Thus, as in the data, the model's Phillips curve is convex.

In Figure 12, I repeat the experiment in a model with rationing in which the price adjustment rate λ_t is endogenized. Now, positive shocks have almost no impact. At least in the vicinity of steady state, this appears to be a model in which monetary policy can do harm but not good! This additional asymmetry is because trend inflation is positive. With positive trend inflation, following a positive shock, conglomerates can pull forward the future positive price adjustments they would have made anyway. This reduces their total cost of a faster price adjustment. By contrast, a negative inflation shock will be undone by future trend inflation, reducing the conglomerate's incentive to perform

[33] This may be easily proven in the limit as $\rho \to \infty$.

price adjustment. Still, for sufficiently large negative shocks, the conglomerate still increases the rate of adjustment on impact, dampening these really large negative shocks. Blanco et al. (2024) find that the quarterly price adjustment probability peaked at about 42% in 1980, when PCEPI inflation was around 12%, and this probability peaked at about 55% in 2022, when PCEPI inflation was around 7%. My +7% shock pushes the quarterly price adjustment probability to around 43%, roughly in the middle of these two data points. This success is despite the fact that my price adjustment cost function only has one parameter, while the Blanco et al. (2024) one has two.

As a final exercise, in Figure 13, I show how the impulse response to an inflation shock varies with the persistence of the driving shock. As I vary the persistence of the shock, I also vary its magnitude so that all shocks produce a 1% increase in the price level at time infinity. Without rationing, this 1% increase in the price level pushes output up 70% with the least persistent shock plotted. This implausibly large response reflects the fact that the price level is a state variable in the model without rationing, so the short-run Phillips curve is horizontal. Once rationing is allowed, the impact is over one hundred times smaller, reflecting the existence of a true short-run Phillips curve in the model with rationing. Allowing for an endogenous rate of price adjustment halves the impact again, as conglomerates pull forward future price adjustments. The model with rationing behaves far more reasonably than the standard model without.

5 Extensions

TODO: WRITE UP.

5.1 Costs of rationing

TODO: WRITE UP.

5.2 Intermediates in production

TODO: WRITE UP.

5.3 Firm specific capital and other partially fixed inputs

TODO: WRITE UP.

5.4 Results from the extended model

TODO: WRITE UP.

6 Conclusion

In this paper I have shown how relaxing one small simplifying assumption from the workhorse model of sticky prices drastically alters the conclusions of that model. Allowing firms to ration removes the welfare costs of moderate inflation and cuts the impact of monetary shocks by two orders of magnitude. The model with rationing also matches the data remarkably well. With just one new parameter, the model roughly matches the amount of stockouts pre-Covid, the amount of stockouts in the high inflation of 2022, the concavity of output over the life of a price and the slope of the short-run Phillips curve derived from high frequency monetary shocks. The model also produces a convex Phillips curve, as we see in the data. Allowing rationing appears essential to understanding the relationship between inflation and output.

References

Abraham, Filip, Yannick Bormans, Jozef Konings & Werner Roeger. 2024. 'Price-Cost Margins, Fixed Costs and Excess Profits'. *The Economic Journal*: ueae037.

Adam, Klaus & Henning Weber. 2019. 'Optimal Trend Inflation'. *American Economic Review* 109 (2): 702–737.

Alessandria, George, Joseph P. Kaboski & Virgiliu Midrigan. 2010. 'Inventories, Lumpy Trade, and Large Devaluations'. *American Economic Review* 100 (5): 2304–2339.

Babb, Nathan R. & Alan K. Detmeister. 2017. 'Nonlinearities in the Phillips Curve for the United States : Evidence Using Metropolitan Data'. *Finance and Economics Discussion Series*. Finance and Economics Discussion Series.

Barro, Robert J. 1977. 'Long-Term Contracting, Sticky Prices, and Monetary Policy'. *Journal of Monetary Economics* 3 (3): 305–316.

Barro, Robert J. & Herschel I. Grossman. 1971. 'A General Disequilibrium Model of Income and Employment'. *The American Economic Review* 61 (1): 82–93.

Bauer, Michael D. & Eric T. Swanson. 2023. 'A Reassessment of Monetary Policy Surprises and High-Frequency Identification'. *NBER Macroeconomics Annual*

37: 87–155.

Bils, Mark. 2016. 'Deducing Markups from Stockout Behavior'. *Research in Economics* 70 (2): 320–331.

Blanco, Andrés, Corina Boar, Callum J. Jones & Virgiliu Midrigan. 2024a. 'Non-Linear Inflation Dynamics in Menu Cost Economies'. Working Paper. Working Paper Series. National Bureau of Economic Research.

———. 2024b. 'The Inflation Accelerator'. Working Paper. Working Paper Series. National Bureau of Economic Research.

Boehm, Christoph E., Aaron Flaaen & Nitya Pandalai-Nayar. 2019. 'Input Linkages and the Transmission of Shocks: Firm-Level Evidence from the 2011 Tōhoku Earthquake'. *The Review of Economics and Statistics* 101 (1): 60–75.

Brave, Scott A., R. Andrew Butters & David Kelley. 2019. 'A New "Big Data" Index of U.S. Economic Activity'. *Economic Perspectives* (1): 1–30.

Brave, Scott A., Ross Cole & David Kelley. 2019. 'A "Big Data" View of the U.S. Economy: Introducing the Brave-Butters-Kelley Indexes'. *Chicago Fed Letter*.

Cameron, A. Colin, Jonah B. Gelbach & Douglas L. Miller. 2011. 'Robust Inference With Multiway Clustering'. *Journal of Business & Economic Statistics* 29 (2): 238–249.

Cavallo, Alberto & Oleksiy Kryvtsov. 2023. 'What Can Stockouts Tell Us about Inflation? Evidence from Online Micro Data'. *Journal of International Economics* 146. NBER International Seminar on Macroeconomics 2022: 103769.

Cooper, Russell W. & John C. Haltiwanger. 2006. 'On the Nature of Capital Adjustment Costs'. *The Review of Economic Studies* 73 (3): 611–633.

Corsetti, Giancarlo & Paolo Pesenti. 2005. 'International Dimensions of Optimal Monetary Policy'. *Journal of Monetary Economics* 52 (2): 281–305.

Dotsey, Michael, Robert G. King & Alexander L. Wolman. 1999. 'State-Dependent Pricing and the General Equilibrium Dynamics of Money and Output*'. *The Quarterly Journal of Economics* 114 (2): 655–690.

Drèze, Jacques H. 1975. 'Existence of an Exchange Equilibrium under Price Rigidities'. *International Economic Review* 16 (2): 301–320.

Forbes, Kristin J., Joseph E. Gagnon & Christopher G. Collins. 2022. 'Low Inflation Bends the Phillips Curve around the World'. *Economia* 45 (89): 52–72.

Gerke, Rafael, Sebastian Giesen, Matija Lozej & Joost Röttger. 2023. 'On Household Labour Supply in Sticky-Wage HANK Models'. SSRN Scholarly Paper. Rochester, NY.

Golosov, Mikhail & Robert E. Lucas Jr. 2007. 'Menu Costs and Phillips Curves'. *Journal of Political Economy* 115 (2): 171–199.

Hahn, Volker. 2022. 'Price Dispersion and the Costs of Inflation'. *Journal of Money, Credit and Banking* 54 (2–3): 459–491.

Holden, Tom D. 2024. *Robust Real Rate Rules*. Working Paper. Kiel, Hamburg: ZBW – Leibniz Information Centre for Economics.

Holden, Tom D., Ales Marsal & Katrin Rabitsch. 2024. 'From Linear to Nonlinear: Rethinking Inflation Dynamics in the Calvo Pricing Mechanism'.

Huo, Zhen & José-Víctor Ríos-Rull. 2020. 'Sticky Wage Models and Labor Supply Constraints'. *American Economic Journal: Macroeconomics* 12 (3): 284–318.

Kabir, Eugene Tan & Ia Vardishvili. 2024. 'Quantifying the Allocative Efficiency of Capital: The Role of Capital Utilization'.

Khan, Aubhik & Julia K. Thomas. 2008. 'Idiosyncratic Shocks and the Role of Nonconvexities in Plant and Aggregate Investment Dynamics'. *Econometrica* 76 (2): 395–436.

Kimball, Miles S. 1995. 'The Quantitative Analytics of the Basic Neomonetarist Model'. *Journal of Money, Credit and Banking* 27 (4): 1241–1277.

Klenow, Peter J. & Oleksiy Kryvtsov. 2008. 'State-Dependent or Time-Dependent Pricing: Does It Matter for Recent U.S. Inflation?*'. *The Quarterly Journal of Economics* 123 (3): 863–904.

Klenow, Peter J. & Benjamin A. Malin. 2010. 'Microeconomic Evidence on Price-Setting☆'. In *Handbook of Monetary Economics*, edited by Benjamin M. Friedman & Michael Woodford, 3:231–284. Elsevier.

Kryvtsov, Oleksiy & Virgiliu Midrigan. 2013. 'Inventories, Markups, and Real Rigidities in Menu Cost Models'. *The Review of Economic Studies* 80 (1): 249–

276.

Kumar, Anil & Pia M. Orrenius. 2016. 'A Closer Look at the Phillips Curve Using State-Level Data'. *Journal of Macroeconomics* 47. What Monetary Policy Can and Cannot Do: 84–102.

Marsal, Ales, Katrin Rabitsch & Lorant Kaszab. 2023. 'From Linear to Nonlinear: Rethinking Inflation Dynamics in the Calvo Pricing Mechanism'. *Department of Economics Working Papers*. Department of Economics Working Papers.

Miranda-Agrippino, Silvia & Giovanni Ricco. 2021. 'The Transmission of Monetary Policy Shocks'. *American Economic Journal: Macroeconomics* 13 (3): 74–107.

Nakamura, Emi & Jón Steinsson. 2008. 'Five Facts about Prices: A Reevaluation of Menu Cost Models*'. *The Quarterly Journal of Economics* 123 (4): 1415–1464.

———. 2010. 'Monetary Non-Neutrality in a Multisector Menu Cost Model*'. *The Quarterly Journal of Economics* 125 (3): 961–1013.

Posch, Olaf. 2018. 'Resurrecting the New-Keynesian Model: (Un)Conventional Policy and the Taylor Rule'. *CESifo Working Paper Series*. CESifo Working Paper Series.

Posch, Olaf, Juan F. Rubio-Ramírez & Jesús Fernández-Villaverde. 2011. 'Solving the New Keynesian Model in Continuous Time'. *2011 Meeting Papers*. 2011 Meeting Papers.

Smets, Frank & Rafael Wouters. 2007. 'Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach'. *American Economic Review* 97 (3): 586–606.

Svensson, Lars EO. 1984. *Sticky Goods Prices, Flexible Asset Prices, and Optimum Monetary Policy*. IIES.

Appendix A The full extended model

TODO: WRITE UP.

Appendix B The full extended model without rationing

TODO: WRITE UP.

Appendix C Further proofs

C.1 Microeconomic welfare under rationing versus meeting demand

Suppose the demand curve is $\hat{p} = Ay^{-\frac{1}{\epsilon}}$ ($\hat{p}$ is the real price, $A > 0$, y is quantity) and the marginal cost curve is $\hat{q} = By^{\frac{\alpha}{1-\alpha}}$ ($\hat{q}$ is real marginal cost, $B > 0$). Then the efficient quantity is $\left(\frac{A}{B}\right)^{1/\left(\frac{\alpha}{1-\alpha}+\frac{1}{\epsilon}\right)}$ and the efficient price is $\hat{p}^* = A^{\frac{\alpha\epsilon}{1+\alpha(\epsilon-1)}}B^{\frac{1-\alpha}{1+\alpha(\epsilon-1)}}$. Assume the demand curve is flatter than the marginal cost curve, so $\frac{1}{\epsilon} < \frac{\alpha}{1-\alpha}$, i.e. $\alpha > \frac{1}{\epsilon+1}$.

Welfare is the difference between the integral under the demand curve and the integral under the marginal cost curve, which is:

$$\frac{\epsilon}{\epsilon-1}Ay^{\frac{\epsilon-1}{\epsilon}} - (1-\alpha)By^{\frac{1}{1-\alpha}}.$$

Suppose that $\hat{p} < \hat{p}^*$. Then output under rationing is $\left(\frac{\hat{p}}{B}\right)^{\frac{1-\alpha}{\alpha}}$ and output without rationing is $\left(\frac{\hat{p}}{A}\right)^{-\epsilon}$. Thus, the welfare gain of rationing over producing the full quantity demanded is:

$$\begin{aligned}
&\frac{\epsilon}{\epsilon-1}A\left(\frac{\hat{p}}{B}\right)^{\frac{1-\alpha\epsilon-1}{\alpha}\frac{}{\epsilon}} - (1-\alpha)B\left(\frac{\hat{p}}{B}\right)^{\frac{1}{\alpha}} - \frac{\epsilon}{\epsilon-1}A\left(\frac{\hat{p}}{A}\right)^{-(\epsilon-1)} + (1-\alpha)B\left(\frac{\hat{p}}{A}\right)^{-\frac{\epsilon}{1-\alpha}} \\
&= \frac{\epsilon}{\epsilon-1}AB^{-\frac{1-\alpha\epsilon-1}{\alpha}\frac{}{\epsilon}}\hat{p}^{\frac{1-\alpha\epsilon-1}{\alpha}\frac{}{\epsilon}} - (1-\alpha)B^{-\frac{1-\alpha}{\alpha}}\hat{p}^{\frac{1}{\alpha}} - \frac{\epsilon}{\epsilon-1}A^{\epsilon}\hat{p}^{-(\epsilon-1)} \\
&\qquad + (1-\alpha)A^{\frac{\epsilon}{1-\alpha}}B\hat{p}^{-\frac{\epsilon}{1-\alpha}} \\
&= (1-\alpha)A^{\frac{\epsilon}{1+\alpha(\epsilon-1)}}B^{-\frac{(1-\alpha)(\epsilon-1)}{1+\alpha(\epsilon-1)}}\left[\frac{1}{1-\alpha}\frac{\epsilon}{\epsilon-1}\left[\left(\frac{\hat{p}}{\hat{p}^*}\right)^{\frac{1-\alpha\epsilon-1}{\alpha}\frac{}{\epsilon}} - \left(\frac{\hat{p}}{\hat{p}^*}\right)^{-(\epsilon-1)}\right]\right. \\
&\qquad \left. - \left[\left(\frac{\hat{p}}{\hat{p}^*}\right)^{\frac{1}{\alpha}} - \left(\frac{\hat{p}}{\hat{p}^*}\right)^{-\frac{\epsilon}{1-\alpha}}\right]\right].
\end{aligned}$$

Let $c := (1-\alpha)\frac{\epsilon-1}{\epsilon} \in (0,1)$, $a := \frac{1}{\alpha} > 1$, $b := \frac{\epsilon}{1-\alpha} > 1$, $z := \frac{\hat{p}}{\hat{p}^*} \in (0,1)$ and let $f\colon (0,1] \to \mathbb{R}$ be defined by:

$$f(x) = \frac{z^{ax} - z^{-bx}}{x} - \left(z^a - z^{-b}\right),$$

for all $x \in (0,1)$. Then the welfare gain of rationing is:

$$(1-\alpha)A^{\frac{\epsilon}{1+\alpha(\epsilon-1)}}B^{-\frac{(1-\alpha)(\epsilon-1)}{1+\alpha(\epsilon-1)}}f(c).$$

Note that since $\alpha > \frac{1}{\epsilon+1}$, $b > a$, so $z^b < z^a < 1$ and hence for $x \in (0,1]$, $z^{ax} = (z^a)^x < 1 < \left(z^b\right)^{-x} = z^{-bx}$.

Next, observe that $f(1) = 0$, so to prove the welfare gain of rationing is strictly positive for $x \in (0,1)$, it is sufficient to prove that $f'(x) < 0$ for all $x \in (0,1]$. Now:[1]

$$\begin{aligned} x^2 f'(x) + z^{ax} - z^{-bx} &= \left(axz^{ax} + bxz^{-bx}\right)(\log z) = \frac{az^{ax} + bz^{-bx}}{a+b}\log\left(z^{ax}z^{bx}\right) \\ &= \left[\frac{z^{ax} + z^{-bx}}{2} - \frac{\left(z^{ax} - z^{-bx}\right)(b-a)}{2(a+b)}\right]\log\left(z^{ax}z^{bx}\right) \\ &< \frac{z^{ax} + z^{-bx}}{2}\log\left(z^{ax}z^{bx}\right) < \frac{z^{ax} + z^{-bx}}{2}\frac{2\left(z^{ax}z^{bx} - 1\right)}{z^{ax}z^{bx} + 1} = z^{ax} - z^{-bx}, \end{aligned}$$

using the fact that $\log(u) < \frac{2(u-1)}{u+1}$ for $u \in (0,1)$. Hence, $f'(x) < 0$ for all $x \in (0,1]$ and so $f(x) > 0$ for all $x \in (0,1)$. Therefore, the welfare gain of rationing over production of the total quantity demanded is strictly positive.

[1] This proof follows the one given here: https://math.stackexchange.com/questions/4989707/.