Excess Capacity and Demand Driven Business Cycles

Tiancheng Sun

The University of Edinburgh, UK*

June 8, 2024

Abstract

I build a macroeconomic model that features chronic excess capacity. Firms can use capacity to compete for buyers who are not fully attentive to prices. If one firm expands capacity while other firms do not, it “steals” or attracts profitable demand from others. Theoretically, I show that this capacity competition can cause an over-accumulation of capacity. In the presence of chronic excess capacity, capital resources can be slack, and demand shocks can have large effects on output. The model is consistent with stylized facts about capacity utilization and survey evidence from Switzerland. Quantitatively, when the model is estimated to match the U.S. macro data, demand shocks turn out to be the main driving forces of business cycles.

* Tiancheng Sun, School of Economics, The University of Edinburgh, 30 Buccleuch, EH8 9JT, Edinburgh, UK. Email: t.sun@ed.ac.uk. Acknowledgments: I thank my supervisor Wouter den Haan for his great support and very detailed comments. I am also deeply indebted to Specha Andrin, who generously helped me to obtain access to the KOF micro data center. I have also benefited from fruitful conversations with John Moore, Ricardo Reis, Nobuhiro Kiyotaki, Mike Elsby, Petr Sedlacek, David Baqaee, Andrew Clausen, Axel Gottfries, Lei Liu, Jiahao Tang, and Hong Wang. I appreciate helpful comments from participants at LSE Macro work-in-progress seminars and Edinburgh research seminar. All errors, of course, are my own.
1 Introduction

This paper develops a theory of chronic excess capacity and studies the role of demand shocks in driving business cycles. Three stylized facts about capacity utilization motivate this focus. First, capacity is never fully utilized at the aggregate level (see Figure 1 for the U.S. capacity utilization rate since 1948).\(^1\) At the micro-level, a large share of plants in the U.S. also operate below capacity (Boehm and Pandalai-Nayar, 2022). Similar results are found with firm-level data for Switzerland (Köberl and Lein, 2011). Second, insufficient demand is considered to be the main reason for capacity underutilization by the vast majority of firms (Boehm and Pandalai-Nayar, 2022). Third, capacity utilization varies substantially over the business cycles and is strongly pro-cyclical.

Like unemployment, the underutilization of capacity suggests that there are always some slack resources in the economy. Intuitively, when demand goes up, more capacity is utilized; when demand goes down, more capacity is left idle. The existence of resource slackness, the observation that demand is important for capacity utilization, and the fact that capacity utilization is strongly pro-cyclical raise the possibility that changes in demand are important for business cycles.

This paper builds a macroeconomic model that is consistent with these three stylized

\(^{1}\)One might note that there is a downward trend in capacity utilization (see Supplementary Appendix K.1 for further discussions). However, the focus of the paper is on the fact that capacity in aggregate is never fully utilized and on the cyclical fluctuations of the capacity utilization rate. Hence, the paper does not aim to explain the causes of the trend. In principle, the trend component of the capacity utilization rate can be removed but is kept as it is in Figure 1 to show that long-term capacity underutilization exists in the unprocessed data.
facts and shows that in an economy where capacity is generally in excess, demand shocks can be the main driving forces of business cycles.

The primary assumption of the model is that buyers are rationally inattentive to prices when they search for capacity to satisfy their demand. In the model, firms supply their capacity to the market, waiting for buyers to show up. The model abstracts away from inventories. Unless buyers show up, capacity would not be utilized, and goods or services would not be produced (e.g., Bai et al., 2012 and Michaillat and Saez, 2015). Hence, when buyers search in the market, they search for capacity, i.e., the production potential, to satisfy their demand. For example, if buyers want to drink coffee, they search for coffee stores, which correspond to the coffee production potential. When buyers search for capacity, they need to process some information in order to direct their search towards the capacity supplied by firms that charge the lowest price. However, buyers are subject to an information processing cost as in the rational inattention literature (e.g., Mattsson and Weibull, 2002 and Matějka and McKay, 2015).

If the unit cost of processing information is zero, buyers will conduct a fully directed search and purchase only from the firms with the lowest price. In this case, buyers pay full attention to prices, and the goods market becomes perfectly competitive. If the unit cost of processing information is infinite, buyers will not process any price information but will conduct an undirected random search for capacity. In this case, buyers pay no attention to prices, and firms that have a larger capacity are more likely to be visited. In general, the unit cost of processing information is positive but finite. In this case, buyers rationally choose to be partially inattentive to prices, and the behavior of the buyers is somewhere between the directed search and the undirected search. Thus, firms that charge a lower price or have a larger capacity are more likely to be visited.

Hence, the demand allocated to each firm is not only a function of the relative price but also a function of the relative capacity. As a result, firms can use capacity to compete for buyers. For a given amount of total demand from buyers, if one firm expands its capacity relative to that of its competitors, it “steals” or attracts demand from its competitors. For example, Starbucks can expand its market share by opening more coffee stores than its competitor Costa.

Using the KOF business tendency survey of the manufacturing industry in Switzerland, I provide empirical evidence of capacity competition in manufacturing industries. I find that, at the firm level, changes in capacity are positively associated with changes in demand. This remains true after controlling for a measure of expected changes in demand, a rich set of other firm-level variables that can be used to anticipate demand shocks, and sector-by-time fixed effects. Hence, the positive relationship between capac-

---

2The KOF business tendency survey is conducted by the KOF Swiss Economic Institute. The survey asks firms about their changes in demand, expected future changes in demand, changes in capacity, changes in prices, capacity utilization rates, etc. See Supplementary Appendix A and also Köberl and Lein (2011) for a description of the data set.
ity and demand cannot be explained away by anticipated demand shocks that not only cause corresponding changes in capacity but also relate positively to the actual changes in current demand. Instead, the evidence is in line with the capacity competition mechanism described above, which says that capacity expansion in one firm can attract demand from others, supporting the notion that capacity competition is an empirically plausible mechanism.\(^3\)

In equilibrium, demand is insufficient for capacity to be fully utilized, but reducing capacity would cause firms to lose demand to their competitors. Each unit of demand is profitable because firms can take advantage of inattentive buyers to charge a markup. Thus, capacity expansion (or reduction) in one firm reduces (or increases) the profits of others. Because of this negative capacity competition externality, firms have an incentive to hold excess capacity in equilibrium.

The existence of chronic excess capacity is important for the model dynamics. Away from full capacity utilization, capital resources are relatively slack, and the real marginal cost curve is relatively flat. Hence, demand shocks can have large effects on output.

To accommodate a clear-cut notion of capacity and to clarify the proposed mechanism, I invoke a simple Leontief technology as in Fagnart et al. (1997) and Boehm and Pandalai-Nayar (2022). Chronic excess capacity thus implies that capital resources are completely slack and that the real marginal cost is completely flat locally around the steady state. When demand increases, more capacity can be utilized without causing an increase in the real marginal cost. Hence, output is highly responsive to demand shocks.

By contrast, in the standard real business cycle (RBC) model, capital resources are tight, and the real marginal cost curve is upward sloping. When demand rises, the increase in output is limited because capital as a production factor is scarce. Moreover, the real wage rate would fall as the marginal productivity of labor falls. However, the evidence documented in the literature suggests that the real wage rate is most likely not counter-cyclical (e.g., Bils, 1985 and Solon et al., 1994). In my model, because capital resources are slack and the real marginal cost curve is locally flat, the real wage rate is acyclical under demand shocks. Thus, by introducing capital resource slackness, the model helps to alleviate the counterfactual implications of the standard RBC model’s response to demand shocks.

Several papers incorporate variable capital utilization into an otherwise standard RBC model, whereby capital may not be fully utilized because firms are subject to a convex capital utilization cost.\(^4\) However, I show that the standard variable capital utilization

---

\(^3\) One caveat is that the KOF data set covers only the manufacturing sector. Although casual observations, such as those in restaurants and coffee shops, suggest the importance of capacity competition in the service sector, providing empirical evidence for the service sector remains a compelling direction for future research.

\(^4\) Greenwood et al. (1988) and Basu and Kimball (1997) assume that the depreciation rate of capital is increasing and convex in terms of the utilization of capital. Kydland and Prescott (1988), Burnside et al. (1993), and Bils and Cho (1994) assume that the overtime premium paid to workers is increasing.
(VU) model does not feature chronic excess capacity or capital resource slackness (see Supplementary Appendix F). First, capacity is still fully utilized in the long run, even though capital may not be.\(^5\) Second, capital resources are tight, as capital is a scarce production factor that lowers the real marginal cost of production by reducing the capital utilization rate. As a result, when demand increases, more capital is utilized, and the real marginal cost increases. The upward sloping real marginal cost curve dampens the response of output to demand shocks and causes a counter-cyclical real wage rate as in the standard RBC model.

In addition to the variable capital utilization model with a convex capital utilization cost, there are several other models of capacity underutilization in the literature. First, Cooley et al. (1995) and Gilchrist and Williams (2000) assume that plants are subject to idiosyncratic productivity shocks. Low productivity plants are left idle to save labor costs.\(^6\) Second, Fagnart et al. (1997) assume that firms are subject to idiosyncratic demand shocks. Extra capacity is held by firms as a precaution to save production costs in case demand is high.\(^7\) Third, Bai et al. (2012) incorporate search and matching frictions into the goods market. They use a competitive search framework as in Moen (1997) where searches are fully directed. Having some capacity underutilized saves buyers the cost of purchasing goods from a tight market.

In all these models, however, capacity is underutilized for a cost-saving reason. Hence, there is no capacity competition externality. In equilibrium, capacity is efficiently utilized and is, therefore, not in excess as the aggregate total costs of producing and acquiring goods are minimized. As a result, capital resources are tight, and the dynamic properties and convex in terms of the utilization of capital. Christiano et al. (2005) and Smets and Wouters (2007) assume that the resources consumed by firms to utilize capital is increasing and convex in terms of the utilization of capital.

\(^5\)There is a conceptual difference between capital and capacity utilization (e.g., Boehm and Pandalai-Nayar, 2022). Capacity is defined as the output level at which the short run average total cost (SRAC) curve achieves its minimum (e.g., Nelson, 1989). Consistent with this definition, capacity in practice is considered the maximum level of output that a firm can produce within a given period of time under a realistic working schedule, taking into account normal downtime (e.g., Corrado and Mattey, 1997). Hence, before capital is fully utilized to work 24 hours a day, the SRAC curve is already upward sloping, and capacity is already over-utilized because of the high marginal cost of capital utilization. Hence, capital underutilization is cost minimizing and efficient in the standard variable capital utilization model. It is not surprising that the model does not feature chronic excess capacity. The model also contradicts survey evidence, which shows that it is mostly insufficient demand, rather than a high marginal cost of utilization due to labor or equipment shortages, that restricts firms’ production (see Section 2).

\(^6\)Like the variable capital utilization model, Cooley et al. (1995) and Gilchrist and Williams (2000) predict that capacity utilization is mainly restricted by the high marginal cost of production rather than by insufficient demand, contradicting the second stylized fact of capacity utilization.

\(^7\)Fagnart et al. (1997) also face difficulty in matching the micro-level data on capacity utilization. To justify a capacity utilization rate of 87% on average, they find that the chance of hitting capacity constraints each quarter has to be as high as 47%. In the data, however, the average capacity utilization rate is about 80%, but less than 20% of the firms or plants surveyed report running at or above full capacity each quarter (e.g., Köberl and Lein, 2011 and Boehm and Pandalai-Nayar, 2022). By contrast, this difficulty can be naturally resolved in my capacity underutilization model with idiosyncratic demand shocks.
of these models are similar to those of the variable capital utilization model where the impact of demand shocks on output is limited.

Michaillat and Saez (2015) also incorporate search and matching frictions into the goods market to explain capacity underutilization, but they abstract away from capital accumulation. Unlike Bai et al. (2012), where searches are fully directed, Michaillat and Saez (2015) use a random matching framework where searches are completely undirected. By contrast, I present a model where the extent to which searches are directed depends on how attentive buyers are. My work, therefore, unifies the two extremes in the literature. In my model, buyers are partially attentive to prices, and firms compete for demand via both capacity and prices. Hence, prices are naturally determined by the profit maximization problem of firms, while in a random matching framework, the bargaining protocol that pins down prices is undetermined. Michaillat and Saez (2015) impose a fixed-price bargaining protocol to make demand shocks important. My model avoids the indeterminacy issue in Michaillat and Saez (2015) and the efficient outcome in Bai et al. (2012), allowing interesting dynamics to show up naturally under demand shocks.

To assess the quantitative importance of demand for business cycles in my capacity underutilization model, I allow for the possibility that business cycles are driven by both demand shocks and labor productivity shocks. In particular, I consider three types of real demand shocks: consumption demand shocks that change the marginal utility of consumption relative to the marginal disutility of labor, investment demand shocks that change the subjective discount factor, and exogenous expenditure shocks that affect the government expenditure.

I use Bayesian estimation techniques to estimate the model. I find that the model attributes most of business cycle fluctuations to demand shocks. Labor productivity shocks account for only 2% of the variation in output and 13% of the variation in hours. Among the three types of demand shocks, consumption demand shocks are dominant and explain more than 72% of the variance in consumption, 78% of the variance in investment, and 60% of the variance in hours and generate the correct business-cycle co-movement among consumption, investment, hours, and the Solow residual. The results resonate well with two recent empirical papers, which suggest that business cycles are mainly driven by a single type of demand shock (e.g., Andrei et al., 2017 and Angeletos et al., 2020). In contrast, the possibility for demand shocks to affect the economy is limited in most existing models. Two notable exceptions are Bai et al. (2012) and Borys et al. (2021).

Bai et al. (2012) find that demand shocks are more important than technology shocks for business cycles. However, they also find it difficult to obtain the correct business cycle co-movement among consumption, investment, and hours using a single type of demand shock. Thus, they have three types of demand shocks to explain business cycle fluctuations. Moreover, in Bai et al. (2012), the matching technology and the production technology combined exhibit increasing returns to scale. This assumption allows demand
shocks to be important for business cycles even though capacity is efficiently utilized in their model. In my model, technology is of constant returns to scale, capacity is inefficiently underutilized, and a single type of demand shock can explain most of business cycle fluctuations.\footnote{In principle, my model could also incorporate increasing returns to scale, which would amplify the importance of demand shocks for business cycles and also help generate pro-cyclical real wage rates under demand shocks. However, to highlight the role played by capacity competition, standard constant returns to scale are assumed.}

Borys et al. (2021) extend Michaillat and Saez (2015) with a dynamic framework and emphasize the importance of demand shocks for business cycles. However, like Michaillat and Saez (2015), Borys et al. (2021) also face the price-indeterminacy issue and rely on the assumed fixed-price rule to make demand important. Furthermore, they do not discuss the business cycle relationship between consumption and investment, as they abstract away from capital accumulation. Finally, in Borys et al. (2021), technology improvements are capacity augmenting but have no effect on labor productivity. Other things being equal, an increase in capacity decreases capacity utilization. But capacity utilization is highly pro-cyclical in the data. Hence, when competing with the technology shocks considered in Borys et al. (2021), not surprisingly, demand shocks play a decisive role. In my work, however, the technology shocks considered are labor productivity shocks as in the standard business cycle literature.\footnote{Capacity utilization can be pro-cyclical under labor productivity shocks. Thus, in terms of explaining business cycles, it is much more difficult for demand shocks to compete with labor productivity shocks than with the technology shocks considered in Borys et al. (2021). For technology improvements that augment capacity only, my model also predicts a decrease in the capacity utilization rate as in Borys et al. (2021).}

Summing up, this paper has three main contributions. First, by extending the standard neoclassical framework with rationally inattentive buyers who search for capacity in a partially undirected way, I build a model that features capacity competition and chronic excess capacity. The model unifies and improves the competitive search framework and the random matching framework in the literature. Second, using the KOF business tendency survey, I provide empirical evidence of capacity competition in manufacturing industries. Third, when viewed through the lens of my model, demand shocks have the potential to be the main driving forces of business cycles, and when the model is estimated to match the U.S. macro data, that turns out to be the case.

The rest of the paper is organized as follows. Section 2 documents the three stylized capacity utilization facts. Section 3 presents the empirical evidence of capacity competition. Section 4 establishes the basic capacity underutilization (CU) model. Section 5 studies the properties of the basic CU model. Section 6 extends the basic CU model to a full CU model and estimates it using Bayesian estimation techniques. Section 7 concludes.
2 Stylized Capacity Utilization Facts

2.1 Long-term Capacity Underutilization

According to the capacity utilization data published by the Federal Reserve Board (FRB), the aggregate capacity of the manufacturing sector is never fully utilized in the U.S. A similar pattern shows up in the euro area according to the Business and Consumer Survey (BCS) database published by the Directorate General for Economic and Financial Affairs (DG ECFIN). The BCS database also contains information about capacity utilization in the service sector. Like the manufacturing sector, the service sector has never seen its aggregate capacity fully utilized. Table 1 summarizes these results.

Table 1: Capacity Utilization Facts

<table>
<thead>
<tr>
<th>Region</th>
<th>Sector</th>
<th>Start Date</th>
<th>Ave. u</th>
<th>Std. u</th>
<th>% of periods with u &lt; 100%</th>
<th>u &lt; 95%</th>
<th>Corr. w. Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>The U.S.</td>
<td>Manufacturing</td>
<td>1948-Q1</td>
<td>79.86</td>
<td>5.22</td>
<td>100</td>
<td>100</td>
<td>0.89 (0.06)</td>
</tr>
<tr>
<td></td>
<td>Non-durable</td>
<td>1967-Q1</td>
<td>80.38</td>
<td>4.18</td>
<td>100</td>
<td>100</td>
<td>0.75 (0.09)</td>
</tr>
<tr>
<td></td>
<td>Durable</td>
<td>1967-Q1</td>
<td>77.38</td>
<td>5.57</td>
<td>100</td>
<td>100</td>
<td>0.88 (0.09)</td>
</tr>
<tr>
<td></td>
<td>Mining</td>
<td>1967-Q1</td>
<td>86.30</td>
<td>4.92</td>
<td>100</td>
<td>100</td>
<td>0.50 (0.13)</td>
</tr>
<tr>
<td></td>
<td>Utilities</td>
<td>1967-Q1</td>
<td>85.75</td>
<td>5.76</td>
<td>100</td>
<td>91</td>
<td>0.45 (0.12)</td>
</tr>
<tr>
<td>Euro Area</td>
<td>Manufacturing</td>
<td>1985-Q3</td>
<td>80.83</td>
<td>2.94</td>
<td>100</td>
<td>100</td>
<td>0.85 (0.16)</td>
</tr>
<tr>
<td></td>
<td>Non-durable</td>
<td>1991-Q1</td>
<td>79.92</td>
<td>2.02</td>
<td>100</td>
<td>100</td>
<td>0.79 (0.11)</td>
</tr>
<tr>
<td></td>
<td>Durable</td>
<td>1991-Q1</td>
<td>81.33</td>
<td>3.65</td>
<td>100</td>
<td>100</td>
<td>0.86 (0.15)</td>
</tr>
<tr>
<td></td>
<td>Services</td>
<td>2011-Q3</td>
<td>88.65</td>
<td>1.40</td>
<td>100</td>
<td>100</td>
<td>0.79 (0.07)</td>
</tr>
<tr>
<td></td>
<td>Transport &amp; related services</td>
<td>2011-Q3</td>
<td>87.16</td>
<td>3.16</td>
<td>100</td>
<td>100</td>
<td>0.76 (0.11)</td>
</tr>
<tr>
<td></td>
<td>Accom., food, &amp; beverage</td>
<td>2011-Q3</td>
<td>84.84</td>
<td>3.19</td>
<td>100</td>
<td>100</td>
<td>0.68 (0.07)</td>
</tr>
<tr>
<td></td>
<td>Computer programming &amp; info.</td>
<td>2011-Q3</td>
<td>89.76</td>
<td>0.99</td>
<td>100</td>
<td>100</td>
<td>0.47 (0.28)</td>
</tr>
<tr>
<td></td>
<td>Business related activities</td>
<td>2011-Q3</td>
<td>88.52</td>
<td>1.50</td>
<td>100</td>
<td>100</td>
<td>0.72 (0.04)</td>
</tr>
</tbody>
</table>

Note: Ave. u, Std. u, and Corr. w. Y stand for average capacity utilization, standard deviation of capacity utilization, and correlation with the logarithm of the real gross domestic product (GDP). Accom. and info. are short for accommodation and information. Numbers in parentheses are the estimated standard deviations of the correlation coefficients, calculated using the VARHAC method proposed by Den Haan and Levin (2000). When the correlation coefficients are calculated, all variables are Hodrick-Prescott (HP) filtered with a smoothing parameter of 1,600. The end date of the data is 2022-Q2. The U.S. data is from the FRB and the Bureau of Economic Analysis (BEA). The euro area data is from the DG ECFIN and the Organization for Economic Co-operation and Development (OECD).

At the micro-level, using the Quarterly Survey of Plant Capacity Utilization (QPC), Boehm and Pandalai-Nayar (2022) find that about 75%-85% of plants in the U.S. manufacturing and publishing sector operate below full capacity. The QPC is conducted by the U.S. Census Bureau and is used by the FRB as the main source to construct its aggregate capacity utilization data.

Similarly, using the KOF business tendency survey of the manufacturing industry in Switzerland, I find that about 78% of firms operated below capacity in 2007, just before

10See Supplementary Appendix I for a description of the euro area sub-sectors included in Table 1.
the financial crisis, and about 91% of firms operated below capacity in 2009, during the financial crisis.

Furthermore, the KOF data set suggests that having some capacity underutilized is often considered “normal” by firms. For firms who report that their capacity is neither “too large” nor “too small”, 87% of them do not fully utilize their capacity, and the average capacity utilization rate for these firms is about 83%.

Finally, the long-term capacity underutilization phenomenon not only exists at the sector level but also shows up at the firm level. According to the KOF business tendency survey, for firms that have at least 5 years of observations in the sample, 38% of them never fully utilize their capacity, 54% of them have their capacity underutilized at least 95% of the time, and 64% of them have their capacity underutilized at least 90% of the time. The same pattern shows up for firms that have at least 10 years of observations in the sample (see Table 2).

<table>
<thead>
<tr>
<th>Sampling Criterion</th>
<th>% of firms whose capacity is underutilized</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100% of the time</td>
</tr>
<tr>
<td>all firms in the sample</td>
<td>63.26</td>
</tr>
<tr>
<td>firms that have at least 3 years of obs.</td>
<td>43.61</td>
</tr>
<tr>
<td>firms that have at least 5 years of obs.</td>
<td>38.25</td>
</tr>
<tr>
<td>firms that have at least 10 years of obs.</td>
<td>35.08</td>
</tr>
</tbody>
</table>

Note: Capacity utilization rates are observed on a quarterly basis. Data source: the KOF business tendency survey of the manufacturing industry from 1983-Q4 to 2021-Q4.

2.2 Insufficient Demand as the Primary Reason for Underutilization

Plant managers in the U.S. manufacturing and publishing sector are asked by the QPC to identify the primary reason why their plants produce at levels below production capacity. Possible answers “include insufficient orders,” “insufficient supply of local labor force/skills,” “insufficient supply of materials,” “equipment limitations,” and others. Multiple answers are allowed. For the period from 2013-Q1 to 2018-Q2, Boehm and Pandalai-Nayar (2022) find that about 78% of plant managers cite “insufficient orders” as the main reason for producing below capacity. The second most cited reason is “insufficient supply of local labor force/skills,” which is chosen by 11% of respondents. Only 6% of respondents choose “equipment limitations,” and 4% of respondents choose “insufficient supply of materials.”

Similarly, in the euro area, the BCS asks firms: “What main factors are currently limiting your production (or business, if in the service sector)?” Possible answers include
“insufficient demand,” “shortage of labor force,” “financial constraints,” “shortage of materials/equipment” (or “shortage of space/equipment,” if in the service sector), none, and others. Multiple answers are allowed. In both the manufacturing sector and the service sector, “insufficient demand” is the most cited reason, chosen by 30%-40% of respondents. Less than 10% of respondents in the manufacturing sector choose “shortage of materials/equipment,” and less than 3% of respondents in the service sector choose “shortage of space/equipment.” Table 3 summarizes the results.

Table 3: Factors Limiting Production/Business - Euro Area

<table>
<thead>
<tr>
<th>Sector</th>
<th>Start Date</th>
<th>% of firms limited by</th>
<th>Demand</th>
<th>Equip./Space/Mat.</th>
<th>Labor</th>
<th>Financial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manufacturing</td>
<td>1985-Q3</td>
<td>average</td>
<td>30.27</td>
<td>8.32</td>
<td>6.00</td>
<td>4.09</td>
</tr>
<tr>
<td></td>
<td></td>
<td>corr. w. u</td>
<td>-0.89 (0.09)</td>
<td>0.56 (0.21)</td>
<td>0.71 (0.11)</td>
<td>-0.70 (1.66)</td>
</tr>
<tr>
<td>Non-durable</td>
<td>1991-Q1</td>
<td>average</td>
<td>34.87</td>
<td>9.67</td>
<td>6.54</td>
<td>3.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td>corr. w. u</td>
<td>-0.91 (0.13)</td>
<td>0.52 (0.24)</td>
<td>0.74 (0.09)</td>
<td>-0.59 (0.10)</td>
</tr>
<tr>
<td>Durable</td>
<td>1991-Q1</td>
<td>average</td>
<td>34.46</td>
<td>1.51</td>
<td>10.88</td>
<td>13.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>corr. w. u</td>
<td>-0.87 (0.03)</td>
<td>0.85 (0.07)</td>
<td>0.91 (0.09)</td>
<td>-0.17 (0.10)</td>
</tr>
<tr>
<td>Services</td>
<td>2005-Q1</td>
<td>average</td>
<td>39.70</td>
<td>6.61</td>
<td>10.78</td>
<td>11.06</td>
</tr>
<tr>
<td></td>
<td></td>
<td>corr. w. u</td>
<td>-0.86 (0.05)</td>
<td>0.64 (0.03)</td>
<td>0.90 (0.04)</td>
<td>-0.16 (0.09)</td>
</tr>
<tr>
<td>Transport &amp; related services</td>
<td>2005-Q1</td>
<td>average</td>
<td>45.15</td>
<td>2.79</td>
<td>12.20</td>
<td>14.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>corr. w. u</td>
<td>-0.73 (0.08)</td>
<td>0.40 (0.11)</td>
<td>0.91 (0.13)</td>
<td>-0.14 (0.10)</td>
</tr>
<tr>
<td>Accom., food, &amp; beverage</td>
<td>2005-Q1</td>
<td>average</td>
<td>35.40</td>
<td>1.18</td>
<td>16.21</td>
<td>11.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>corr. w. u</td>
<td>-0.58 (0.19)</td>
<td>0.21 (0.10)</td>
<td>0.53 (0.22)</td>
<td>-0.11 (0.09)</td>
</tr>
<tr>
<td>Computer programming &amp; info.</td>
<td>2005-Q1</td>
<td>average</td>
<td>37.61</td>
<td>1.60</td>
<td>9.93</td>
<td>9.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>corr. w. u</td>
<td>-0.81 (0.13)</td>
<td>0.67 (0.10)</td>
<td>0.81 (0.11)</td>
<td>-0.21 (0.12)</td>
</tr>
<tr>
<td>Business related activities</td>
<td>2005-Q1</td>
<td>average</td>
<td>37.61</td>
<td>1.60</td>
<td>9.93</td>
<td>9.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>corr. w. u</td>
<td>-0.81 (0.13)</td>
<td>0.67 (0.10)</td>
<td>0.81 (0.11)</td>
<td>-0.21 (0.12)</td>
</tr>
</tbody>
</table>

Note: Equip. and matl. are short for equipment and materials. Corr. w. u stands for correlation with the capacity utilization rate at the sector level. Numbers in parentheses are the estimated standard deviations of the correlation coefficients, calculated using the VARHAC method proposed by Den Haan and Levin (2000). When the correlation coefficients are calculated, all variables are HP-filtered with a smoothing parameter of 1,600. The end date of the data is 2022-Q2. The data is from the DG ECFIN.

The KOF business tendency survey confirms that the same pattern also exists in the Swiss manufacturing sector. For the period from 1999-Q1 to 2021-Q4, of all respondents, 45% choose “insufficient demand,” 6% choose “insufficient technical capacity,” 9% choose “shortage of labor force,” 1% choose “shortage of materials/intermediate products,” and 6% choose “financial restrictions” as the “main factors currently limiting our business.”

Consistent with the observation that insufficient demand is the primary reason for capacity underutilization, in the euro area, a higher capacity utilization rate is associated with a smaller fraction of firms limited by insufficient demand and a larger fraction of firms limited by shortage of equipment, space, materials, and/or labor (see Table 3). This correlation also exists in the Swiss manufacturing sector according to the KOF data set (see Table 4).
Table 4: Factors Limiting Production - Switzerland

<table>
<thead>
<tr>
<th>Indicator of whether a firm is limited by</th>
<th>Demand</th>
<th>Capacity</th>
<th>Materials</th>
<th>Labor</th>
<th>Financial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacity utilization rate</td>
<td>-1.282***</td>
<td>0.290***</td>
<td>0.019***</td>
<td>0.320***</td>
<td>-0.026***</td>
</tr>
<tr>
<td></td>
<td>(0.032)</td>
<td>(0.018)</td>
<td>(0.004)</td>
<td>(0.018)</td>
<td>(0.016)</td>
</tr>
<tr>
<td>Size, age, &amp; age squared</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Firm &amp; sector-by-time FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>F-statistic</td>
<td>529.94</td>
<td>89.58</td>
<td>7.24</td>
<td>105.53</td>
<td>0.94</td>
</tr>
<tr>
<td>Observations</td>
<td>63,395</td>
<td>63,395</td>
<td>63,395</td>
<td>63,395</td>
<td>63,395</td>
</tr>
</tbody>
</table>

Note: Firm-level clustered standard errors in parentheses. * p ≤ 0.1; ** p ≤ 0.05; *** p ≤ 0.01. The indicator is equal to 1 if the firm surveyed is limited by the corresponding factor and is equal to 0 if otherwise. Firm size is measured by the number of employees. FE stands for fixed effect. Data source: the KOF business tendency survey of the manufacturing industry from 1983-Q4 to 2021-Q4.

2.3 Pro-cyclicality of Capacity Utilization

The last column of Table 1 shows that there is a strong positive correlation between capacity utilization rate and the real gross domestic product (GDP) in both the U.S. and the euro area. Similarly, the correlation coefficient between the capacity utilization rate of the Swiss manufacturing sector and the real GDP of Switzerland is about 0.58.\textsuperscript{11}

The pro-cyclicality of capacity utilization, together with the fact that insufficient demand is the primary reason for capacity underutilization, raises the possibility that demand shocks are important for business cycles.

3 Evidence of Capacity Competition

Using data from the KOF business tendency survey of the Swiss manufacturing industry, this section presents evidence of capacity competition in manufacturing industries. The advantage of the KOF data set is that it allows us to observe a rich set of firm-level characteristics including changes in demand, capacity, and prices, as well as expected future changes in demand and prices.\textsuperscript{12}

3.1 Empirical Framework

To capture the possible role of capacity competition, I assume that the relative demand faced by firm $j$ in sector $s$ at time $t$ is not only a function of the firm’s relative price but

\textsuperscript{11}The capacity utilization rate of the Swiss manufacturing sector is based on the KOF business tendency survey and is released by the OECD. The real GDP of Switzerland is from the State Secretariat for Economic Affairs (SECO). When the correlation coefficient is calculated, all variables are HP-filtered with a smoothing parameter of 1,600. The estimated standard deviation of the correlation coefficient is 0.10 and is calculated using the VARHAC method proposed by Den Haan and Levin (2000).

\textsuperscript{12}See Supplementary Appendix A and also Köberl and Lein (2011) for a description of the data set.
also a function of the firm’s relative capacity: \(^1\)

\[
\ln \left( \frac{y_{j,s,t}}{y_{s,t}} \right) = \varepsilon \ln \left( \frac{\bar{y}_{j,s,t}}{\bar{y}_{s,t}} \right) - \varepsilon P \ln \left( \frac{P_{j,s,t}}{P_{s,t}} \right) + a_{j,s,t},
\]

where \(y_{j,s,t}\) is the demand for firm \(j\), \(y_{s,t}\) is the aggregate demand for sector \(s\), \(\bar{y}_{j,s,t}\) is the capacity held by firm \(j\), \(\bar{y}_{s,t}\) is the aggregate capacity of sector \(s\), \(P_{j,s,t}\) is the price charged by firm \(j\), \(P_{s,t}\) is the aggregate price of sector \(s\), \(a_{j,s,t}\) is everything else that affects the relative demand of firm \(j\) and can be interpreted as relative attractiveness, \(\varepsilon P\) is the price elasticity of demand, and \(\varepsilon \bar{y}\) is the capacity elasticity of demand.

The existence of a capacity competition element corresponds to a positive capacity elasticity of demand \(\varepsilon \bar{y} > 0\). For a given amount of sector level demand, if a firm expands or reduces its capacity relative to that of its competitors, the firm gains demand from or loses demand to its competitors. If there is no capacity competition, \(\varepsilon \bar{y} = 0\), the above demand function is reduced to the standard demand function as implied by the Dixit-Stiglitz monopolistic competition model.

The KOF business tendency survey data contains qualitative information about quarterly changes in demand, prices, and capacity. For example, quarterly changes in demand in the KOF data set can take three possible values, i.e., -1, 0, 1, which correspond to “decreased,” “unchanged,” and “increased” respectively. Hence, it is convenient to re-write the demand equation (1) in changes

\[
\Delta_q \ln \left( \frac{y_{j,s,t}}{y_{s,t}} \right) = \varepsilon \bar{y} \Delta_q \ln \left( \frac{\bar{y}_{j,s,t}}{\bar{y}_{s,t}} \right) - \varepsilon P \Delta_q \ln \left( \frac{P_{j,s,t}}{P_{s,t}} \right) + F_{s,t} + \Delta_q a_{j,s,t},
\]

where \(\Delta_q\) represents a quarterly difference, each period is a quarter, and \(F_{s,t} \equiv \Delta_q \ln \left( y_{s,t} \right) + \varepsilon P \Delta_q \ln \left( P_{s,t} \right) - \varepsilon \bar{y} \Delta_q \ln \left( \bar{y}_{s,t} \right)\) captures the sector-by-time fixed effect.

This section uses a variety of linear regressions with a sector-by-time fixed effect to test whether the coefficient \(\varepsilon \bar{y}\) is positive or not. The null hypothesis is that there is no capacity competition \((\varepsilon \bar{y} = 0)\).

If changes in capacity are affected by anticipated changes in demand, there could be an endogeneity issue. For example, if firm \(j\) expects at time \(t - 1\) that its relative attractiveness \(a_{j,s,t}\) will increase in the next 3 months (or quarter), the firm is more likely to increase its capacity, and the realized change in the relative attractiveness \(\Delta_q a_{j,s,t}\) is also more likely to be high. Hence, \(\Delta_q \ln \left( \bar{y}_{j,s,t} \right)\) and \(\Delta_q a_{j,s,t}\) can be positively correlated.

Fortunately, the data set allows us to observe the firm’s expectation of future changes in demand and a rich set of other firm-level variables that can be used to predict future changes in the relative attractiveness \(\Delta_q a_{j,s,t}\). \(^1\) The endogeneity issue can be addressed

\(^1\)It will become clear later that the demand function (1) assumed here is consistent with the demand function derived from the basic capacity underutilization (CU) model proposed in Section 4.

\(^1\)The firm’s expected change in demand in the KOF data set can take three possible values, i.e., -1, 0, 1, which correspond to “decreased,” “unchanged,” and “increased” respectively. It is, thus, an imperfect
by controlling for these variables. Based on this idea, the following identification strategy is developed.

3.2 Identification Strategy

Capacity takes time to respond because of the time needed to build, to deliver, and/or to install. This is known as the “gestation period” in the literature and is estimated to be between 3 and 36 months. Thus, I assume that the capacity change observed at time \( t \) is based on the information received at least 3 months ago.

Let \( I_{j,s,t-1} \) be the information set that firm \( j \) used at time \( t - 1 \) to predict future changes in its relative attractiveness \( \Delta_q a_{j,s,t} \). By controlling for \( I_{j,s,t-1} \), equation (2) can be re-written as

\[
\Delta_q \ln (y_{j,s,t}) = \varepsilon \Delta_q \ln (\bar{y}_{j,s,t}) - \varepsilon \Delta_q \ln (P_{j,s,t}) + F_{s,t} + L (\Delta_q a_{j,s,t} | I_{j,s,t-1}) + \nu_{j,s,t},
\]

where \( L \) is the linear projection of \( \Delta_q a_{j,s,t} \) on the information set \( I_{j,s,t-1} \). The residual \( \nu_{j,s,t} \) is a linearly unpredictable change in the relative attractiveness of the firm:

\[
\nu_{j,s,t} \equiv \Delta_q a_{j,s,t} - L (\Delta_q a_{j,s,t} | I_{j,s,t-1}),
\]

which took place in the past quarter (between \( t - 1 \) and \( t \)) and, thus, is uncorrelated with the predetermined change in capacity \( \Delta_q \ln (\bar{y}_{j,s,t}) \), which was based on the information received at least 1 quarter ago (before \( t - 1 \)).

The change in capacity \( \Delta_q \ln (\bar{y}_{j,s,t}) \) is now uncorrelated with \( \nu_{j,s,t} \), but the change in price \( \Delta_q \ln (P_{j,s,t}) \) may still be correlated with \( \nu_{j,s,t} \), as both took place in the past quarter. If \( \Delta_q \ln (\bar{y}_{j,s,t}) \) and \( \Delta_q \ln (P_{j,s,t}) \) are also correlated after controlling for \( I_{j,s,t-1} \), the identification of \( \varepsilon \) could be affected.

To address this issue, I use the expected price change in the next 3 months surveyed a quarter ago as an instrumental variable (IV) for the actual price change in the past 3 months. Like other variables in the KOF data set, the expected price change can take three possible values, i.e., -1, 0, 1, which correspond to “decreased,” “unchanged,” and

---

15 See, for example, Mayer and Sonenblum (1955), Mayer (1960), Abel and Blanchard (1986), Montgomery (1995), Koekoek (2000), Del Boca et al. (2008), and Meier (2020).

16 This one quarter lag is a common assumption in the business cycle literature and is also assumed in the basic capacity underutilization model, which will be introduced later in Section 4. See also Supplementary Appendix A.4 for a robustness analysis.
“increased” respectively. Since the expectation was surveyed a quarter ago (before $t - 1$), it is uncorrelated with the linearly unpredictable change in the relative attractiveness $\nu_{j,s,t}$, which took place in the past quarter (between $t - 1$ and $t$). Hence, with the help of the instrumental variable, $\varepsilon_y$ can be identified.

To ensure a proper implementation of the strategy, a large number of firm-level variables are included in the information set $I_{j,s,t-1}$ (see Supplementary Appendix A.6 for details). First, I control for the firm’s expected changes in its future business situation, incoming orders (i.e., demand), production, purchase of intermediate products, number of employees, and purchasing price. Second, I control for the firm’s historical changes in its incoming orders, production, order backlog, intermediate products inventory, finished products inventory, capacity, profitability, competitive position, and selling price. Finally, I control for the firm’s assessment of its business situation and the assessment of its order backlog level, intermediate products inventory level, finished products inventory level, employment level, and capacity level.

### 3.3 Results and Discussions

First, Table 5 shows the results of ordinary least square (OLS) regressions, which indicate that there is a strong positive correlation between capacity changes and demand changes at the firm-level. Because quarterly demand changes are observed only from 2019 but monthly demand changes have been recorded in the KOF data set since 1983, to increase the sample size, I also use quarterly demand changes implied by monthly demand changes in regressions.

![Table 5](https://example.com/table5)

<table>
<thead>
<tr>
<th>change in demand in the past 3 months</th>
<th>change in demand (implied by monthly changes in demand)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta q \ln (\bar{y}_{j,s,t})$</td>
<td>0.338***</td>
</tr>
<tr>
<td></td>
<td>(0.019)</td>
</tr>
<tr>
<td>Observations</td>
<td>9,447</td>
</tr>
</tbody>
</table>

Note: Default OLS standard errors in parentheses. * $p \leq 0.1$; ** $p \leq 0.05$; *** $p \leq 0.01$. $\Delta q \ln (\bar{y}_{j,s,t})$ stands for change in capacity in the past 3 months. Data source: the KOF business tendency survey of the manufacturing industry from 1983-Q4 to 2021-Q4.

---

17 Monthly demand changes can take three possible values, i.e., -1, 0, 1, which correspond to “decreased,” “unchanged,” and “increased” respectively. The implied quarterly demand change is -1 (0 or 1), if all the monthly demand changes in the past 3 months are non-positive (0 or non-negative) and if at least one of them is -1 (0 or 1). In all other cases, a missing value is assigned. Alternatively, Supplementary Appendix A.5 uses the average of the monthly demand changes in the past 3 months to measure quarterly demand changes. The results presented in Section 3.3 are robust to this alternative measure.
Second, Table 6 presents the regression results where the information set $I_{j,s,t-1}$ is controlled for and where the expected quarterly price change a quarter ago is used as an instrumental variable (IV) for the actual quarterly price change. The IV estimations are obtained using standard two-stage least-squares (2SLS). The null hypothesis that there is no capacity competition among firms in the Swiss manufacturing sector can be rejected at the 1% significance level.

Table 6: Evidence of Capacity Competition - IV Estimation

<table>
<thead>
<tr>
<th></th>
<th>change in demand</th>
<th>change in demand</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>in the past 3 months</td>
<td>(implied by monthly changes in demand)</td>
</tr>
<tr>
<td>$\Delta_q \ln (\bar{y}_{j,s,t})$</td>
<td>$0.265^{***}$</td>
<td>$0.263^{***}$</td>
</tr>
<tr>
<td></td>
<td>(0.069)</td>
<td>(0.070)</td>
</tr>
<tr>
<td>$\Delta_q \ln (\bar{y}<em>{j,s,t}) \times \text{assess_cap}</em>{j,s,t-1}$</td>
<td>0.010</td>
<td>0.097</td>
</tr>
<tr>
<td></td>
<td>(0.109)</td>
<td>(0.088)</td>
</tr>
<tr>
<td>First stage F-statistic</td>
<td>53.37</td>
<td>53.73</td>
</tr>
<tr>
<td>Observations</td>
<td>1,066</td>
<td>1,066</td>
</tr>
</tbody>
</table>

Note: Firm-level clustered standard errors in parentheses. * $p \leq 0.1$; ** $p \leq 0.05$; *** $p \leq 0.01$. $\Delta_q \ln (\bar{y}_{j,s,t})$ is the change in capacity in the past 3 months. $\text{assess\_cap}_{j,s,t-1}$ is the assessment of capacity level a quarter ago. Information set $I_{j,s,t-1}$, firm size measured by the number of employees, firm age, firm age squared, firm fixed effect, and sector-by-time fixed effect are controlled for. Data source: the KOF business tendency survey of the manufacturing industry from 1983-Q4 to 2021-Q4.

Since all variables are qualitative, the magnitudes of the regression coefficients are best understood in terms of probabilities of reporting demand changes. In the sample, for firms who see an increase in capacity compared to those who see no change in capacity, the probability of reporting a decrease in demand falls from 28% to 20%, and the probability of reporting an increase in demand rises from 27% to 50%. The corresponding regression coefficient is the sum of these two changes in probabilities $0.08 + 0.23 = 0.31$, consistent with the OLS regression coefficient in Table 5.\textsuperscript{18} The magnitudes of the regression coefficients in Table 6 are somewhat smaller than those in Table 5 but are still substantial.

Finally, it is interesting to see how the firm’s assessment of its capacity level may affect the sensitivity of demand to capacity. I, therefore, include an interaction term between the change in capacity and the assessment of capacity level to see if the corresponding regression coefficient is significant. The firm’s assessment of capacity level is surveyed

\textsuperscript{18}Similarly, for firms who see a decrease in capacity compared to those who see no change in capacity, the probability of reporting a decrease in demand rises from 28% to 60%, and the probability of reporting an increase in demand falls from 27% to 20%. The sum of these two changes is 0.39. The actual OLS regression coefficient, 0.338, is between 0.31 and 0.39.
quarterly and is equal to 1, 0, or -1, if the assessment is “too large,” “adequate,” or “too small” respectively. Hence, the assessment captures the firm’s actual level of capacity relative to the firm’s desired level of capacity. Table 6 shows that the effect of capacity on demand is not sensitive to whether firms find their capacity too large or too small.

### 4 A Basic Model of Capacity Underutilization

This section presents a basic capacity underutilization (CU) model that incorporates capacity competition. The model has two features that are different from standard RBC models. First, buyers search for capacity to satisfy their demand, but they have a limited capability to process information. Second, the production technology is Leontief. The first assumption allows firms to use capacity to compete for demand, while the second assumption gives us a clear-cut notion of capacity.

#### 4.1 Technology

There is a unit mass of identical firms indexed by $j \in [0, 1]$. All goods produced are perfect substitutes and can be used either as consumption or investment. At the beginning of each period, firm $j$ has some capital $k_{j,t}$ inherited from the last period. The law of motion for capital is standard:

$$k_{j,t+1} = k_{j,t}(1 - \delta) + i_{j,t},$$  

(5)

where $i_{j,t}$ is the investment made by firm $j$ at time $t$ and parameter $\delta \in (0, 1]$ is the depreciation rate.

The production technology is assumed to be Leontief:

$$y_{j,t} = \min \left\{ \frac{l_{j,t}}{\alpha_v}, Ak_{j,t} \right\},$$  

(6)

where $y_{j,t}$ is the amount of goods produced by firm $j$, $l_{j,t}$ is the variable labor hired by firm $j$, parameter $\alpha_v > 0$ is the variable labor required per unit of output, and parameter $A > 0$ is the productivity of capital.

Capacity is defined as the output level at which the short run average total cost (SRAC) curve is tangent to the long run average total cost (LRAC) curve (e.g., Morrison, 1985). Along the LRAC curve, one minimizes the average total cost by adjusting both variable factors, such as labor, and quasi-fixed factors, such as capital.\(^{19}\) Along the SRAC curve, one minimizes the same average total cost, adjusting only variable factors. When output is at the point of tangency, full capacity utilization is achieved, and the average total cost could not be further minimized even when quasi-fixed factors can be adjusted.

\(^{19}\)Quasi-fixed factors are those that do not fluctuate with output in the short run but are adjustable in the long run.
Capacity defined in this way is also known as the economic measure of capacity in the literature (e.g., Nelson, 1989).

In practice, capacity is considered the maximum level of output that a firm can produce within a given period of time under a realistic working schedule, taking into account normal downtime (e.g., Corrado and Mattey, 1997). With constant returns to scale, the LRAC curve is flat. In this case, the economic measure of capacity is the output level at which the SRAC curve achieves its minimum, capturing partly the practical notion that extraordinary efforts are required to produce beyond capacity. Eiteman and Guthrie (1952) conduct a survey and find that the SRAC curve is typically downward sloping until a point near or at capacity. This suggests that the practical notion of capacity is roughly consistent with the theoretical definition of capacity when production technology exhibits constant returns to scale.

With a Leontief technology, the SRAC curve of firm $j$ is given by

$$SRAC(y_{j,t}, k_{j,t}; w_t) = w_t \alpha_v + (r + \delta) \frac{k_{j,t}}{y_{j,t}}, \quad (7)$$

where $w_t$ is the real wage rate, $r > 0$ is the steady state real interest rate\(^{20}\), and $y_{j,t} \in (0, Ak_{j,t}]$ must be positive but no larger than the production limit. Note that the SRAC curve is downward sloping in output until it reaches the production limit.

The LRAC curve is the minimum of the SRAC when capital can be adjusted:

$$LRAC(y_{j,t}; w_t) = w_t \alpha_v + \frac{r + \delta}{A}, \quad (8)$$

which is flat, as the Leontief technology exhibits constant returns to scale.

Capacity $\bar{y}_{j,t}$ is the output level at which the SRAC curve reaches its minimum and is tangent to the flat LRAC curve. We have

$$\bar{y}_{j,t} = Ak_{j,t}, \quad (9)$$

which shows that capacity is simply the maximum output that can be produced within a given period of time. Because the amount of capital stock is predetermined, capacity may not be fully utilized when demand is not high enough.\(^{21}\)

I find that it is both theoretically appealing and empirically plausible to start with a simple Leontief technology. Theoretically, the concept of capacity is naturally clear with

\(^{20}\)Nelson (1989) uses the yield on the firm’s latest issue of long-term debt to measure the cost of capital in practice. I instead use the steady state real interest rate for simplicity. Because the technology is Leontief, the properties of the model, however, will not be affected by the measure used.

\(^{21}\)If demand is permanently low and if capacity is sluggish to adjust, capacity utilization can remain low for quite a while. But underutilization caused by this mechanism is still temporary, as firms have the ability to gradually reduce their unused capacity over years. Hence, the mechanism cannot explain why capacity is underutilized for over 70 years since 1949 (see Supplementary Appendix K.2 for further discussions).
Leontief technology, consistent with both the theoretical definition and the practical notion. Thus, researchers who explicitly model capacity utilization often assume that firms produce with some Leontief technology at least in the short run (see, e.g., Fagnart et al., 1997 and Boehm and Pandalai-Nayar, 2022). This is also a standard assumption in the management science and operations research literature.22

Empirically, Leontief technology is not uncommon. For example, drivers and cars are perfect complements in a taxi company; assembly workers and assembly lines are close to perfect complements in mass production firms. At the micro-level, firms are likely to operate with fixed input-output coefficients especially in the short run because these coefficients are much dictated by the technologies embodied in capital and are carefully designed by modern engineers (Eiteman, 1947). The observation that input-output coefficients are fixed in the short run also motivates the putty-clay technology introduced by Johansen (1959). Indeed, empirical evidence based on accounting, engineering, or questionnaire data suggests that marginal cost at the micro-level is typically constant at least up until some point close to capacity (see Walters, 1963 for a literature survey).

4.2 Buyers and the Goods Market Structure

The purchasing process takes two steps. First, a household or a firm decides how many goods shall be consumed or invested in period $t$ based on the aggregate price $P_t$. Second, the household or the firm sends out buyers to purchase the goods for them. Each buyer purchases a unit of goods. All buyers are identical.

Each firm $j$ supplies its capacity $\bar{y}_{jt}$ to the market, waiting for buyers to show up. The model abstracts away from inventories. Unless buyers show up, capacity would not be utilized, and goods or services would not be produced (e.g., Bai et al., 2012 and Michaillat and Saez, 2015). Hence, when buyers go to the market, they search for capacity, i.e., the production potential, to satisfy their demand.

If a buyer’s demand is matched with a unit of capacity supplied by firm $j$, the buyer would purchase from firm $j$. The payoff for the buyer is assumed to be a strictly decreasing

---

function of the real price $\frac{P_{j,t}}{P_t}$ charged by the firm:\footnote{It will become clear later that the logarithmic functional form helps generate a demand curve of constant elasticity. However, the results of this paper will not be affected as long as the payoff function is strictly decreasing in the real price of the goods purchased. In addition, although buyers purchase on behalf of households and firms, the buyers have their own payoffs. This assumption is convenient as it simplifies the aggregation problem by allowing buyers to be homogeneous across potentially different households and firms. In Supplementary Appendix B, I show that a homogeneous purchasing problem can be obtained directly as a sub-problem of households and firms, if the information processing cost is proportional to the amount of goods purchased and is paid in terms of goods or labor. In this case, the payoff of purchasing a unit of goods from firm $j$ decreases linearly rather than log-linearly in the real price charged by the firm. However, the results of this paper remain unchanged.}

$$v_t(j) = -\ln \left( \frac{P_{j,t}}{P_t} \right).$$

(10)

To yield the highest payoff, in principle, buyers want to purchase only from firms that charge the lowest price. However, as in Mattsson and Weibull (2002), I assume that it is costly for buyers to implement the best possible outcome because they need to process some price information in order to direct their actions towards the best possibility.\footnote{In general, the payoff function (10) can also be affected by qualities (see Supplementary Appendix C for how the model can be extended to incorporate quality differences). The difficulty of processing heterogeneous quality information is arguably an important reason why it is costly for buyers to find the best option. However, for simplicity, the model abstracts away from quality differences. Perhaps surprisingly, despite recent improvements in information technology, price dispersion on online platforms is large and persistent even for near homogeneous goods or in a controlled laboratory environment, suggesting that even the information processing costs of prices have not been fully removed by information technology (see Pan et al., 2004, Baye and Morgan, 2004, and Baye et al., 2006 for a literature survey and Supplementary Appendix K.3 for further discussions).}

Without exerting any information processing effort, buyers can purchase only blindly and randomly. The purchasing behavior in this case is called the default purchasing behavior, which is the most inattentive behavior of buyers. I assume that the default purchasing behavior is an undirected search for capacity, a behavior that results in a random matching between the demand from buyers and the capacity supplied by firms.

The rationale behind the above assumption is as follows. Consider a buyer with a unit of demand that needs to be matched with a unit of capacity. The matching process can be described by a probability density function $f_t$. $f_t(x)$ gives the likelihood that a unit of capacity indexed by $x \in [0, \bar{y}_t]$ is matched with the buyer, where $\bar{y}_t \equiv \int_0^1 \tilde{y}_{j,t} dj$ is the total amount of capacity supplied by firms. If the buyer pays no attention to prices, the matching process should be the most disordered. The disorder of the matching process can be measured by the entropy of $f_t$, which is given by

$$-\int_0^{\bar{y}_t} \ln (f_t(x)) f_t(x) dx.$$  

(11)

For all distributions with a support limited to the interval $[0, \bar{y}_t]$, the maximum entropy distribution is the uniform distribution: $f_t(x) = 1/\bar{y}_t$. Hence, the most disordered matching process is that each unit of capacity has an equal probability to be matched.
Thus, without exerting any information processing effort, the probability density that a buyer purchases from firm $j$ is proportional to the capacity held by the firm:

$$n_t^* (j) \equiv \frac{y_{j,t}^{-}}{y_t}, \quad (12)$$

where $n_t^*$ is called the default probability density function as it describes the default purchasing behavior.

If a buyer wants to deviate from the default $n_t^*$, she has to process some price information and incurs an information processing cost. Let $n_t$ be the probability density function eventually obtained by a buyer. $n_t (j)$ gives the likelihood that the buyer purchases from firm $j$. Following the rational inattention literature, I assume that the information processing cost is proportional to the amount of information processed, measured by the relative entropy of $n_t$ with respect to $n_t^*$. Relative entropy, also known as Kullback–Leibler (KL) divergence, is non-negative, convex, and obtains its minimum value zero if $n_t = n_t^*$:

$$D_{KL} (n_t || n_t^*) = \int_0^1 n_t (j) \ln \left( \frac{n_t (j)}{n_t^* (j)} \right) dj. \quad (13)$$

Intuitively, the more different $n_t$ is from $n_t^*$, the more information needs to be processed.

The buyer’s problem is to maximize the expected payoff net of the information processing cost:

$$\max_{n_t \geq 0} \int_0^1 v_t (j) n_t (j) dj - \Lambda \left( \int_0^1 n_t (j) \ln \left( \frac{n_t (j)}{n_t^* (j)} \right) dj \right), \quad (14)$$

subject to

$$\int_0^1 n_t (j) dj = 1, \quad (15)$$

where $\Lambda > 0$ is the unit cost of processing information. Note that $n_t^*$ is exogenous to the above buyer’s problem.

Mattsson and Weibull (2002) show that the unique solution to the above problem is

$$\forall j \in [0, 1] : \quad n_t (j) = \frac{n_t^* (j)e^{v_t (j)}}{\int_0^1 n_t^* (j)e^{v_t (j)} dj}. \quad (16)$$

Equation (16) says that the optimal probability of purchasing from firm $j$ is proportional to the default probability $n_t^* (j)$ and moderated by the payoff of purchasing from firm $j$. The higher the payoff, the higher the probability that firm $j$ would be chosen by the buyer. We can substitute the buyer’s payoff function (10) for $v_t (j)$ in equation (16) and obtain

$$\forall j \in [0, 1] : \quad n_t (j) = \frac{n_t^* (j)P_{j,t}^{-\frac{1}{2}}}{\int_0^1 n_t^* (j)P_{j,t}^{-\frac{1}{2}} dj}. \quad (17)$$
If all prices are equal, buyers will be indifferent choosing between any two firms and will have no incentive to do any costly information processing. In this case, buyers will simply follow the undirected search for capacity and will be distributed across firms according to the default probability density function \( n^*_t \).

If prices are different, firms that charge a relatively low price attract additional buyers. However, because of a positive information processing cost, not all buyers will purchase the goods at the lowest price. Even the goods with the highest price will still be purchased by some buyers. We can see from equation (17) that \( n_t(j) \) is always positive as long as \( P_{j,t} \) remains finite. Intuitively, since buyers have a limited capability of processing price information, they rationally choose to be partially inattentive to prices rather than always purchasing the cheapest goods. Hence, buyers allow themselves to make some “mistakes” with positive probabilities in order to save the information processing cost.

Since all buyers are identical, demand aggregation is easy. The demand for the goods produced by firm \( j \) is the multiplication of the total demand \( y_t \) from buyers and the share of buyers who purchase from firm \( j \): \( y_{j,t} = n_t(j) y_t \). Substituting the buyer’s solution (17) for \( n_t(j) \), we have

\[
y_{j,t} = \frac{n_t^*(j) P_{j,t}^{-\varepsilon}}{\int_0^1 n_t^*(j) P_{j,t}^{-\varepsilon}dj} y_t \cdot \frac{\int_0^1 y_{j,t} P_{j,t}^{-\varepsilon}dj}{\int_0^1 y_t P_{j,t}^{-\varepsilon}dj} y_t,
\]

where \( \varepsilon \equiv \Lambda^{-1} \) is the price elasticity of demand and the second equality comes from equation (12). Note that the demand function derived here is consistent with the demand function (1) used in the empirical framework of Section 3.25

The aggregate price \( P_t \) must satisfy the following aggregation condition:

\[
P_t y_t = \int_0^1 P_{j,t} y_{j,t} dj,
\]

which ensures that the money spent on aggregate goods is equal to the money earned by firms.

With a positive information processing cost, buyers act as if the same goods produced by different firms are imperfect substitutes. The lower the information processing cost is, the more competitive the market is. In one extreme where \( \Lambda \to 0 \), buyers are fully

---

25The property that the price elasticity of demand is \( \varepsilon = \Lambda^{-1} \) and the property that the capacity elasticity of demand is \( \partial \ln y_{j,t}/\partial \ln y_{j,t} = 1 \) are both special results of the assumption that all goods are technically perfect substitutes. If goods produced by different firms are imperfect substitutes, the demand function could be more generally written as \( y_{j,t} \propto \frac{\Lambda^{\rho^{-1}}}{\Lambda^{\rho^{-1}} + a_{j,t}} \frac{1}{P_{j,t}^{-\varepsilon} (\Lambda^{\rho^{-1}} + a_{j,t})} \), where \( \rho^{-1} > 0 \) is the elasticity of substitution among goods and \( a_{j,t} \) is the quality of goods \( j \) (see Supplementary Appendix C). The extended demand function is the same as the demand function (1) used in Section 3. In general, both the elasticity of substitution \( \rho^{-1} \) and the unit cost of processing information \( \Lambda \) affect the price or capacity elasticity of demand. The main properties of the model, however, will not be affected. To keep the model simple and to highlight the role played by \( \Lambda \), the paper focuses on the case where goods are perfect substitutes (\( \rho = 0 \)) and firms are homogeneous.
attentive to prices, the search process is fully directed, and the goods market is perfectly competitive \((\varepsilon \to \infty)\). In the other extreme where \(\Lambda \to \infty\), buyers pay no attention to prices, the search process is fully undirected, and the demand that goes to each firm no longer depends on relative prices \((\varepsilon \to 0)\). Hence, the model spans both directed search and undirected search as well as the intermediate case where the search is partially directed. The parameter \(\Lambda\) controls the degree to which the search is directed.

In addition to prices, the default purchasing behavior \(n^*_{t}\) plays an important role in determining the relative size of demand. \(n^*_{t}\) is endogenously affected by the relative capacity of the firm (see equation (12)). For a given amount of total demand from buyers, if a firm expands its capacity while others do not, it “steals” or attracts demand from others. Thus, in addition to the usual price competition, there is capacity competition among firms.

For example, Starbucks can expand its market share by opening more brick-and-mortar coffee stores than its competitor Costa. A printing store can expand its market share by installing more printing machines than other printing stores. The empirical evidence discussed in Section 3 also supports the presence of capacity competition. Intuitively, firms with a larger capacity are more likely to be matched with buyers who are not fully attentive to prices. Demand is thus affected by capacity. In the standard Dixit-Stiglitz monopolistic competition model, capacity has no effect on the allocated demand. Therefore, the demand curve in my model is fundamentally different from that in the standard Dixit-Stiglitz setup, though both imply that demand is affected by \(P_{j,t}\).

### 4.3 Households

There is a unit mass of identical households. Consider a representative household who maximizes her expected lifetime utility

\[
\max_{\{c_t, l_t\}_{t=0}^{\infty}} \mathbb{E}_0 \left( \sum_{t=0}^{\infty} \beta^t u(z_{c,t}, c_t, l_t) \right),
\]

subject to the budget constraint of the household

\[
c_t = w_t l_t + d_t,
\]

where \(\beta \in [0, 1)\) is the subjective discount factor, \(z_{c,t}\) is a preference parameter that varies with time, \(c_t\) is the amount of consumption goods that the household asks her buyer to purchase, \(l_t\) is the labor supply, \(w_t\) is the real wage rate, and \(d_t\) is the amount of dividends received from firms. I assume that the labor market is perfectly competitive. Hence, the household takes the real wage rate \(w_t\) as exogenous.
The functional form of the one-period utility is assumed to be
\[
    u(z_{c,t}, c_t, l_t) = \begin{cases} \\
    \phi e^{z_{c,t}} \left( c_t / (\phi e^{z_{c,t}}) \right)^{1-\gamma} - \bar{\omega} l_t, & \gamma \neq 1 \\
    \phi e^{z_{c,t}} \ln \left( c_t / (\phi e^{z_{c,t}}) \right) - \bar{\omega} l_t, & \gamma = 1 
\end{cases}
\] (22)

where \(z_{c,t}\) is a consumption demand disturbance, \(\phi > 0\) is a scaling parameter, \(\gamma^{-1} > 0\) is the elasticity of inter-temporal substitution, and \(\bar{\omega} > 0\) is the marginal dis-utility of labor. Assume that \(z_{c,t}\) follows an AR(1) process \(z_{c,t} = \rho z_{c,t-1} + e_{c,t}\), where \(\rho \in [0,1)\) is a persistence parameter and \(e_{c,t}\) is an i.i.d. innovation. The assumption that the marginal dis-utility of labor is a constant follows from the indivisible labor theory proposed by Hansen (1985) and Rogerson (1988).

The first-order conditions (FOCs) of the household’s problem are
\[
    \lambda_t = \left( \frac{c_t}{\phi e^{z_{c,t}}} \right)^{-\gamma},
\]
(23)
\[
    \bar{\omega} = \lambda_t w_t,
\]
(24)
where \(\lambda_t\) is the Lagrangian multiplier for the household’s budget constraint (21) and can be interpreted as the shadow price of goods measured in terms of utils.

4.4 Firms

Each firm aims to maximize its firm value, which is the present value of the firm’s dividend flows \(d_{j,t}\):
\[
    \max_{\{P_{jt}, y_{j,t}, i_{j,t}, k_{j,t+1}\}_{t=0}^{\infty}} \mathbb{E}_0 \left( \sum_{t=0}^{\infty} \beta^t \frac{\lambda_t}{\lambda_0} d_{j,t} \right),
\]
(25)
subject to the demand curve (18), the resource constraint of the firm
\[
    d_{j,t} + i_{j,t} = \left( \frac{P_{jt}}{P_t} - w_t \alpha_v \right) y_{j,t},
\]
(26)
the capacity constraint
\[
    y_{j,t} \leq A k_{j,t},
\]
(27)
and the law of motion for capital (5). Because firms are owned by households, the stochastic discount factor is given by \(\beta^{\lambda_{t+1}/\lambda_t}\), which also implies that the real interest rate in steady state is \(r = \beta^{-1} - 1\).

To prevent firms from charging an infinitely high price, the unit cost of processing information \(\Lambda\) should not be too large. I assume that \(\Lambda\) is less than one; thus, the price elasticity of demand is larger than one: \(\varepsilon = \Lambda^{-1} > 1\).
The FOCs of the firm’s problem are

\[
\frac{P_{j,t}}{P_t} = \frac{\varepsilon}{\varepsilon - 1} (w_t \alpha_v + \mu_{j,t}),
\]

(28)

\[
1 = \beta \mathbb{E}_t \left( \frac{\lambda_{t+1}}{\lambda_t} \left( \frac{P_{j,t+1}}{P_{t+1}} \varepsilon A u_{j,t+1} + A \mu_{j,t+1} + 1 - \delta \right) \right),
\]

(29)

where \( \mu_{j,t} \geq 0 \) is the Lagrangian multiplier of the capacity constraint (27) and \( u_{j,t+1} \equiv y_{j,t+1}/y_{j,t+1} = y_{j,t+1}/Ak_{j,t+1} \) is the capacity utilization rate of the firm.

Equation (28) says that the firm sets its price according to a constant markup rule. If the firm operates below its capacity limit, the price charged by the firm will be proportional to the firm’s marginal cost: \( w_t \alpha_v \). Once demand hits the firm’s capacity limit, the price will be above \( \frac{\varepsilon}{\varepsilon - 1} w_t \alpha_v \) so as to equate demand to capacity: \( y_{j,t} = y_{j,t} = Ak_{j,t} \).

Equation (29) shows that there are two reasons for a firm to invest in capital. First, capital investment relaxes the capacity constraint. This aspect is captured by the term \( A\mu_{j,t+1} \), which is the shadow value of relaxing the capacity constraint. Second, capital investment makes the goods produced by the firm more likely to be purchased by buyers. This aspect is captured by the multiplication of the demand attracted per unit of capital \( Au_{j,t+1} \) and the profit contributed by each unit of demand attracted \( (P_{j,t+1}/P_{t+1}) \varepsilon^{-1} \).

### 4.5 Symmetric Equilibrium

Since all firms are identical, we can omit the subscripts that index a variable to a particular firm after obtaining the FOCs. The symmetric equilibrium is a stable stochastic process of nine variables \( (c_t, \lambda_t, w_t, i_t, u_t, \mu_t, l_t, k_{t+1}, y_t) \) that satisfies the household’s FOCs (23)-(24), the firm’s pricing condition

\[
1 = \frac{\varepsilon}{\varepsilon - 1} (w_t \alpha_v + \mu_t),
\]

(30)

the firm’s investment condition

\[
1 = \beta \mathbb{E}_t \left( \frac{\lambda_{t+1}}{\lambda_t} \left( \frac{1}{\varepsilon} A u_{t+1} + A \mu_{t+1} + 1 - \delta \right) \right),
\]

(31)

the measure of the capacity utilization rate \( u_t = y_t/Ak_t \), the complementary slackness condition for the capacity constraint \( \mu_t (Ak_t - y_t) = 0 \), where \( \mu_t \geq 0 \) and \( Ak_t - y_t \geq 0 \), the amount of labor hired according to the Leontief production function \( l_t = \alpha_v y_t \), the law of motion for capital

\[
k_{t+1} = k_t (1 - \delta) + i_t,
\]

(32)
and the aggregate resource constraint

\[ c_t + i_t = y_t. \] (33)

5 Properties of the Basic Capacity Underutilization Model

In this section, I show the following properties of the basic capacity underutilization (CU) model. First, the decentralized equilibrium is generally inefficient because buyers are not fully attentive to prices. Second, if the degree of inattention is large enough, capacity will be underutilized in steady state and the economy will exhibit chronic excess capacity. Third, locally around the steady state where capacity is in excess, capital resources are slack. Finally, because of the capital resource slackness, the real wage rate is acyclical and output is highly and much more responsive to demand shocks than in the standard RBC model.

5.1 Inefficiency and Capacity Competition Externality

The efficient allocation can be obtained by solving a corresponding social planner’s problem:

\[
\max_{\{c_t, i_t, y_t, k_t, l_t\}} \mathbb{E}_0 \left( \sum_{t=0}^{\infty} \beta^t u(z_{c,t}, c_t, l_t) \right),
\]

subject to the Leontief production function (6), the law of motion for capital (32), and the aggregate resource constraint (33).

The conditions that characterize the efficient allocations are the same as those that characterize the decentralized equilibrium, except for the condition that describes the trade-off between the benefit and the cost of production

\[ \lambda_t = \bar{\omega}_v + \lambda_t \mu_t, \] (35)

and the condition that gives the investment

\[ 1 = \beta \mathbb{E}_t \left( \frac{\lambda_{t+1}}{\lambda_t} (A \mu_{t+1} + 1 - \delta) \right). \] (36)

In the decentralized equilibrium, the trade-off between the benefit and the cost of production is given by the firm’s pricing condition (30) combined with the household’s labor supply condition (24):

\[ \lambda_t = \frac{\varepsilon}{\varepsilon - 1} \left( \bar{\omega}_v + \lambda_t \mu_t \right); \] (37)

and the investment condition is given by equation (31).
Hence, the equilibrium of the basic CU model is efficient if and only if $\varepsilon \to \infty$ or equivalently $\Lambda \to 0$. In this case, the basic CU model is reduced to a standard RBC model with a Leontief production technology, or a Leontief-RBC model for short, where buyers are fully attentive to prices and the goods market is perfectly competitive.

In general, as long as buyers are not fully attentive to prices, the equilibrium of the basic CU model is inefficient.

First, the rational inattention of buyers allows firms to enjoy some market power, depressing the real wage rate and inflating the price in utils $\lambda_t$. As a result, consumption tends to be inefficiently low.

Second, when buyers are not fully attentive to prices, they search for capacity in a partially undirected way. Hence, firms can expand their relative capacity to attract more demand than their competitors. Each unit of demand is profitable because firms can take advantage of the inattentive buyers to charge a markup. Thus, capacity expansion by one firm has a negative externality on others. From an individual firm’s perspective, capacity expansion not only relaxes the capacity constraint, but also attracts profitable demand (see equation (31)). From a social planner’s perspective, however, the return on capacity is derived only from relaxing the capacity constraint (see equation (36)). Because of this capacity competition externality, firms tend to hold excess capacity.

The first mechanism also shows up in the standard Dixit-Stiglitz setup. The second mechanism characterizes the basic CU model and is important for us to understand long-term capacity underutilization and chronic excess capacity.

## 5.2 Long-term Capacity Underutilization and Chronic Excess Capacity

Let us focus on the steady state of the basic CU model. If a variable is of the form $x_{j,t}$ or $x_t$, its value in steady state is denoted by $x_j$ or $x$.

If the unit cost of processing information is large enough, i.e., $\Lambda \in \left( \frac{r+\delta}{\varepsilon}, 1 \right)$, we have,

---

26 In a symmetric equilibrium, since all firms charge the same price, the behavior of the buyers is completely undirected as there is no need to process any price information. However, this does not mean that the presence of information processing cost is not important. Without the information processing cost, buyers would purchase only the cheapest goods, the market is perfectly competitive, and firms are unable to use capacity to compete for demand and are unable to charge a markup.

27 In a somewhat similar spirit, the literature on industrial organization documents the possibility of established firms holding excess capacity to deter entry (e.g., Spence, 1977, Dixit, 1980, and Bulow et al., 1985). Established firms invest in capacity to protect demand from being “stolen” by potential entrants. In my model, firms invest in capacity to “steal” demand from others. Both mechanisms allow capacity to have a positive effect on demand. Because of the complex strategic interactions involved, the entry deterrence mechanism is rarely incorporated into macroeconomic models. However, the capacity competition mechanism proposed in this paper can be easily incorporated into the existing macroeconomic framework.
according to the firm’s investment condition (31),
\[ u = \frac{r + \delta}{A\Lambda} - \frac{\mu}{A} \leq \frac{r + \delta}{A\Lambda} < 1, \tag{38} \]
which says that the capacity utilization rate \( u \) in steady state is less than one. In other words, there is long-term capacity underutilization.

If \( \Lambda \to 0 \), however, the basic CU model is reduced to the Leontief-RBC model, and the allocation of the economy becomes efficient. In this case, according to the socially optimal investment condition (36), we have
\[ \mu = \frac{r + \delta}{A} > 0, \tag{39} \]
which means that capacity must be fully utilized in steady state: \( y = Ak \).

The comparison between the basic CU model and the Leontief-RBC model highlights how a single modification on the assumption of the behavior of buyers can cause a substantial change.

In the Leontief-RBC model, buyers pay full attention to prices. Thus, capacity has no effect on demand. The only reason for firms to invest in capital is to relax the capacity constraint. If the capacity constraint is not binding in steady state, the marginal value of relaxing the capacity constraint is zero: \( A\mu = 0 \). The only value of capital disappears, but holding capital is costly. Hence, the firm will reduce its capital holding until its capacity constraint is binding in steady state.

In the basic CU model, however, buyers are not fully attentive to prices and search for capacity in a partially undirected way. Hence, there is capacity competition among firms. The limited capability of buyers to process price information also provides firms with a market power, which allows them to earn a profit for each unit of goods sold: \( \varepsilon^{-1} = \Lambda > 0 \). This profit is a lure for firms to use capacity to compete for demand. Even though capacity is not fully utilized, as long as the profit rate is large enough, firms will not want to reduce their capacity to lose demand to others, justifying the existence of long-term capacity underutilization in the basic CU model.

In addition, from an individual firm’s perspective, capacity utilization is restricted by demand. If demand was sufficient such that firm \( j \) could always sell its products at the price \( P_j/P > w\alpha \), firm \( j \) would produce at full capacity to maximize its profit. The model, therefore, is consistent with the fact that insufficient demand is the primary reason cited by firms for capacity underutilization.

I now examine whether the underutilized capacity at the firm level is also in excess.

**Definition 1.** (Excess capacity) Capacity is in excess if firms hold more capital than what could be justified by minimizing the present value of aggregate total costs. If capacity is in excess in steady state, we say that the economy exhibits chronic excess capacity.
It is possible that some firms do not fully utilize their capacity, but the capital held by firms minimizes the present value of aggregate total costs. In this case, capacity is not in excess. For example, having some capacity underutilized at the firm level saves aggregate production costs in Fagnart et al. (1997) and reduces purchasing costs in Bai et al. (2012). In both cases, capacity is efficiently utilized and the present value of aggregate total costs is minimized. In the basic CU model, however, long-term capacity underutilization at the firm level also implies that the economy exhibits chronic excess capacity.

Proposition 1. The basic CU model exhibits chronic excess capacity if and only if capacity is not fully utilized at the firm level in steady state.

Proof. See Supplementary Appendix D for details.

Intuitively, although it is valuable to accumulate extra capacity to attract demand from an individual firm’s perspective, from a social planner’s perspective, the only value of capacity is to reduce the aggregate total costs by relaxing capacity constraints. Therefore, if firms in steady state hold more capacity than output, the capital held by firms will be too large to be justified by cost minimization. It will become clear later that the existence of chronic excess capacity allows capital resources to be slack and the economy to be highly responsive to demand shocks.

In the rest of the paper, I assume that $\Lambda$ is sufficiently large and consumption demand shocks are not too large such that the capacity constraint (27) never binds.\(^\text{28}\) With this assumption, I can ignore the occasionally binding capacity constraint to focus on the local dynamic properties of the basic CU model around the steady state where capacity is in excess.

5.3 Capital Resource Slackness

To explain how the existence of chronic excess capacity affects the model dynamics, I introduce the concept of capital resource tightness (or slackness), which captures the scarcity of capital as a production factor.

Definition 2. (Capital resource tightness) If a marginal decrease in aggregate capital makes the current aggregate output level infeasible or leads to a decrease in the marginal product of labor, capital resources are tight. If a marginal decrease in aggregate capital has no effect on the marginal product of labor, capital resources are slack.

Capital resource tightness can be measured by the decrease in the real wage rate that keeps the real marginal cost unchanged after a decrease in aggregate capital, where the

\(^{28}\) It might be interesting to note that if the capacity constraint never binds, one can regard the model as equivalent to a standard monopolistic competition model where the production function is simply linear in labor, unaffected by any physical capital, and firms accumulate a “marketing capital” that affects demand. This analogy breaks down, of course, if the capacity constraint binds occasionally.
decrease in the real wage rate measures the decrease in the marginal product of labor. Let \( \zeta \) be the magnitude of capital resource tightness at the steady state:

\[
\zeta \equiv -\frac{C_{Y,K}K}{C_{Y,w}w},
\]

(40)

where \( Y \) is the aggregate output, \( K \) is the aggregate capital, and \( C \) is the aggregate variable cost function \( C(Y,K;w) \). \( C_{Y,K} \) and \( C_{Y,w} \) are the second order partial derivatives of \( C \). If the only component of the aggregate variable cost is labor cost, capital resource tightness can be expressed as the negative capital elasticity of the real marginal cost.\(^{29}\)

With constant returns to scale, capital resource tightness can be further expressed as the output elasticity of the real marginal cost, i.e., the steepness of the real marginal cost curve:

\[
\zeta = -\frac{C_{Y,K}K}{C_Y} = -\frac{\partial \ln C_Y(Y,K;w)}{\partial \ln K} = \frac{\partial \ln C_Y(Y,K;w)}{\partial \ln Y}.
\]

(41)

Intuitively, the steeper the real marginal cost curve is, the tighter capital resources are.

In the basic CU model, the aggregate variable cost function is linear in \( Y \) until aggregate capacity is fully utilized: \( C(Y,K;w) = w\alpha Y \) for \( Y \leq \bar{Y} \), where \( Y = \int_0^1 y_j dj \), \( K = \int_0^1 k_j dj \), and \( \bar{Y} = AK \) is the aggregate capacity. Therefore, the shape of the real marginal cost curve \( C_Y \) is flat when aggregate capacity is underutilized \( Y < AK \) and is vertical when aggregate capacity is fully utilized \( Y = AK \).

If \( \Lambda \) is sufficiently large, i.e., \( \Lambda \in (r+\delta A, 1) \), the economy exhibits chronic excess capacity, and aggregate capacity is underutilized in steady state \( Y < AK \).\(^{30}\) Hence, the real marginal cost curve is flat and capital resources are completely slack \( (\zeta = 0) \) locally around the steady state.

If \( \Lambda = 0 \), the economy is reduced to the Leontief-RBC model, where aggregate capacity is fully utilized in steady state \( Y = AK \). Hence, the real marginal cost curve is vertical and capital resources are infinitely tight \( (\zeta = \infty) \) locally around the steady state.

From now on, I’ll simply use \( y \) (or \( k \)) to denote aggregate output \( Y \) (or aggregate capital \( K \)) because \( Y = y \) and \( K = k \) in the symmetric equilibrium.

### 5.4 Acyclical Real Wage Rate and Large Responses to Demand

Capital resource slackness implies an acyclical real wage rate and allows output to be highly responsive to demand shocks, while capital resource tightness implies a countercyclical real wage rate and dampens the response of output to demand shocks.

\(^{29}\)This is true in a large number of macroeconomic models including the standard RBC models and the basic CU model. If \( C(Y,K;w) = wL(Y,K) \), where \( L \) is the aggregate variable labor required to produce the aggregate output \( Y \) given the aggregate capital \( K \), we have \( C_{Y,w} = C_Y \) and \( \zeta = -C_{Y,K}K/C_Y \).

\(^{30}\)In general, if the aggregate production technology is of constant returns to scale, one can show that capacity is in excess if and only if aggregate capacity is underutilized (see Supplementary Appendix D for details).
In the Leontief-RBC model ($\Lambda = 0$), for example, capital resources are infinitely tight. When demand increases, output cannot increase, but the marginal value of relaxing the capacity constraint $\mu_t$ increases, creating a pressure for the real marginal cost to increase. According to the firm’s pricing condition (30), this pressure is transmitted to a decrease in the real wage rate. The counter-cyclical real wage rate will then cause a pro-cyclical price in utils $\lambda_t$, reducing the responses of consumption and investment to demand shocks (see equations (23) and (31)).

By contrast, in the basic CU model ($\Lambda > r + \delta + k\Lambda$), capital resources are slack. If there is a higher demand, more output can be produced without causing any pressure for the real marginal cost to increase. The real wage rate $w_t$ and the price in utils $\lambda_t$ are both locally constant under demand shocks.\(^{31}\) Hence, the responses of consumption and investment to demand shocks are not reduced.

To see this clearly, I log-linearize the household’s consumption condition (23) and the firm’s investment condition (31) of the basic CU model around the steady state where capacity is in excess. Let $\hat{x}_t \equiv \ln x_t - \ln x$ be the log-deviation of $x_t$ from its steady state $x$. The log-linearized consumption condition is

$$0 = z_{c,t} - \hat{c}_t,$$

and the log-linearized investment condition is

$$0 = E_t (\hat{u}_{t+1}) = E_t (\hat{y}_{t+1} - \hat{k}_{t+1}).$$

Equation (42) shows that consumption responds one-for-one to changes in consumption demand. Equation (43) shows that if there is an expected increase in future demand, investment shoots up so that capacity can match with the increased demand in the next period. Although capital is not scarce as a production factor, firms still have a strong incentive to invest because the amount of demand that can be attracted by each unit of capacity invested increases with the expected aggregate demand $E_t (\hat{y}_{t+1})$. Since capital resources are slack, all the induced investment can and will be made in the first period. Therefore, the response of investment to persistent demand shocks is very large in the basic CU model.

### 5.5 Calibration

To illustrate the above results quantitatively, I first calibrate the basic CU model at a quarterly frequency and then conduct two numerical experiments with a completely transitory ($\rho_c = 0$) and a highly persistent ($\rho_c = 0.99$) consumption demand disturbance.

---

\(^{31}\)The real wage rate $w_t$ will be acyclical as long as capital resources are slack. But to have an acyclical price in utils $\lambda_t$, the dis-utility of labor also has to be linear.
respectively. In both cases, the standard deviation of the consumption demand innovation \( \sigma_c \) is set to 1%.

Table 7: Parameters and Calibration Targets – Basic CU Model

<table>
<thead>
<tr>
<th>Parameters ( \delta )</th>
<th>Values</th>
<th>Targets Matched</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quarterly depreciation rate</td>
<td>0.0210</td>
<td>0.021</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>1.0000</td>
<td>Elasticity of inter-temporal substitution 1</td>
</tr>
<tr>
<td>( \phi )</td>
<td>0.8300</td>
<td>Price in utils normalized to 1</td>
</tr>
<tr>
<td>( \bar{\omega} )</td>
<td>0.6798</td>
<td>Output normalized to 1</td>
</tr>
<tr>
<td>( \beta )</td>
<td>0.9747</td>
<td>Capacity utilization rate 0.8</td>
</tr>
<tr>
<td>( A )</td>
<td>0.1544</td>
<td>Investment to output ratio 0.17</td>
</tr>
<tr>
<td>( \alpha_v )</td>
<td>0.9120</td>
<td>Employment rate 0.912</td>
</tr>
<tr>
<td>( \Lambda )</td>
<td>0.3800</td>
<td>Labor share of income 0.62</td>
</tr>
</tbody>
</table>

We have eight parameters to calibrate. The depreciation rate \( \delta \) is calibrated to match the average ratio of gross private domestic investment to private fixed assets from 1947 to 2016 in the National Income and Product Accounts (NIPA) and the Fixed Assets Accounts (FAA) prepared by the U.S. Bureau of Economic Analysis (BEA). The inverse of the elasticity of inter-temporal substitution \( \gamma \) is chosen to be 1, a value that implies a commonly used log utility. One can see from equation (42) that the value of \( \gamma \) actually does not affect the dynamic properties of the basic CU model.

The rest of the six parameters are jointly calibrated to achieve the following six targets in steady state. The price in utils in steady state \( \lambda \) is normalized to 1. The size of output in steady state \( y \) is normalized to 1. The employment rate in steady state \( l \) is matched to one minus the average of U-5 and U-6 from 1994 to 2016 published by the Bureau of Labor Statistics (BLS).\(^{33}\) The investment to output ratio in steady state \( i/y \) is matched to the average ratio of gross private domestic investment to gross domestic product (GDP) from 1947 to 2016 in NIPA. The capacity utilization rate in steady state \( u \) is matched to the average manufacturing capacity utilization rate from 1947 to 2016 reported by the FRB. The labor share of income in steady state \( w/l \) is matched to the average labor share of income estimated by the BLS from 1947 to 2016.

Table 7 summarizes the calibrated parameter values and their mostly associated targets. Given this set of parameter values, all targets are exactly matched in steady state. For comparison, I also calibrate the Leontief-RBC model and a standard RBC model with a Cobb-Douglas (CD) production technology, or the CD-RBC model for short (see Supplementary Appendix E for details).

\(^{32}\)Equilibrium conditions are log-linearized to obtain numerical solutions. The magnitude of \( \sigma_c \) affects only the level but not the shape of the impulse response functions obtained in this way.

\(^{33}\)According to the BLS, U-5 is total unemployed, plus all persons marginally attached to the labor force, as a percent of the civilian labor force plus all persons marginally attached to the labor force; and U-6 is U-5 plus total employed part time for economic reasons, as a percent of the civilian labor force plus all persons marginally attached to the labor force.
5.6 Impulse Responses and Discussion

Figure 2 plots the impulse response functions (IRFs) for the basic CU model together with those for the Leontief-RBC model and those for the CD-RBC model. The first two rows of Figure 2 show the case where the change in consumption demand is completely transitory ($\rho_c = 0$), while the last two rows of Figure 2 show the case where the change in consumption demand is highly persistent ($\rho_c = 0.99$).

Impulse Responses for the Basic CU Model and the RBC Models

![Impulse Responses Graph](https://example.com/impulse_responses.png)

Figure 2: Responses to a 1 percentage point increase in demand shock $e_c$. The solid lines are for the basic CU model. The dashed lines are for the Leontief-RBC model. The dotted lines are for the CD-RBC model. All variables are expressed as log deviations from steady state.

Indeed, although the only difference between the Leontief-RBC model and the basic CU model is the assumption on the behavior of buyers, the local dynamics of the two models are drastically different.

In the Leontief-RBC model, the capacity constraint is binding and capital resources are infinitely tight. Output is restricted by capacity and cannot be changed immediately. Hence, consumption and investment must move in opposite directions. If there is a one-off increase in consumption demand ($\rho_c = 0$), a 1% increase in consumption demand can lead only to a 0.84% increase in consumption, and the increase in consumption must be satisfied by a sharp decrease in investment. If the increase in consumption demand is...
highly persistent \((\rho_c = 0.99)\), it is worthwhile to invest so that households can enjoy a higher consumption in the future. However, the induced investment must be satisfied by a temporary decrease in consumption. Besides, the real wage rate correlates negatively with demand shocks and the Solow residual is acyclical. In the U.S. data, however, consumption and investment co-move, the real wage rate is most likely pro-cyclical or acyclical, and the Solow residual is strongly pro-cyclical. These illustrate the typical difficulties of demand driven business cycles in an economy where capacity is fully utilized.

In the basic CU model, capacity is in excess and capital resources are slack. Consumption and investment now move in the same direction when \(\rho_c > 0\). If changes in consumption demand are completely transitory \((\rho_c = 0)\), a 1% increase in consumption demand leads to a 1% increase in consumption, and no investment has to be sacrificed. If changes in consumption demand are persistent \((\rho_c = 0.99)\), a 1% increase in consumption demand leads to a 1% increase in consumption and also a huge increase in investment. Firms want to increase their investment because the amount of demand that can be attracted by each unit of capacity invested increases. A 1% increase in future consumption encourages roughly a 1% increase in capacity, which, in turn, requires roughly a \((1/\delta)\)% = 48% increase in investment. Since capital resources are slack, investment can be produced immediately so that capacity increases quickly to compete for the increased demand. Because of the induced investment, a 1% persistent increase in consumption demand is amplified to more than an 8% increase in output. Capital resource slackness also allows the real wage rate to be acyclical. Finally, as a result of a pro-cyclical capacity utilization, the implied Solow residual is highly pro-cyclical under demand shocks.

In the CD-RBC model, capital resources are tight but not as tight as in the Leontief-RBC. Hence, the response of output to demand shocks is larger than that in the Leontief-RBC model but still much smaller than in the basic CU model. Particularly, the response of investment is as limited as in the Leontief-RBC model. In addition, the real wage rate correlates negatively with demand shocks and the Solow residual is acyclical.

---

\(^{34}\)Throughout this paper, the Solow residual is calculated by assuming a CD production function for the final product and a CD labor share of 0.62.

\(^{35}\)The one-off response of consumption does not mean that the consumption smoothing motive is absent. Consumption smoothing motive says that a household wants her marginal utility of consumption to be smooth. Since the demand shock is a preference shock that increases the marginal utility of consumption \(\phi e^{\gamma z_{c,t}} e^{-\gamma t}\) only in the first period, to smooth out the marginal utility of consumption across time, the best thing to do is to consume more in the first period, offsetting the effect of the increased \(z_{c,t}\), and then to consume normally as \(z_{c,t}\) goes back to normal. This is exactly what happens in the basic CU model.

\(^{36}\)This lack of persistence in the response of investment in the basic CU model follows from capital resource slackness but does not depend on the linear relationship between capacity and capital. Suppose that the production function of the firm is given by \(y_{j,t} = \min \left\{ \frac{l_{j,t}}{\alpha c}, Ak_{j,t}^{\alpha} \right\}\), where \(\alpha \in (0, 1)\). In this example, even though capacity is given by a nonlinear function of capital, \(y_{j,t} = Ak_{j,t}^{\alpha}\), the response of investment to demand shocks is still concentrated in the first period.
5.7 Comparison to a Standard Variable Capital Utilization Model

In the business cycle literature, the standard way to include variable capital utilization is to assume that firms are subject to a convex utilization cost. Hence, capital may not be fully utilized because of a high marginal cost of utilization.

However, unlike the basic CU model, the standard variable capital utilization (VU) model does not feature capacity competition. There is only price competition among firms, and the only role of capital is to save aggregate total costs. Therefore, despite capital underutilization, capacity is not in excess, and capital resources are tight, as capital is a scarce production factor that can reduce the real marginal cost by lowering the capital utilization rate. Hence, the properties of the standard VU model are similar to those of the standard RBC models. When demand increases, more capital is utilized, and the real marginal cost increases. As in the standard RBC models, the upward sloping real marginal cost curve dampens the response of output to demand shocks and causes a counter-cyclical real wage rate under demand shocks (see Supplementary Appendix F for a detailed comparison).

To sum up, the basic CU model exhibits a positive co-movement among consumption, investment, capacity utilization, and hours; a large fluctuation in investment; an acyclical real wage rate; and a pro-cyclical Solow residual, even though all fluctuations are driven by demand shocks. These results are difficult to obtain in standard models.

6 Estimating the Capacity Underutilization Model

In this section, I extend the basic CU model to a full CU model and estimate it using Bayesian estimation techniques. When the model is estimated to match the U.S. macro data, demand shocks turn out to be the main driving forces of business cycles.

6.1 Model Extensions

The full CU model is obtained by extending the basic CU model with home production, indirect (overhead) labor, capital and investment adjustment costs, and exogenous expenditures.

---

37 A drawback of the basic CU model is that the response of investment lacks persistence. This issue, however, can be resolved by including capital and investment adjustment costs (see Section 6.1).

38 The capacity competition mechanism, however, is compatible with and complementary to various existing models or mechanisms in the literature. For example, the basic CU model could be extended to include a small convex utilization cost as in the VU model to generate some smoothing effects on capital adjustments, to include some idiosyncratic demand shocks as in Fagnart et al. (1997) to generate a distribution of capacity utilization rates across firms, or to include some extrapolative demand expectation as in Hirshleifer et al. (2015) to generate over-estimation of demand that amplifies capacity underutilization in the mid-run.
6.1.1 Home Production

I assume that households also use their time to produce goods or services at home. The total number of hours available is normalized to one. The amount of home produced goods or services is given by

\[ c_{h,t} = Z_{h,t} (1 - l_t) \],

where \( Z_{h,t} \) is the productivity of time spent at home (e.g., Gronau and Hamermesh, 2006). The one-period utility function of the representative household is now given by

\[
u(z_{c,t}, c_t, c_{h,t}) = \begin{cases} 
\phi e^{z_{c,t}} \left( \frac{c_t}{\phi z_{c,t}} \right)^{1-\gamma} + \bar{c}_{h,t}, & \gamma \neq 1 \\
\phi e^{z_{c,t}} \ln \left( \frac{c_t}{\phi z_{c,t}} \right) + \bar{c}_{h,t}, & \gamma = 1
\end{cases}
\]

which replaces the utility function (22) in the household’s problem (e.g., Boppart and Ngai, 2021 and Rachel, 2021). The inclusion of home production enriches the model dynamics: an increase in the productivity at home increases the opportunity cost of supplying labor and thus reduces the labor supply.

In general, the productivity at home \( Z_{h,t} \) is related to labor productivity \( z_{l,t} \):

\[ Z_{h,t} = e^{\phi_l z_{l,t}}. \]

If the parameter \( \phi_l \) is set to zero, an increase in labor productivity would have no effect on the opportunity cost of supplying labor. However, if \( \phi_l \) is positive, an increase in labor productivity also increases the productivity at home and thus the opportunity cost of supplying labor. Hence, the magnitude of \( \phi_l \) affects the ability of labor productivity shocks in driving pro-cyclical working hours and will be estimated later in Section 6.2.

6.1.2 Indirect Labor

I assume that the Leontief production function of firm \( j \in [0, 1] \) at time \( t \) is now given by

\[
y_{j,t} = \min \left\{ A k_{j,t}, \frac{l_{f,j,t}}{\alpha_{f,t}}, \frac{l_{v,j,t}}{\alpha_{v,t}} \right\},
\]

where \( l_{v,j,t} \) is the amount of direct labor, \( l_{f,j,t} \) is the amount of indirect labor, and the inverse of \( \alpha_{f,t} \) is the productivity of the indirect labor.

Direct labor produces goods or services directly. The hours of direct labor fluctuate with output and can be easily adjusted within a period. Examples of direct labor positions are machine operators, assembly line operators, and cleaners.

Indirect labor, by contrast, supports the production process of firms but is not directly involved in the active conversion of materials into finished products. Like capital, indirect labor is a quasi-fixed factor. Examples of indirect labor positions are production supervisors, managers, and administrators.

By definition, the production capacity of firm \( j \) at time \( t \) is determined by the quasi-
fixed factors, i.e., indirect labor and capital stock:

$$\bar{y}_{j,t} = \min \left\{ A k_{j,t}, \frac{l_{j,t}}{\alpha_{f,t}} \right\}, \quad (46)$$

where $$\alpha_{f,t}$$ is also the amount of indirect labor required per unit of capacity. The introduction of indirect labor allows output to be more volatile than total hours worked. Hence, output to labor ratio can be pro-cyclical under demand shocks.

### 6.1.3 Capital and Investment Adjustment Costs

As is common in the literature, I introduce capital and investment adjustment costs to get a persistent and hump-shaped investment response. The capital stock of firm $$j$$ is accumulated according to

$$k_{j,t+1} = k_{j,t} (1 - \delta) + i_{j,t} (1 - S (i_{j,t}, i_{j,t-1}, k_{j,t})), \quad (47)$$

where $$S$$ is the adjustment cost function.

Following Hayashi (1982) and Christiano et al. (2005), I assume that the adjustment cost function is given by

$$S (i_{j,t}, i_{j,t-1}, k_{j,t}) = \phi_i \left( \frac{i_{j,t}}{i_{j,t-1}} - 1 \right)^2 + \phi_k \left( \frac{i_{j,t}}{k_{j,t}} - \delta \right)^2, \quad (48)$$

where $$\phi_k \geq 0$$ and $$\phi_i \geq 0$$ are parameters that capture the curvature of capital and investment adjustment costs respectively.

### 6.1.4 Exogenous Expenditure

Following Smets and Wouters (2007), government spending and net exports are treated as an exogenous expenditure. I abstract away from the crowd-out and crowd-in effects of government spending by assuming that goods or services demanded by the exogenous expenditure are produced by an independent sector that requires direct labor only. Let $$g_t$$ be the exogenous expenditure and $$l_{g,t}$$ be the corresponding labor hired. We have

$$l_{g,t} = \alpha_{g,t} g_t$$

where $$\alpha_{g,t}$$ is the direct labor per unit of exogenous expenditure. The final product of the economy is

$$y_t = c_t + i_t + g_t,$$

and the total number of hours worked is

$$l_t = l_{f,t} + l_{v,t} + l_{g,t}.$$

### 6.1.5 Exogenous Shocks

There are four types of exogenous disturbances.

First, the exogenous expenditure disturbance $$z_{g,t}$$ pins down $$g_t = ge^{z_{g,t}}$$. $$z_{g,t}$$ follows an AR(1) process with i.i.d. normal shocks:

$$z_{g,t} = \rho_g z_{g,t-1} + e_{g,t}.$$
Second, the labor productivity disturbance $z_{l,t}$ determines the productivity of all types of labor and the productivity at home: $\alpha_{v,t} = \alpha_v e^{-z_{l,t}}$, $\alpha_{f,t} = \alpha_f e^{-z_{l,t}}$, $\alpha_{g,t} = \alpha_g e^{-z_{l,t}}$, and $Z_{h,t} = e^{\phi_{l,t}}$. $z_{l,t}$ follows an AR(1) process with i.i.d. normal shocks: $z_{l,t} = \rho_{l} z_{l,t-1} + e_{l,t}$.

Third, the investment demand disturbance $z_{i,t}$ affects the subjective discount factor of the representative household, whose lifetime utility is given by

$$E_0 \left( \sum_{t=0}^{\infty} \beta^t e^{-z_{i,t} u(z_{c,t}, c_t, c_{h,t})} \right).$$

(49)

$z_{i,t}$ follows an AR(2) process with i.i.d. normal shocks: $z_{i,t} = \rho_{i1} z_{i,t-1} + \rho_{i2} (z_{i,t-2} - z_{i,t-1}) + e_{i,t}$. This assumption allows us to have both a persistent movement and a zigzag pattern in investment.

Finally, the consumption demand disturbance $z_{c,t}$ follows an AR(1) process with i.i.d. normal shocks and is also affected by shocks to investment demand as follows: $z_{c,t} = \rho_{c} z_{c,t-1} + e_{c,t} + \rho_{ci} e_{i,t}$. The parameter $\rho_{ci}$ allows us to capture the negative correlation between consumption and capital in the U.S. from 1968 to 1993.39

All shocks, $e_{c,t}$, $e_{i,t}$, $e_{g,t}$, and $e_{l,t}$, are mutually uncorrelated.

6.2 Estimation

I use Bayesian estimation techniques to estimate the full CU model. Having four exogenous shocks, I am able to match four detrended U.S. macro time series: real consumption, real investment, real exogenous expenditure, and the number of hours worked.40

To alleviate the burden on the estimation problem, I choose nine calibration targets that must be matched throughout this estimation procedure. First, the elasticity of intertemporal substitution (EIS) is calibrated to 0.5. In the literature, the estimated EIS ranges from 0 to 2 with a mean of about 0.5 (e.g., Hall, 1988, Gruber, 2013, Havránek, 2015, and Best et al., 2020). Second, the exogenous expenditure to output ratio in steady state is matched to the average ratio of the exogenous expenditure to GDP from 1947 to 2016 in the BEA NIPA. The other seven calibration targets are the same as described in Section 5.5. All targets are exactly matched in steady state. Because of these nine targets, nine parameters are pinned down as functions of other parameters. The other parameters and their posterior distributions are estimated using Bayesian estimation techniques (see Supplementary Appendix G for details).41

---

39 See Supplementary Appendix J for a further discussion.
40 See Supplementary Appendix I for a description of the macro-data used for estimation.
41 The model is log-linearized around the steady state and the estimation is done on the software platform Dynare.
6.3 Variance Decomposition

What are the main driving forces of business cycles? Table 8 gives the forecast error variance decomposition of output, consumption, investment, hours, and the capacity utilization rate at a 10-year horizon based on the posterior modes of the parameters of the full CU model.

<table>
<thead>
<tr>
<th></th>
<th>Full CU Model</th>
<th></th>
<th></th>
<th></th>
<th>Full VU Model</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$e_c$</td>
<td>$e_i$</td>
<td>$e_g$</td>
<td>$e_l$</td>
<td>$e_c$</td>
<td>$e_i$</td>
<td>$e_g$</td>
<td>$e_l$</td>
</tr>
<tr>
<td>Output $Y$</td>
<td>56.53</td>
<td>15.00</td>
<td>26.24</td>
<td>2.23</td>
<td>33.41</td>
<td>4.30</td>
<td>22.09</td>
<td>40.20</td>
</tr>
<tr>
<td>Consumption $c$</td>
<td>72.67</td>
<td>26.43</td>
<td>0.00</td>
<td>0.90</td>
<td>64.71</td>
<td>15.89</td>
<td>0.00</td>
<td>19.40</td>
</tr>
<tr>
<td>Investment $i$</td>
<td>78.69</td>
<td>12.75</td>
<td>0.00</td>
<td>8.57</td>
<td>12.58</td>
<td>0.74</td>
<td>0.00</td>
<td>86.68</td>
</tr>
<tr>
<td>Hours $l$</td>
<td>60.86</td>
<td>14.80</td>
<td>10.93</td>
<td>13.41</td>
<td>66.59</td>
<td>11.59</td>
<td>11.38</td>
<td>10.44</td>
</tr>
<tr>
<td>CU Rate $u$</td>
<td>71.50</td>
<td>22.02</td>
<td>0.00</td>
<td>6.47</td>
<td>27.44</td>
<td>10.35</td>
<td>0.00</td>
<td>62.21</td>
</tr>
</tbody>
</table>

Note: The forecast error variance decomposition for the estimated models at a 10-year (40-quarter) horizon. CU Rate stands for capacity utilization rate.

According to the estimated full CU model, business cycle movements are primarily driven by three types of demand shocks, i.e., the consumption demand, the investment demand, and the exogenous expenditure shocks. Labor productivity shocks account for only 2.23% of the variation in output and 13.41% of the variation in hours. Among these three types of demand shocks, consumption demand shocks turn out to be the main driving force of business cycles as most of the variations in consumption, investment, hours, and the capacity utilization rate are explained by innovations to the consumption demand disturbance.

For comparison, I also extend the standard variable capital utilization (VU) model discussed in Section 5.7 by including home production, capital and investment adjustment costs, exogenous expenditures, and four exogenous shocks in the same way as I did for the basic CU model. The extended VU model is called the full VU model. The estimation procedure of the full VU model is set to be as close as possible to that of the full CU model.42

The variance decomposition results of the full VU model are listed in Table 8. Consistent with the standard findings in the business cycle literature, a large fraction of the variation in output and most of the variation in investment are driven by labor productivity shocks. However, labor productivity shocks cannot explain most of the variation in hours (see Smets and Wouters, 2007 for a related discussion). In addition, consumption is not volatile enough under labor productivity shocks (see Bai et al., 2012 for similar re-

42See Supplementary Appendix H for a detailed description of the estimation of the full VU model.
results found in a standard RBC model). Therefore, standard models often rely on multiple types of shocks to explain business cycles.

6.4 Consumption Demand Drives Business Cycles

To highlight the ability of consumption demand shocks to drive business cycles, I shut down all the other types of shocks and feed only the estimated consumption demand shocks into the models. The structural shocks of the estimated models are backed out from the U.S. data using the Kalman smoother technique.

Table 9: Main Business Cycle Statistics

<table>
<thead>
<tr>
<th></th>
<th>The U.S. Data</th>
<th>Full CU Model ($e_c$)</th>
<th>Full VU Model ($e_c$)</th>
<th>Full VU Model ($e_l$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Std Dev</td>
<td>Cov ($\cdot$, Y)</td>
<td>Std Dev</td>
<td>Cov ($\cdot$, Y)</td>
</tr>
<tr>
<td>Output</td>
<td>1.62</td>
<td>1.00</td>
<td>1.78</td>
<td>1.00</td>
</tr>
<tr>
<td>Consumption</td>
<td>1.16</td>
<td>0.57</td>
<td>1.05</td>
<td>0.57</td>
</tr>
<tr>
<td>Investment</td>
<td>7.55</td>
<td>3.95</td>
<td>6.81</td>
<td>3.78</td>
</tr>
<tr>
<td>Hours</td>
<td>1.68</td>
<td>0.89</td>
<td>1.48</td>
<td>0.83</td>
</tr>
<tr>
<td>Solow Resid</td>
<td>1.00</td>
<td>0.39</td>
<td>0.90</td>
<td>0.48</td>
</tr>
<tr>
<td>Real Wage Rate</td>
<td>0.90</td>
<td>0.08</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Note: Std Dev, Cov ($\cdot$, Y), and Resid stand for standard deviation, covariance with output relative to the variance of output, and residual respectively. $e_c$ and $e_l$ denote consumption demand shocks and labor productivity shocks respectively. The U.S. data is from the BEA and the BLS. All variables are HP-filtered logarithms of the original series with a smoothing parameter of 1,600.

Table 9 compares the main business cycle statistics of the fitted values with those of the U.S. data. The business cycle properties of the full CU model under consumption demand shocks are quite close to those of the data. Consumption demand shocks not only generate sizable business cycle fluctuations but also generate the correct business cycle co-movement among the key aggregate variables. Particularly, investment is highly volatile. Consumption, investment, hours, and the Solow residual are as pro-cyclical as in the U.S. data.

In the full VU model, however, investment and the Solow residual are not volatile enough under consumption demand shocks, while hours and consumption are not volatile enough under labor productivity shocks.

Therefore, in terms of explaining the business-cycle fluctuations of consumption, investment, output, and hours, the full CU model under consumption demand shocks not only outperforms the full VU model under consumption demand shocks but also outperforms the full VU model under labor productivity shocks.
6.5 Discussion on the Cyclical Property of the Real Wage Rate

The last row of Table 9 shows that the aggregate real wage rate in the U.S. is acyclical. Since Bils (1985) and Solon et al. (1994), economists using longitudinal data often find pro-cyclical real wage rates, and they attribute the acyclical pattern of the aggregate real wage rate to a composition bias: the share of low-wage workers in booms is larger than in recessions. Although the cyclical property of real wage rates is still a subject of debate, most of the existing literature suggests that real wage rates are either acyclical or pro-cyclical.\footnote{Recently, several researchers using micro-level data find that time-based real pay rates that exclude overtime premium and performance payment are much less pro-cyclical or acyclical (e.g., Devereux, 2001, Shin and Solon, 2007; Swanson, 2007, Hart and Roberts, 2013, and Makridis and Gittleman, 2022). The cyclicality of real wage rates is also found to be affected by the sample used (e.g., Abraham and Haltiwanger, 1995, Messina et al., 2009, Gu et al., 2020, and Otrok and Pourpourides, 2019).}

Compared to the full VU model under demand shocks, the full CU model is able to generate an acyclical real wage rate under demand shocks, greatly alleviating the counter-cyclical real wage rate issue, a long-lasting issue since Dunlop (1938), while at the same time improving the ability of demand shocks to drive business cycles.\footnote{One might be concerned that the real wage rate shows little volatility in the full CU model under demand shocks. However, if the economic variable is acyclical in the data and if one focuses only on a single type of shocks, a lack of volatility in that variable is exactly what a good model should predict. Supplementary Appendix J shows that when all four types of shocks are present, the real wage rate predicted by the full CU model is as volatile as the aggregate real wage rate in the U.S. data.}

Given a perfectly competitive labor market, a constant returns to scale production technology, and a constant markup, an acyclical real wage rate under demand shocks is the most pro-cyclical result one can get from a model of these benchmark assumptions. Nevertheless, the CU model could be extended in several ways to get pro-cyclical real wage rates if needed. First, a counter-cyclical markup can be introduced to the CU model to generate a pro-cyclical real wage rate (e.g., Rotemberg and Woodford, 1991). Second, one can allow technology to exhibit increasing returns to scale in the area where capacity is underutilized. Third, one can extend the labor market to allow profit sharing due to collective bargaining or incentive payment so that workers can enjoy a pro-cyclical real wage rate under demand shocks. These extensions could be interesting future directions of research.

7 Conclusion

This paper develops a business cycle model with chronic excess capacity. Buyers are not fully attentive to prices as they are subject to an information processing cost. Because of this friction, buyers search for capacity in a partially undirected way. Hence, firms with a larger capacity are more likely to be visited by buyers. If a firm expands its capacity relative to that of others, it attracts buyers from its competitors. Furthermore,
each unit of demand is profitable because firms can take advantage of inattentive buyers and charge a markup. Hence, from an individual firm’s perspective, capacity expansion not only relaxes the capacity constraint but also attracts profits from others. From a social planner’s perspective, however, the only value of capacity is to relax the capacity constraint. This capacity competition externality encourages firms to over-accumulate capital and causes the economy to exhibit chronic excess capacity. When capacity is in excess, capital resources are slack, and output is highly responsive to demand shocks.

The model is consistent with the three stylized facts about capacity utilization, i.e., long-term capacity underutilization, insufficient demand viewed by firms as the primary reason for capacity underutilization, and a highly pro-cyclical capacity utilization rate. Capacity competition is also consistent with the firm-level evidence from the KOF business tendency survey.

Quantitatively, I estimate a version of the capacity underutilization model. It turns out that consumption demand shocks can explain most of the business cycle fluctuations, generating large responses in consumption, investment, and hours, an acyclical real wage rate, and a pro-cyclical Solow residual, whereas labor productivity shocks explain only a small fraction of the variations in output and hours. Hence, when viewed through the lens of the model, demand shocks are the main driving forces of business cycles.

Supplementary Data

Supplementary data are available at *Review of Economic Studies* online. And the replication package is available at [https://doi.org/10.5281/zenodo.11094818](https://doi.org/10.5281/zenodo.11094818).

Data Availability Statement

The microdata on the KOF business tendency survey of the manufacturing industry provided by the KOF Swiss Economic Institute cannot be shared publicly, but the instructions to obtain this data are available on Zenodo at [https://doi.org/10.5281/zenodo.11094818](https://doi.org/10.5281/zenodo.11094818). All the other data and all the code underlying this research are available on Zenodo at [https://doi.org/10.5281/zenodo.11094818](https://doi.org/10.5281/zenodo.11094818).
References


