



Universidad
Carlos III de Madrid
www.uc3m.es

TESIS DOCTORAL

Frameworks for Evaluating Macroeconomic Policies

Autor:

Robert Kirkby

Director:

Javier Díaz-Giménez

DEPARTAMENTO DE ECONOMÍA

Getafe, Mayo 2014



TESIS DOCTORAL

***Frameworks for Evaluating
Macroeconomic Policies***

Autor: Robert Kirkby

Director: Javier Díaz-Giménez

Firma del Tribunal Calificador:

Firma

Presidente:

Vocal:

Secretario:

Calificación:

Getafe, de de

PhD Thesis: Frameworks for Evaluating Macroeconomic Policies

by Robert Kirkby

supervised by Javier Díaz-Giménez

This thesis brings together the three chapters that together form my PhD thesis. As indicated by the title, *Frameworks for Evaluating Macroeconomic Policies*, the common theme linking the three is a focus on the development of modeling frameworks that can be used for the evaluation of Macroeconomic policies. Ways in which these models can be compared with each other and with the data are recurrent themes.

The first chapter *How to Model Money? Racing Monetary Frameworks against the Quantity Theory of Money* is about finding frameworks for evaluating monetary policies. Currently three main approaches exist: Cash-in-Advance, New Keynesian, and Search-Money. Using empirical facts on the Quantity Theory of Money as a yardstick we compare these three frameworks. It results that all three frameworks display the Quantity Theory of Money over the long-run, as in the data. But all three frameworks display way too much of the Quantity Theory of Money over the short-run. The race thus ends in a draw, but one illustrative of the strengths and weaknesses of all three frameworks. The results suggest that better modeling of other causes of inflation, and of heterogeneity, are important to improving monetary models.

The second chapter *Evaluating a Flat-Tax Reform* is a quantitative modelling of a flat-tax reform for the US. The modeling focuses on replicating the details of current US taxation and inequality. This later is important as the effects on inequality of such a tax reform are one of the main arguments given against it.

The third chapter *Estimation of Bewley-Huggett-Aiyagari Models: Theory and Implementation* present in-progress work developing theory relevant to simulated moment and simulated likelihood estimation of a class of heterogeneous agent models. Theory focuses on developing the required assumptions directly from model fundamentals, and from accounting for the dependence of the estimation on numerical solution and simulation of the models. Attention is also given to implementation of the estimators, in particular which algorithms work computationally.

The three chapters are presented here in the form of three separate articles. However the common thread of developing frameworks for the evaluation of macroeconomic policies is clearly evident throughout. I hope they may be of interest to the reader.

How to Model Money? Racing Monetary Frameworks against the Quantity Theory of Money

Javier Díaz-Giménez and Robert Kirkby

May 5, 2014

Abstract

We show that, between 1960 and 2009, the Quantity Theory of Money held in the United States in the long run and that it failed to hold in the short run. We ask whether standard monetary model economies from the Cash-in-Advance, the New-Keynesian, and the Search-Money frameworks can replicate these results, and we find that they do in the long run, but that they fail in the short run because prices respond too quickly to changes in the growth rates of money.

Keywords: Monetary; Quantity Theory of Money; Cash-in-Advance; New-Keynesian; Search-Money.

In the monetary economics literature there co-exist three main frameworks for modeling money: Cash-in-Advance, New-Keynesian, and Search-Money. Is one framework better than the others? To answer this we need to directly compare the frameworks, to see which one would win in a race. Such direct comparison of these different frameworks to model money is rare. Often an argument is made for, say, New-Keynesian models on the basis that we observe price stickiness. Or for Search-Money on the grounds that money should arise in the model to solve some problem, rather than be preordained. But such arguments really just repeat the main assumptions of each framework. Here we directly compare the three main approaches to modeling money in terms of their ability to replicate empirical facts related to the Quantity Theory of Money.

The empirical facts of the Quantity Theory of Money — that it holds in the long-run, but fails in the short-run — can be shown for the United States economy using the Lucas Illustrations; as we explain shortly. It is in reproducing these empirical facts on which we evaluate the different approaches to modelling money. The choice of the Quantity Theory of Money and the Lucas Illustrations is motivated on two grounds. First, as a major theorem of monetary economics, present in undergraduate textbooks, it is clearly considered as important in and of itself. Second, the Lucas Illustrations of the Quantity Theory of Money do not depend on structural assumptions, and so can be used to make genuine comparisons across the monetary frameworks. This is in contrast to more commonly used tools in monetary economics, such as impulse response functions, which depend upon structural assumptions. Since different monetary frameworks involve different structural assumptions such tools cannot be used to genuinely compare the different frameworks. The Lucas Illustrations of the Quantity Theory of Money thus provide a test based on a major theorem of monetary economics which can be used to make genuine comparisons between monetary frameworks.

So what does the Quantity Theory of Money look like the the data? The first and most striking feature of data is the complete absence of the Quantity Theory of Money in the short-run. This

short-run failure of the Quantity Theory of Money to hold, in the United States and elsewhere, can be illustrated by simply plotting the rate of inflation—plus the rate of growth of output—against the rate of growth of money. In this article we start by plotting those two series for the United States for the 1960–2009 period using quarterly data. If the Quantity Theory of Money relationship held in the short run, the data points would trace a 45 degree line. Instead, we show that their scatter plot forms a vague cloud, out of which no clear pattern emerges (see Figure 1).

Whether the Quantity Theory of Money holds in the long-run is harder to establish, because we must make operative the meaning of “long run”. To do this, we replicate the method used in Lucas (1980). In that article, Lucas identifies the long-run with the low frequency fluctuations of the rate of inflation and rate of growth of money. He then uses a two-sided moving average filter to extract the low frequency movements from those series, and he shows that the plots of the filtered series move closer to the 45 degree line as he filters out the higher frequencies. Lucas considers the period between 1955 and 1975 and he does not include the rate of growth of output in his calculations. He concludes that the Quantity Theory of Money held in the long-run in the United States during that period.

Lucas’ results are telling, but they are qualitative. Whiteman (1984) shows that one way to quantify Lucas’ findings is to estimate the slope of the ordinary least squares linear regression of the rate of growth of prices plus the rate of growth of output on the rate of growth of money. Whiteman argues that, when the Quantity Theory of Money relationship holds, the slope of this regression will be close to one.

We replicate Whiteman’s calculations, and we also compute an additional measure of closeness to the 45 degree line: the average Cartesian distance of the data points from the 45 degree line that goes through the grand mean of the sample. When the Quantity Theory of Money relationship holds, the data points will be close to the 45 degree line and this average distance will be close to zero. We compute these two measures for the United States in the 1960–2009 period and we conclude that the Quantity Theory of Money relationship did not hold in the short-run in the United States, but that it held in the long-run during that period. Our results confirm quantitatively both Hume (1742a)’s thought experiment and Lucas (1980)’s findings.

Next, we ask whether the Quantity Theory of Money holds in three monetary model economies, which we consider to represent the standard frameworks that economists currently use to evaluate the implications of monetary policy—the Cash-in-Advance framework, the New-Keynesian framework, and the Search-Money framework. Our research should be understood as part of the search for a monetary model economy whose predictions we can trust. The methodological idea is that the more dimensions along which model economies replicate the known behavior of the monetary time series of real economies, the more we trust their predictions along other, harder to test, dimensions.

Monetary economists often test their model economies, for example using impulse response functions. But these methods do not allow them to compare model economies from different frameworks because impulse response functions depend on structural and modelling assumptions that usually differ across frameworks. Meaningful comparisons of alternative ways to model money are harder to perform because they force researchers to use evaluation methods that do not depend on the specific modelling assumptions of each framework. Using Lucas’ illustrations of The Quantity Theory of Money relationship and our measures to quantify this relationship are meaningful ways to evaluate and compare these three leading monetary frameworks. As one of the major theorems of monetary economics the Quantity Theory of Money also seems an inherently desirable property for monetary models.

That the Quantity Theory of Money should hold in monetary model economies is both compelling and quite easily achieved formally. In the standard neoclassical models, the Quantity Theory of Money holds every period. This result is mathematically known as the “zero-degree homogeneity of real decisions in the price level”. Instead, the challenge for monetary model economies is to break away from the Quantity Theory of Money relationship *in the short-run*. Or, in other words, to find a way of modelling money that makes prices respond *sluggishly* to changes in the money supply.

The three monetary frameworks that we study here represent different ideas about how to model money: the Cash-in-Advance framework focuses on the transaction role of money, the New-Keynesian framework on the role of money as a nominal anchor around which prices are sticky, and the Search-Money framework on the role of money as a way to solve problems created by the absence of double-coincidence of wants in barter. To find out whether any of these three ways of modelling money succeed in making prices react sluggishly to changes in the money supply, we ask whether the Quantity Theory of Money holds in these three frameworks both in the short-run and in the long-run. For each one of these frameworks we have chosen a canonical model economy. These model economies were not designed with the specific aim of delivering the Quantity Theory of Money relationship. They were designed to capture some of the sluggishness of nominal variables, the question here is how much they deliver.

The Cash-in-Advance framework makes the use of money in exchange compulsory by forcing households to buy consumption goods using money carried over from the previous period. In general, this cash-in-advance constraint is inefficient, but it solves the informational problem that would arise when trying to coordinate all the simultaneous trades; a problem that non-monetary economies ignore. In the words of Lucas (1980) the cash-in-advance framework “is an attempt to study the transaction demand for money in as simple as possible a general equilibrium setting”. In this article, we use Cooley and Hansen (1989)’s cash-in-advance business cycle model as our canonical example of the cash-in-advance framework. We chose this model economy because it combines a cash-in-advance constraint and the standard neoclassical model of business cycles.

The New-Keynesian framework models the relationship between money and interest rates using a money demand equation. To break away from the short-run neutrality of money, New-Keynesian models assume that prices are sticky. In this framework, prices either cannot be changed every period by assumption, or doing so is costly, also by assumption. This price stickiness assumption is justified using empirical evidence that prices do not change often in the real world. In this article, we use the New-Keynesian monetary model economy described in Chapter 3 of Galí (2008) as our canonical example of the New-Keynesian framework. We chose this model economy because it includes money explicitly.¹

The Search-Money framework is a successful attempt to satisfy Wallace (1998)’s dictum that “money should not be a primitive in monetary theory”; that is, that there must be an endogenous reason that justifies the existence of money. In the Search-Money framework this reason is to enable trade. Search-Money models assume that people meet in pairs and exchange goods using barter. But this means that trade only occurs when both trading partners have a good that the other one wants. This is the well-known problem of barter: trade is often limited by the absence of a double-coincidence of wants. Money solves this problem because everyone always wants money,

¹Many New-Keynesian models often omit money entirely and they use a Taylor rule on interest rates instead. They rationalize this modelling choice on the grounds that modern central banks tend to focus on interest rates, and not so much on monetary aggregates. When they model money explicitly their standard approach is to use of a money-demand equation. See, for example, Christiano, Eichenbaum, and Evans (2005) and Sargent and Surico (2011).

at least as an enabler of trade. In this article, we use a stochastic extension of the Search-Money model described in Aruoba, Waller, and Wright (2011) as our canonical example of the Search-Money framework. We chose this model economy because it includes capital. Capital accumulation plays an important role in these economies because it reduces the number of monetary trades and, consequently, it amplifies the effects of monetary innovations on the rest of the economy, where monetary exchanges take place.^{2,3}

We choose standard implementations of all three frameworks. There are three reasons for this: (i) we suspect that using more complicated models would not solve the difficulties of the models, (ii) most of these complications, such as capital-adjustment costs or consumption habit formation, could be just as easily applied to any of the three frameworks and so would not tell us about how best to model money, and (iii) by using standard implementations the models are easier to follow and understand, and we are better able to provide intuition and explanation of the results.

The first of these three points deserves further explanation. In the body of this paper we use the Lucas Illustrations to describe the performance of the three frameworks to reproduce the empirical facts of the Quantity Theory of Money — our preferred method due to its easy to understand and intuitive nature. As an alternative Appendix C provides a Bayesian estimation view of the same issue. We find that, when Bayesian estimated, the models explain the short-run movements of inflation as arising mostly from large and poorly identified shocks to inflation that connect the underlying inflation predicted by the model to the inflation in the data. This is the Bayesian view of our finding that the models contain too much of the Quantity Theory of Money in the short-run. This same issue, that short-run inflation in a Bayesian estimation is mostly explained by a shocks connecting the underlying inflation of the model to the inflation in the data, is also found to be the case by King and Watson (2012) for the model of Smets and Wouters (2007).⁴ The model of Smets and Wouters (2007) represents the current gold-standard in terms of advanced New-Keynesian models incorporating many business cycle frictions. That King and Watson (2012) report a similar, Bayesian-version, of our finding that the the models contain too much of the Quantity Theory of Money in the short-run, suggests to us that using such more advance models would not resolve the issues that we illustrate for all three frameworks in this article, and would simply obscure the underlying problem, that the frameworks contain too much of the Quantity Theory of Money in the short-run. [Note: The Bayesian Appendix is currently under construction. So for I have only finished Bayesian estimation of the Cash-in-Advance model. Given the likeness between the results of the CIA model and those of King & Watson it seems likely that the New-Keynesian model will give the same results, and this paragraph is based on that assumption, but I have not completed the estimation of this model yet. The Bayesian estimation of these models is part of the response to the referees comments on the earlier version of this paper submitted as part of my thesis.]

²Specifically, in Lagos and Wright (2005) model economy which does not include capital, monetary trades account for 20.6% of real output on average. In contrast, when capital is added to a Search-Money model economy, as in Aruoba, Waller, and Wright (2011), monetary-trades account for 1.6% of real output on average.

³Berentson, Menzio, and Wright (2011) also solve a stochastic extension of the original Search-Money model described in Lagos and Wright (2005). Their extension differs slightly from ours in the timing of the money shock. It also differs because they impose an AR(1) process on interest rates, which implies a Markov process on money, while we impose an AR(1) process on (log) money, which implies a Markov process on interest rates.

⁴These shocks between underlying inflation predicted by the model, and the inflation observed in the data are often referred to as 'price mark-up shocks'. The same conclusion, albeit made less comprehensively than in King and Watson (2012) can be drawn from the variance decomposition of Inflation to be found in Figure 1 of Smets and Wouters (2007).

To ensure that the comparison of the three monetary frameworks is meaningful, we choose their functional forms and parameters so that they are as similar as possible to each other. Moreover, we make the stochastic processes on the monetary shocks identical in the three model economies, and we simulate them using exactly the the same sequences of realizations of the shocks.

First, we plot Lucas' illustrations and we find that the Cash-in-Advance, the New-Keynesian, and the Search-Money frameworks all display the Quantity Theory of Money relationship in the long-run and, therefore, that they replicate the long-run behavior of the United States. In all three model economies the filtered points lie along the 45 degree line that goes through the grand mean of the sample, exactly as the Quantity Theory of Money predicts (see Figure 4).

Next, we simulate 100 stochastic realizations of the equilibrium processes of the three model economies and we find that Whiteman's regression coefficient is close to one, and that the Cartesian distances of the filtered points from the 45 degree line are close to zero. This confirms our qualitative results. We conclude that the differences between the three frameworks are tiny, if any, and that all three of them pass the long-run Quantity Theory of Money test with flying colors.⁵

In sharp contrast, the three monetary frameworks fail to replicate our finding that the Quantity Theory of Money relationship does not hold in the United States in the short run. While in the United States the graphs of the unfiltered data contain no suggestion of the Quantity Theory of Money, our simulations of the three model economies produce data points that lie very close to the 45 degree line. This suggests that the three model economies display a short-run Quantity Theory of Money relationship that is too tight, even though it is not exact.

Whiteman's regression coefficients and the Cartesian distances of the data points from the 45 degree lines confirm our qualitative results. We conclude that prices respond too quickly to changes in the rate of growth of money in the three frameworks that we study, and that the search for a model economy in which the response of prices to monetary innovations replicates the sluggishness found in the data still remains an important challenge for monetary economics.

0.1 Literature Review

The Quantity Theory of Money was first formulated by David Hume using a thought experiment which he summarized in the following quote, Were all the gold in England annihilated at once, and one and twenty shillings substituted in the place of every guinea, would money be more plentiful or interest lower? No surely: We should only use silver instead of gold. (Hume, 1742, Of Interest).

Missing, at least explicitly, from Humes formulation is the role played by changes in real output. When it was first explicitly developed is not clear, but it plays a prominent role in the works of Fisher (1911) and Friedman and Schwartz (1963). Writing on the Greenback period, 1867-1879, Friedman & Schwartz observe that prices decreased slightly, despite an increase in the money supply — attributing the difference as being substantially due to the large increase in real output that occurred during this period.

⁵Sargent and Surico (2011) study the Quantity Theory of Money in the United States between 1900 and 2005 and they argue that it is not a stable relationship. They use the Lucas Illustrations approach, they find that the slope of the Quantity Theory of Money relationship changes over time, and they attribute these changes to changes in the monetary policy regime followed by the Federal Reserve —for example, around 1980 the Fed changed from targeting monetary aggregates to targeting inflation rates. We discuss Sargent and Surico's findings in Section 2 and in Appendix B.

Most evidence in support of the Quantity Theory of Money at that the time came from comparing, eg., multi-decade averages of money growth vs. price growth. Plotting these for various countries one would get a roughly forty-five degree line. The contribution of Lucas (1980) was to provide a time-series view of the Quantity Theory of Money, by associating the long-run with low-frequency and then filtering the time series to obtain this long-run view. More recently, Benati (2005, 2009) has extended this approach for longer time periods, for the UK, and for alternative filters.

1 The Quantity Theory of Money

Multiply the supply of money by m and prices will become m times larger —this is a rough but useful characterization of the Quantity Theory of Money. More precisely, the Quantity Theory of Money claims that the rate of growth of nominal prices plus the rate of growth of output is equal to the rate of growth of the money supply.

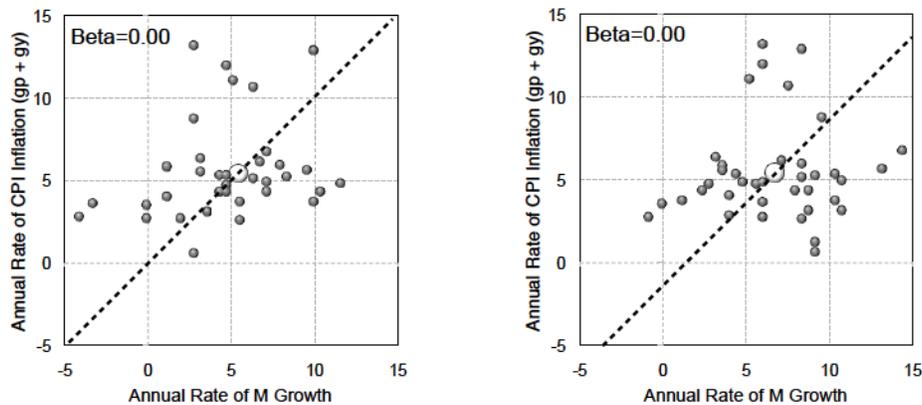
The formal expression of the Quantity Theory of Money is the following

$$MV = PY \tag{1}$$

where M is the nominal money supply, V is the velocity of circulation of money, P is the price level, and Y is real output. Let g_x be the growth rate of variable x . Then, if we assume that V is relatively constant, it follows that

$$g_M \simeq g_P + g_Y \tag{2}$$

Therefore, when the Quantity Theory of Money holds, if we graph $g_P + g_Y$ against g_M , we will get a 45 degree line. And, when the Quantity Theory of Money does not hold, we will get a meaningless bird-shot scatter plot. This is the central idea in Lucas (1980).



*The coordinates of the center of the white circle in each panel are the grand mean of the unfiltered sample.

Figure 1: The Quantity Theory in United States in the Short Run (1960:Q1–2009:Q4)

2 The Quantity Theory of Money in the United States

In this section we discuss whether the Quantity Theory of Money relationship holds in the United States both in the short run and in the long run.

2.1 The Quantity Theory of Money in the United States in the Short Run

In his 1980 article, Lucas plots the quarterly growth rate of money against the quarterly growth rate of prices for the 1955–1975 period using M1 as the monetary aggregate and he obtains a bird-shot scatter plot that shows that the Quantity Theory of Money does not hold in the short run in the United States during that period.

Here we use Lucas’ idea but we make three changes: we start our sample period in 1960 and we extend it to 2009, we use both M1 and M2 as our monetary aggregates, and, while Lucas plots the rate of growth of prices against the rate of growth of money, we follow the text-book description of the Quantity Theory of Money *exactly* and plot the rate of growth of prices *plus the rate of growth of output* against the rate of growth of money. We implement these three changes, we plot the resulting time series, and we obtain the bird-shot scatter plots that we represent in Figure 1. Our scatter plots illustrate that the Quantity Theory of Money relationship did not hold in the United States in the short run between 1960 and 2009 either for M1 or for M2, and they confirm Lucas (1980)’s findings.⁶

2.2 The Quantity Theory of Money in the United States in the Long Run

To illustrate whether the Quantity Theory of Money holds in the long run, Lucas (1980) associates the short-run with the high-frequency fluctuations of the quantity theory time series expressed in growth rates, and the long-run with the low-frequency fluctuations of those series. To remove the high-frequency fluctuations and to obtain the low-frequency signal, Lucas transforms the original series using the following two-sided, exponentially-weighted, moving-average filter

$$x_t(\beta) = \alpha \sum_{k=1}^T \beta^{|t-k|} x_k \quad (3)$$

where

$$\alpha = \frac{(1 - \beta)^2}{1 - \beta^2 - 2\beta^{(T+1)/2}(1 - \beta)} \quad 0 \leq \beta < 1 \quad (4)$$

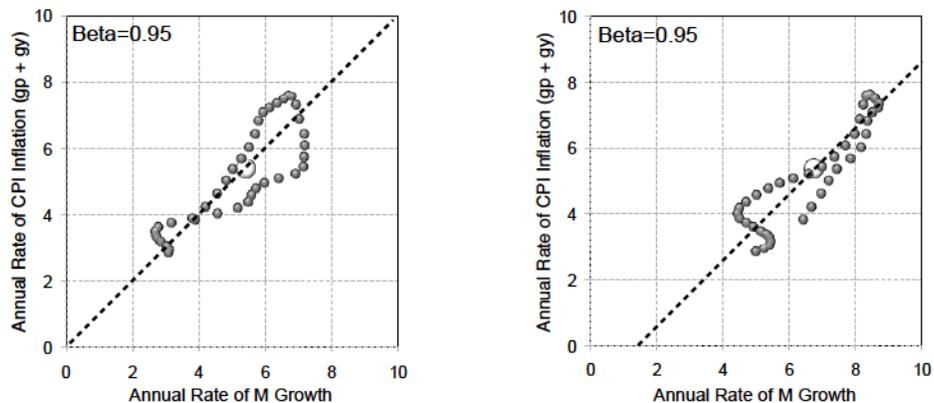
and where T is the number of observations in the time series.⁷

⁶We have taken all the data from FRED2 (<http://research.stlouisfed.org/fred2/>). The time series that we have used are GNPC96, M1SL, M2SL, and CPIAUCNS. Our results are robust to using three alternative measures for inflation: CPIAUCSL, CPILFENS, CPILFESL. We measure the growth rates as the percentage changes on the same quarter of the previous year.

⁷Parameter α guarantees that the means of the original and the filtered time series coincide. In fact, Lucas (1980) uses a slightly different definition of this parameter. He makes $\alpha = (1 - \beta)/(1 + \beta)$. His definition guarantees that the means coincide assuming that the lengths of the unfiltered series are infinite. Instead, we use Sargent and Surico (2011)’s small-sample correction to the value of α . This correction preserves the means of the series, but assuming that the lengths of the unfiltered series are finite.

A value of $\beta = 0.0$ returns the original time series, and increasingly higher values of β filter out the higher frequency fluctuations from the original time series and leave only the increasingly lower frequency fluctuations in the transformed series. Figure 5 illustrates how our version of Lucas' filter transforms the original U.S. time series as we increase the value of parameter β .⁸ The filter is two-sided because the behavior of households is likely to be affected both by what happened to them in the past and by their expectations of what might happen to them in the future.⁹

One important advantage of using Lucas' methods to find out whether the Quantity Theory of Money holds in the long run is that his filter is atheoretical. This means that its results do not depend on any modelling assumptions. In contrast, other methods that are more sophisticated econometrically, such as structural VARs, require identifying assumptions that are model-dependent. Those methods are less useful to compare model economies that are fundamentally different, like those that we consider in this article.¹⁰



A: M1 in the United States $\beta = 0.95$

B: M2 in the United States $\beta = 0.95$

*The coordinates of the center of the white circle in each panel are the grand mean of the unfiltered sample.

Figure 2: The Quantity Theory in United States in the Long Run (1960:Q1–2009:Q4)

Higher values of β extract the higher frequency fluctuations from the original series. Therefore, if the Quantity Theory of Money relationship holds in the long-run, as we increase the value of β , the plots of the filtered time series should look increasingly like the 45 degree line that runs through the grand mean of the unfiltered series. And, if it does not hold, we have no theory to account for the relationship between those variables and we expect the filtered data to become a blob around the grand mean of the unfiltered data. In fact, Lucas (1980) shows that this is precisely what happens when he plots the unemployment rate against the rate of growth of money, for the 1955–1975 period.

In Figure 2 we plot the Quantity Theory of Money relationship in the United States in the long-run or, more precisely, when $\beta = 0.95$. In both panels of that figure we see that, when we filter out the high-frequency fluctuations, the original bird-shot scatters displayed in Figure 1 disappear and

⁸To prevent clutter, in all our figures we follow Lucas (1980) exactly and plot only the fourth quarter of every year. To prevent end-of-sample distortions, we drop the first two and last two years from each graph, even though we use them in the filter.

⁹The choice of filter is not important. For example, Benati (2005, 2009) reports similar conclusions using a band-pass filter.

¹⁰See Lucas (1980) for a discussion of his filter and of its frequency interpretation, and see Whiteman (1984) for further details on this discussion.

the observations approach the 45 degree line that runs through the grand mean of the unfiltered sample. Therefore, the scatter plots displayed in Figure 2 illustrate that the Quantity Theory of Money held in the United States in the long-run both for M1 and for M2 during the 1960–2009 period, and they confirm Lucas (1980)’s findings.

The long-run scatter plot for M1 is interesting from the perspective of the monetary history of the United States. In Panel A of Figure 2 the observations start near the bottom left-hand-side corner of the graph and they march roughly up the 45 degree line during the late 1960s and the 1970s. When they reach the top-right-hand corner of the graph, they suddenly drop down almost vertically. This period of sharply falling average growth rates of prices represents the beginning of the 1980s when the Federal Reserve, under Paul Volcker, started tightening monetary policy to fight inflation—and eventually defeat it. Then, in the 1990s and 2000s the points return to the 45 degree line as the U.S. economy transitions to a new monetary regime with a lower inflation rate and lower money growth rate.

2.3 Quantifying Lucas’ Illustrations

There are two relatively straight-forward methods to quantify Lucas’ illustrations. The first one is to compute the average Cartesian distance of the points in the plots from the 45 degree line that runs through the grand mean of the unfiltered observations.¹¹ The other one is to compute the slope of an ordinary least squares (OLS) linear regression of the growth rate of prices plus the growth rate of real output on the growth rate of money. This second method was proposed by Whiteman (1984).

The formal definition of the first statistic is the following

$$D45 = \frac{1}{\sqrt{2T}} \sum_i |x_i - y_i + (\bar{y} - \bar{x})| \tag{5}$$

where y_i is the value of the i -th observation of the growth rate of prices plus the growth rate of output, either of the original or of the filtered time series; x_i is the corresponding observation of the growth rate of money and \bar{x} and \bar{y} are the average values of the unfiltered x_i and y_i . Obviously, if the Quantity Theory of Money relationship holds, the value of the $D45$ statistic will be small and, if it does not hold, it will be large.

In Whiteman (1984)’s regression, the value of the OLS coefficient will be close to unity when the Quantity Theory of Money relationship holds, and it can take any value when it does not hold. Obviously, chances are that it will be different from unity in this case.

In Table 1 we report the values of these two statistics for the United States both in the short run, when $\beta = 0.00$, and in the long run, when $\beta = 0.95$, for $M1$ and for $M2$. Our numerical results confirm what we learnt from Lucas’ Illustrations. The Quantity Theory of Money did not hold in the short run in the United States, between 1960 and 2009, either for M1 or for M2, but it held in the long run during the same period for both monetary aggregates. Moreover, according our two statistics, the Quantity Theory of Money relationship was tighter for M2 than for M1, both in the short run and in the long run. The distances from the 45 degree line were smaller for M2 than for M1 in both instances, and the slopes of the linear regressions were higher for M2, also in both instances (see Table 1).

¹¹The Cartesian distance of a point, (x_i, y_i) , from a line, $ax + by + c = 0$, is $d = |ax_i + by_i + c|/\sqrt{a^2 + b^2}$.

Table 1: The Quantity Theory of Money Statistics in the United States

	Short Run ($\beta=0.0$)		Long Run ($\beta=0.95$)	
	M1	M2	M1	M2
Distance from 45 Degree Line ($D45$)	2.9850	2.5420	0.5003	0.3953
OLS Regression Coefficient	0.0189	0.0723	0.8179	0.9164

2.4 An Apparent Conflict with the Literature

Our finding that the Quantity Theory of Money holds in the long run in the United States is somewhat in conflict with those of Sargent and Surico (2011). They use M2 as the monetary aggregate, they divide the 1900–2005 period into four subperiods, which they identify with different monetary policy regimes, and they find that the long-run slopes of Lucas’ Illustrations differ in these four regimes.

They use their results to argue that the 1984–2005 subperiod corresponds to an inflation targeting regime, and that this delivers a flatter slope. We contend that Sargent and Surico’s result is not due to a breakdown in the long-run Quantity Theory of Money relationship. Instead, we think that Sargent and Surico (2011)’s slopes vary for two reasons: first, because they follow Lucas (1980) literally and leave out the growth rate of output from their calculations and, second, because of the specific subperiods in which they choose to divide their sample.

Leaving out the growth rate of output, as Lucas (1980) did when he studied the 1950–1975 period, has little effect on his illustrations because the growth rate of output was small relative to the growth rate of prices during that period—recall that in the 1970s there was a hump in the U.S. inflation rate time series. Therefore, this omission affects Lucas’ original illustrations only slightly. But this is not the case for the post-1984 period, when the inflation rate was moderate and output growth was relatively high.¹²

The importance of the starting points of Sargent and Surico’s subperiods is highlighted by our earlier comments that the inflation rate decreased around 1980 and that this reduction was followed by a lower growth rate of money, but with a delay. As we have already mentioned, this is illustrated by the temporary dip below the 45 degree line of the filtered points that correspond to the early 1980’s in Panel A of Figure 2, and the subsequent return to the 45 degree line during the 1990’s.

In Appendix B we provide an econometric test of our claim that the slopes of Lucas’ Illustrations remained unchanged during the 1960–2009 period when we include output growth in the Quantity Theory of Money relationship. This period runs across 1984, which is the year which Sargent and Surico (2011) identify as the year when the monetary regime, and hence the slope of Lucas’ illustrations, supposedly changed.

To do this, we exploit the econometric interpretation of the Quantity Theory of Money as a cointegration relationship between the logs of the price level, real output, and the money supply. We test for a structural break at an unknown point in time in the cointegrating vector—which would be the econometric interpretation of a change in the slope of Lucas’ Illustrations—and we reject that such a break occurred for M2 at any time during the 1960–2009 period.

¹²The ratio of annual inflation to GDP growth fell from roughly 2 in the 1960–1983 period, to approximately half that amount in the 1984–2009 period.

3 The Quantity Theory of Money in the Model Economies

In this section we explore the extent to which the Quantity Theory of Money relationship holds in three of the modelling frameworks most frequently used by economists to think about monetary policy: the Cash-in-Advance framework, the New-Keynesian framework, and the Search-Money framework. As we did for the United States, to answer this question we use two methods: Lucas’ illustrations of the Quantity Theory of Money relationship and the two statistics that we have used in the previous section to quantify this relationship.

As we have already mentioned, for each one of these three frameworks we choose a representative model economy: for the Cash-in-Advance framework, we use the model economy described in Cooley and Hansen (1989); for the New-Keynesian framework, the model economy described in Chapter 3 of Galí (2008); and, for the Search-Money framework, a stochastic extension of the model economy described in Aruoba, Waller, and Wright (2011). Since these three model economies are standard in the literature we relegate their detailed description to Appendix A of this article.

We also describe in detail our calibration procedure in that appendix. To make our comparisons meaningful, we use the same functional forms and parameter values for the utility functions and for the processes on the technology and the monetary shocks, whenever possible.¹³ We also use the same methods to characterize the equilibrium processes of our three model economies and to find their solutions.

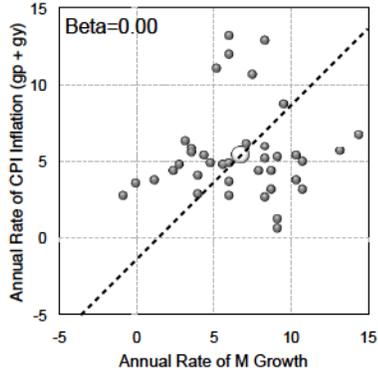
In all three cases, we describe the equilibrium processes as systems of stochastic difference equations and we solve these systems using the default perturbation methods of Dynare that allow us to obtain quadratic laws-of-motion. Then we simulate the three model economies and we obtain samples of 204 quarterly observations to replicate the number of observations in our United States sample. To obtain these samples, we use the same seeds for the random number generators. Consequently, the sequences of realizations of the random processes are identical in the three model economies.

3.1 The Quantity Theory of Money in the Model Economies in the Short Run

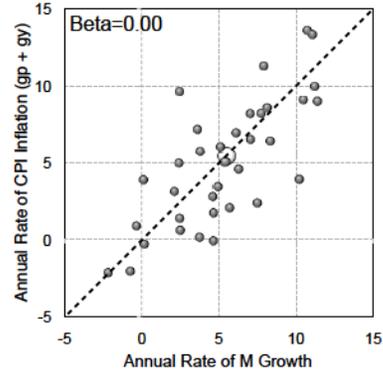
Figure 3 represents Lucas’ Illustrations in the short run—that is for $\beta = 0$ —for M2 in the United States and for the monetary aggregates of our three model economies. We observe that the Quantity Theory of Money relationship is much stronger in the three model economies than in the United States. In the three model economies, the points lie close to the 45 degree line as predicted by Quantity Theory of Money. And in the United States data it is hard discern any pattern.

We have only one United States time series from which to compute the Quantity Theory of Money statistics, but we can simulate many stochastic realizations of the equilibrium processes of our model economies. To reduce the size of the sampling error, we compute the D45 statistics and the slopes of the Quantity Theory of Money OLS linear regressions using 100 independent random samples. In Table 2 we report the sample means and the sample standard deviations of those statistics. We also reproduce the results for M2 for the United States economy to facilitate the comparisons.

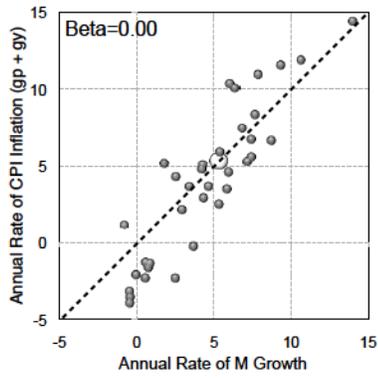
¹³We repeated our calculations with the functional forms and the calibration targets used in the original articles, and we found that this does not change our results qualitatively. This is partly due to the fact that the original articles target similar data moments and study similar time periods.



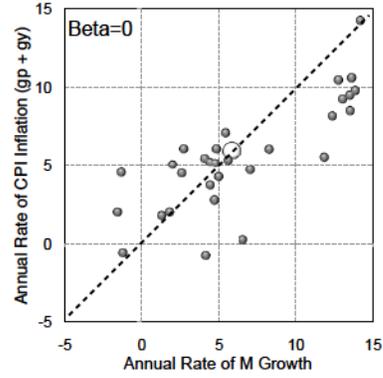
A: United States (M2)



B: Cash-in-Advance Model Economy



C: New-Keynesian Model Economy



D: Search-Money Model Economy

*The coordinates of the center of the white circle in each panel are the grand mean of the unfiltered sample.

Figure 3: Lucas' Illustrations in the Short Run ($\beta = 0.00$)

Table 2: The Quantity Theory of Money Statistics in the Short Run

	US (M2)	Cash-in-Advance	New-Keynesian	Search-Money
D45	2.5420	0.9338	1.5611	1.9278
(std dev)	—	(0.0890)	(0.1185)	(0.1408)
OLS coeff.	0.0723	1.0612	1.3366	1.4219
(std dev)	—	(0.0593)	(0.0501)	(0.0652)

Both sets of statistics confirm what we found using Lucas' Illustrations, and they establish that our graphs are not the result of a sampling oddity. The D45 statistic for the United States is 2.54, while in all three model economies it is below 2.0. This result arises from the fact that the points in the graph for the United States form a shapeless cloud, while those in the graphs for the three model economies form clouds which are closer to the 45 degree line. Thus, the D45 statistic confirms that our three model economies display too much of the Quantity Theory of Money relationship in the short run, when compared with the United States economy.

The slopes of the Quantity Theory of Money regressions tell pretty much the same story. They are much closer to unity in the three model economies than in the United States, and the slope of the Quantity Theory of Money regression line is closest to unity in the Cash-in-Advance model economy. We interpret these results to mean that in our three model economies the rate of growth of prices responds too quickly to changes in the rate of growth of money, relative to the United States, or that our three model economies do not display enough short-run sluggishness in the response of prices.

3.2 The Departures from the Quantity Theory of Money in the Short Run

In this subsection we describe how the three model economies depart from the Quantity Theory of Money in the short run using only one equation for each one of them. Specifically, we provide an expression for the equilibrium values of the term PY/M for each model economy.¹⁴ We provide the derivation of these equations in Appendix A, together with the full descriptions of the model economies. If the Quantity Theory of Money held exactly $(M/P)/Y$ would be constant. Therefore, these single equation expressions thus help to understand how each model economy departs from the Quantity Theory of Money in the short-run.

(a) *The Cash-in-Advance Model Economy*

In the Cash-in-Advance model economy we obtain that

$$\frac{PY}{M} = \frac{P(C + X)}{M} = 1 + \frac{PX}{M} \quad (6)$$

where C is consumption, and X is investment. So the Cash-in-Advance framework succeeds in departing from the Quantity Theory of Money in as far as monetary policy distorts investment decisions. These distortions take place on the cash-good (consumption) and credit-good (investment) margin.

(b) *The New-Keynesian Model Economy*

In the New-Keynesian model economy when we rewrite PY/M in logs we obtain that

$$p_t + y_t - m_t = \eta i_t = \eta r_t^n + \eta E_t \{ f(\pi_t, \pi_{t+1}, \pi_{t+2}) \} \quad (7)$$

where i_t is the nominal interest rate, η is the elasticity of money demand, r_t^n is the natural real interest rate, and $f(\pi_t, \pi_{t+1}, \pi_{t+2})$ is a linear function of current and future inflation. It is evident from expression (7) that the elasticity of money demand and the changing values of the nominal

¹⁴These expressions are also related to the issue of money demand as discussed in Lucas (2000). Lucas defines money demand as the relationship between nominal interest rates and the ratio of real money holdings to real output, or $(M/P)/Y$. This ratio is the inverse of the PY/M term which we consider here.

interest rates play an important role in allowing the New-Keynesian model to get away from the Quantity Theory of Money in the short run.

What role do sticky prices play in this? The natural real interest rate, r_t^n , is independent of both monetary variables and the parameters that determine the degree of price stickiness. So, for sticky prices to be part of the story, they must operate through the inflation rate which evolves according to

$$\pi_t = (1 - \theta)(p_t^* - p_{t-1}) \quad (8)$$

where p_t^* is the price level chosen by the firms that get to reset their prices, and $(1 - \theta)$ is the share of those firms. So sticky prices affect the rate of inflation and, therefore, the nominal interest rates and they contribute to the short-run departure from the Quantity Theory of Money relationship. In practice, however, this effect is small.¹⁵ This is because the nominal interest rate does not change immediately as predicted by Fisher's equation, $i = r + \pi$, after a monetary shock because these shocks generate a liquidity effect. But this liquidity effect is not very long-lasting and it all but disappears, when we filter out the high frequency fluctuations of the nominal time series. We conclude that the short-run behavior of the New-Keynesian model arises directly from the money demand equation, and that the role played by the degree of price stickiness is small.

(c) *The Search-Money Model Economy*

In the Search-Money model economy we obtain that,

$$\frac{PY}{M} = \frac{1}{z(q, K)} \frac{\gamma Y}{F_N(K, N)} \quad (9)$$

where γ is a parameter that quantifies the disutility of labor. In the simulations of this model economy total output, Y , the stock of capital, K , and the marginal product of labour, $F_N(K, N)$, are almost constant. Consequently, they do not account for the departure from the Quantity Theory of Money relationship in the short run. They are almost constant because most of the trades are non-monetary, the centralized night market is much bigger than the decentralized day market, and this market is almost unaffected by changes in the money supply. Almost all the variability in expression (9) comes from changes in $z(q, K)$, which represents the terms of trade in monetary exchanges and, more specifically, from changes in q —the amount produced and traded in the monetary exchanges that take place in the decentralized market. These changes in q are caused by the unexpected changes in the amount of money and by changes in the inflation rate, which is the cost of holding money.

In summary, the Search-Money framework succeeds in departing from the Quantity Theory of Money relationship in the short run because of the effects of changes in the money supply on the value of money. People want to hold money because it is useful for monetary trades. The value of money is jointly determined by this demand for money and by the money supply. Since the nominal price of consumption goods is the inverse of the cost of acquiring money, changes in the value of money result in changes in the price level.¹⁶ Therefore, changes in the money supply affect the value of money, and thereby the price level. Moreover, the resulting inflation also affects output

¹⁵We experimented with various parameter values, and we found that the Quantity Theory of Money relationship was largely unaffected by the degree of price-stickiness, θ , but that it was very sensitive to the value of the elasticity of money demand, η .

¹⁶The price of consumption goods is the number of units of money that agents exchange for one unit of the consumption good. The cost of acquiring money is the number of units of the consumption good that agents give up to obtain one unit of money.

because of a holdup problem.¹⁷ This effect is magnified because monetary trades account only for a small fraction of total exchanges and, therefore, changes in the supply of money are large relative to the size of the total amount of monetary trades. Consequently, the effect of a given change in money supply on the value of money and, hence, on prices in the Search-Money framework, is larger than in the other two frameworks.

3.3 The Quantity Theory of Money in the Model Economies in the Long Run

Figure 4 represents Lucas' Illustrations in the long run, that is, for $\beta = 0.95$, for M2 in the United States and for the monetary aggregates of our three model economies. Given that the Quantity Theory of Money relationship was clearly present in the three models economies in the short run, it comes as no surprise that it also holds in all three of them in the long run. In fact, the Quantity Theory of Money relationship is so tight in every model economy that we are hard put to say in which one of them it is tightest. For that purpose, we must turn to the statistics that we report in Table 3.

The D45 statistic shows that the Quantity Theory of Money relationship is tightest in the New-Keynesian model economy, followed by the Search-Money model economy and by the Cash-in-Advance model economy. But the differences between them are small. And in all three cases the Quantity Theory of Money relationship is much tighter than in the United States. The values of the slopes of the OLS linear regressions confirm these findings.

We interpret these results to mean that in the long run the Quantity Theory of Money relationship is present both in the United States and in our three model economies. But, once again, it is sizably tighter in the model economies.

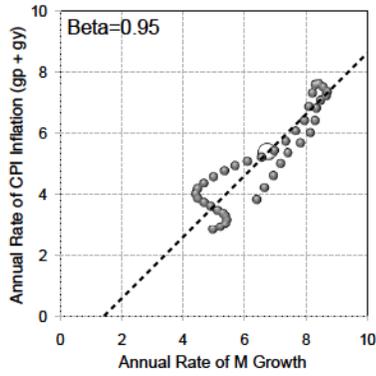
Table 3: The Quantity Theory of Money Statistics in the Long Run

	US (M2)	Cash-in-Advance	New-Keynesian	Search-Money
D45	0.3953	0.0555	0.0149	0.0188
(std dev)	–	(0.0147)	(0.0031)	(0.0039)
OLS coeff.	0.9164	0.9922	1.0015	1.0015
(std dev)	–	(0.0690)	(0.0116)	(0.0126)

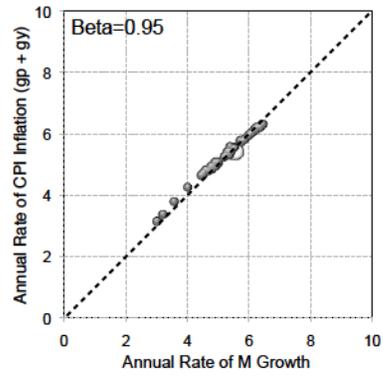
3.4 Who Wins?

We have shown that all three of our model frameworks display the Quantity Theory of Money relationship in the long run and, therefore, succeed in replicating the long-run behavior of the United States economy. But we have also shown that the Quantity Theory of Money relationship is much tighter in all three model economies than in the United States in the short run. Given the

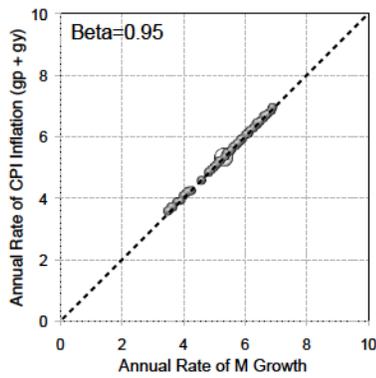
¹⁷In every Search-Money model inflation decreases output. Sellers in single-coincidence meetings know that they can increase the price of the consumption good because the outside option of the buyer—to hold onto the money until next period—is less attractive when the inflation rate is higher. These increased prices—known as the holdup problem—decrease economic efficiency and output. See, eg., Lagos and Wright (2005) for a discussion of the holdup problem.



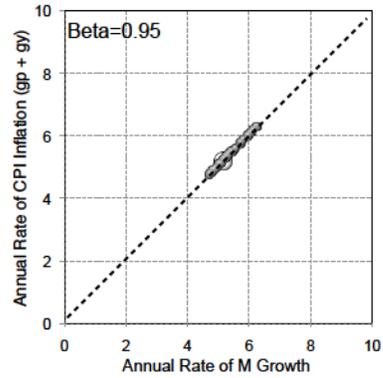
A: United States (M2)



B: Cash-in-Advance Model Economy



C: New-Keynesian Model Economy



D: Search-Money Model Economy

*The coordinates of the center of the white circle in each panel are the grand mean of the sample.

Figure 4: Lucas' Illustrations in the Long Run ($\beta = 0.95$)

importance of departures from the Quantity Theory of Money for the behavior of money, prices, and hence monetary policy, we think that this is an important shortcoming for the model economies.

The difficulties in capturing the short-run departures from the Quantity Theory of Money have been known to afflict the Cash-in-Advance framework since the work of Hodrick, Kocherlakota, and Lucas (1991). While the New-Keynesian and Search-Money frameworks have cast light on a number of other issues in monetary economics, they have not resolved these difficulties. Perhaps further research within these frameworks will succeed in enabling them to depart from the Quantity Theory of Money relationship in the short run.

Progress within the Cash-in-Advance framework in the attempt to slow down the response of prices to monetary shocks—a problem closely related to departing from the Quantity Theory of Money relationship in the short-run—appears to have stalled. Early attempts to solve this problem use constructs such as portfolio-adjustment costs (see, for instance, Christiano, 1991, and Christiano and Eichenbaum, 1995). But this line of research has trailed off since then, after being only moderately successful. Perhaps a partial exception to this rule can be found in the recent work of Alvarez, Atkeson, and Edmond (2009) who allow for cash-in-advance constraints that last for multiple periods.

Progress within the New-Keynesian framework seems to be more promising. For example, Christiano, Eichenbaum, and Evans (2005) develop a New-Keynesian model capable of reproducing the slower reaction of the economy to monetary shocks observed in empirical studies that use impulse-response functions. This suggests that these more advanced New-Keynesian models might offer better hopes for departing from the Quantity Theory of Money relationship in the short run.

The Search-Money framework is a more recent construct, and bringing productive capital into that framework is a very recent achievement. Therefore, future refinements within this framework might enable it to depart from the Quantity Theory of Money relationship in the short run. Time will tell.

4 Conclusion

In this article we show that the Quantity Theory of Money held in the long-run in the United States between 1960 and 2009. And we also show that it failed to hold in the short-run during that period. Given the prominence of the Quantity Theory of Money in monetary theory, we argue that monetary model economies should replicate both the long-run success and short-run failure of the Quantity Theory of Money observed in the United States, if we are to trust their prescriptions for monetary policy.

Our analysis, based on the Lucas (1980) Illustrations, shows that every one of the three main frameworks that are currently used to study monetary policy—the Cash-in-Advance framework, the New-Keynesian framework, and the Search-Money framework—display the Quantity Theory of Money relationship both in the long-run and in the short-run. This failure of all three frameworks to depart from the Quantity Theory of Money in the short-run casts some doubts on their usefulness for the analysis of monetary policy—which most monetary theorists consider to be an inherently short-run phenomenon.

To break away from the Quantity Theory of Money in the short-run, the three monetary frameworks that we study here need a more sluggish response of the growth rate of prices to changes in

the growth rate of money. We are not sure about what causes this sluggish response of prices to changes in money in the real world. But the generally accepted conjecture is that the way money is introduced into the economy most probably makes a difference.

When money changes are universal and simultaneous—that is, when they affect every agent at the same time—the rate of growth of prices responds immediately to changes in the rate of growth on money. But, as we mentioned in the introduction, when money enters the economy at a specific point, it has to spread around from there. The time it takes in this spreading around probably creates the sluggishness. Like the Quantity Theory of Money itself, this idea can also be traced back to David Hume; “[T]he money in its progress through the whole commonwealth...first quicken[s] the diligence of every individual before it encrease the price of labour.” (Hume, 1742, *Of Money*).

In representative agent model economies there is only one point at which money can enter the economy. Once it reaches this agent, it has nowhere to spread around, and the sluggish response of prices is very hard to achieve. In this type of model economies every change in the growth rate of money is both universal and simultaneous by construction. This reasoning allows us to conjecture that agent heterogeneity may very well turn out to be a necessary condition for model economies to display the needed sluggishness.

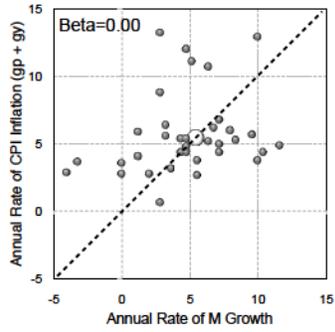
Díaz-Giménez, Prescott, Alvarez, and Fitzgerald (1992) model the role of money as an asset in a heterogeneous household setup, and they give an early quantitative step in what could turn out to be the correct direction. The findings of Alvarez, Atkeson, and Edmond (2009), Telyukova and Visschers (2013), and Williamson (2008), each of which includes a degree of agent heterogeneity, suggest that agent heterogeneity may indeed be key in replicating the sluggishness observed in the data. As far as the Quantity Theory of Money relationship is concerned, the explicit modeling of agent heterogeneity is probably one of the best bets for future research.

References

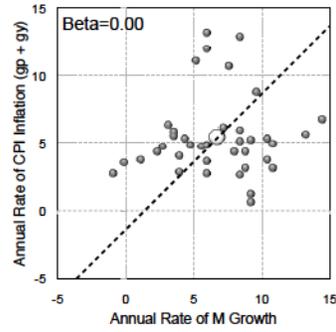
- George Akerlof. Irving fisher on his head: The consequences of constant threshold-target monitoring of money holdings. Quarterly Journal of Economics, 93(2):169–187, 1979.
- Fernando Alvarez, Andrew Atkeson, and Chris Edmond. Sluggish responses of prices and inflation to monetary shocks in an inventory model of money demand. Quarterly Journal of Economics, 124(3):911–967, 2009.
- S. Boragan Aruoba and Randall Wright. Search, money and capital: A neoclassical dichotomy. Journal of Money, Credit, and Banking, 35:1086–1105, 2003.
- S. Boragan Aruoba, Christopher Waller, and Randall Wright. Money and capital. Journal of Monetary Economics, 58:98–116, 2011.
- Luca Benati. Long-run evidence on money growth and inflation. Quarterly Bulletin of the Bank of England, Autumn:349–355, 2005.
- Luca Benati. Long-run evidence on money growth and inflation. European Central Bank, Working Paper Series, No. 1027, 2009.
- Aleksander Berentson, Guido Menzio, and Randall Wright. Inflation and unemployment in the long run. American Economic Review, 101:371–398, 2011.

- Guillermo Calvo. Staggered prices in a utility-maximizing framework. Journal of Monetary Economics, 12:383–398, 1983.
- Lawrence Christiano. Modeling the liquidity effect of a money shock. Quarterly Review of the Federal Reserve Bank of Minneapolis, pages 3–34, 1991.
- Lawrence Christiano and Martin Eichenbaum. Liquidity effects and the monetary transmission mechanism. American Economic Review, 82:346–353, 1992.
- Lawrence Christiano and Martin Eichenbaum. Liquidity effects, monetary policy, and the business cycle. Journal of Money, Banking, and Credit, 27:1113–1136, 1995.
- Lawrence Christiano, Martin Eichenbaum, and Charles Evans. Nominal rigidities and the dynamic effects of a shock to monetary policy. Journal of Political Economy, 113:1–45, 2005.
- Richard Clarida, Jordi Galí, and Mark Gertler. Monetary policy rules and macroeconomic stability: Evidence and some theory. The Quarterly Journal of Economics, 115(1):147–180, 2000.
- Robert Clower. A reconsideration of the microfoundations of monetary theory. Western Economic Journal, 6:1–9, 1967.
- Thomas Cooley and Gary Hansen. The inflation tax in a real business cycle model. American Economic Review, 79:733–748, 1989.
- Thomas Cooley and Gary Hansen. Money and the business cycle. In Cooley T., editor, Frontiers of Business Cycle Research, chapter 6. Princeton University Press, 1995.
- Javier Díaz-Giménez, Edward Prescott, F. Alvarez, and T. Fitzgerald. Banking in computable general equilibrium economies. Journal of Economic Dynamics and Control, 16:533–559, 1992.
- Irving Fisher. The Purchasing Power of Money. Macmillan, 1911.
- Milton Friedman and Anna J. Schwartz. A Monetary History of the United States, 1867-1960. Princeton University Press, 1963.
- Timothy Fuerst. Liquidity, loanable funds, and real activity. Journal of Monetary Economics, 29:3–24, 1992.
- Jordi Galí. Monetary Policy, Inflation, and the Business Cycle: An Introduction to the New Keynesian Framework. Princeton University Press, 2008.
- Allen Head, Lucy Qian Liu, Guido Menzio, and Randall Wright. Stick prices: A new monetarist approach. Journal of the European Economic Association, 10(5):939–973, 2012.
- Robert Hodrick, Narayana Kocherlakota, and Deborah Lucas. The variability of velocity in cash-in-advance models. Journal of Political Economy, 99(2):358–384, 1991.
- David Hume. Of interest. In Eugene F. Miller, editor, Essays, Moral, Political, and Literary, chapter Part II, Chapter IV. Liberty Fund, Inc., 1987 edition, 1742a. <http://www.econlib.org/library/LFBooks/Hume/hmMPL27.html>.
- David Hume. Of money. In Eugene F. Miller, editor, Essays, Moral, Political, and Literary, chapter Part II, Chapter III. Liberty Fund, Inc., 1987 edition, 1742b. <http://www.econlib.org/library/LFBooks/Hume/hmMPL27.html>.

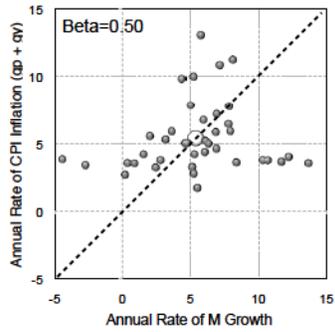
- Robert King and Mark Watson. Testing long-run neutrality. Economic Quarterly, 1997.
- Robert King and Mark Watson. Inflation and unit labor cost. Journal of Money, Credit, and Banking, 44:111–149, 2012.
- Nobuhiro Kiyotaki and Randall Wright. On money as a medium of exchange. Journal of Political Economy, 97-4:927–54, 1989.
- Ricardo Lagos and Randall Wright. A unified framework for monetary theory and policy analysis. Journal of Political Economy, 113:463–484, 2005.
- Robert Lucas. Two illustrations of the quantity theory of money. American Economic Review, 70: 1005–1014, 1980.
- Robert Lucas. Money demand in the united states: A quantitative review. Carnegie-Rochester Conference Series on Public Policy, 29(1):137–167, 1988.
- Robert Lucas. Liquidity and interest rates. Journal of Economic Theory, 50:237–264, 1990.
- Robert Lucas. Inflation and welfare. Econometrica, 68(2):247–274, 2000.
- James Nason and Timothy Cogley. Testing the implications of long-run neutrality for monetary business cycle models. Journal of Applied Econometrics, 9:S37–70, 1994.
- Thomas Sargent and Paolo Surico. Two illustrations of the quantity theory of money: Breakdowns and revivals. American Economic Review, 101(1):109–128, 2011.
- Byeongseon Seo. Tests for structural change in cointegrated systems. Econometric Theory, 14: 222–259, 1998.
- Frank Smets and Rafael Wouters. Shocks and frictions in us business cycles: A bayesian dsge approach. American Economic Review, 97(3):586–606, 2007.
- James Stock and Mark Watson. A simple estimator of cointegrating vectors in higher order integrated systems. Econometrica, 61(4):783–820, 1993.
- Nancy Stokey and Robert Lucas. Optimal fiscal and monetary policy in an economy without capital. Journal of Monetary Economics, 12:55–93, 1983.
- Nancy Stokey and Robert Lucas. Money and interest in a cash-in-advance economy. Econometrica, 66:491–513, 1987.
- Pedro Teles and Ruilin Zhou. A stable money demand: Looking for the right monetary aggregate. Economic Perspectives (Federal Reserve Bank of Chicago), Q1:50–63, 2005.
- Irina Telyukova and Ludo Visschers. Precautionary money demand in a business-cycle model. Journal of Monetary Economics, Forthcoming, 2013.
- Neil Wallace. A dictum for monetary theory. Quarterly Review of the Federal Reserve Bank of Minneapolis, 22(1):20–26, 1998.
- Charles Whiteman. Lucas on the quantity theory: Hypothesis testing without theory. American Economic Review, 74:742–49, 1984.
- Stephen Williamson. Monetary policy and distribution. Journal of Monetary Economics, 55:1038–1053, 2008.



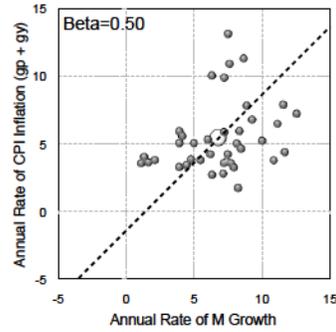
Panel A: M1 ($\beta = 0.00$)



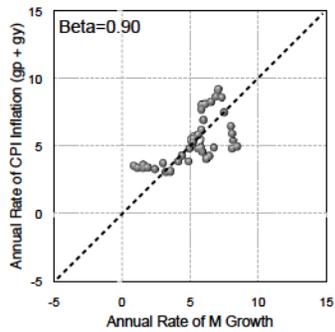
Panel B: M2 ($\beta = 0.00$)



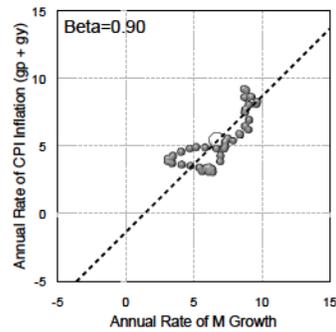
Panel C: M1 ($\beta = 0.50$)



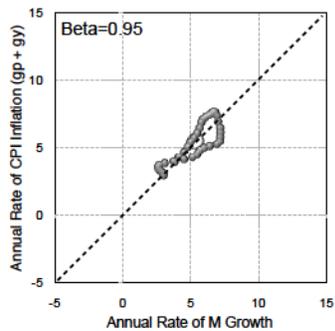
Panel D: M2 ($\beta = 0.50$)



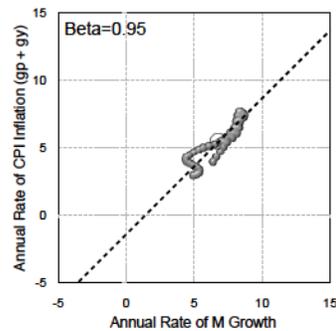
Panel E: M1 ($\beta = 0.90$)



Panel F: M2 ($\beta = 0.90$)



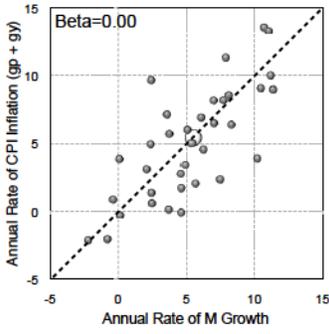
Panel G: M1 ($\beta = 0.95$)



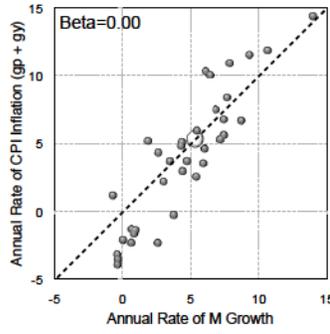
Panel H: M2 ($\beta = 0.95$)

*The coordinates of the center of the white circle in each panel are the grand mean of the sample.

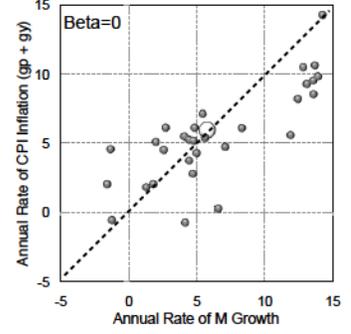
Figure 5: Lucas' Illustrations in the United States (1960:Q1–2009:Q4)



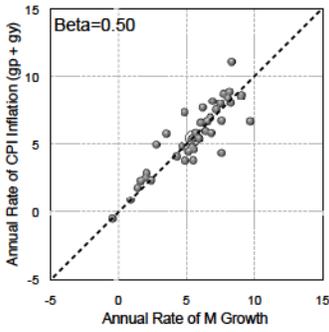
A: Cash-in-Advance ($\beta = 0.00$)



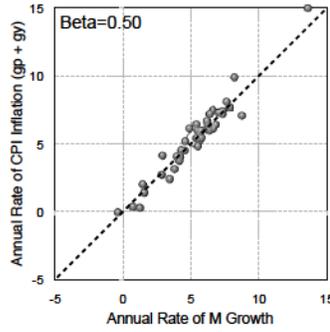
B: New-Keynesian ($\beta = 0.00$)



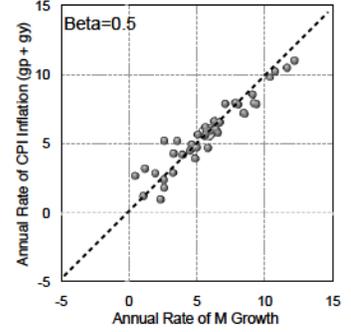
C: Search ($\beta = 0.00$)



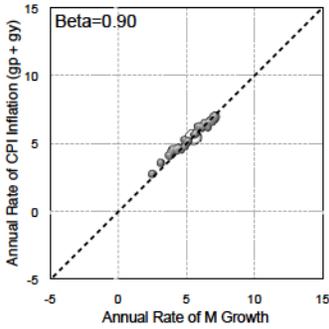
D: Cash-in-Advance ($\beta = 0.50$)



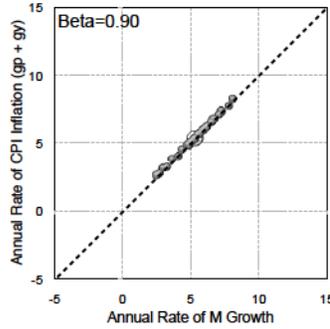
E: New-Keynesian ($\beta = 0.50$)



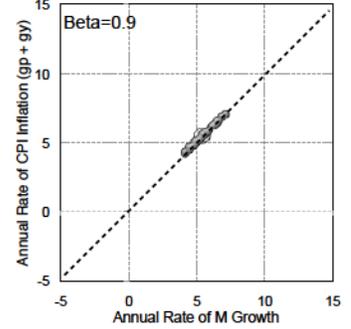
F: Search ($\beta = 0.50$)



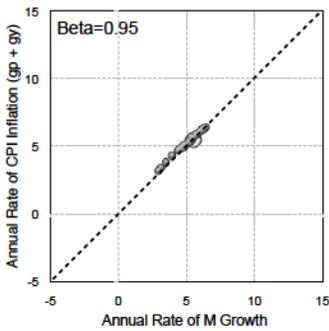
G: Cash-in-Advance ($\beta = 0.90$)



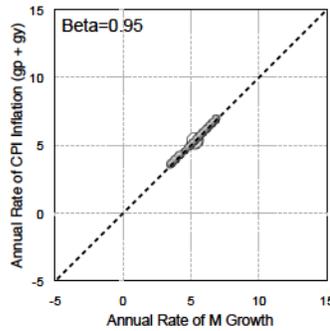
H: New-Keynesian ($\beta = 0.90$)



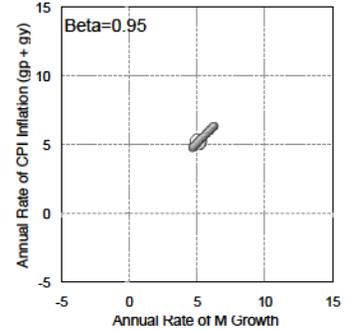
I: Search ($\beta = 0.90$)



J: Cash-in-Advance ($\beta = 0.95$)



K: New-Keynesian ($\beta = 0.95$)



L: Search ($\beta = 0.95$)

*The coordinates of the center of the white circle in each panel are the grand mean of the sample.

Figure 6: Lucas' Illustrations in the Model Economies

A The Monetary Model Economies

In this appendix we describe in detail each of the three model economies used as canonical examples for the three monetary frameworks: Cooley and Hansen (1989) for the Cash-in-Advance framework; Galí (2008) Chapter 3 for the New-Keynesian framework; and Aruoba, Waller, and Wright (2011) for the Search-Money framework. We then describe the details of the calibration and computation of all three models.

A.1 The Cash-in-Advance Model Economy

The cash-in-advance abstraction is an explicit way to model the transactions function of money by requiring that at least some goods have to be purchased with cash. This abstraction was first developed and analyzed in Lucas (1980), and more generally by Stokey and Lucas (1983, 1987), although the idea to model frictions in this way dates back to Clower (1967). Quantitative explorations of the business cycle implications of this abstraction can be found in Cooley and Hansen (1989, 1995). In this article, to represent the cash-in-advance abstraction, we use a minor variation of the model economy described in Cooley and Hansen (1989), but in its actual description we follow Nason and Cogley (1994).

In this model economy there are three goods: a consumption good, an investment good, and leisure. We assume that only the consumption good must be bought with cash carried over from the previous period, while the investment good and leisure can be purchased on credit. Model economies with this type of cash-in-advance constraint attempt to account for the distortionary effects of inflation on real activity. These distortions create an incentive for people to substitute away from activities that require cash —from consumption, in our case— towards activities that are exempt from this requirement —towards investment and leisure, in our case.

As shown by Hodrick, Kocherlakota, and Lucas (1991), one of the shortcomings of the cash-in-advance abstraction is that the model economies react too quickly to monetary shocks. Numerous extensions have attempted to deal with this shortcoming by adding liquidity effects via portfolio adjustment costs (see Lucas (1990); Fuerst (1992); Christiano and Eichenbaum (1992); and Christiano and Eichenbaum (1995), amongst others). But these extensions, while they have succeeded in addressing the issue of the liquidity effects, have had very limited success in generating a sluggish response of prices. See Christiano (1991) for an interesting discussion of the motivations, strengths, and weaknesses of the cash-in-advance approach to modelling money.

A.1.1 Households

The economy is inhabited by a continuum of identical households of measure one who order their preferences over stochastic processes of consumption and labor according to the following utility function:

$$\max E \sum_{t=0}^{\infty} \beta^t \left(\frac{C_t^{1-\sigma}}{1-\sigma} - \gamma \frac{N_t^{1+\varphi}}{1+\varphi} \right) \tag{10}$$

where $0 < \beta < 1$ is the discount factor, C_t is consumption, and N_t is labor¹⁸. Households in this model economy are endowed with one unit of time which they can allocate to the supply of labour services to the firm or to the enjoyment of leisure, that is $N_t \in [0, 1]$ for all t . The households face a budget constraint given by

$$P_t C_t + P_t X_t + M_t \leq P_t W_t N_t + P_t R_t K_t + M_{t-1} + T_t \quad (11)$$

where P_t is the price level, X_t is investment in capital, M_t are money holdings, W_t is the real wage, R_t is the real interest rate, K_t is capital holdings, and T_t is the lump-sum transfer of the cash injections made by monetary authorities.

The stock of capital evolves according to

$$K_{t+1} = (1 - \delta)K_t + X_t \quad (12)$$

where $0 < \delta < 1$ is the depreciation rate.

The innovation of the cash-in-advance abstraction that makes money necessary is to add a cash-in-advance constraint. This constraint requires that the consumption good must be purchased with money, in particular with money that must be ‘held in advance’. That is, with money holdings that are chosen one period ahead plus the money injected into the economy in the current period. This cash-in-advance constraint is

$$P_t C_t \leq M_{t-1} + T_t \quad (13)$$

The process on money defined later, following Cooley and Hansen (1989), will make the cash-in-advance constraint always binding¹⁹.

Therefore, the problem of the representative household is to choose C_t , N_t , M_t , X_t and K_t in order to maximize (10) subject to (11), (12), (13), and $N_t \in [0, 1]$.

A.1.2 Firms

Firms in the economy operate in competitive factor and product markets and produce output according to a constant returns-to-scale production function. These assumptions allow us to use a representative firm with a production function that takes the following form

$$Y_t = A_t K_{f,t}^\alpha N_{f,t}^{1-\alpha} \quad (14)$$

where Y_t is output, $K_{f,t}$ and $N_{f,t}$ are the capital and labour inputs, and A_t is a technology shock. Each period t the firms decision problem written in real terms is

$$\max_{Y_t, K_{f,t}, N_{f,t}} Y_t - W_t N_{f,t} - R_t K_{f,t} \quad (15)$$

The technology shock follows an exogenous AR(1) process in logs, given by

$$a_t = \rho_a a_{t-1} + \varsigma_t \quad (16)$$

where $a_t \equiv \log(A_t)$, ς_t is an identically and independently distributed process that follows a normal distribution with zero mean and variance σ_ς^2 .

¹⁸The utility function in Cooley and Hansen (1989) is $\log(C_t) - \gamma N_t$, which is a subcase of ours.

¹⁹This assumption, that the cash-in-advance constraint always binds, was shown to be unsequential by Hodrick, Kocherlakota, and Lucas (1991), since when it is allowed to be occasionally binding it remains the case that for quantitatively plausible calibrations it will bind almost all of the time anyway.

A.1.3 Money

The monetary authority of this economy issues non-interest bearing currency, M^s , according to the following rule

$$M_{t+1}^s = e^{\nu_t} M_t^s \quad (17)$$

where the stochastic money growth rate, ν_t , is revealed at the beginning of period t and evolves according to

$$\nu_t = (1 - \rho_m)\bar{\nu} + \rho_m\nu_{t-1} + \xi_t \quad (18)$$

where $0 < \rho_m < 1$ and where ξ is an identically and independently distributed process that follows a normal distribution with zero mean and variance σ_ξ^2 .

Given the money supply rule, the government makes the required money injections to implement it each period. These injections take the following form

$$T_t = M_{t+1}^s - M_t^s \quad (19)$$

and are given as lump-sum payments to the households, adding directly to their money holdings.

A.1.4 Prices and Market Clearance

Prices in this model economy are completely flexible and they adjust instantaneously so that labor, capital and money markets always clear. That is,

$$\begin{aligned} N_t &= N_{f,t} \\ K_t &= K_{f,t} \\ M_t &= M_t^s \end{aligned} \quad (20)$$

A.1.5 Equilibrium

To solve the model it must first be made stationary. The first step to achieve this is to divide equations (11) and (13) by the price level, P_t . The second step is to replace M_t and P_t in those two equations with $\hat{M}_t = M_t/M_t^s$ and $\hat{P}_t = P_t/M_t^s$, this allows us to remove the trending variables M_t, P_t .

Once the problem is stationary, the equilibrium of the cash-in-advance model economy can be characterized by the following system of equations that combines optimality conditions, budget and

technology constraints, and market clearing conditions.

$$K_{t+1} + \hat{M}_t/\hat{P}_t = W_t N_t + (R_t + 1 - \delta)K_t \quad (21)$$

$$C_t = \frac{\hat{M}_{t-1} + e^{\nu_t} - 1}{e^{\nu_t} \hat{P}_t} \quad (22)$$

$$W_t = (1 - \alpha)e^{a_t} K_t^\alpha N_t^{-\alpha} \quad (23)$$

$$R_t = \alpha e^{a_t} K_t^{\alpha-1} N_t^{1-\alpha} \quad (24)$$

$$\frac{N_t^\varphi}{W_t} = \beta E \left\{ \frac{N_{t+1}^\varphi}{W_{t+1}} (R_{t+1} + 1 - \delta) \right\} \quad (25)$$

$$\frac{W_t}{N_t^\varphi} = \frac{\gamma}{\beta} E \left\{ C_{t+1} e^{\nu_{t+1}} \frac{\hat{P}_{t+1}}{\hat{P}_t} \right\} \quad (26)$$

$$\hat{M}_t = 1 \quad (27)$$

$$a_t = \rho_a a_{t-1} + \varsigma_t \quad (28)$$

$$\nu_t = (1 - \rho_m)\bar{\nu} + \rho_m \nu_{t-1} + \xi_t \quad (29)$$

A.1.6 The Quantity Theory of Money in a Single Equation

We now describe with a single equation the quantity theory of money in way way that makes it easier to see how the Cash-in-Advance framework temporarily escapes from the quantity theory of money. Specifically, we give an expression for the term PY/M . Were the Quantity Theory of Money to hold exactly this term would be equal to a constant.

Using the cash-in-advance constraint, $P_t C_t = M_{t-1} + T_t$, with the aggregate resource constraint, $Y_t = C_t + X_t$, we have

$$\frac{P_t Y_t}{M_t} = \frac{P_t (C_t + X_t)}{M_t} = \frac{M_{t-1} + T_t}{M_t} + \frac{P_t X_t}{M_t} = 1 + \frac{P_t X_t}{M_t} \quad (30)$$

So the Cash-in-Advance framework succeeds in breaking away from the Quantity Theory of Money in-so-far as monetary policy distorts investment decisions (distorts the cash goods vs. credit goods margin).

A.2 The New-Keynesian Model Economy

To represent New-Keynesian abstraction we use the model economy described in Chapter 3 of Galí (2008). If money were absent, both the cash-in-advance model economy described above and the New-Keynesian model economy described below would simplify to similar versions of the standard real business cycle model economy.

The main purpose of New-Keynesian model economies is to analyze monetary policy. These model economies use sticky prices, which they justify with a mixture of theoretical justifications like rational inattention with empirical evidence that prices change infrequently. Sticky prices allow money to have short-run effects, while remaining long-run neutral. The New-Keynesian approach is perhaps more interested in modelling the effects of monetary policy on the economy, than in the modelling of money itself.

In the subsections below we discuss a version of the text-book description of the basic New Keynesian model economy which we have taken from Chapter 3 of Galí (2008). Even though this model economy can be characterized fully by a system of equations obtained by log-linearization about the steady-state of an explicit model economy, we provide the details of the full model economy to highlight its similarities with the cash-in-advance economy that we have just described.

This model economy has a representative household and it assumes that prices are sticky and that they change according to a Calvo rule, (Calvo, 1983). From these micro-foundations we derive the New Keynesian Phillips curve and the dynamic Investment-Savings equation. To close the model we add a process on nominal interest rates and a money demand function that define the monetary policy rule and the relationship between the money supply and the interest rate.

A.2.1 Households

The model has a representative household who chooses consumption, labor, and savings so as to maximize an expected discounted utility function

$$\max E \sum_{t=0}^{\infty} \beta^t \left(\frac{C_t^{1-\sigma}}{1-\sigma} - \gamma \frac{N_t^{1+\varphi}}{1+\varphi} \right) \quad (31)$$

where $0 < \beta < 1$ is the discount factor, C_t is consumption, and N_t is labor. Note that this is identical to the utility function in expression (10) for the cash-in-advance model.

However, in this model economy we assume that the household consumes a continuum of goods indexed by $i \in [0, 1]$. These goods are transformed into a composite good according to the following equation

$$C_t = \left[\int_0^1 C_t(i)^{\frac{\epsilon-1}{\epsilon}} di \right]^{\frac{\epsilon}{\epsilon-1}} \quad (32)$$

In this model economy the maximization of expected discounted utility is subject to the following series of budget constraints,

$$\int_0^1 P_t(i) C_t(i) di + I_t B_t \leq B_{t-1} + P_t W_t N_t + T_t \quad (33)$$

where B_t are purchases of nominal one-period bonds which have gross rate of return I_t , W_t is the wage, T_t is a lump-sum component of income, which may include dividends from firm ownership, and P_t is the aggregate price level which is given by

$$P_t = \left[\int_0^1 P_t(i)^{1-\epsilon} di \right]^{\frac{1}{1-\epsilon}} \quad (34)$$

The representative household demands money according to a money demand function that depends on the nominal interest rates. However it is more convenient to write this demand function in logs and we provide it in expression (38) below.

A.2.2 Firms

Each differentiated consumption good is produced by a different firm. All firms have the same production technology given by $Y_t(i) = A_t N_t(i)^{1-\alpha}$, where $Y_t(i)$ is the production of firm i , A_t is a

common technology level, and $N_t(i)$ is the labour used by firm i . The firms set prices a la Calvo, that is, each period firms are allowed to change prices only with probability $1 - \theta$. Firms set prices to maximize their expected discounted future profits for the period in which that price is in place. Thus, problem for firm setting price in period t is

$$\max_{P_t^*} \sum_{k=0}^{\infty} \theta^k E_t \{ I_{t,t+k} (P_t^* Y_{t+k|t} - \Psi_{t+k}(Y_{t+k|t})) \} \quad (35)$$

subject to a demand function

$$Y_{t+k|t} = \left(\frac{P_t^*}{P_{t+k}} \right)^{-\epsilon} C_{t+k} \quad (36)$$

which comes from the first-order conditions of the representative agents problem. Where $Y_t = \left(\int_0^1 Y_t(i)^{\frac{\epsilon-1}{\epsilon}} di \right)^{\frac{\epsilon}{\epsilon-1}}$ is production of the final (composite) good, P_t^* is the price being set, $\Psi_{t+k}(\cdot)$ is the cost function, $Y_{t+k|t}$ is the production at time $t+k$ of a firm that last changed price in period t , $I_{t,t+k}$ is the stochastic discount factor for nominal payoffs, and $P_t = \left[\int_0^1 P_t(i)^{1-\epsilon} di \right]^{\frac{1}{1-\epsilon}}$ is the aggregate price level.

The technology process, A_t , follows an AR(1) process in logs, a_t ,

$$a_t = \rho_a a_{t-1} + \varsigma_t \quad (37)$$

where $\rho_a \in [0, 1)$, ς_t is iid $\mathcal{N}(0, \sigma_\varsigma^2)$.

A.2.3 Money

The money demand function in logs is

$$m_t - p_t = y_t - \eta^i t \quad (38)$$

where m_t is (log) money, and p_t are (log) prices.

Monetary policy, in keeping with all of the models covered in this paper is given by an exogenous AR(1) process,

$$\nu_t = (1 - \rho_m) \bar{\nu} + \rho_m \nu_{t-1} + \xi_t \quad (39)$$

where $\nu_t \equiv \Delta m_t$, $\rho_m \in [0, 1)$, ξ_t is white noise. Note that the process on money in the New Keynesian model (equation (39)) is exactly the same one as was used the Cash-in-Advance model (equations (17) & (18)), just that here we write the process with m_t in logs.

A.2.4 Prices and Market Clearance

The evolution of the aggregate consumer price level is given by

$$P_t = \left[\theta P_{t-1}^{1-\epsilon} + (1 - \theta) (P_t^*)^{1-\epsilon} \right]^{\frac{1}{1-\epsilon}} \quad (40)$$

Thus consumer price inflation is

$$\Pi_t^{1-\epsilon} = \theta + (1 - \theta) \left(\frac{P_t^*}{P_{t-1}} \right)^{1-\epsilon} \quad (41)$$

where $\Pi_t = P_t/P_{t-1}$ is the consumer price inflation rate.

The remaining component of the model is the requirement for market clearing. The market clearing conditions are given by, $\forall t$: that the markets for each consumption good clear, $C_t(i) = Y_t(i)$, $\forall i \in [0, 1]$, and that the labour market clears, $N_t = \int_0^1 N_t(i) di$.

A.2.5 Equilibrium

The system of equations that constitute the reduced form of the basic New Keynesian model are now given²⁰. They are derived from the microfoundations listed previously. The difference in notation, with lowercase letters replacing the uppercase letters, is that all of the variables listed here are now in log-linear form rather than the levels represented by the uppercase letters, eg. y_t is log-deviation of output while Y_t is output. The New Keynesian Phillips curve is given by

$$\pi_t = \beta E_t\{\pi_{t+1}\} + \kappa \tilde{y}_t \quad (42)$$

where $\kappa \equiv (\sigma + \frac{\varphi+\alpha}{1-\alpha}) \frac{(1-\theta)(1-\beta\theta)}{\theta} \Theta$, and $\Theta \equiv \frac{1-\alpha}{1-\alpha+\alpha\epsilon} \leq 1$; π_t is the inflation rate, and $\tilde{y}_t = y_t - y_t^n$ is the output gap, that is the difference between current output, y_t , and the natural level of output y_t^n which would occur if prices were flexible. The dynamic IS equation is

$$\tilde{y}_t = -\frac{1}{\sigma}(i_t - E_t\{\pi_{t+1}\} - r_t^n) + E_t\{\tilde{y}_{t+1}\} \quad (43)$$

where i_t is the nominal interest rate, r_t^n is the natural interest rate (again, that which would result if prices were flexible). Both of these two equations are derived from the models micro-foundations.²¹

Letting $l_t = m_t - p_t$ be real money holdings and rewriting the money market equilibrium condition as $\tilde{y}_t - \eta i_t = l_t - y_t^n$, we can substitute out for i_t and get the following system of equations from the three above,

$$\pi_t = \beta E_t\{\pi_{t+1}\} + \kappa \tilde{y}_t \quad (44)$$

$$(1 + \sigma\eta)\tilde{y}_t = \sigma\eta E_t\{\tilde{y}_{t+1}\} + l_t + \eta E_t\{\pi_{t+1}\} + \eta \hat{r}_t^n - y_t^n \quad (45)$$

$$l_{t-1} = l_t + \pi_t - \Delta m_t \quad (46)$$

where \hat{r}_t^n is the deviation from steady-state of the natural rate of interest.

The two other formulae necessary to complete the model are those for the natural level of output and the natural rate of interest expressed in terms of deviation from steady-state, both of which depend on the technology level.

$$y_t^n = \phi_{ya}^n a_t + \vartheta_y^n \quad (47)$$

$$\hat{r}_t^n = -\sigma \phi_{ya}^n (1 - \rho_a) a_t \quad (48)$$

where $\vartheta_y^n = -\frac{(1-\alpha)(\mu - \log(1-\alpha))}{\sigma(1-\alpha) + \varphi + \alpha} > 0$ and $\psi_{ya}^n = \frac{1+\varphi}{\sigma(1-\alpha) + \varphi + \alpha}$. The model is thus the system of equations given by (44)-(48) together with the processes on the changes in the money supply (39) and technology shocks (37).

²⁰This system of equations already incorporates the parametrization of $\gamma = 1$, to which the models are later calibrated.

²¹See Galí (2008) for a step-by-step derivation.

A.2.6 The Quantity Theory of Money in a Single Equation

We now describe with a single equation the Quantity Theory of Money in way way that makes it easier to see how the New-Keynesian framework temporarily escape from the Quantity Theory of Money. Specifically, we give an expression for the term PY/M . Since the New-Keynesian model is log-linearized we will look at the log of this term, namely $p + y - m$. Were the Quantity Theory of Money to hold exactly this term would be equal to a constant.

First observe that simply rewriting the money demand equation, (38), we get

$$p_t + y_t - m_t = \eta i_t \tag{49}$$

where i_t is the nominal interest rate and η is the elasticity of money demand. Combining the New-Keynesian Phillips Curve, equation (44), with the dynamic IS, equation (43), we get that

$$i_t = r_t^n - \frac{\sigma}{\kappa} \pi_t + \left(1 + \frac{\beta}{\kappa} - \frac{\sigma}{\kappa}\right) E_t\{\pi_{t+1}\} + \frac{\beta}{\kappa} E_t\{\pi_{t+2}\} \tag{50}$$

So the nominal interest rate depends on the natural real rate of interest r_t^n (which depends on the current technology shock) and current and future expected inflation. Thus we have that

$$p_t + y_t - m_t = \eta r_t^n + \eta E_t\{f(\pi_t, \pi_{t+1}, \pi_{t+2})\} \tag{51}$$

Now, r_t^n is independent of monetary factors and the parameters relating to sticky prices. So for sticky prices to be part of the story they must be operating through inflation. From equation, (41), we have inflation evolves as

$$\pi_t = (1 - \theta)(p_t^* - p_{t-1}) \tag{52}$$

where p_t^* is the price level being chosen by those firms that get to reset their prices. So in principle, sticky prices may affect the rate of inflation, and thus the nominal interest rates — helping to break away from the Quantity Theory of Money. In practice however the effect quantitatively negligible.

A.3 The Search-Money Model Economy

The aim of Search-Money models is to provide structural reasons that justify the existence of money. This abstraction focuses on money as a facilitator of exchange based on the idea that money exists mainly to solve problems related to the presence of single-coincidence of wants. Search-Money models go a step deeper than the other two abstractions that we consider here, in which money exists simply because the modeler assumes that it does, rather than to solve an explicit problem; like the absence of a double coincidence of wants in exchange. For this reason the model is the only one of the three we consider that satisfies Wallace’s Dictum for monetary economics, that “Money should not be a primitive in monetary theory — in the same way that firm should not be a primitive in industrial organization theory or bond a primitive in finance theory” (Wallace, 1998).

Search-Money models have become more popular in recent years as they have begun to overcome some teething problems that plagued them in their earlier days: for instance in Kiyotaki and Wright (1989) money holdings were restricted to being 0 or 1 units per agent. Lagos and Wright (2005) overcame these issues by introducing the concept of a centralized (Arrow-Debreu) night-market alongside the decentralized (Kiyotaki-Wright) day-market. The use of the night-market remains integral to the latest generation of Search-Money models such as Head, Liu, Menzio, and Wright

(2012) and Berentson, Menzio, and Wright (2011). To represent the Search-Money abstraction we use a stochastic extension of the model economy described in Aruoba, Waller, and Wright (2011) — the stochastic extension is necessary to allow us to use the same process on money growth as in the other models²² The model of Aruoba, Waller, and Wright (2011) uses the same combination of decentralized day-market and competitive night-market as Lagos and Wright (2005) and incorporates physical capital.

In this model economy there are continuum of agents, a decentralized day-market, and a centralized night-market. Money is essential in the day-market because meetings are anonymous, and credit is precluded in a fraction of these meetings because there is no possibility of credibly promising to repay at a later date. As a result exchange must be quid pro quo and so without money some trades would never take place — namely, those in which there was no double-coincidence of wants. Capital investments are made during the competitive night-market, and capital is used in production during both markets²³. The model of Aruoba, Waller, and Wright (2011) includes a government sector, we eliminate this, which requires some recalibration of the model^{24,25}.

A.3.1 Households

There is a continuum of households indexed by i who live forever and whose measure we normalize to 1. Time is discrete and households discount the future at rate $\beta \in (0, 1)$. Each period is divided into two subperiods which are commonly referred to as “day” and “night”. Households consume and supply labour in both subperiods, and their preferences over sequences of consumption and labor are ordered according to the following period utility function

$$\mathcal{U}(c, n, C, N) = u(c) - h(n) + U(C) - N \quad (53)$$

where c and C denote consumption and n and N denote labour in the day and night subperiods. Assume that u , h , and U are twice continuously differentiable with $u' > 0, h' > 0, U' > 0, u'' < 0, h'' \geq 0$ and $U'' \leq 0$. Also, $u(0) = c(0) = 0$, and suppose that there exists $q^* \in (0, \infty)$ such that $u'(q^*) = h'(q^*)$ and $C^* \in (0, \infty)$ such that $U'(C^*) = 1$ with $U(C^*) > C^*$.

Aruoba et al. (2011) propose to use the following functional form to take the model to the data

$$U(c, n, C, N) = \left\{ \left[(c + \chi)^{(1-\sigma)} - \chi^{(1-\sigma)} \right] / (1 - \sigma) - \gamma n \right\} + \{ \Xi \log(C) - N \} \quad (54)$$

With the exception of the inclusion of parameter χ , $u(c) = [(c + \chi)^{(1-\sigma)} - \chi^{(1-\sigma)}] / (1 - \sigma)$ is the same constant elasticity of substitution utility of consumption as the ones we have used in the other two models economies; the utility of consumption is $U(C) = \Xi \log(C)$ and the disutility of labor

²²An earlier version of this paper used the model of Lagos and Wright (2005). This model failed to break away from the Quantity Theory of Money in the short-run, performing much worse than the other models presented here.

²³The appearance of capital in both markets is important. Earlier work by Aruoba and Wright (2003) to introduce capital, with capital appearing in only one market, led to the results that the day and night markets could be solved for separately, and thus money had no effect on consumption, investment, or anything else in the competitive night-market.

²⁴Our results are robust to leaving the government sector in the model of Aruoba, Waller, and Wright (2011).

²⁵Aruoba, Waller, and Wright (2011) actually present three models. Here we follow their model 2. Their model 1 is the same model, but with a slightly different calibration. Their model 3 uses ‘competitive search’, setting prices in the decentralized day market by price taking, rather than Nash bargaining. For robustness we tried out using their model 3 and it makes no real difference to the results we found using their model 2.

is $h(n) = \gamma n$ in the day market. The assumption that utility is quasi-linear in labour is used by Aruoba et al. (2011) and is necessary to keep the model analytically tractable²⁶.

A.3.2 Production and Trade

The day-good, c , comes in many differentiated varieties indexed by i . Each household consumes only a subset of these goods. Each household can transform its own labour into one of these goods that the household itself does not consume by the production function $F(K_i, N_i)$, namely household i produces good i which it does not consume. The production function is given by a standard Cobb-Douglas formulation $F(K, N) = K^\alpha N^{1-\alpha}$. Trade during the day is decentralized and anonymous and households are matched randomly in a typical search setup.

For two households i and j drawn randomly, there are three possible trading situations. The probability that one consumes what the other produces, but not vice-versa —and, therefore, there is a single coincidence of wants is ω , and we assume that it is symmetric. Then, the probability that neither one of them consumes what the other one produces is $1 - 2\omega$. In a single-coincidence meeting, if i wants the good that j produces we call i the buyer and j the seller. In a fraction ϖ of single-coincidence meetings the buyer can only pay with money, in the remaining fraction, $1 - \varpi$, the buyer has access to credit, l . By assumption capital can not be used for transaction purposes.

The night good, C , comes in a single and homogeneous variety, which is consumed by every household. Each household can transform its own labour into income at the market wage. Trade, during, the night occurs in a centralized Walrasian market. Consequently, the night-good can be purchased on credit. Since money is a good, it can be traded in the night market just like any other good. Investments in capital are also made during the night market.

All the differentiated day-goods and the night-good are perfectly divisible and non-storable, with the exceptions of money and capital which are storable.

A.3.3 Money

In this model economy there is an object called *money* that is perfectly divisible and storable in any non-negative quantity. The total money stock at time t is M_t , and it evolves according to

$$M_{t+1} = e^{\nu_{t+1}} M_t \tag{55}$$

The monetary injections, $(e^{\nu_{t+1}} - 1)M_t$, are made after the night market closes and they are distributed lump-sum and equally to every household. The rate of growth of money, ν , follows an AR(1) process given by

$$\nu_t = (1 - \rho_m)\bar{\nu} + \rho_m\nu_{t-1} + \xi_t \tag{56}$$

where $0 < \rho_m < 1$ and where ξ is an identical and independently distributed process with zero mean and variance σ_ξ^2 . Although for the equilibrium proofs below we only need to assume that the

²⁶ Quasi linearity means that there are no wealth effects in the demand for money, so all agents in the centralized night markets choose the same money holdings. As a robustness test we simulated both the New-Keynesian and Cash-in-Advance models setting the parameters so that utility was quasi-linear in labour (ie. $\varphi = 0$, $\gamma = 1$; note that $\gamma = 1$ is the value to which this parameter is calibrated in those models anyway.). The effect on the results was negligible.

rate of growth of money follows a first-order Markov process. So the process on money, as given by equations (55) and (56), is identical to that used in the New-Keynesian and Cash-in-Advance models.

A.3.4 Prices and Market Clearance

Let $1/p_t$ be the price of money in the centralized night-market, that is, p_t is the nominal price of night good C .

In the deterministic version of the model economy described in Aruoba, Waller, and Wright (2011), the only uncertainty comes from the random matching. In the stochastic extension that we use here, the rate of growth of the money supply, ν_t , is also uncertain. Consequently, in our model economy the decisions of each household at each point in time depend on its current money holdings, m , on its capital holdings k , during the centralized night market on its earlier borrowing during the day l , and on the aggregate state which is the rate of growth of money, ν . Therefore, the households' choices at time t can be characterized with a value function that has m , k , and ν as its arguments; as well as l in the centralized night-market.

Let $V(m, k, \nu)$ be the value function for a household when it enters the decentralized day-market, and $W(m, k, l, \nu)$ its value function when it enters the centralized night-market. Since trade is bilateral in the day-market and the day-good is non-storable, the seller's production, n , must be equal to the buyer's consumption, c .

Let m be money holdings. The the value of trading at the day-market is

$$V_t(m, k, \nu) = \omega V_t^b(m, k, \nu) + \omega V_t^s(m, k, \nu) + (1 - 2\omega)W_t(m, k, 0, \nu) \quad (57)$$

where $V_t^b(m, k, \nu)$ and $V_t^s(m, k, \nu)$ denote the values to being a buyer and being a seller, as given by

$$\begin{aligned} V_t^b(m, k, \nu) &= \varpi[u(q_b) + W_t(m - d_b, k, 0, \nu)] + (1 - \varpi)[u(\tilde{q}_b) + W_t(m, k, l_b, \nu)] \\ V_t^s(m, k, \nu) &= \varpi[-c(q_s, k) + W_t(m + d_s, k, 0, \nu)] + (1 - \varpi)[-c(\tilde{q}_s, k) + W_t(m, k, -l_s, \nu)] \end{aligned} \quad (58)$$

In these expressions q_b and d_b (q_s and d_s) denote the quantity of goods and money exchanged when buying (selling) for money, while \tilde{q}_b and l_b (\tilde{q}_s and $-l_s$) denote the quantity and the value of the loan for the buyer (seller) when trading on credit.

At the centralized night-market agents solve the following problem

$$W_t(m, k, l, \nu) = \max_{C, N, m', k'} [U(C) - N + \beta E_t\{V_{t+1}(m' + (e^{\nu'} - 1)M, \nu')\} | \nu] \quad (60)$$

subject to $C = wN + (1 + r - \delta)k - k' + \frac{m - m' - l}{p}$, $C \geq 0$, $0 \leq N \leq \bar{N}$, and $m' \geq 0$, where \bar{N} is the endowment of night-hours, w is the wage, and r is the interest rate on capital.²⁷

It is assumed that the markets for capital and labour in the night-market are competitive, thus $w = F_N(K, N)$ and $r = F_K(K, N)$.

²⁷Notice that $(e^{\nu'} - 1)M$ is the transfer of money that is added lump-sum to the households' holdings after they exit the night-market.

Now that we have defined the value functions, we consider the terms of trade in the decentralized day-market. In single-coincidence meetings, we use the generalized Nash solution in which the buyer has bargaining power $\zeta > 0$ and threat points which are given by the continuation values. In the fraction ϖ of meetings where money is used (q, d) is the consumption for money exchange pair that maximizes the following problem

$$(q, d) = \underset{q, d}{\operatorname{argmax}} \{ [u(q) + W_t(m_b - d, k_b, 0, \nu) - W_t(m_b, k_b, 0, \nu)]^\zeta [-c(q, k_s) + W_t(m_s + d, k_s, 0, \nu) - W_t(m_s, k_s, 0, \nu)]^{1-\zeta} \} \quad (61)$$

subject to $d \leq m_b$ and $q \geq 0$. In the remaining fraction, $1 - \varpi$, of meetings where credit is available, (\tilde{q}, l) is determined just like (q, d) , except that the Nash bargaining problem is no longer any constraint on l , the way $d \leq m_b$ had to hold in monetary trades.

As Aruoba, Waller, and Wright (2011) observe, the solution to the bargaining problem in 61 will involve $d = m_b$. Substituting this into the bargaining problem and taking the first order condition with respect to q we have

$$\frac{m_b}{p} = \frac{z(q, k_s)w}{\gamma} \quad (62)$$

where

$$z(q, k) \equiv \frac{\zeta c(q, k)u'(q) + (1 - \zeta)u(q)c_q(q, k)}{\zeta u'(q) + (1 - \zeta)c_q(q, k)} \quad (63)$$

reflects the terms of trade in the bargaining meetings.

Real output, $Y = Y_D + Y_N$, is the combination of real output in the decentralized day market, $Y_D = \omega\varpi M/p + \omega\varpi\omega l/p$, and real output in the centralized night market $F(K, N)$.

Following Aruoba, Waller, and Wright (2011) we measure inflation in terms of the price level in the centralized market p_t ²⁸.

A.3.5 Equilibrium

The system of equations that defines an equilibrium is now given. To make the model stationary we define $\hat{m}_t = m_t/M_t$ and $\hat{p}_t = p_t/M_t$; observe that in equilibrium it follows that $\hat{m}_t = 1$ for all t . The derivation of this system of equations follows almost exactly as described in Aruoba, Waller, and Wright (2011). The first three equations are related to the first-order conditions of the household.

$$z(q_t, K_t) = \beta E \left\{ \frac{z(q_{t+1}, K_{t+1})}{\exp(\nu_{t+1})} \left(1 - \omega\zeta + \omega\zeta \frac{u'(q_{t+1})}{z(q_{t+1}, K_{t+1})} \right) \right\} \quad (64)$$

$$U'(C_t) = \beta E \{ U'(C_{t+1}) [1 + F_K(K_{t+1}, N_{t+1}) - \delta] - \omega [\varpi \Gamma(q_{t+1}, K_{t+1}) + (1 - \varpi)(1 - \zeta)c_k(\hat{q}_{t+1}, K_{t+1})] \} \quad (65)$$

$$U'(C_t) = \frac{1}{F_N(K_t, N_t)} \quad (66)$$

²⁸We also tried using a Laspeyres measure of inflation that included prices in the decentralized markets. But since in the calibrated model the decentralized market accounts for only about 3% of total real output this made no noticeable difference.

The fourth equation is aggregate resource constraint

$$C_t = F(K_t, N_t) + (1 - \delta)K_t - K_{t+1} \quad (67)$$

The next two equations determine the price level in the competitive night market, and the real value of the credit loans made in the decentralized day market (in the fraction ϖ of meetings where credit is available)²⁹.

$$\hat{p}_t = \frac{\gamma}{z(q_t, K_t)F_N(K_t, N_t)} \quad (68)$$

$$l_t/p_t = F_N(K_t, N_t)[(1 - \zeta)u(\tilde{q}) + \zeta c(\tilde{q}, K)] \quad (69)$$

The next four equations are related to the terms of trade in the decentralized day market ($z(q, K)$, as defined in (63)), and some related derivatives and quantities.

$$z(q_t, K_t) = \frac{\zeta c(q_t, K_t)u'(q_t) + (1 - \zeta)u(q_t)c_q(q_t, K_t)}{\zeta u'(q_t) + (1 - \zeta)c_q(q_t, K_t)} \quad (70)$$

$$z_q(q_t, K_t) = \frac{u'(q)c_q[\zeta u'(q) + (1 - \zeta)c_q] + \zeta(1 - \zeta)(u(q_t) - c)(u'(q_t)c_{qq} - c_q u''(q_t))}{[\zeta u'(q_t) + (1 - \zeta)c_q]^2} \quad (71)$$

$$z_K(q_t, K_t) = \frac{\zeta u'(q_t)c_k[\zeta u'(q_t) + (1 - \zeta)c_q] + \zeta(1 - \zeta)(u(q_t) - c)u'(q_t)c_{qK}}{[\zeta u'(q_t) + (1 - \zeta)c_q]^2} \quad (72)$$

$$\Gamma(q_t, K_t) = c_K(q_t, K_t) - c_q(q_t, K_t) \frac{z_K(q_t, K_t)}{z_q(q_t, K_t)} \quad (73)$$

where c is shorthand for $c(q_t, K_t)$, c_q for $c_q(q_t, K_t)$, c_K for $c_K(q_t, K_t)$, c_{qq} for $c_{qq}(q_t, K_t)$, and c_{qK} for $c_{qK}(q_t, K_t)$. The next equation is simply the definition of real output,

$$Y = F(K, N) + \omega\varpi M/p + \omega\varpi\omega l/p \quad (74)$$

The final equation is that defining the money growth rate,

$$\nu_t = (1 - \rho_m)\bar{\nu} + \rho_m\nu_{t-1} + \xi_t \quad (75)$$

The Search-Money model with capital is thus given by the system of stochastic difference equations, (64)-(75).

A.3.6 The Quantity Theory of Money in a Single Equation

We now describe with a single equation the Quantity Theory of Money in way way that makes it easier to see how the Search-Money framework temporarily escapes from the Quantity Theory of Money. Specifically, we give an expression for the term PY/M . Were the Quantity Theory of Money to hold exactly this term would be equal to a constant.

²⁹While neither l_t nor p_t are stationary, by treating l_t/p_t as a single variable the equation is stationary.

In the Search-Money model, by equation (68), we have that

$$\frac{PY}{M} = \frac{1}{z(q, K)} \frac{\gamma Y}{F_N(K, N)} \quad (76)$$

In the simulation results total output (Y), capital stock (K), and the marginal product of labour ($F_N(K, N)$) are almost constant, and thus not related to the ability of the Search-Money model to get away from the Quantity Theory of Money. They are almost constant because most of the economy is based on non-monetary trades, the centralized night market is much bigger than the decentralized day market, and so unaffected by changes in the money supply. All of the movement occurs in the $z(q, K)$ term, specifically from changes in q — the amount produced/traded in the exchanges involving money in the decentralized market. The amount produced in monetary exchanges varies with the amount of money and inflation (the cost of holding money).

A.4 Calibration and Computation

For our comparisons of the three model economies to be meaningful, we choose their functional forms and parameters so that they are as similar as possible. This use of identical parameter values wherever the models coincide, of identical exogenous processes, and of identical functional forms for the utility of consumption, as well as the fact that we have solved the three model economies using identical solution methods allows us to make a genuine comparison between them. Since we have removed all other possible sources of variation, we can safely attribute any differences in their outputs with respect to the Quantity Theory of Money relationship to the different ways in which these three frameworks model money.

Parameter Choices

We have decided to use Galí (2008) as our main reference for our parameter choices, with the obvious exceptions of the parameters and functions of the Cash-in-Advance and Search-Money model economies that do not exist in the New-Keynesian framework, such as the parameters related to the search for trading partners in the Search-Money model economy.³⁰

Since Galí (2008) exploits the certainty equivalence principle in his solution method, he does not define the shocks to either the technology or the money supply. Instead, we take the processes for those shocks from Cooley and Hansen (1989). We report our chosen parameter values in Table 4.³¹ Our results are robust to using the original parameter calibrations of each model, that is those parameter values given in the papers from which the models are taken. Importantly, the original parameterizations of the models (in the papers from which they are taken) are all calibrated to similar postwar periods. Since the frameworks are quite different using exactly the same calibration targets for the different frameworks is not possible, although some common calibration targets, such as interest rates and capital-output ratios were used by a number of the the original papers.

Simulation

³⁰The calibrations reported in Aruoba, Waller, and Wright (2011) are annual and so had to be adjusted. This was done using the same methodology and targets they report — some targets, such as the capital-output ratio, have to be adjusted to quarterly values.

³¹Galí (2008) pg. 52 says that $\epsilon_p = 6$, however this appears to be a typo. When we use this value, we fail to replicate his results. Therefore, we use $\epsilon_p = 6/5$ instead following <http://www.dynare.org/phpBB3/tviewtopic.php?f=1andt=2978>. In this case we replicated Galí's results successfully.

Table 4: Parameter Values

		Cash-in-Advance	New Keynesian ^a	Search-Money
<i>Preferences</i>				
Time Discount factor	β	0.99	0.99	0.99
Curvature of Consumption	σ	1	1	1
Weight on Labour ^b	γ	1	1	2.1
Curvature of Labour ^c	φ	1	1	0
<i>Technology</i>				
Returns to Capital ^{d,e}	α	0.33	n.a.	0.33
Depreciation Rate ^e	δ	0.025	n.a.	0.025
Autocorrelation	ρ_a	0.9	0.9	n.a.
Variance of Shock	σ_ζ	0.007	0.007	n.a.
<i>Money</i>				
Elasticity of Money Demand	η	n.a.	4	n.a.
Autocorrelation	ρ_m	0.5	0.5	0.5
Variance of Shock	σ_ξ	0.009	0.009	0.009
Constant Term	$\bar{\nu}$	0.014	0.014	0.014
Dist. of Shock	ξ	log-normal	log-normal	log-normal
<i>Price Setting</i>				
Market Power	ϵ	n.a.	6/5	n.a.
Calvo Stickiness	θ	n.a.	0.66	n.a.
<i>Search</i>				
Prob. of Single Coincidence ^g	ω	n.a.	n.a.	0.08
Bargaining Power	ζ	n.a.	n.a.	0.92
Night weight on consumption	Ξ	n.a.	n.a.	0.8
Make $u(0) = 0$	χ	n.a.	n.a.	0.001
Prob. of credit availability	ϖ	n.a.	n.a.	0.85

^aEvery other parameter that appears in the equations that characterize the equilibrium of the New-Keynesian model economy can be derived from the parameters that we have identified in this table using the following system of equations: $\mathcal{M} = \epsilon/(1 - \epsilon)$; $\mu = \log \mathcal{M}$ $\rho = -\log \beta$; $\Theta = (1 - \alpha)/(1 - \alpha + \alpha\epsilon)$; $\lambda = \Theta(1 - \theta)(1 - \beta\theta)/\theta$; $\kappa = \lambda[\sigma + (\varphi + \alpha)/(1 - \alpha)]$; $\vartheta_y^n = \{(1 - \alpha)[\mu - \log(1 - \alpha)]\}/[\sigma(1 - \alpha) + \varphi + \alpha]$; $\psi_{ya}^n = (1 + \varphi)/[\sigma(1 - \alpha) + \varphi + \alpha]$.

^bIn the Search-Money model this parameter is calibrated 2.1, as this is needed as part of the the calibration procedure of Aruoba, Waller, and Wright (2011) (setting this parameter to one in the Search-Money with capital model, while messing up the calibration, does not affect the results).

^cIn the Search-Money model the disutility of labor is linear in both the day-market and in the night-market.

^dAbbreviation “n.a.” means “not applicable”.

^eThere is no ‘returns to capital’ or ‘depreciation’ in the New-Keynesian economy as there is no capital.

To simulate our model economies we have used identical seeds for the random number generator so that the sequences of the realizations of the random shocks are identical in all three model economies. To obtain the model economy time series we discard the first 200 periods of each equilibrium realization to purge away the initial conditions, and then we draw a sample of 204 quarterly observations to replicate the number of observations in our United States time series. Whenever we need to obtain multiple samples, we repeat this process as necessary.

Computation

The equilibria of the three models economies that we have described above can be reduced to systems of stochastic equations. We have solved these systems using the default perturbation methods of Dynare to calculate quadratic approximations to the decision rules.³²

³²We have run every code with Dynare Version 4.2.1-2 using Octave 3.2.4.

B Is the Quantity Theory of Money a Stable Relationship?

Using Lucas' Illustrations, Sargent and Surico (2011) argue that the relationship between the price level, real GDP, and the money supply (measured as M2) given by the Quantity Theory of Money varies with the monetary regime. In this appendix we look at this issue using the interpretation of the long-run Quantity Theory of Money as being a cointegration relationship. Under this interpretation, a change in the Quantity Theory of Money would involve a change in the cointegrating vector that defines the long-run relationship. Using a test for a structural break in the cointegrating vector at an unknown time period developed by Seo (1998) we reject that such a break has occurred in our 1960–2009 period. The Quantity Theory of Money appears to be a stable relationship.

As we described in Section 2, Sargent and Surico (2011) look at the Quantity Theory of Money during the period 1900–2005 using the Lucas Illustrations. They divide the period into four subperiods, coinciding with different monetary policy regimes. They find that the long-run slopes of the Lucas Illustrations differ in these four regimes, using M2 as the monetary aggregate. Of particular relevance to us they identify 1984, a date in the middle of our sample, as corresponding to a change to an inflation targeting regime which delivers a flatter long-run slope. Their findings imply that a structural break in the cointegrating vector defining the Quantity Theory of Money occurs around 1984 — it is this implication that we aim to test here.

The Quantity Theory of Money defines a relationship between the price level, real GDP, and the money supply — all variables that are nonstationary. This relationship holds in the long-run, but short lived departures from this relationship are common. In the language of modern econometrics the Quantity Theory of Money defines a cointegration relationship. Cointegration theory thus gives us an alternative method to the Lucas Illustrations by which to look at whether the Quantity Theory of Money holds in the US economy in the long-run. However since the models we deal with are stationary it is not appropriate for our main purpose of evaluating different approaches to modelling money.

Here we look at whether the Quantity Theory of Money is a stable relationship — as opposed to the argument of Sargent and Surico (2011) that it varies with the monetary regime. The stability of the Quantity Theory of Money can be seen as a question about the stability of the cointegrating vector. If the Quantity Theory of Money relationship changes between different monetary regimes this would appear as a structural break in the cointegrating coefficients. For the Lucas Illustrations the stability of the Quantity Theory of Money means that the slope of the Lucas Illustrations is independent of the monetary regime. Thus the validity of our use of the Lucas Illustration to analyze the period 1960–2009, which covers two different monetary regimes, requires that Quantity Theory of Money be stable.

While cointegration theory has not been applied much to the Quantity Theory of Money it has often been used to evaluate the related issue of money demand equations (Lucas, 1988; Stock and Watson, 1993; Seo, 1998).³³ The Johansen test suggests that the Quantity Theory of Money represents a cointegrating relationship.

To look at the stability of the Quantity Theory of Money we use the test for a structural break at an unknown date in the cointegrating vector developed by Seo (1998). We find that there has

³³Lucas (1988) does not use any cointegration theory, but lays some theory and evidence on money demand equations which the later two papers then extend and evaluate with cointegration theory.

not been a break in the Quantity Theory of Money for M2³⁴, and we conclude that the relationship is stable. That the Quantity Theory of Money is a stable relationship means that the slope of Lucas' Illustration has remained unchanged throughout the 1960–2009 period.

The log of the Quantity Theory of Money can be expressed in Vector Error Correction Model representation as

$$\begin{pmatrix} \Delta \log Y_t \\ \Delta \log P_t \\ \Delta \log M_t \end{pmatrix} = \alpha \beta' \begin{pmatrix} \log Y_{t-1} \\ \log P_{t-1} \\ \log M_{t-1} \end{pmatrix} + \sum_{i=1}^{L-1} \Gamma_i \begin{pmatrix} \Delta \log Y_{t-1} \\ \Delta \log P_{t-1} \\ \Delta \log M_{t-1} \end{pmatrix} + u_t \quad (77)$$

where α is a 3×1 vector, β is a 3×1 vector, u_t is a 3×1 vector of independently and identically distributed shocks with mean zero and covariance matrix Σ , L is the lag-operator, and Γ_i is the matrix of i^{th} -order autocorrelation coefficients.

Vector β represents the long-run relationship between the price level, real GDP, and the money supply—the Quantity Theory of Money relationship. For the purpose of identification the first element of β is normalized to 1. Vector α captures how the system adjusts to transitory departures and returns back to the long-run relationship captured by β . We apply tests for a structural break in one or both vectors α and β .

Seo (1998)'s test is based on the Maximum Likelihood Estimate of the cointegrating vectors from the Vector Error Correction Model representation. For a structural break at a known date a simple Lagrange Multiplier test could be applied—comparing the likelihood under the null hypothesis of no break to the alternative hypothesis of a break at time t . Three different simple tests can be constructed for a break in β , in α , or in both vectors. To test for a break at an unknown date we calculate the test statistics for a break at each possible date, and the 'largest' of these individual statistics becomes itself a test statistic for a break at an unknown date.³⁵

We use three definitions to determine the value of the 'largest' statistic: the average (*Avg-LM*), the exponential average (*Exp-LM*), and the supremum (*Sup-LM*). Seo (1998) derives the asymptotic properties of each of these metrics under the assumptions of (i) no drift, (ii) no trend in the data generating process, and (iii) a trend in the data generating process and he provides critical values for all these tests.

We apply the test for a structural break in the cointegrating vectors to our quarterly data for the 1960–2009 period³⁶. We use the tests based on a trend in the data generating process, because real GDP, M2, and the price level all show clear upward trends. We use logs of all of these variables to linearize the Quantity Theory of Money relationship.

The Akaike information criterion recommends to use as many lags as possible, Hannan-Quinn suggests 13, the Bayesian information criterion suggests 3. We choose 8 because this represents 2 years of data, which keeps us in line with the literature that estimates money demand equations. In particular, Stock and Watson (1993) use cointegration to test for a relationship between real money holdings (money supply divided by the price index), real GDP, and the interest rate. Our

³⁴That is, we cannot reject the null hypothesis of no break in the Quantity Theory of Money for M2.

³⁵The unknown date is assumed to lie in the interval $[0.15, 0.85]T$, where T is the total number of periods. That is, we assume that the break does not occur at the ends of the time period.

³⁶All the data comes from FRED2 (<http://research.stlouisfed.org/fred2/>) and is described in full in Section 2. The regression results shown here are those based on GNPC96, CPIAUCNS, and M2SL. Our results are robust to using our other measures for inflation.

results are robust to the use of 12 lags.³⁷

As a robustness test, and to compare with the literature on money demand equations, we tried adding the interest rate to our cointegration regressions: for quarterly data the presence of the interest rate in the cointegrating relationship was consistently rejected³⁸; both for 8 and for 12 lags, and using the interest rate on both 3-month and 10-year Treasury Bills (Stock and Watson (1993) also use commercial paper rates, but this series was discontinued in 1997).

All the test results that we report below are based on a 5% level of significance. The Dicky-Fuller tests confirm that all the variables contain unit roots. Johansen's test for cointegration confirms the presence of a single cointegrating relationship using both 8 and 12 lags. All the estimation and test results described up to this point were found using Gretl Version 1.9.5. The structural break test itself is performed using Gauss³⁹, the following results are the estimation results.

We now summarize the estimation and testing results using real GDP, Consumer Price Index, and M2 money supply,

$$\begin{pmatrix} \Delta \log Y_t \\ \Delta \log P_t \\ \Delta \log M_t \end{pmatrix} = \begin{pmatrix} 0.003_{(0.002)} \\ -0.003_{(0.002)} \\ 0.003_{(0.073)} \end{pmatrix} \begin{pmatrix} 1 \\ 5.089_{(1.302)} \\ -4.167_{(0.903)} \end{pmatrix}' \begin{pmatrix} \log Y_{t-1} \\ \log P_{t-1} \\ \log M_{t-1} \end{pmatrix} + \dots \quad (78)$$

$$LR(H_0 : rank(\Pi) = 0) = 34.599*$$

$$LR(H_0 : rank(\Pi) = 1) = 17.349$$

$$LR(H_0 : rank(\Pi) = 2) = 5.204$$

$$Avg - LM_n^\alpha = 6.872*,$$

$$Exp - LM_n^\alpha = 4.337*,$$

$$Sup - LM_n^\alpha = 11.933$$

$$Avg - LM_n^\beta = 3.737,$$

$$Exp - LM_n^\beta = 2.399,$$

$$Sup - LM_n^\beta = 9.306$$

$$Avg - LM_n^{\alpha\beta} = 10.609*,$$

$$Exp - LM_n^{\alpha\beta} = 7.340*,$$

$$Sup - LM_n^{\alpha\beta} = 20.845*$$

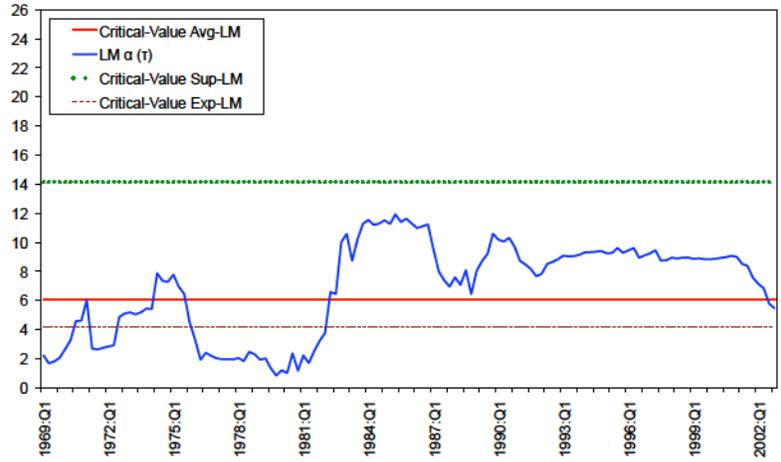
where standard errors are in parentheses and * indicates significance at the 5% level. Figure 7 shows the evolution of the LM statistic over time. Using 12 lags, we further reject the possibility of a structural break in α (or $\alpha\beta'$).

To interpret these results first recall that our main interest is in the vector, β , representing the cointegrating relationship. The tests also consider the adjustment vector, α , that represents how the system reacts to deviations from the cointegrating relationship — how the economy goes about returning to the cointegrating relationship given by the Quantity Theory of Money . For each of α and β (as well as for a joint-test of α and β) we have three test statistics ($Exp - LM$, $Avg - LM$, $Sup - LM$).

³⁷When we use 12 lags we have to restrict t to be in the interval $[0.2, 0.8]$ because otherwise the $\alpha\beta$ matrix was too close to being singular. When we placed the same restriction on t with 8 lags, the results did not change.

³⁸From a theoretical viewpoint whether the interest rate is a stationary or nonstationary variable is an open question in econometrics. So theoretically it is unclear whether including interest rates in a cointegration relationship makes sense. In any case our data rejected it's presence in the cointegrating relationship. We note in passing that Stock and Watson (1993) use annual data.

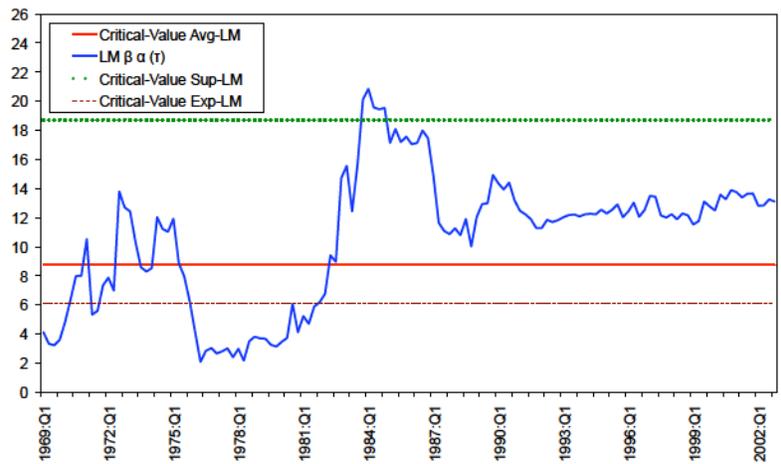
³⁹Version 10. We thank Byeongseon Seo for a copy of his code implementing the tests.



Panel A: $LM^\alpha(\tau)$



Panel B: $LM^\beta(\tau)$



Panel C: $LM^{\alpha\beta}(\tau)$

Figure 7: Values of the LM Statistics for the Structural Break Tests in the Cointegration Vectors

In the printed estimation outputs the first part reports the estimated values for the vectors α and β , with the standard errors for the coefficients in parentheses. Then come the *LR* statistics of the Johansen cointegration test, which in both cases reject the first null hypothesis of no cointegrating relationship, and then accept the second null that there is a single cointegrating relationship. The third part gives the results of the tests for a structural break in the cointegrating vectors α and β (as well as for a joint-test of α and β). In the cases where these statistics are significant, indicated by *, we reject the null of no structural break in that vector. The results suggest that there is no structural break in β , while there is one in α .

The panels of Figure 7 show the evolution of the three statistics (*Exp - LM*, *Avg - LM*, *Sup - LM*) over time. The dotted horizontal lines show the critical values for the three statistics (*Exp - LM*, *Avg - LM*, *Sup - LM*), while the value of the statistic is shown by the jagged line. In the cases where we rejected the null of no structural break a sudden increase in the statistic at time t to above the critical values (the horizontal lines) suggests that this is the point in time at which it is most likely that the structural break occurred.

The results point towards the existence of a cointegrating relationship, β , which is stable over the sample period. Therefore we conclude that change of monetary policy that took place around 1984 is *not* associated with a break in the cointegrating relationship. That is, we conclude that the Quantity Theory of Money is stable over the period 1960–2009.

There is evidence of a change in the adjustment process for returning to the long-run relationship (a break in α). Figure 7 shows that this appears to take around 1984 (when the test statistic for α shoots up above the critical values (horizontal lines)). This might be understood as the effects of the change in the monetary policy regime on the short-run effects of monetary policy, as in Clarida, Galí, and Gertler (2000). That the estimated coefficients in the adjustment vector, α , are insignificant is likely related to the evidence of a structural break in this vector. Note that this does not represent a break in the cointegration relationship represented by the Quantity Theory of Money in the long-run, just in how the economy goes about returning to the Quantity Theory of Money when current events cause the economy to be away from the Quantity Theory of Money in the short-run.

We conclude that we do not find evidence of a change in the long-run relationship embodied by the Quantity Theory of Money, and this implies that the slope of the Lucas Illustration has not changed. There is however evidence that the adjustment of the economy back to the Quantity Theory of Money changes around 1984, when the monetary policy regime changes.

C A Bayesian View of the Quantity Theory of Money (In Progress)

We Bayesian estimate the models to target the three time-series of data that make up the Quantity Theory of Money: growth of the money supply (MISL), real output growth (GNPC96), and inflation (CPIAUCSL). The models cannot simply be estimated as is: they contain two shocks (money and technology) to target three time-series, so the likelihood would be minus infinity. An extra shock is therefore needed. Following the standard modeling approach for Bayesian estimation of Monetary DSGE model, eg. Smets and Wouters (2007), we include a shock between the 'fundamental' inflation predicted by the model and the observed inflation in the data. That is, $\pi_t^{observed} = \pi_t^{fundamental} + \epsilon_t^p$. This is commonly referred to as a 'price mark-up disturbance'. In Smets and Wouters (2007), for example, this shock follows an ARMA(1,1) process. The addition of these shocks between fundamental and observed inflation is the only modification of the models relative to their description in Appendix A. We model these as $\epsilon_t^p = \bar{\epsilon}^p + \tilde{\epsilon}_t^p$, where $\tilde{\epsilon}_t^p$ is i.i.d with distribution $N(0, \sigma_{\epsilon^p})$.

Our models now have three shock processes — money supply, output, and prices — to explain the three time-series — growth of the money supply, real output growth, and inflation. We are ready to proceed with the Bayesian estimation. The first step in Bayesian estimation of the models is to define the prior distributions of the parameters. The prior distributions are given in Table 5, and are chosen to be centered around the baseline calibrated parameter values used in the body of the paper.

Table 5: Prior and Posterior Distributions for the Cash-in-Advance Model

	Prior distribution			Posterior distribution			
	Dist.	Param1	Param2	Mean	Std. Dev.	5%	95%
γ	Normal	1	0.5	1.46	0.50	0.69	2.22
ρ_a	Uniform	0.5	0.99	0.59	0.14	0.50	0.71
ρ_m	Uniform	0.3	0.99	0.31	0.20	0.30	0.32
$\bar{\nu}$	Uniform	0	0.05	0.22	0.01	0.01	0.03
\bar{a}	Uniform	0	0.1	0.02	0.03	0.00	0.03
$\bar{\epsilon}^p$	Uniform	-3	3	-1.02	1.73	-2.30	0.93
σ_ϵ	Uniform	0	0.05	0.01	0.01	0.01	0.02
σ_ξ	Uniform	0	0.05	0.05	0.01	0.05	0.05
σ_{ϵ^p}	Uniform	0	3	2.99	0.87	2.98	3.00

For Uniform distributions, Param1 and Param2 are the maximum and minimum values, $U(Param1, Param2)$. For Normally distributed variables they are the mean and standard deviation, $N(Param1, Param2)$.

We can now see the same problem of the Cash-in-Advance framework displaying way to much of the Quantity Theory of Money that we have seen with the Lucas Illustrations. From a Bayesian viewpoint it must be possible for the model to reproduce the variance in short-run prices seen in the data, otherwise the likelihood of the model would be minus infinity. However, for the Bayesian estimate of the Cash-in-Advance framework to capture the short-run movements in inflations — to capture the failure of the Quantity Theory of Money in the short-run — it requires very large and poorly identified transitory shocks to inflation. The model itself is simply unable to explain the short-run failure of the Quantity Theory of Money, and so it requires large transitory and

unexplained shocks to inflation to be able to replicate the empirical fact that the Quantity Theory of Money fails in the short-run.

D The Quantity Theory of Money (In Progress)

We start with a simple accounting identity,

$$\text{Money supply } \times \text{ Velocity} = \text{Price } \times \text{ Real Output}$$

The Quantity Theory of Money is that Velocity is constant. Thus, the Quantity Theory of Money tells us that the price level is fully determined once we know output and the money supply. The Quantity Theory of Money first emerges in modern form in the work of David Hume (Of Money, 1753; Of Interest, 1753).⁴⁰ Hume characterizes it as,

Were all the gold in England annihilated at once, and one and twenty shillings substituted in the place of every guinea, would money be more plentiful or interest lower? No surely: We should only use silver instead of gold. Were gold rendered as common as silver, and silver as common as copper; would money be more plentiful or interest lower? We may assuredly give the same answer. Our shillings would then be yellow, and our halfpence white; and we should have no guineas. No other difference would ever be observed; no alteration on commerce, manufactures, navigation, or interest; unless we imagine, that the colour of the metal is of any consequence.

Now, what is so visible in these greater variations of scarcity or abundance in the precious metals, must hold in all inferior changes. If the multiplying of gold and silver fifteen times makes no difference, much less can the doubling or tripling them. All augmentation has no other effect than to heighten the price of labour and commodities; and even this variation is little more than that of a name. In the progress towards these changes, the augmentation may have some influence, by exciting industry; but after the prices are settled, suitably to the new abundance of gold and silver, it has no manner of influence.

[...] Money having chiefly a fictitious value, the greater or less plenty of it is of no consequence, if we consider a nation within itself; and the quantity of specie, when once fixed, though ever so large, has no other effect, than to oblige every one to tell out a greater number of those shining bits of metal, for clothes, furniture or equipage, without encreasing any one convenience of life.

Here one already sees the two main empirical facts that characterize the Quantity Theory of Money today. Namely that the Quantity Theory of Money holds in the long-run, and equally important that it fails in the short run.⁴¹

What is missing however from these early characterizations of the Quantity Theory of Money is mention of the important role played by real output growth in the relationship between money and prices.⁴² When this role was first explicitly noted we are not sure, but it is present in Fisher (1911) and certainly well appreciated by the time of Milton Friedman and Anna Schwartzs 1963 book A Monetary History of the United States, 1867-1960. Writing on the Greenback period, 1867-1879,

⁴⁰The idea that the abundance of gold and silver was somehow related to the price level being already in existence. Its roots are in observations from the Salamanca School of Economics in the 16th century in response to the influx of precious metals from the Americas.

⁴¹We can also see the closely related idea that money is long-run neutral — that changes in the money supply cannot affect living standards in the long-run. We describe the similarities and differences between these two concepts later in this section.

⁴²Whether in Hume's work this is a sin of commission, or simply of omission, is open to interpretation.

they observe that prices decreased slightly, despite an increase in the money supply — attributing the difference as substantially due to the large increase in real output that occurred during this period.

Most evidence on the Quantity Theory of Money holding in the long-run came, at this time came, from plotting, say, four decade averages of money, price and output growth for a cross-section of countries and observing that the point lay close to the forty-five degree line predicted by the Quantity Theory of Money. Lucas (1980) provided a time series view of this same issue. Associating the long-run with low-frequency movements in the time series for money and inflation he showed that the Quantity Theory of Money holds in the long-run for the United States over the period 1955-1975. More recently Benati (2005, 2009) has extended these results for longer time periods, for the UK, and for other filters.

The Quantity Theory of Money has often been considered indirectly using two alternative ways to look at the issue, namely the 'velocity of money' (ie. PY/M), and the 'demand for money' (the size of money holdings measured as months of nominal income, ie. M/PY). The idea that the Quantity Theory of Money holds in the long-run but fails in the short-run thus become that the velocity of money changes in the short-run but is constant in the long-run; and similarly for the demand for money.

The perspective of the demand for money was often used as the basis for microeconomic models. Fisher (1911) presented an model in which people make payments and monitor the level of their bank accounts. This model, as well as many other inventory-based models, of money holdings implied that the velocity of money would be constant even in the short-run (Fisher was aware that empirically velocity fluctuates in the short-run). This occurred as these models implied a theory of money demand in which real money holdings are unit-elastic to changes in income. This difficulty persisted for some time. Akerlof (1979), in an article entitled "Irving Fisher on His Head: ...", showed that with a small change in assumptions — a change in the rule used to monitor real money holdings — Fisher's model would instead predict a (short-run) elasticity of real money holding with respect to income of zero. Thus in the short-run all changes in money would appear as fluctuations in velocity, and the Quantity Theory of Money will fail in the short-run.

Alongside attempts to model the demand for money there has been an empirical literature that aims to estimate the demand for money. Lucas (1988) for instance presents some theory and complementary estimations for money demand equations. This was extended to a cointegration analysis by Stock and Watson (1993)⁴³, and for the stability of the cointegration relationship by Seo (1998).

Sargent & Surico

Teles and Zhou (2005) provide another approach to the Quantity Theory of Money. They take a reverse-engineering view to the question of which is the correct monetary aggregate. Since the Quantity Theory of Money holds in the long-run, they suggest that the 'correct' monetary aggregate is the one with the most stable demand for money. Using US data to compare a number of different monetary aggregates they conclude that MZM (money zero maturity) is the 'correct' monetary aggregate.

⁴³Recognizing that the variables in question are non-stationary, it follows that the estimation results of Lucas (1988) are spurious and cointegration is the correct way to proceed. Of course, the concept of cointegration was unknown until 1987.

D.1 Relation to Other Monetary Theories

We now discuss the relation of the Quantity Theory of Money to two closely related, but fundamentally different, concepts: the long-run neutrality of money, and the fiscal theory of the price level. We do this as the relationship between the three is subtle, and is often confused. We start simply with the observation that the empirical fact established in this paper — that the Quantity Theory of Money holds in the long-run, and fails in the short-run — is logically independent from whether or not money is long-run neutral, and from whether or not fiscal factors determine the price level (the fiscal theory of the price level). That said, our personal reading of the literature is that money is long-run neutral, and that fiscal factors can be important in determining the price level.

D.1.1 Long-Run Neutrality of Money

The long-run neutrality of money says that in the long-run, changes in the money supply will have no effect on the level of real output. This is closely related to the Quantity Theory of Money on a very intuitive level. Let's assume that money is long-run neutral and pretend for a moment that real output is constant, then if we double the money supply we might expect prices to double — that is, we might expect the Quantity Theory of Money to hold.

However long-run neutrality of money and the Quantity Theory of Money are two different concepts. To see the difference we now develop a toy model. In the model, money fails to be long-run neutral but the Quantity Theory of Money holds in the long-run. If the model seems somewhat silly or forced this is simply because of the strongly intuitive link between the long-run neutrality of money and the Quantity Theory of Money in realistic settings.

Model: Let t denote the current time period, starting the model from period 0. Let the growth rate of output be zero percent if money growth is negative, and one percent if money growth is positive — obviously money is not going to be long-run neutral. The money supply is exactly determined by the government who decide how much money to print. Since the government knows about the effect of money growth on the growth rate of output they decide to increase the money supply at a rate of two percent per year ($g_M = 2$). Then the growth rate of output is one percent ($g_Y = 1$). If the growth rate of prices is one percent ($g_P = 1$), then by the definition of velocity we have that $g_V = g_P + g_Y - g_M = 1 + 1 - 2 = 0$. That is, velocity is constant and the Quantity Theory of Money holds in all periods. Thus we have a model in which the Quantity Theory of Money holds, but money is not long-run neutral.

It is theoretically possible to create a model in which money is long-run neutral but the Quantity Theory of Money fails to hold (in both the short-run and long-run).⁴⁴ But to do this requires the velocity of money to go to either infinity or zero in the long-run. If we imposed that the velocity of money be bounded then there must exist some fixed time horizon — some sufficiently long-run — over which the Quantity Theory of Money holds. Since both of these cases, velocity of infinity

⁴⁴*Model:* Let t denote the current time period, starting the model from period 0. Let real output be exogenous and grow at a constant rate of one percent per year ($g_Y = 1$) — so money is long-run neutral. Now let money supply also be completely exogenous and assume that the government has perfect control of the printing press which it uses to increase the money supply be t -percent per year ($g_M = t$). Admittedly, not the most perfect use of its perfect control. Now assume that the growth in prices is two- t -percent per year ($g_P = 2t$). Then by the definition of velocity we have that $g_V = g_P + g_Y - g_M = 2t + 1 - t = t + 1$ — clearly velocity is not constant and the Quantity Theory of Money fails to hold. But velocity is going to go to infinity, and this seems fairly silly.

or zero, seem completely implausible we conclude that it is fair to say that the long-run neutrality of money implies that the Quantity Theory of Money holds over the long-run.

From this we conclude that long-run neutrality of money implies that the Quantity Theory of Money holds in the long-run, but that the Quantity Theory of Money holding in the long-run *does not* imply the long-run neutrality of money.

So since the empirical fact that the Quantity Theory of Money holds in the long-run, and by extension the Lucas Illustrations, cannot be considered as evidence that money is long-run neutral^{45,46}, how might we test whether money is long-run neutral? One approach is to test long-run restrictions in Vector Autoregressions. For an explanation of how to do so, as well as evidence from this test suggesting that money is long-run neutral, see King and Watson (1997).⁴⁷

D.1.2 Fiscal Theory of the Price Level

As money is long-run neutral, then the level of real output is given, independent of changes in money and prices. The Quantity Theory of Money can then be considered as predicting what will be the price level, given the money supply. So one could have a steady level of prices simply by keeping the money supply stable. Why then do we see some countries printing lot's of money, with the resultant high inflation? Why would anyone want to do this? It is this question: What determines the money supply, which the Fiscal Theory of the Price Level addresses. Countries will print money if they need it to pay for spending (or pay off debt). Thus expectations about current and future government budget deficits can be considered as equivalent to expectations about how the money supply will change in the future, and thus about future prices. The fiscal behaviour of the government determines the money supply, and so in the long-run also determines the price level — this is the Fiscal Theory of the Price Level.

⁴⁵It is merely the absence of evidence against long-run neutrality of money.

⁴⁶In their original form of g_P against g_M it is doubly true that the Lucas Illustrations cannot be considered as evidence regarding the long-run neutrality of money. Consider the following two examples, in both of which money is long-run neutral.

Example 1: Let real output be exogenous and let it grow at one percent per year ($g_Y = 1$). Let money growth be exogenously determined by the government as $g_M = 2$. Let the Quantity Theory of Money hold ($g_V = 0$), then prices must grow at one percent ($g_P = 1$). So a plot of g_P against g_M would give a slope of 26.6 degrees ($= \tan^{-1}(g_P/g_M) = \tan^{-1}(1/2)$). Thus we have both that money is long-run neutral and that the Quantity Theory of Money holds, and yet a graph of g_P against g_M has a slope far from the forty-five degree line.

Example 2: Let real output be exogenous and let it grow at zero percent per year ($g_Y = 0$). Let money growth be exogenously determined by the government as $g_M = 2$. Let the Quantity Theory of Money hold ($g_V = 0$), then prices must grow at one percent ($g_P = 2$). So a plot of g_P against g_M would give a slope of 45 degrees ($= \tan^{-1}(g_P/g_M) = \tan^{-1}(2/2)$). Thus we have both that money is long-run neutral and that the Quantity Theory of Money holds, and a graph of g_P against g_M has a slope of exactly forty-five degrees.

What is going on here? For a graph of g_P against g_M to have a slope of exactly forty-five degrees it must be the case that $g_P = g_M$. Start from the accounting identity that $PY = MV$. Then it is tautologically true that $g_P + g_Y = g_P + g_V$. Thus $g_P = g_M$ is logically equivalent to saying that $g_Y = g_V$. That $g_Y = g_V$ is patently untrue.

⁴⁷Using the same methods they also provide a test for long-run super-neutrality of money (a.k.a. long-run neutrality of inflation), and reject that money is long-run super-neutral.

Evaluating a Flat Tax Reform

Javier Díaz-Giménez, Robert Kirkby, and Josep Pijoan-Mas

May 5, 2014

Abstract

In this article we quantify the aggregate and distributional consequences of investment expensing and progressivity in flat-tax reforms of the United States economy. We find that investment expensing as in the Hall and Rabushka type of reform brings about sizable output gains and a non-trivial increase in after-tax income inequality. But we also find that it results in large aggregate welfare gains in steady-state. Two additional flat-tax reforms with full investment expensing and varying degrees of progressivity reveal that the distributional role of the tax-exemption in the labor income tax is limited.

Keywords: Flat-tax reforms; Progressivity; Efficiency; Inequality.

JEL Classification: D31; E62; H23

Contents

1	Introduction	4
2	The Model Economy	7
2.1	Population and endowment dynamics	7
2.2	Preferences	8
2.3	Production	9
2.4	The government sector	9
2.5	Market arrangements	11
2.6	The households' decision problem	11
2.7	Equilibrium	12
3	Calibration	13
3.1	Model period	13
3.2	Normalization conditions	13
3.3	Macroeconomic and demographic targets	13
3.4	Government policy	14
3.5	The distributions of earnings and wealth	16
3.6	Calibration results	17
4	The Flat-Tax Reforms	18
4.1	Investment expensing in flat-tax reforms	20
4.1.1	Macroeconomic aggregates and factor ratios	20
4.1.2	Expenditure ratios	22
4.1.3	Fiscal policy ratios	22
4.1.4	Earnings, income, and wealth inequality	23
4.2	Progressivity in consumption-based flat-tax reforms	24
4.2.1	Macroeconomic aggregates and factor ratios	24
4.2.2	Expenditure ratios	25
4.2.3	Fiscal policy ratios	25
4.2.4	Earnings, income, and wealth inequality	26

5	Concluding Comments	26
A	Hall and Rabushka	30
B	The Transition Matrix on Exogenous Shocks	32
C	Non-convexities	33
D	Calibration	34
E	Computation	35
E.1	Value Functions and Stationary Distribution	35
E.2	Model Moments/Statistics	36
E.3	General Equilibrium	36
E.4	Calibration	37
E.4.1	Calibration Weights	38
E.5	General Equilibrium for a Revenue Neutral Reform	38
E.6	Transition Paths	38

1 Introduction

The main arguments for the use of flat-taxes are that the positive effects on output and efficiency. The main argument against flat-taxes is that they will increase inequality. We provide a quantitative assessment for the United States of the effects of the flat-tax reform proposed by Hall and Rabushka (1995) on output, efficiency, and inequality. We further evaluate the roles of two of the main components of the reform; the tax-deductibility of investment expenditures, and a tax-free threshold.

Importantly, we start from a model that includes the details of the current US tax system — the variety of different taxes and rates — and is able to replicate the facts relating to current US output and inequality in income and wealth. Given the concentrations of wealth in the top few percentiles (in 2007 the wealthiest one-percent owned 33.6 percent of all wealth), and the importance reaction of aggregate capital in determining the effects of the flat-tax, an accurate modeling of the concentration of wealth is quantitatively important. Accurate modelling of the inequality of wealth is of the contributions of this paper relative to the previous effort to quantify the effects of a flat-tax reform undertaken by Ventura (1999).

That modeling behavioural responses of taxpayers is important to understanding the distributional effects of tax-reform is not just a theoretical issue. Kasten, Sammartino, and Toder (1994) find that accounting for behavioural responses changes findings on whether the tax rate on the top one percent of income earners increased or decreased with the US Tax Reform Act of 1986. In assessing a fundamental tax-reform such as that of Hall and Rabushka (1995) it is clear that approaches modeling behavioural responses based on estimates of elasticities are inappropriate — the elasticities are only locally valid and by their very nature fundamental tax reforms are definitionally *not* local. Thus structural modeling of the tax-reforms is necessary. The kind of approach to quantifying the effects of taxes based on elasticities represented by, eg., Diamond and Saez (2011) is simply not applicable.

One important question is whether fundamental tax reforms should tax a broad definition of income, or whether they should tax exclusively consumption expenditures.¹ The key distinction between these two families of tax reforms is the tax treatment of investment expenditures. Income-based taxes tax both consumption and investment. This yields a broader tax base, but it generates distortions in capital accumulation. Consumption-based taxes do not tax investment expenditures and they have a smaller tax base. But they do not distort the capital accumulation decision. The classical optimality results of Chamley (1986) and Judd (1985), prescribe only consumption-based taxes because in the long run the distortions in capital accumulation are more severe than the distortions in the labor choices.²

Against this argument in favor of taxing consumption expenditures exclusively, there are other arguments that defend taxing a broad definition of income. Aiyagari (1995) points out that when labor income is uncertain and uninsurable the aggregate stock of capital may be too large, and some capital income taxation may be needed in order to bring it back to the *modified golden rule* of the textbook representative household growth model. Additionally, taxing capital income instead of labor income is a way to insure households against idiosyncratic labor market risks. This is because poor households own few assets and most of their income comes from labor sources, therefore a

¹See for instance Hubbard (1997) or Lazear and Porterba (2005) for a discussion of this issue.

²Lucas (1990), for instance, uses a representative household model to measure the welfare gains associated to the elimination of the current capital income taxation and he finds them to be large, of the order of a 5% equivalent variation in consumption every period.

shift of taxation from labor to capital makes four times better for the poor. In a quantitative exercise that supports this argument, Domeij and Heathcote (2004) show that the welfare losses brought about by eliminating capital income taxes can be large. In their model economy, even though eliminating capital income taxation increases output by 10 percent, it reduces aggregate welfare by about 1.5 percent.

In addition, implementing a consumption tax as a sales tax or a value added tax raises concerns about fairness because it is believed that it cannot be made a function of income. However this is not entirely true, since there are ways to design fundamental tax reforms that are both consumption-based and progressive. A famous example is the flat-tax reform originally suggested by Hall and Rabushka (1995). These authors propose to abolish the personal income tax and the corporate income tax and to substitute them with a unified flat tax on labor and business income. This tax scheme is equivalent to a consumption tax because it makes investment expenditures deductible from the business income tax base. The average tax rates on labor income are progressive because a fixed amount of labor income is tax-exempt. Simulation results such as those reported by Ventura (1999) or by Altig, Auerbach, Kotlikoff, Smetters, and Walliser (2001) find that a reform along these lines generates large increases in the accumulation of productive capital, and non-trivial increases in income and wealth inequality. However, other authors such as Gentry and Hubbard (1997) argue that this type of tax reforms may not generate more inequality than similar reforms that tax all income.

In this article we contribute to the debate on whether fundamental tax reforms should be income-based or consumption-based, and we find that revenue-neutral consumption-based tax reforms should be preferred because they result in larger welfare gains. We also study the role played by the size of the labor income tax deductions in consumption-based reforms, and we find that the welfare gains are increasing in the progressivity of the reforms.

To do so we calibrate a heterogeneous household, general equilibrium model economy to United States data and we use it to compare the steady-state aggregate, distributional, and welfare consequences of four fundamental tax reforms. Our model economy is an extension of the model economy described in Castaneda, Díaz-Giménez, and Ríos-Rull (2003). In essence it is a variation of the neoclassical growth model with heterogeneous households and uninsurable idiosyncratic risk that combines life-cycle and dynastic features.

We introduce the heterogeneity in labor market opportunities and wealth using an uninsurable process on the endowment of efficiency labor units. We model the life-cycle features using stochastic aging and retirement as in Gertler (1999). We model the dynastic links making households altruistic towards their descendants. Once our model economy is properly calibrated, these features guarantee that households in our model economy save for precautionary reasons, for life-cycle reasons, and for altruistic reasons. This property is important because capital accumulation is one of the main channels through which investment expensing and progressivity affect the economy.

Another important feature of our benchmark model economy is that our households choose their work effort. This is important for two reasons. First, because it allows us to quantify the direct effect of tax distortions on labor supply. Second, because as Pijoan-Mas (2006) shows, when labor market opportunities are uncertain, the labor supply becomes a quantitatively important self-insurance mechanism that allows households to reduce their precautionary savings.³ Given that changes in the progressivity of the tax code will change the uncertainty of after tax income, the interaction of labor and savings decisions can have sizable aggregate, distributional, and welfare

³Swanson (2012) provides complementary theoretical results on how labor supply affects measures of risk-aversion.

consequences.

A final distinguishing feature of our model economy is that it replicates the United States marginal distributions of labor earnings, income, and wealth in very much detail. And, in contrast with the model economies that focus exclusively on life-cycle features, it does a particularly good job in replicating the very top tails of those distributions. This feature is crucial for the quantitative evaluation of tax reforms because the tax burdens and the incentives to work and save that a tax code creates are very different at different points of the earnings and wealth distributions, and their effects are largest on the very income-rich and wealthy. Moreover, as Mirrlees (1971) points out, the distributional details are fundamental in measuring the trade-offs involved in choosing between efficiency and equality of tax reforms, because both the aggregate and the welfare changes depend critically on the number of households of each type that populate the economy.

In this article we start by evaluating the consumption-based flat-tax reform originally proposed by Hall and Rabushka (1995). To do so, we substitute the current personal and corporate income taxes with an integrated 19 percent flat tax on labor income and capital income—which we use as a proxy for business income. We deduct investment expenditures from the capital income tax base and we choose the personal deduction on the labor income tax to make the reform revenue neutral. We find that aggregate output and labor productivity increase by 11.3 and 12.6 percent, and that both after-tax income and wealth inequality increase substantially. Specifically the Gini index of after-tax income increases from 0.51 to 0.55, and the Gini index of wealth increases from 0.82 to 0.84. These results are consistent with most findings in the literature.

To measure the quantitative importance of not taxing the capital accumulation decision at the margin, we compare the allocations that obtain in the reformed economy with those that obtain when we simulate an alternative income-based flat-tax reform where the tax rate remains at 19 percent, but where investment is not expendable, and where we adjust the deduction in the labor income tax deduction to make the reform revenue neutral. We find that the aggregate gains brought about by this reform are more modest—output increases by only 4 percent—and that the increases in income and wealth inequality are also smaller—the steady state Gini index of after tax income that obtains after the reform is 0.53.

Our second contribution to the fundamental tax reform debate is to measure the quantitative importance of the labor income tax exemptions in consumption-based flat-tax reforms. To this purpose, we compare the steady-state allocations that obtain in the 19 percent flat-tax reform with the steady-state allocations of two other reformed economies which differ in the sizes of their flat-tax rates and of their labor income tax exemptions. Specifically, we study a proportional flat-tax reform in which all labor income is taxed at an integrated flat-tax rate of 15.3 percent, and a very progressive flat-tax reform in which we double the labor income tax deduction and in which the integrated flat-tax rate is 24.7 percent. In the model economy with the more progressive flat tax, output, consumption, aggregate hours, and the capital stock are all smaller. These results were to be expected. But more surprisingly, we also find that labor productivity is increasing in the progressivity of the reform and that the inequality of income after-taxes is very similar across the three consumption-based tax reforms.

These novel results are justified by a better allocation of household labor hours. It turns out that in the more progressive reforms, household hours are more correlated with labor market productivity. This is because the more progressive tax code provides more insurance against labor market uncertainty, and this allows households to improve their inter-temporal allocation of labor, and to make better use of their labor market opportunities—essentially the more progressive tax

reforms allow the households to work less when the times are bad. Consequently, since more progressive tax reforms increase the correlation of labor hours with the idiosyncratic labor shocks, they make the distribution of labor earnings before taxes more unequal and the average productivity per hour worked higher. This increased inequality in the distribution of before-tax earnings partly offsets the increased redistribution brought about by the higher labor income tax exemption and the higher flat-tax rate. And they result in similar concentrations of after-tax income.

2 The Model Economy

Our model economy is inhabited by a measure one continuum of heterogeneous dynastic households. Households make decisions about consumption, savings, and hours worked.

2.1 Population and endowment dynamics

The households are endowed with ℓ units of disposable time each period, and they are either workers or retirees. Workers face an uninsured idiosyncratic stochastic process that determines their endowment of efficiency labor units. They also face an exogenous probability of retiring. Retirees have zero labor efficiency units, so they do not work, and they face an exogenous probability of dying. When a retiree dies, it is replaced by a working-age descendant who inherits the retiree's estate and, possibly, some of its earning abilities. We use the one-dimensional shock, s , to denote the household's random age and random endowment of efficiency labor units jointly.⁴

The process on s is independent and identical across households, and follows a finite state Markov chain with conditional transition probabilities given by $\Gamma = \Gamma(s' | s) = Pr\{s_{t+1} = s' | s_t = s\}$, where s and $s' \in S$. We assume that s takes values in one of two possible J -dimensional sets, \mathcal{E} and \mathcal{R} . Therefore the formal description of set S is $S = \mathcal{E} \cup \mathcal{R} = \{1, 2, \dots, J\} \cup \{J+1, J+2, \dots, 2J\}$. When a household draws shock $s \in \mathcal{E}$, it is a worker and its endowment of efficiency labor units is $e(s) > 0$. When a household draws shock $s \in \mathcal{R}$ it is a retiree. When a household's shock changes from $s \in \mathcal{E}$ to $s' \in \mathcal{R}$, we say that it has retired and when it changes from $s \in \mathcal{R}$ to $s' \in \mathcal{E}$, we say that it has died and has been replaced by a working-age descendant. When a household dies, its estate is liquidated, and its descendant inherits a fraction $1 - \tau_e(\tilde{a})$ of the estate, where \tilde{a} denotes the value of the household's stock of wealth at the end of the period, and $\tau_e(\tilde{a})$ represents estate taxes.

This specification of the joint age and endowment process implies that the transition probability matrix, Γ , controls the demographics of the model economy, the life-cycle profile of earnings, and their intergenerational persistence (in combination with hours worked choices). When we come to calibrating this markov process it is done based on these issues; demographics, life-cycle profile of earnings, and intergenerational persistence of earnings.

To specify the process on s (and the values for $e(s)$) we must choose the values of $(2J)^2 + J$ parameters, of which $(2J)^2$ are the conditional transition probabilities and the remaining J are the values of the endowment of efficiency labor units. To reduce this large number of parameters, we

⁴ To ease interpretation, we use $e(s)$ to denote the endowment of efficiency labour units, and simply have s take integer values. In principle, s and $e(s)$ could be combined to be one object. By separating the efficiency labor units, $e(s)$, from the actual values taken by s , it is easier to see how earnings ability is transferred from one generation to the next, the transitions of s , without being confused by the fact that $e(s) = 0$ in all of the retirement states.

impose some additional restrictions on matrix Γ . To understand these restrictions better, it helps to consider the following partition of matrix Γ :

$$\Gamma = \begin{bmatrix} \Gamma_{\mathcal{E}\mathcal{E}} & \Gamma_{\mathcal{E}\mathcal{R}} \\ \Gamma_{\mathcal{R}\mathcal{E}} & \Gamma_{\mathcal{R}\mathcal{R}} \end{bmatrix}$$

Submatrix $\Gamma_{\mathcal{E}\mathcal{E}}$ contains the transition probabilities of working-age households that are still of working-age one period later. Since we impose no restrictions on these transitions, to characterize $\Gamma_{\mathcal{E}\mathcal{E}}$ we must choose the values of J^2 parameters.

Submatrix $\Gamma_{\mathcal{E}\mathcal{R}}$ describes the transitions from the working-age states into the retirement states. The value of this submatrix is $\Gamma_{\mathcal{E}\mathcal{R}} = p_{e\varrho}I$, where $p_{e\varrho}$ is the probability of retiring and I is the identity matrix. This is because we assume that every working-age household faces the same probability of retiring, and because we use only the last realization of the working-age shock to keep track of the earnings ability of retirees. Consequently, to characterize $\Gamma_{\mathcal{E}\mathcal{R}}$ we must choose the value of only one parameter.

Submatrix $\Gamma_{\mathcal{R}\mathcal{E}}$ describes the transitions from the retirement states into the working-age states that take place when a retiree exits the economy and is replaced by a working-age descendant. The rows of this submatrix contain a two parameter transformation of the stationary distribution of $s \in \mathcal{E}$, which we denote by $\gamma_{\mathcal{E}}^*$. This transformation allows us to control both the life-cycle profile of earnings and its intergenerational correlation. Intuitively, the transformation amounts to shifting the probability mass from $\gamma_{\mathcal{E}}^*$ towards both the first row of $\Gamma_{\mathcal{R}\mathcal{E}}$ and towards its diagonal.⁵ Consequently, to characterize $\Gamma_{\mathcal{R}\mathcal{E}}$ we must choose the value of the two shift parameters.

Finally, submatrix $\Gamma_{\mathcal{R}\mathcal{R}}$ contains the transition probabilities of retired households that are still retired one period later. The value of this submatrix is $\Gamma_{\mathcal{R}\mathcal{R}} = p_{\varrho\varrho}I$, where $(1 - p_{\varrho\varrho})$ is the probability of exiting the economy. This is because the type of retired households never changes, and because we assume that every retired household faces the same probability of exit. Therefore, to identify this submatrix we must choose the value of only one parameter.

To keep the dimension of the process on s as small as possible while still being able to achieve our calibration targets, we choose $J = 4$. Therefore, to characterize the process on s (and the values of $e(s)$), we must choose the values of $(J^2 + 4) + J = 24$ parameters.⁶

2.2 Preferences

We assume that households derive utility from consumption, $c_t \geq 0$, and from non-market uses of their time, and that they care about the utility of their descendants as if it were their own utility. Consequently, the households' preferences can be described by the following standard expected utility function:

$$E \left\{ \sum_{t=0}^{\infty} \beta^t u(c_t, \ell - h_t) \mid s_0 \right\},$$

where function u is continuous and strictly concave in both arguments; $0 < \beta < 1$ is the time-discount factor; ℓ is the endowment of productive time; and $0 \leq h_t \leq \ell$ is labor. Consequently,

⁵The exact definitions of the two shift parameters, ϕ_1 and ϕ_2 , can be found in Appendix B.

⁶Notice that we have not yet imposed that Γ must be a Markov matrix. When we do this, the number of free parameters is reduced to 20.

$\ell - h_t$ is the amount of time that the households allocate to non-market activities. Our choice for the households' common utility function is

$$u(c, l) = \frac{c^{1-\sigma_1}}{1-\sigma_1} + \chi \frac{(\ell - l)^{1-\sigma_2}}{1-\sigma_2}$$

We make this choice because the households in our model economies face very large changes the market value of their time. And if we had chosen the more standard non-separable specification for preferences, these changes would have resulted in extremely large variations in hours worked.

2.3 Production

We assume that aggregate output, Y_t , depends on aggregate capital, K_t , and on the aggregate labor input, L_t , through a constant returns to scale aggregate production function, $Y_t = f(K_t, L_t)$. We choose a standard Cobb-Douglas aggregate production function with capital share θ .⁷ Aggregate capital is obtained aggregating the wealth of every household, and the aggregate labor input is obtained aggregating the efficiency labor units supplied by every household. We assume that capital depreciates geometrically at a constant rate, δ , and we use r and w to denote the prices of capital and of the efficiency units of labor before all taxes.

2.4 The government sector

The government in our model economies taxes capital income, labor income, consumption, and estates, and it uses the proceeds of taxation to make real transfers to retired households and to finance an exogenous amount of government consumption.

Social security in our model economy takes the form of transfers to retired households, which we denote by ω , and which are financed with a payroll tax. The inclusion of a social security system has important implications for our research questions. First, it reduces the size of the steady-state aggregate stock of capital.⁸ Second, it plays an important role in helping us to replicate the large fraction of households who own very few or zero assets in the United States.⁹ Third, since public pensions are paid as life-time annuities, it insures the households against the risk of living for too long, and therefore it reduces their incentives to save.

Our calibration procedure allows us to match the size of the average public retirement pension paid in the United States and it ensures that the motives for saving in our model economy are quantitatively realistic. But pensions in our model economy are independent of contributions to social security and this feature qualifies the precision of our analysis in two ways. First, the overall amount of idiosyncratic risk in our model economy diminishes because the labor market history does not condition the retirement benefits. Second, we abstract from a potentially important reason

⁷ie. $Y_t = K_t^\theta L_t^{1-\theta}$. In the post-WWII U.S. real wages have grown, while factor income shares have displayed no clear trend. To replicate this behavior, the elasticity of substitution between capital and labor of the aggregate production function must be 1, as is the case in Cobb-Douglas functions. This is related to the 'Kaldor's stylized facts of growth', often referred to in models as 'balanced growth' — a nice discussion of the historical development of this issue can be found in the introduction of Cooley and Prescott (1995).

⁸Samuelson (1975) proves this result in a pure overlapping generations model. Our model economy is a dynastic model, so the pay-as-you-go social security system is isomorphic to a transfer system that reduces uncertainty in income. Therefore, the social security system reduces aggregate capital by reducing the need for precautionary savings.

⁹See Hubbard, Skinner, and Zeldes (1995).

to work, since in real world economies increasing the labor effort entitles the households to receive larger pension benefits.¹⁰

The capital income taxes in the economy are described by the function:

$$\tau_k(y_k) = a_1 y_k \tag{1}$$

where y_k denotes capital income. Of course, in the U.S. economy different types of capital are taxed at different rates and receive different types of deductions. In order to simplify our model economy we consider just one type of capital good.¹¹

Labor income taxes are described by function $\tau_l(y_a)$, where y_a denotes the labor income tax base. This tax is not used in the current United States tax system. But it is part of the flat-tax reforms which we describe in Section 4.

Payroll taxes paid by firms are described by function $\tau_{sf}(y_l)$, where y_l denotes labor income, and payroll taxes paid by households are described by function $\tau_{sh}(y_l)$. Our choice for the payroll tax function is

$$\tau_{sf}(y_l) = \tau_{sh}(y_l) = \begin{cases} a_2 y_l & \text{for } 0 \leq y_l \leq a_3 \\ a_2 a_3 & \text{otherwise} \end{cases} \tag{2}$$

This function approximates the cap on U.S. payroll taxes.¹² To replicate the U.S. Social Security tax code, we assume that the payroll taxes paid by the model economy households and firms are identical.

Household income taxes are described by the function:

$$\tau_y(y_b) = a_4 \left[y_b - (y_b^{-a_5} + a_6)^{-1/a_5} \right] \tag{3}$$

where the definition of the tax base is $y_b = y_k + y_l - \tau_k - \tau_{sf}$. This is the function chosen by Gouveia and Strauss (1994, 1999) to model the U.S. effective federal personal income taxes¹³. Notice that both capital income taxes and payroll taxes paid by firms are excluded from the household income tax base both in the United States personal income taxes and in our model economy household income taxes.

We assume that consumption taxes are proportional and that they are described by the function:

$$\tau_c(c) = a_9 c \tag{4}$$

And finally, we assume that the estate tax function is

$$\tau_e(\tilde{a}) = \begin{cases} 0 & \text{for } \tilde{a} < a_7 \\ a_8(\tilde{a} - a_7) & \text{otherwise} \end{cases} \tag{5}$$

¹⁰We make this assumption for technical reasons. Namely, because discriminating between households according to their past contributions to a social security system requires a second asset-type state variable and this would make our computational costs unmanageable. See Appendix E for the details on our computational algorithm.

¹¹To be consistent with this assumption, we calibrate the value of the tax rate on capital, a_1 , as the average tax rate levied on all capital income. By doing this we are implicitly assuming that every household in the economy owns varying amounts of shares of an identical portfolio of assets.

¹²In our model economy this cap on payroll taxes creates a non-convexity in the choice set of the households. We discuss this non-convexity in Section B of the Appendix.

¹³Observe that with this functional form a_4 defines the top (asymptotic) marginal tax rate, while a_5 and a_6 control the curvature and initial steepness. In the notation of Gouveia and Strauss (1999) $a_4 = b$, $a_5 = p$, $a_6 = s$.

This function replicates the main features of the current effective estate taxes in the United States.¹⁴

Therefore, in our model economies, a government policy rule is a specification of $\{\tau_k(y_k), \tau_l(y_a), \tau_{sf}(y_l), \tau_{sh}(y_l), \tau_y(y_b), \tau_c(c), \tau_e(\tilde{a}), \omega\}$ and of a process on government consumption, $\{G_t\}$. Since we also assume that the government balances its budget every period, these policies must satisfy the following restriction: $G_t + Z_t = T_t$, where Z_t and T_t denote aggregate transfers and aggregate tax revenues.

2.5 Market arrangements

We assume that there are no insurance markets for the household-specific shock. Instead, to buffer their streams of consumption against the shocks, the households in our model economy can accumulate wealth in the form of real capital. We assume that these asset holdings, a_t , belong to a compact set \mathcal{A} . The lower bound of this set can be interpreted as a form of liquidity constraints, or as a solvency requirement.¹⁵ The existence of an upper bound for the asset holdings is guaranteed as long as the after-tax rate of return to savings is smaller than the households' common rate of time preference. This condition is always satisfied in equilibrium.¹⁶

We also assume that firms rent factors of production from households in competitive spot markets. This assumption implies that factor prices are given by the corresponding marginal productivities.

2.6 The households' decision problem

The individual state variables are the realization of the household-specific shock, s , and the value of the stock of assets, a .¹⁷ The Bellman equation of the household decision problem is the following:

$$V(a, s) = \max_{\substack{c \geq 0 \\ \tilde{a} \in \mathcal{A} \\ 0 \leq h \leq \ell}} u(c, \ell - h) + \beta \sum_{s' \in S} \Gamma_{ss'} V[a'(\tilde{a}), s'], \quad (6)$$

$$s.t. \quad c + z = y - \tau + a, \quad (7)$$

$$y = ar + e(s)hw + \omega, \quad (8)$$

$$\tau = \tau_k(y_k) + \tau_l(y_a) + \tau_{sf}(y_l) + \tau_{sh}(y_l) + \tau_y(y_b) + \tau_c(c), \quad (9)$$

$$a'(\tilde{a}) = \begin{cases} \tilde{a} - \tau_e(\tilde{a}) & \text{if } s \in \mathcal{R} \text{ and } s' \in \mathcal{E}, \\ \tilde{a} & \text{otherwise.} \end{cases} \quad (10)$$

where function v is the households' common value function. Notice that household income, which we denote by y , includes three terms: capital income, $y_k = ar$, labor income, $y_l = e(s)hw$, and retirement pensions, ω . Every household can earn capital income. Only workers can earn labor income. And only retirees receive retirement pensions. The household policy that solves this

¹⁴See, eg., Aaron and Munnell (1992).

¹⁵Given that leisure is an argument in the households' utility function, this borrowing constraint can be interpreted as a solvency constraint that prevents the households from going bankrupt in every state of the world.

¹⁶Bewley (1983) and Huggett (1993) prove this proposition.

¹⁷In our model economy there are no aggregate state variables because we abstract from aggregate uncertainty and we restrict our analysis to the steady states of the economies.

problem is a set of functions that map the individual state into the optimal choices for consumption, end-of-period savings, and labor hours. We denote this policy by $\{c(a, s), \tilde{a}(a, s), h(a, s)\}$.

2.7 Equilibrium

Each period the economy-wide state is a probability measure, μ , defined over the appropriate σ -algebra on $S \times \mathcal{A}$ that counts the households of each type, and that we denote by \mathcal{B} . In the steady-state this measure is time invariant, even though the individual state variables and the decisions of the individual households change from one period to the next.¹⁸

Definition 1 *A steady state equilibrium for this economy is a household value function, $V(a, s)$; a household policy, $\{c(a, s), \tilde{a}(a, s), h(a, s)\}$; a government policy, $\{\tau_k(y_k), \tau_l(y_a), \tau_{sf}(y_l), \tau_{sh}(y_l), \tau_y(y_b), \tau_c(c), \tau_e(\tilde{a}), \omega, G\}$; a stationary probability measure of households, μ ; factor prices, (r, w) ; and macroeconomic aggregates, $\{K, L, T, Z\}$, such that:*

- (i) *Given factor prices and the government policy, the household value function and the household policy solve the households' decision problem described in expressions (6)-(10).*
- (ii) *Firms behave as competitive maximizers. That is, their decisions imply that factor prices are factor marginal productivities $r = f_1(K, L) - \delta$ and $w = f_2(K, L)$.*
- (iii) *Factor inputs, tax revenues, and transfers are obtained aggregating over households:*

$$\begin{aligned} K &= \int a \, d\mu \\ L &= \int h(a, s) \, e(s) \, d\mu \\ T &= \int [\tau_k(y_k) + \tau_l(y_a) + \tau_{sf}(y_l) + \tau_{sh}(y_l) + \tau_y(y_b) + \tau_c(c)] \, d\mu + \int \gamma_{s\mathcal{E}} \mathbf{I}_{\{s \in \mathcal{R}\}} \tau_e(\tilde{a}) \tilde{a}(a, s) \, d\mu \\ Z &= \int \omega \mathbf{I}_{\{s \in \mathcal{R}\}} \, d\mu. \end{aligned}$$

where \mathbf{I} denotes the indicator function, the definition of parameter $\gamma_{s\mathcal{E}}$ is $\gamma_{s\mathcal{E}} \equiv \sum_{s' \in \mathcal{E}} \Gamma_{ss'}$ and, consequently, $(\gamma_{s\mathcal{E}} \mathbf{I}_{s \in \mathcal{R}})$ is the probability that a retiree of type s exits the economy (i.e. dies). And where every integral in the four definitions above is defined over the state space $S \times \mathcal{A}$.

(iv) *The goods market clears: $\int [c(a, s) + \tilde{a}(a, s)] \, dx + G = f(K, L) + (1 - \delta) K$.*

(v) *The government budget constraint is satisfied: $G + Z = T$*

(vi) *The measure of households is stationary:*

$$x(B) = \int_B \left\{ \int_{S \times \mathcal{A}} [\mathbf{I}_{a' = \tilde{a}(a, s)} \mathbf{I}_{s \notin \mathcal{R} \vee s' \notin \mathcal{E}} + \mathbf{I}_{a' = [1 - \tau_e(\tilde{a})] \tilde{a}(a, s)} \mathbf{I}_{s \in \mathcal{R} \wedge s' \in \mathcal{E}}] \Gamma_{ss'} \, d\mu \right\} da' \, ds'$$

for all $B \in \mathcal{B}$, where \vee and \wedge are the logical operators “or” and “and”. This equation counts the households, and the cumbersome indicator functions and logical operators are used to account for estate taxation. We describe the procedure that we use to compute this equilibrium in Appendix E.

¹⁸See Hopenhayn and Prescott (1992) and Huggett (1993).

3 Calibration

Calibration of the model was based on the year 2007, with the main data sources being the 2007 Survey of Consumer Finances (SCF) and Federal Reserve Economic Database (FRED). Our model economy is characterized by 43 parameters.¹⁹ Of these parameters, 5 describe the preferences of the households, 2 the production technology, 11 the government policy, and 25 the joint process on the age of the households and on the endowments of efficiency labor units (including the choice of J itself). Six of the parameters are decided by normalization conditions, and the remaining 37 are determined by statistics that describe relevant features of the United States economy. Eight of the remaining parameters are directly identified by eight target statistics. To determine the values of the remaining 29 parameters we use the Simulated Method of Moments, picking the 29 parameters to minimize the weighted distance between 29 (simulated) moments of the model economy and the corresponding 29 data moments of the United States economy for the year 2007. We here describe the data moments of the United States economy used and discuss why they were chosen. Further details on the calibration can be found in Appendix D, and on the computational steps involved in Appendix E.

3.1 Model period

The U.S. tax code defines tax bases in annual terms. Since the income tax, the payroll tax and the estate tax are not proportional taxes, the obvious choice for our model period is one year. Moreover, the Survey of Consumer Finances, which is our main source of micro-data, is also yearly.

3.2 Normalization conditions

As discussed in 2.1 we set $J = 4$. The household endowment of disposable time is an arbitrary constant and we choose it to be $\ell = 1$. Finally, since matrix Γ is a Markov matrix, its rows must add up to one. This property imposes four additional normalization conditions on the rows of $\Gamma_{\mathcal{E}\mathcal{E}}$.²⁰

3.3 Macroeconomic and demographic targets

Ratios: We target a capital to output ratio, K/Y , of 4.67, a capital income share of 0.376, and an investment to output ratio, I/Y , of 22.0 percent. We obtain our target value for the capital output share dividing \$555,400, which was average household wealth in the United States in 2007 according the 2007 Survey of Consumer Finances, by \$118,953, which was per household Gross Domestic Product for the United States in 2007.²¹ Our target for the capital income share is the value that obtains when we use the methods described in Cooley and Prescott (1995) and we

¹⁹46 parameters if we were to include the three parameters that are needed to characterize the tax reform experiments.

²⁰Note that our assumptions about the structure of matrix Γ imply that once submatrix $\Gamma_{\mathcal{E}\mathcal{E}}$ has been appropriately normalized, every row of Γ adds up to one without imposing any further restrictions.

²¹Calculated as nominal GDP (FRED: GDP) divided by number of households. Where number of households is US population/2.56. US population is 295 million (FRED: POP), and 2.56 is average Household size according to SCF, Díaz-Giménez, Glover, and Ríos-Rull (2011).)

exclude the public sector from the computations.²² To calculate the value of our target for I/Y , we define investment as the sum of gross private fixed domestic investment, change in business inventories, and 75 percent of the private consumption expenditures in consumer durables using data for 2007.²³

Allocation of time and consumption: We target a value of $H/\ell = 33$ percent for the average share of disposable time allocated to working in the market.²⁴ For the curvature of consumption we choose a value of $\sigma_1 = 1.5$. This value falls within the range (1–3) that is standard in the literature.²⁵ We do not calibrate the curvature of leisure, σ_2 , but instead allow it to be determined as part of the Simulated Methods of Moments estimation. Interestingly this delivers a value similar to those in the literature.²⁶

The age structure of the population: We target the expected durations of working-lives and retirement of the model economy households to be 45 and 18 years. These targets replicate the average durations of working-lives and retirement in the United States.

The life-cycle profile of earnings: To replicate the life-cycle profile of earnings of the United States in our model economy, we target the ratio of the average earnings of households aged 46 and 50 to the average earnings of households aged 26 and 30; these two age ranges are those least affected by choices of whether or not to work. This ratio was 1.73 in 2007 according to the Survey of Consumer Finances.²⁷

The intergenerational transmission of earnings ability: To replicate the intergenerational correlation of earnings of the United States in our model economy, we target the cross-sectional correlation between the average life-time earnings of one generation of households and the average life-time earnings of their immediate descendants. Solon (1992) and Zimmerman (1992) measure this statistic for fathers and sons in the United States, and they report that it is 0.4, approximately.

3.4 Government policy

In Table 1 we report the revenues obtained by the combined Federal, State, and Local Governments in the United States in the 2007 fiscal year. To choose the parameter values of the tax functions in our model economy we must first allocate the United States tax revenues to the tax instruments of

²²See Castaneda, Díaz-Giménez, and Ríos-Rull (2003) for details about this number.

²³This definition of investment is approximately consistent with the 2007 Survey of Consumer Finances definition of household wealth, which includes the value of vehicles, but does not include the values of other consumer durables. Data are from FRED: FPIA \$2266.1 billion, PCEDG \$1188.4 billion (annual average), and change in BUSINV \$-18.9 billion (=74.5-93.3, annual averages). Thus we get total investment as \$3138.5 billion. GDP was \$14253.2 billion.

²⁴See Juster and Stafford (1991) for details about this number.

²⁵Recent calibration exercises find very similar values for σ_1 . For example, Heathcote, Storesletten, and Violante (2010) report a value of 1.44 and Pijoan-Mas (2006) reports a value of 1.46 for this parameter.

²⁶Ríos-Rull, Schorfheide, Fuentes-Albero, Kryshko, and Santaaulalia-Llopis (2012) contain a comprehensive discussion of this parameter, both from the perspective of calibration and from the perspective of bayesian estimation. Since $\ell = 1$, the Frish elasticity — the elasticity of hours worked with respect to wages — is given by *Frisch elasticity* = $\frac{1}{\sigma_2} \frac{\ell-h}{h} = \frac{2/3}{1/3} \frac{1}{\sigma_2}$.

²⁷Table 11 of Díaz-Giménez, Glover, and Ríos-Rull (2011) provides the averages for these two age groups: \$52,300 for ages 26-30, and \$90,700 for ages 46-50.

Table 1: Federal, State, and Local Government Receipts

Fiscal Year	2007	
	\$Billion	%GDP
Gross Domestic Product (GDP)	13861.4	100.00
Total Federal, State and Local Gvt Receipts	4197.0	30.28
Individual Income Taxes	1468.4	10.59
Social Insurance and Retirement	869.6	6.27
Sales and Gross Receipts Taxes	449.9	3.25
Property Taxes	409.5	2.95
Corporate Profit Taxes	427.2	3.08
Excise Taxes	65.1	0.47
Estate and Gift Taxes	26.0	0.19
Custom Duties and Fees	26.0	0.19
Other Taxes	47.5	0.34

Source: Tables B78, B81, B82, and B86 of the Economic Report of the President 2013.

our benchmark model economy. We choose the parameters of the model economy household income tax so that they collect the revenues levied by the U.S. personal income tax, the parameters of the model economy capital income tax so that it collects the revenues levied by the U.S. corporate income tax, and with the model economy payroll and estate taxes we do likewise. The remaining sources of government revenues in the United States are sales and gross receipts taxes, property taxes, excise taxes, custom duties and fees, and other taxes. Added together, these tax instruments collected 7.2 percent of U.S. GDP in 2007. In our model economy we allocate these revenues to the consumption tax.²⁸

To choose the parameters of the expenditure side of the government budget, we do the following: First, since the government of our model economy must balance its budget, we require that the output shares of government consumption and government transfers—the two expenditure items in our model economy—add up to 30.28 percent, which was the GDP share of total tax revenues in the United States in 2007. Then we target a value for the transfers to output ratio in the model economy of 5.53 percent. This value corresponds to the share GDP accounted for by Medicare and by two thirds of Social Security transfers in the United States in 2007. We chose this target because transfers in our model economies are lump-sum, and Social Security transfers in the U.S. economy are mildly progressive. This choice leaves us with a residual share for government expenditures to GDP of 24.75(= 30.28 – 5.53) which is our target for the G/Y ratio in our model economy.²⁹ We discuss the details of our choices for the various model economy tax function parameters in the paragraphs below.

Capital income taxes: We choose a_1 , the capital income tax rate of function (1), so that the revenues collected by this tax in the benchmark model economy match the revenues collected by the corporate profit tax in the U.S. economy as a fraction of GDP.

²⁸Since we also target government transfers and government expenditures (see below), the model economy’s consumption tax rate is determined residually to balance the government budget.

²⁹The size of the government measured by expenditures as a percentage of GDP in this model is slightly smaller than in the data. This is because we target the revenues as a measure of the size of government, 30.28 percent of GDP, and then require the government to run a balanced budget. In 2007, government measured by Government Total Expenditures was 32.95 percent of GDP (FRED: W068RC1A027NBEA). The Federal deficit alone, –1.11 percent of GDP, accounts for most of the difference (FRED: FYFSGDA188S). Since our model does not contain government debt we also miss interest payments on Federal government debt.

Payroll taxes: To characterize the payroll tax function described in expression (2), we must choose the values of parameters a_2 and a_3 . In 2007 in the U.S. the payroll tax rate paid by both households and firms was 7.65 percent each and it was levied only on the first \$97,500 of gross labor earnings. This value was approximately equal to 82 percent of the U.S. per household GDP. To replicate these values, in our model economy we make $a_2 = 0.0785$ and $a_3 = 0.82\bar{y}$, where \bar{y} denotes output per household. These choices imply that the payroll tax collections in our model economy are endogenous, and that we can use them as an overidentification restriction.

Household income taxes: To characterize the income tax function described in expression (3), we must choose the values of parameters a_4 , a_5 and a_6 . Since a_4 and a_5 are unit-independent, we use the values reported by Flynn (2009)³⁰ for these parameters, namely, $a_4 = 0.296$ and $a_5 = 0.596$. To determine the value of a_6 , we equate the tax rate levied on a value of income equal to average output per household in our model economy to the effective tax rate on GDP per household levied in the U.S. economy. Again, these choices imply that the household income tax collections in our model economy are endogenous, and that we can use them as another overidentification restriction.

Estate taxes: To characterize the estate tax function described in expression (5), we must choose the values of parameters a_7 and a_8 . During 2007 in the United States the first \$2,000,000 of the value of estates were tax exempt. This value was approximately equal to 20 times the average value of GDP per household.³¹ In our model economy we make $a_7 = 20\bar{y}$, to replicate this feature of the United States estate tax code. Finally, we choose the value of a_8 so that the estate tax in our model economy collects the same revenues as the estate tax in the United States.

Consumption taxes: We choose the value of parameter a_9 in the consumption tax function described in expression (4) residually, so that the government in our model economy balances its budget. Therefore, the consumption tax collections in our model economy are also endogenous, and they can be interpreted as a third overidentification restriction.

3.5 The distributions of earnings and wealth

We use 18 targets relating to the distributions of earnings and wealth: the 2 Gini indexes and 16 additional points from the Lorenz curves of the United States distributions of earnings and wealth, namely the shares of the various quintiles and of the top percentiles. We report these targets in Table 4. The values are taken from Díaz-Giménez, Glover, and Ríos-Rull (2011) who calculate them based on the Survey of Consumer Finances.

³⁰These are the values reported by Flynn (2009) for 2005, the closest year to our target of 2007. Flynn (2009) presents estimates of the parameters for the effective tax function used by Gouveia and Strauss (1994, 1999) for the years 1979 to 2005 based on the IRS Public Use File dataset. The estimates for the parameters are largely stable for the period 2003-2005 (ie. post the tax reform of 2001) and it seems reasonable to assume that this would continue through to 2007, since no further tax reforms occur inbetween.

³¹The estate tax thresholds for all years since 1934 can be found in footnote 5 of [this website](#) from the Tax Policy Center.

3.6 Calibration results

Our calibration procedure allow us to characterize the stochastic process on the endowment of efficiency labor units. This process is not to be taken literally, since it is a black box that represents everything that we do not know about our model economy. In particular, we cannot compare our process with the panel data estimates of wage processes for prime-age males, such as those reported in Blundell and MaCurdy (1999), or in the more recent Heathcote, Storesletten, and Violante (2010). This is because our process is a measure of household labor market opportunities and not of individual labor market opportunities. In our model economy labor market opportunities result in household labor supply decisions, which include participation decisions of the members of the household.³² Also, panel data sets, such as the Panel Study of Income Dynamics (PSID) or the National Longitudinal Survey of Youth (NLSY), miss the upper tail of the wage distribution, both because of top-coding and because they are not explicitly designed to measure the earnings of the very rich. The upper tail of the earnings distribution is very important if we want our model economy to be consistent with the upper tail of the wealth distribution of the United States as reported by the Survey of Consumer Finances which does not have either one of these two problems.

Table 2: The stochastic process for the endowment of efficiency labor units

	$e(s)$	γ_s^* (%)	$\Gamma_{\mathcal{E}\mathcal{E}}$ (%) From s To s'			
			$s' = 1$	$s' = 2$	$s' = 3$	$s' = 4$
$s = 1$	0.40	54.82	98.02	1.01	0.96	0.01
$s = 2$	1.36	42.04	0.01	90.42	9.55	0.01
$s = 3$	6.77	3.09	7.67	8.89	83.43	0.01
$s = 4$	198.88	0.05	9.93	7.04	0.01	83.02

Note: $e(s)$ denotes the relative endowments of efficiency labor units; γ_s^* denotes the stationary distribution of working-age households (note that this is not the stationary distribution of $\Gamma_{\mathcal{E}\mathcal{E}}$, it is taken from the stationary distribution of Γ and renormalized); $\Gamma_{\mathcal{E}\mathcal{E}}$ denotes the transition probabilities of the process on the endowment of efficiency labor units for working-age households that are still workers one period later.

In the second column of Table 2 we report the relative endowments of efficiency labor units, and in the third column the invariant measures of each type of working-age households. The endowments of workers of $s = 1$, $s = 2$, $s = 3$, and $s = 4$ are, approximately, 0.4, 1.4, 6.8, and 199. This means that, in our model economy, the luckiest workers are 497.5 times as lucky as the unluckiest ones. The stationary distribution shows that each period 97 percent of the workers are unlucky and draw states $s = 1$ or $s = 2$, and that only one out of every 2,000 workers is extremely lucky and draws state $s = 4$.

In the last four columns of Table 2 we also report the transition probabilities between the working-age states. These probabilities are conditional on their not retiring; hence they sum to one hundred percent. The states are of decreasing persistency. Conditional on not retiring their expected durations are 34.3, 6.9, 3.8, and 3.7 years.³³

As far as the transitions are concerned, we find that a worker whose current state is $s = 1$ is more likely to move to state $s = 2$ than to any of the other states. Likewise, a worker whose current

³²See Guner, Kaygusuv, and Ventura (2012) for an evaluation of tax reforms modeling two-member households explicitly.

³³Let p_{ii} be the probability of remaining in state i . The expected duration is the number of periods T such that the probability of remaining in state i for T periods is equal to one half. Since p_{ii}^T is the probability of remaining in state i after T periods the expected duration is given by $p_{ii}^T = 0.5$, or rewriting, $T = \ln(0.5)/\ln(p_{ii})$.

state is either $s = 2$ or $s = 3$ is most likely to move back to state $s = 1$. Only very rarely workers whose current state is either $s = 1$ or $s = 2$ will make a transition either to state $s = 3$ or to state $s = 4$. Finally, when a worker draws state $s = 4$, it is most likely that she will draw either state $s = 2$ or state $s = 1$ shortly afterwards.

Table 3: Parameter values for the benchmark model economy

<i>Preferences</i>		
Time discount factor	β	0.95
Curvature of consumption	σ_1	1.50
Curvature of leisure	σ_2	1.83
Relative share of consumption and leisure	χ	0.81
Endowment of discretionary time	ℓ	1.00
<i>Technology</i>		
Capital income share	θ	0.38
Capital depreciation rate	δ	0.05
<i>Age and endowment process</i>		
Probability of retiring	p_{ee}	0.02
Probability of dying	$1 - p_{ee}$	0.06
Life cycle earnings profile	ϕ_1	0.85
Intergenerational persistence of earnings	ϕ_2	0.86
<i>Fiscal policy</i>		
Government consumption	G	0.48
Retirement pensions	ω	0.17
Capital income tax function	a_1	0.20
Payroll tax function	a_2	0.08
	a_3	0.71
Household income tax function	a_4	0.26
	a_5	0.77
	a_6	1.00
Estate tax function	a_7	18.28
	a_8	0.12
Consumption tax function	a_9	0.09

We report the values of every other parameter of our model economy in Table 3, and in Table 4 we report the statistics that describe the main aggregate and distributional features of the United States and the benchmark model economies.³⁴ These numbers confirm that overall our model economy succeeds in replicating the most relevant features of the United States in very much detail. We are particularly encouraged by our model economy's ability to replicate the U.S. fiscal policy ratios and the U.S. distributions of earnings, income and wealth, since these two sets of targets are the main focus of this article. Recall that in our calibration exercise we have not targeted either the payroll tax collections, the household income tax collections, the consumption tax collections, or the statistics that describe the income distribution, and that all of these statistics can be considered to be overidentification restrictions.

4 The Flat-Tax Reforms

We study the consumption-based flat tax reform originally proposed by Hall and Rabushka (1995). Among the key features of the Hall-Rabushka flat-tax are the tax-deductability of investment, and a

³⁴For a listing of the values of targets used in the Simulated Method of Moments see Table 15 in the Appendix.

Table 4: The Benchmark Model Economy (E_B) and the United States

Macroeconomic Ratios									
	C/Y	I/Y	G/Y	K/Y	—	—			
U.S.	67.30	22.00	24.75	4.67	—	—			
E_B	56.45	22.95	20.81	4.55	—	—			
Fiscal Policy Ratios									
	G/Y	Z/Y	T/Y	T_y/Y	T_l/Y	T_k/Y	T_s/Y	T_c/Y	T_e/Y
U.S.	24.8	5.5	30.3	10.6	—	3.1	6.3	7.2	0.19
E_B	20.8	5.1	25.9	12.0	—	3.0	5.5	5.2	0.21
The Distributions of Earnings									
Gini	Quintiles (%)					Top groups (%)			
Economy		1st	2nd	3rd	4th	5th	90–95	95–99	99–100
United States	0.640	0.0	4.2	11.7	20.8	63.5	11.7	16.6	18.7
E_B	0.646	0.0	3.0	11.6	21.3	61.2	9.0	16.7	20.2
The Distributions of Income (before all taxes and after transfers)									
United States	0.575	2.8	6.7	11.3	18.3	60.9	10.2	15.9	21.0
E_B	0.528	4.3	7.6	8.8	21.8	54.7	7.4	17.1	17.4
The Distributions of Wealth									
United States	0.820	0.0	1.1	4.5	11.2	83.4	11.1	26.7	33.6
E_B	0.824	0.0	0.0	3.4	12.3	83.0	14.4	23.7	34.1

Note: Many of these statistics are targeted as part of the calibration and simulated moment estimation of the model. Those that are not targeted are C/Y , T/Y , T_y/Y , T_s/Y , T_c/Y , and all those relating to the distribution of income (the distributions of earnings and wealth are targeted). Consumption is measured as Personal Consumption Expenditures (FRED: PCE). The data sources for all the other statistics are described in the text.

tax-free threshold. In this section we also consider an income-based flat-tax reform, which drops the tax-deductability of investment. In the following section we consider varying the tax-free threshold. The Hall-Rabushka flat-tax is only intended to replace the personal income tax and the corporate income tax; the payroll taxes, estate tax, and state-level taxes are therefore left untouched. Hall and Rabushka (1995) propose a marginal tax rate of 19 percent, which the tax-free threshold set to ensure budget neutrality: that tax-revenues, government spending, and government transfers are left untouched.³⁵

To implement both the consumption-based flat tax reform of Hall and Rabushka (1995), and the income-based flat-tax reform, we replace the household income tax with a flat tax on all labor income above a tax-exempt level, and the calibrated capital income tax with an integrated flat tax on capital income. The function that describes the labor income tax is

$$\tau_l(y_a) = \begin{cases} 0 & \text{for } y_a < a_{10} \\ a_{11}(y_a - a_{10}) & \text{otherwise} \end{cases} \quad (11)$$

where the tax base is labor income net of social security taxes paid by firms, $y_a = y_l - \tau_{sf}(y_l)$, parameter a_{10} is the tax-exempt level of labor income, and parameter a_{11} is the flat-tax rate. The capital income tax function in the reformed economies is the same as the capital income tax function defined in Expression (1) above. The only difference is that in the consumption-based tax

³⁵If the tax reforms cause GDP to increase (decrease) this will imply that tax-revenues as a percentage of GDP decrease (increase).

reforms investment expenditures are tax-exempt and, consequently, the capital income tax base is capital net of depreciation income minus savings. Therefore, in these reforms $y_k = r a - (a' - a)$.³⁶ Since in every reform capital and labor income are taxed at the same marginal tax rate, we impose that $a_1 = a_{11}$. Finally, every flat-tax reform is designed to be revenue neutral and none of them changes the composition of public outlays. Therefore, in every flat-tax reform the values of T , G , and Z remain unchanged, and they are equal to their values in the benchmark model economy.

4.1 Investment expensing in flat-tax reforms

In this section we compare the allocations that obtain in the steady-states of two flat-tax model economies that differ only in the fiscal treatment of investment expenditures. In the consumption-based flat-tax economy investment expenditures are fully deductible, and in the income-based flat tax economy they are fully taxed.

The consumption-based flat-tax economy, which in this section we call E_C , is the standard flat-tax reform originally proposed by Hall and Rabushka (1995). As these authors suggested, its marginal tax rate on capital and labor income is 19 percent. And we choose the size of its labor income tax deduction to make the reform revenue neutral. This requires a labor income tax deduction $a_{10} = 0.0489$, which corresponds to 5.3% percent of output per household in the benchmark model economy or \$6,300, approximately.

In the income-based flat-tax economy, which we call E_Y , we keep the 19 percent marginal integrated flat tax rate, but since investment expenditures are not deductible, we change the labor income tax deduction to make the reform revenue neutral. In principle, the direction of this change could go either way. Taxing investment increases the base of the capital income tax. In partial equilibrium this would increase the capital income tax revenues and it would require a larger deduction in the labor income tax to make the reform revenue neutral. But in general equilibrium taxing investment reduces the capital stock. Therefore it also reduces aggregate output and the bases of both the capital and the labor income flat taxes.

It turns out that this second effect dominates. And we find that the value of the labor income tax deduction that makes this reform revenue neutral is $a_{10} = 0$, which trivially corresponds to \$0. This is because steady-state output in the income-based flat-tax reform only slightly larger than in the baseline economy, while the marginal tax rate is lower: under the flat-tax reform the marginal rate is 19%, while pre-reform the top marginal rate on capital was 20% and the top rate on income was 26%.³⁷

4.1.1 Macroeconomic aggregates and factor ratios

In Table 5 we report the main macroeconomic aggregates and factor ratios of our model economies. We find that the steady-state aggregate changes brought about by the consumption-based flat tax reform are substantial. Steady state output in model economy E_C is 13 percent larger than in the benchmark economy, E_B . This increase in output is brought about by a very large increase

³⁶Taxing capital income at the household level is equivalent to the proposed business income tax of Hall and Rabushka (1995), which is applied to firms by taxing business income net of wages, depreciation expenses and net investment.

³⁷In fact, as can be seen in Table 7 even a zero threshold is not enough to achieve perfect revenue neutrality and tax revenues are forced to fall 0.2% of baseline output as a result.

in aggregate capital, of 37 percent. In contrast, the changes brought about in the aggregate labor input are small. Both total labor hours and the total labor input almost unchanged. We also find that the productivity of labor hours increases by approximately 12 percent, as a result of capital deepening.

Table 5: Production, inputs and input ratios in the model economies

	Y	K	L	H/ℓ	K/L	L/H	Y/H	K/Y
E_B	0.92	4.19	0.37	33.10	11.34	1.12	2.78	4.55
$E_C/E_B(\%)$	12.7	37.2	0.0	0.5	37.2	-0.4	12.1	21.8
$E_Y/E_B(\%)$	4.9	14.2	-0.3	-0.4	14.5	0.1	5.4	8.8

^aVariable L denotes the aggregate labor input.

^bVariable H denotes the share of the endowment of time allocated to the market.

There are two reasons that justify the increase in the capital stock. First, the consumption-based flat tax reform eliminates the distortion in the intertemporal allocation of consumption, which encourages the households to save and to accumulate capital. Second, as we discuss below, the distribution of after tax income becomes more unequal. This increases the need for precautionary savings and therefore it increases the size of the capital stock even further.

We find that the changes brought about by the income-based flat tax reform are much smaller. As expected, taxing investment has large implications for the capital accumulation decision. Compared to the benchmark model economy, in the income-based flat-tax reform aggregate capital increases by 14 percent, which is only slightly more than one third of the increase that obtains in the consumption-based reform. Still, the increase in the capital stock is not small.

There are two reasons for this increase. First, as in the consumption-based reform, there is an increase in the precautionary motive for savings. Second, the income-based flat-tax reform reduces the marginal capital income tax rates faced by the wealthy, and they increase the marginal capital income tax rates faced by the wealth poor. This is because in our benchmark economy capital income is taxed twice—once by the capital income tax and a second time by the household income tax—and because the rates on capital income of the household income tax are progressive. The aggregate effect of these changes is to increase capital accumulation because wealthy households are more concerned with after tax returns and less concerned with precautionary motives than poor households.

The income-based flat-tax reform also brings about changes in the aggregate labour input that are very small. Consequently, aggregate output in this model economy is only 5 percent larger than in the benchmark economy, while in the consumption-based flat-tax reform it is 13 percent larger.

These findings can be compared with those reported in Altig, Auerbach, Kotlikoff, Smetters, and Walliser (2001), albeit in a somewhat indirect way. Altig, Auerbach, Kotlikoff, Smetters, and Walliser (2001) study a sequence of reforms. First, they look at a purely proportional flat tax on all income. Second, they allow for full expensing of new investment, which makes their income tax equivalent to a consumption tax. And third, they add a labor income tax exemption.³⁸ They find that these three reforms increase aggregate output in the long run. A strictly proportional income tax increases output by 5 percent. Allowing for the expensing of new investment increases output by an additional 4 percent. And adding a fixed deduction to labor income requires a higher marginal tax rate that brings the output increase back to 4.5 percent. Therefore, their results are more modest than ours.

³⁸They consider two additional reforms with different tax breaks for capital holders during the transition.

This is partly because our model economy extends Altig, Auerbach, Kotlikoff, Smetters, and Walliser (2001) in several important dimensions. First, we consider uninsurable labor market uncertainty. This brings into the analysis of flat-tax reforms the partial insurance role played by the various tax codes, which is absent from Altig, Auerbach, Kotlikoff, Smetters, and Walliser (2001). Second, we allow for earnings and wealth mobility. This feature of our model economy should reduce the welfare consequences of the reforms because our income process is mean reverting, at least at the dynastic level. Third, earnings, income and wealth are more concentrated in our model economy than in Altig, Auerbach, Kotlikoff, Smetters, and Walliser (2001), and they match their counterparts in the data. Finally, our households are altruistic towards their descendants and our model economy displays some of the intergenerational correlation of earnings observed in the United States. We think that this feature is important because the bequest motive is arguably one of the main determinants of wealth accumulation (see Nardi (2004), for example). Meaningful evaluations of the distributional consequences of tax reforms require realistic wealth distributions, but this realism should be achieved through the appropriate margins.

4.1.2 Expenditure ratios

In Table 6 we report the key expenditure ratios in the benchmark model economy and in the two reformed flat-tax model economies. Since the level of public expenditure, G , is the same in the three model economies, the G/Y ratios fall whenever output increases. Since both reforms bring about sizable increases in aggregate capital, the decreasing marginal returns to capital make the investment to output ratios increase and the consumption to output ratios fall. However, even though the C/Y ratios fall in both flat-tax reforms, it is important to highlight that in both of them aggregate consumption increases (see Column 4 in Table 6). This increase is about 2 percent larger in the consumption-based flat-tax reform than in the income-based flat-tax reform.

Table 6: Expenditure ratios in the model economies (%)

	C/Y	I/Y	G/Y	C/Y_B	I/Y_B	G/Y_B
E_B	56.4	23.0	20.8	56.4	23.0	20.8
E_C	53.8	28.0	18.5	60.6	31.6	20.8
E_Y	55.6	25.0	19.6	58.3	26.2	20.6

Note: Columns 1, 2 and 3 report aggregate consumption, investment and government expenditure as a fraction of each economy's output. Columns 4, 5 and 6 report these same magnitudes as a fraction of output in the benchmark economy.

4.1.3 Fiscal policy ratios

In Table 7 we report the main fiscal policy ratios of the model economies. We have already mentioned that in both reformed model economies the government expenditures to output ratios are smaller than in the benchmark model economy. Moreover, the tax revenue to output ratios and the transfers to output ratios of these model economies are reduced in the same proportion.

The bottom 2 rows of Table 7 display the tax ratios relative to the output in the benchmark model economy. We observe that, contrary to what Hall and Rabushka (1995) had guessed, the labor income tax in the consumption-based flat-tax reform collects much the same revenues as the personal income tax of the benchmark model economy. This is also the case in the income-based

Table 7: The Fiscal Policy Ratios in the Model Economies (%)

	G/Y	Z/Y	T/Y	T_y/Y	T_l/Y	T_k/Y	T_s/Y	T_c/Y	T_e/Y
E_B	20.8	5.1	25.9	12.0	0.0	3.0	5.5	5.2	0.21
E_C	18.5	4.5	23.0	0.0	10.8	1.8	5.2	4.9	0.31
E_Y	19.6	4.8	24.5	0.0	11.4	2.4	5.4	5.1	0.25
E_C/Y_B	20.8	5.1	25.9	0.0	12.1	2.0	5.8	5.6	0.35
E_Y/Y_B	20.6	5.1	25.7	0.0	11.9	2.5	5.6	5.4	0.26

Note: Rows 1, 2 and 3 report aggregate magnitudes as a fraction of each economy’s output. Rows 4 and 5 report these same magnitudes as a fraction of output in the benchmark economy. In interpreting this table recall that the reforms are designed to generate the same total tax revenue as existed pre-reform; thus G/Y_B , Z/Y_B , and T/Y_B are constant across the different tax systems.

flat-tax reform. The slight revenue losses of the capital income tax are compensated by the higher revenues collected by all the other tax instruments.

4.1.4 Earnings, income, and wealth inequality

In Table 8 we report the Gini indexes and the Lorenz curves of earnings, after-tax income, and wealth in the benchmark and in the reformed model economies. We find that the effects of the flat tax reforms on earnings inequality are very small, but that both reforms bring about sizable increases in after-tax income inequality and in wealth inequality. The first result is not surprising since the three model economies have identical processes on the endowments of efficiency labor units and changes in the distribution of hours worked are very small.

The higher inequality in wealth is easy to understand because the marginal taxes on capital income for the wealthy are lower after the flat-tax reforms. And this gives rich households stronger incentives to accumulate capital. The inequality in after-tax income is larger in the flat-tax economies because of the increase in the inequality in the wealth distribution and because of the lower redistributive power of flat taxes.

Table 8: The Distributions of Earnings, Income, and Wealth in the Model Economies

<i>The Distribution of Earnings</i>									
<i>Economy</i>	<i>Gini</i>	<i>Quantiles (%)</i>					<i>Top Groups (%)</i>		
		1st	2nd	3rd	4th	5th	90-50	95-99	99-100
E_B	0.646	0.0	3.0	11.6	21.3	61.2	9.0	16.7	20.2
E_C	0.643	0.0	3.1	11.5	21.6	60.9	8.9	16.7	20.0
E_Y	0.648	0.0	2.7	11.6	21.5	61.4	9.2	16.8	20.1
<i>The Distribution of Incomes (before all taxes and after transfers)</i>									
E_B	0.528	4.3	7.6	8.8	21.8	54.7	7.4	17.1	17.4
E_C	0.535	4.0	7.3	9.4	21.2	55.2	7.6	16.9	17.4
E_Y	0.532	4.2	7.4	9.1	21.4	55.0	7.6	17.1	17.3
<i>The Distribution of Wealth</i>									
E_B	0.824	0.0	0.0	3.4	12.3	82.9	14.4	23.7	34.1
E_C	0.834	0.0	0.0	2.9	12.0	83.8	13.6	24.5	36.1
E_Y	0.826	0.0	0.0	3.4	12.1	83.2	14.2	23.7	34.8

4.2 Progressivity in consumption-based flat-tax reforms

In this section we compare the allocations that obtain in the steady-states of three consumption-based flat-tax reforms. The first flat-tax reform is the least progressive of the three. In this model economy all labor income is taxed. Therefore the value of the labor income tax deduction is zero and $a_{10} = 0.0$. The integrated flat-tax rate that makes this reform revenue neutral is 18.1 percent. To keep in mind that this reform allows no deduction we call this economy E_{ND} .

The second consumption-based flat-tax reform is the standard flat-tax reform proposed by Hall and Rabushka (1995) which we have discussed in the previous section. Its marginal tax rate on capital and labor income is 19 percent, and the labor income tax deduction that makes this reform revenue neutral is $a_{10} = 0.0489$, which corresponds to 5.3 percent of output per household in the benchmark model economy or \$6,300, approximately. We continue to refer to this model economy as E_C .

The third consumption-based flat-tax reform is the most progressive of the three. In this model economy, we double the labor income tax deduction of model economy E_C . Therefore, in this model economy $a_{10} = 0.0978$, which corresponds to approximately \$12,600. The value of the integrated flat-tax rate that makes this reform revenue neutral is 19.8 percent, and we call this model economy with double the deduction E_{DD} .

4.2.1 Macroeconomic aggregates and factor ratios

In Table 9 we report the main macroeconomic aggregates and factor ratios of our three consumption-based flat-tax reforms. Relative to the benchmark model economy, we find that the three flat-tax reforms are expansionary. We also find that reforms generate large increases in the stock of capital—between 33 and 38 percent—and that the three reforms generate small changes in the labor decision. But while in model economies E_{ND} and E_C both aggregate hours and the aggregate labor input increase, in model economy E_{DD} , these two variables decrease.

Table 9: Production, inputs and input ratios in the model economies

	Y	K	L	H/ℓ	K/L	L/H	Y/H	K/Y
E_B	0.92	4.19	0.37	33.10	11.34	1.12	2.78	4.55
$E_{ND}/E_B(\%)$	11.6	33.5	0.2	0.6	33.2	-0.4	11.0	19.6
$E_C/E_B(\%)$	12.7	37.2	0.0	0.5	37.2	-0.4	12.1	21.8
$E_{DD}/E_B(\%)$	11.3	34.2	-0.7	-1.2	35.1	0.6	12.6	20.7

^aVariable L denotes the aggregate labor input.

^bVariable H denotes the share of the endowment of time allocated to the market.

But we find that the increases in the productivity of labor hours are larger in the reformed economies with higher flat-tax rates: 11.0 percent in model economy E_{ND} , 12.1 percent in model economy E_C , and 12.6 percent in model economy E_{DD} . These increases in labor productivity are due to increases both in the labor to hours ratios, L/H , and in the capital to labor ratios, K/L —see the fifth, sixth, seventh, and eighth columns of Table 9.

By definition, the increases in the L/H ratios are the result of household hours being more correlated with the endowment of efficiency labor units. This tells us that as we move towards a more progressive flat-tax system households need to provide less self-insurance. Consequently, they accumulate less precautionary savings—and hence the stock of capital is lower—and they use

less precautionary hours —and hence people work less, but labor hours become more correlated with productivity. This makes the allocations more similar to the ones that would obtain under complete markets.³⁹

The increases in the K/L ratios are ultimately due to the same reason: the reduction in hours reduces the aggregate labor input and, therefore, capital per efficiency unit of labor increases. These results lead us to conclude that the fixed deduction in labor income makes labor hours more productive, and that it improves the allocation of the work effort. In economies with higher flat-tax rates, people end up working less on average, but they work more when they are more productive.

4.2.2 Expenditure ratios

In Table 10 we report the key expenditure ratios in the benchmark model economy and in the three reformed flat-tax model economies. Since the flat tax reforms are expansionary and the levels of government expenditures do not change, the G/Y ratios in the flat-tax model economies are smaller than in the benchmark model economy. The lower G/Y shares are compensated with large increases in the investment to output ratios, because the consumption to output ratios also decrease. These results are consistent with the large increases in the capital stock which we have discussed in Section 4.2.1 above.

Table 10: Expenditure Ratios in the Model Economies (%)

	C/Y	I/Y	G/Y	C/Y_B	I/Y_B	G/Y_B
E_B	56.4	23.0	20.8	56.4	23.0	20.8
E_{ND}	54.5	27.5	18.3	60.9	30.7	20.4
E_C	53.8	28.0	18.5	60.6	31.6	20.8
E_{DD}	54.3	27.8	18.2	60.5	30.9	20.3

Note: Columns 1, 2 and 3 report aggregate consumption, investment and government expenditure as a fraction of each economy’s output. Columns 4, 5 and 6 report these same magnitudes as a fraction of output in the benchmark economy.

We also find that aggregate consumption increases in the three consumption-based flat-tax reforms. We find that the differences in consumption and investment —and in their ratios to output— brought about by differences in the progressivity of the flat-tax reforms are small.

4.2.3 Fiscal policy ratios

In Table 11 we report the main fiscal policy ratios in the benchmark model economy and in the three reformed flat-tax model economies. In the three reforms total government revenues, T , government consumption, G , and total transfers, Z , do not change, and hence their ratios to output fall.

When we compare the changes in the composition of government revenues, we confirm that in every consumption-based flat-tax reform the labor income tax and collects less revenues than the personal income tax of the benchmark model economy (see bottom 3 rows in Table 11). Likewise, capital income taxes collect less revenues in the three flat-tax model economies. In contrast, payroll and consumption taxes collect more revenues in the three flat-tax model economies,

³⁹See Pijoan-Mas (2006) for an analysis of the interaction of work effort and savings as self-insurance mechanisms, and for a comparison of capital and labor allocations in complete and incomplete-market economies.

Table 11: The Fiscal Policy Ratios in the Model Economies (%)

	G/Y	Z/Y	T/Y	T_y/Y	T_l/Y	T_k/Y	T_s/Y	T_c/Y	T_e/Y
E_B	20.8	5.1	25.9	12.0	0.0	3.0	5.5	5.2	0.21
E_{ND}	18.3	4.6	22.8	0.0	10.9	1.4	5.2	5.0	0.30
E_C	18.5	4.5	23.0	0.0	10.8	1.8	5.2	4.9	0.31
E_{DD}	18.2	4.6	22.8	0.0	10.6	1.7	5.2	5.0	0.30
E_{ND}/Y_B	20.4	5.1	25.5	0.0	12.2	1.6	5.8	5.6	0.33
E_C/Y_B	20.8	5.1	25.9	0.0	12.1	2.0	5.8	5.6	0.35
E_{DD}/Y_B	20.3	5.1	25.3	0.0	11.8	1.9	5.7	5.5	0.34

Note: Rows 1 to 4 report aggregate magnitudes as a fraction of each economy’s output. Rows 5 to 7 report these same magnitudes as a fraction of output in the benchmark economy.

Revenues from the consumption tax decrease with the progressivity of the reform because the consumption tax rate remains unchanged and aggregate consumption—which is the tax base—decreases as the flat-tax rates increase (see Table 10). The same is true for the payroll tax: the tax rate does not change, and the tax base, which is essentially aggregate labor income, decreases with the flat-tax rate.⁴⁰

The capital income tax raises less revenue in all three of the flat-tax reforms. The substantial increase in the capital stock is not enough to compensate for the lower tax rates and the investment-expenditures deduction.

4.2.4 Earnings, income, and wealth inequality

In Table 12 we report the Gini indexes and the Lorenz curves of earnings, income, and wealth of the benchmark model economy and of the consumption-based flat-tax reforms. The Gini index of earnings falls under all three flat-tax reforms. By contrast the Gini indexes of wealth and income rise under all three flat-tax reforms. Interestingly the Gini indexes for earnings, income, and wealth are all increasing in the progressivity of the reforms — their direct effects on progressivity are outweighed by a decreased use of precautionary labor and savings. We conclude that the flat-tax reforms bring about increases in inequality and that, overall, the distributional role played by the progressivity of consumption-based flat taxes is small.

Earnings inequality increases as the flat tax reform becomes more progressive because higher flat-tax rates and higher deductions increase the correlation between wages and work effort (see the discussion in Section 4.2.1). Since this implies that labor income becomes more volatile, households transfer income between periods using larger buffer stocks of precautionary savings to smooth their consumption profiles. This implies that wealth inequality also increases.

5 Concluding Comments

Hall and Rabushka (1995) claimed that revenue-neutral consumption-based flat-tax reforms would be expansionary and that the tax exemption in their proposed labor income tax could be used to

⁴⁰Indeed, the tax base of the payroll tax is not exactly the aggregate labor income as labor income above the threshold a_3 is exempt and changes in the distribution of labor earnings make the exact fraction of untaxed labor income different in different economies.

Table 12: The Gini Indexes and the Lorenz Curves in the Model Economies

<i>The Distribution of Earnings</i>									
	<i>Gini</i>	<i>Quantiles (%)</i>					<i>Top Groups (%)</i>		
<i>Economy</i>		1st	2nd	3rd	4th	5th	90-50	95-99	99-100
E_B	0.646	0.0	3.0	11.6	21.3	61.2	9.0	16.7	20.2
E_{ND}	0.644	0.0	3.1	11.5	21.5	61.0	9.0	16.7	20.1
E_C	0.643	0.0	3.1	11.5	21.6	60.9	8.9	16.7	20.0
E_{DD}	0.646	0.0	3.2	11.2	21.5	61.2	8.9	16.8	20.2
<i>The Distribution of Incomes (before all taxes and after transfers)</i>									
E_B	0.528	4.3	7.6	8.8	21.8	54.7	7.4	17.1	17.4
E_{ND}	0.535	4.0	7.3	9.4	21.1	55.3	7.7	16.8	17.5
E_C	0.535	4.0	7.3	9.4	21.2	55.2	7.6	16.9	17.4
E_{DD}	0.537	4.1	7.3	9.1	21.2	55.5	7.6	17.0	17.5
<i>The Distribution of Wealth</i>									
E_B	0.824	0.0	0.0	3.4	12.3	82.9	14.4	23.7	34.1
E_{ND}	0.833	0.0	0.0	3.0	12.0	83.8	13.7	24.5	35.7
E_C	0.834	0.0	0.0	2.9	12.0	83.8	13.6	24.5	36.1
E_{DD}	0.835	0.0	0.0	2.9	11.9	83.9	13.6	24.5	36.1

achieve certain distributional targets. Our results confirm that consumption-based flat-tax reforms can indeed generate large gains in output, but that they do so at the expense of increases in the inequality of after-tax income and wealth. These findings are consistent with those reported in the flat-tax literature.

We find that the differences in the allocations that obtain in consumption-based and in income-based flat-tax reforms can be large. This tells us that the role played by the expensing of investment is important. Indeed, we show that it accounts for two thirds of the output increases brought about by the reforms, and that it would increase the Gini index of the after-tax income distribution well beyond the value that would obtain in a purely income-based flat-tax reform.

Work is currently underway on this project to calculate the transition paths between steady-states and use to perform welfare analysis of the winners and losers from flat-tax reforms.

References

- Henry Aaron and Alicia Munnell. Reassessing the role for wealth transfer taxes. *National Tax Journal*, 45:119–143, 1992.
- S. Rao Aiyagari. Uninsured idiosyncratic risk and aggregate saving. *Quarterly Journal of Economics*, 109(3):659–684, 1994.
- S. Rao Aiyagari. Optimal capital income taxation with incomplete markets, borrowing constraints, and constant discounting. *Journal of Political Economy*, 103(6):1158–1175, 1995.
- Altig, Auerbach, Kotlikoff, Smetters, and Walliser. Simulating fundamental tax reform in the united states. *American Economic Review*, 91(3):574–595, 2001.
- Martin Andreasen. How to maximize the likelihood function for a dsge model. *Computational Economics*, 35(2):127–154, 2010.

- Truman Bewley. A difficulty with the optimum quantity of money. *Econometrica*, 51:1485–1504, 1983.
- Blundell and MaCurdy. Labor supply of men: A review of alternative approaches. In Ashenfelter and David Card, editors, *Handbook of Labor Economics*, volume 3. 1999.
- Ana Castaneda, Javier Díaz-Giménez, and Jose Victor Ríos-Rull. Accounting for the u.s. earnings and wealth inequality. *Journal of Political Economy*, 111(4):818–857, 2003.
- Chamley. Optimal taxation of capital income in general equilibrium with infinite lives. *Econometrica*, 54:607–622, 1986.
- Thomas Cooley and Ed Prescott. Economic growth and business cycles. In Thomas Cooley, editor, *Frontiers of Business Cycle Research*, chapter 1. Princeton University Press, 1995.
- Peter Diamond and Emmanuel Saez. The case for a progressive tax: From basic research to policy recommendations. *Journal of Economic Perspectives*, 25(4):165–190, 2011.
- Javier Díaz-Giménez, Andy Glover, and Jose Victor Ríos-Rull. Facts on the distributions of earnings, income, and wealth in the united states: 2007 update. *Quarterly Review of the Federal Reserve of Minneapolis*, 34(1):2–31, 2011.
- David Domeij and Jonathan Heathcote. On the distributional effects of decreasing capital taxes. *International Economic Review*, 45(2):523–544, 2004.
- Sean Flynn. Us effective tax functions 1979-2005, 2009. URL https://files.nyu.edu/smf354/public/flynn_2012_eff_tax_fun.pdf.
- Gentry and Glenn Hubbard. Distributional implications of introducing a broad-based consumption tax. *NBER Working Paper 5482*, 1997.
- Mark Gertler. Government debt and social security in a life-cycle economy. *Carnegie-Rochester Series on Public Policy*, 50:61–110, 1999.
- Gouveia and Strauss. Effective federal individual income tax functions: An exploratory empirical analysis. *National Tax Journal*, 47(2):317–339, 1994.
- Gouveia and Strauss. Effective functions for the us individual income tax: 1966-89. In *Proceedings of the 1999 National Tax Association*, 1999.
- Nezih Guner, Remzi Kaygusuv, and Gustavo Ventura. Taxation and household labor supply. *Review of Economic Studies*, 79(3):1113–1149, 2012.
- Robert E. Hall and Alvin Rabushka. *The Flat Tax (Second Edition)*. Hoover Institution Press, 1995.
- Jonathan Heathcote, Kjetil Storesletten, and Giovanni Violante. The macroeconomic implications of rising wage inequality in the united states. *Journal of Political Economy*, 118(4):681–722, 2010.
- Hugo Hopenhayn and Edward C. Prescott. Stochastic monotonicity and stationary distributions for dynamic economies. *Econometrica*, 60(6):1387–1406, 1992.
- Glenn Hubbard. How different are income and consumption taxes? *American Economic Review Papers and Proceedings*, 87(2):138–142, 1997.

- Glenn Hubbard, Jonathan Skinner, and Stephen Zeldes. Precautionary saving and social insurance. *Journal of Political Economy*, 103(2):360–399, 1995.
- Mark Huggett. The risk-free rate in heterogenous-agent incomplete-insurance economies. *Journal of Economic Dynamics and Control*, 17:953–969, 1993.
- Kenneth Judd. Redistributive taxation in a simple perfect foresight model. *Journal of Public Economics*, 28:59–83, 1985.
- Juster and Stafford. The allocation of time: Empirical findings, behavioral models, and problems of measurement. *Journal of Economic Literature*, XXIX(2):471–522, 1991.
- Richard Kasten, Frank Sammartino, and Eric Toder. Trends in federal tax progressivity — 1980-93. In Joel Slemrod, editor, *Tax Progressivity and Income Inequality*, pages 9–50. Cambridge University Press, 1994.
- Edward Lazear and James Porterba. Reforming taxes to promote economic growth. *The Economists' Voice*, 3(1):1–7, 2005.
- Robert E. Lucas. Supply-side economics: An analytical review. *Oxford Economic Papers*, 42(2):292–316, 1990.
- James Mirrlees. An exploration in the theory of optimum income taxation. *Review of Economic Studies*, 38(2):175–208, 1971.
- De Nardi. Wealth inequality with intergenerational linkages. *Review of Economic Studies*, 71(3):743–768, 2004.
- Josep Pijoan-Mas. Precautionary savings or working longer hours? *Review of Economic Dynamics*, 9(2):326–352, 2006.
- José-Víctor Ríos-Rull, Frank Schorfheide, Cristina Fuentes-Albero, Maxym Kryshko, and Raul Santaeulalia-Llopis. Methods versus substance: Measuring the effects of technology shocks. *Journal of Monetary Economics*, 59(8):826–846, 2012.
- Paul Samuelson. Optimum social-security in a life-cycle growth model. *International Economic Review*, 16(3):539–544, 1975.
- Gary Solon. Intergenerational income mobility in the united states. *American Economic Review*, 82(3):393–406, 1992.
- Eric Swanson. Risk aversion and the labor margin in dynamic equilibrium models. *American Economic Review*, 102(4):1663–91, 2012.
- Gustavo Ventura. Flat tax reform: A quantitative exploration. *Journal of Economic Dynamics and Control*, 23:1425–1458, 1999.
- Zimmerman. Regression towards mediocrity in economic stature. *American Economic Review*, 3:409–429, 1992.

A Hall and Rabushka

The flat-tax reform modeled in this paper is based on that laid out by Hall and Rabushka (1995). Here we describe the main differences between their reform and the one modelled here. We also describe a number of issues they address that we do not cover here. The flat-tax proposal of Hall & Rabushka is fully detailed in their book, including a discussion of the issues of efficiency and fairness, and with an appendix detailing a full legislative proposal for the flat-tax. We recommend anyone interested in more detail on how such a reform would work and further discussion of the issues — economic, fairness, and legal — to read Hall and Rabushka (1995) *The Flat-Tax Reform*.

The flat-tax reform of Hall and Rabushka (1995) is only intended to replace the personal income tax and the corporate income tax. Thus, in modelling the tax reform we leave unchanged the (social security and medicare) payroll tax, the estate tax, various excise taxes, and state-level consumption taxes. We observe that, (i) the state-level consumption taxes are similar to the flat-tax with which they are being replaced⁴¹, and (ii) other than the payroll tax, the others (estate and excise taxes), are a tiny fraction of total tax revenue.

Hall and Rabushka (1995) also discuss a number of other advantages to introducing a flat-tax that, for various reasons, are not addressed in our analysis. Let's quickly describe some of them, again the interested reader can find more in their book.

- First, it is estimated that the Americans spend well over 1 billion hours per year on tax compliance, at a cost of over \$100 billion; the flat-tax involves a one-page tax return and is likely to save almost all of this time and money.⁴²
- Second, we consider neither tax evasion (illegal) nor tax avoidance (legal). Tax evasion is thought to cost over \$100 billion, mostly due to people not declaring income, and leads to the Internal Revenue Service spending over \$10 billion per year on detecting and prosecuting tax evasion. Hall & Rabushka claim that by reducing tax rates the flat-tax reform would reduce (illegal) tax evasion. Tax avoidance is mostly about claiming tax deductions, the cost to taxpayers is also known as tax expenditures. The main tax expenditures are for homeownership (the mortgage interest deduction), charitable donations, and state-and-local taxes; others include the tax-exemptions given to employer-provided health insurance, medicare income, accelerated depreciation of investments, and the imputed rent homeowners receive from living in their own home. The total cost of tax-expenses is over \$2 trillion.⁴³ By eliminating all tax deductions the flat-tax reform abolishes tax avoidance, and removing this distortion will result in a more efficient economy; we do not model this aspect.
- Third, under the flat-tax reform there is horizontal equity of taxation — two people in exactly the same situation will have to pay exactly the same amount in taxes. This is not the case under the current US tax system in which the amount of tax paid depends on things like which tax deductions are claimed, and, eg., whether it is possible to disguise some labor income as capital income. In our model there is perfect horizontal equity both before and after the flat-tax reform.

⁴¹They differ in the tax rate, in not having a tax-exempt threshold, and in not having investment-expensing.

⁴²These numbers are based on studies that pre-date the use of software to file tax returns, how much difference that makes is not known.

⁴³An up-to-date introduction to tax-expenditures is provided by [this newspaper article](#) by Bruce Barlett, and in a [related series of articles](#) on some of the largest tax-expenditures.

- It would no longer be possible to disguise income sources; eg. to get paid in stock options (taxed as capital) instead of being paid a wage (taxed as income). The same applies for shifting the year in which income is declared.

In short, the adoption of a flat-tax reform may lead to benefits relating to being, (i) easier for taxpayers to comply with and tax authorities to administer, (ii) reducing tax-evasion, (iii) improve economic efficiency by eliminating distortions related to tax-avoidance, (iv) improve the horizontal equity of taxation. None of these advantages appear in our analysis.

Some other issues

One thought on Hall & Rabushka's flat-tax reform that they do not address in their book: it appears to bias towards investing physical capital, rather than in human- or intangible-capital.

Dealing with existing capital depreciation deductions during the transition? Hall & Rabushka suggest that existing deductions may be allowed for during the transition, we do not consider this issue. They suggest something similar with regard to the home mortgage interest deduction.

Variants of the Flat-Tax (different threshold-rate combinations, different amounts of investment expensing): Hall and Rabushka (1995), locations 1059-57.

"Capital gains on owner-occupied houses are not taxed under our proposal" (as far as I can tell owner imputed rent is not either)

"Because it is high-income taxpayers who have the biggest incentive and the best opportunity to use special tricks to exploit tax rate differentials, applying the same tax rate to these taxpayers for all their income in all years is the most important goal of flat-rate taxation."

"A total of \$1,709 billion in business income was earned in the United States in 1991, but only \$791 billion in business income was reported on individual returns that year. The chance that a dollar of business income would actually be reported was less than half." Under the flat-tax it would all be taxed. We do not capture this in the model.

Hall & Rabushka on the basic idea of the flat-tax:

"Here is the logic of our system, stripped to basics: We want to tax consumption. The public does one of two things with its income — spends it or invests it. We can measure consumption as income minus investment. A really simple tax would just have each firm pay tax on the total amount of income generated by the firm less that firm's investment in plant and equipment. The value-added tax works just that way. But a value-added tax is unfair because it is not progressive. That's why we break the tax in two. The firm pays tax on all the income generated at the firm except the income paid to its workers. The workers pay tax on what they earn, and the tax they pay is progressive."

To measure the total amount of income generated at a business, the best approach is to take the total receipts of the firm over the year and subtract the payments the firm has made to its workers and suppliers. This approach guarantees a comprehensive tax base. The successful value-added taxes in Europe work this way. The base for the business tax is the following:

Total revenue from sales of goods and services less purchases of inputs from other firms less wages salaries, and pensions paid to workers less purchases of plant and equipment

The other piece is the wage tax. Each family pays 19 percent of its wage, salary, and pension

income over a family allowance (the allowance makes the system progressive). The base for the consumption tax is total wages, salaries, and retirement benefits less the total amount of family allowances.”

B The Transition Matrix on Exogenous Shocks

This appendix explains the definition of parameters ϕ_1 and ϕ_2 , and how they affect the transition matrix. Let p_{ij} denote the transition probability from $i \in \mathcal{R}$ to $j \in \mathcal{E}$, let γ_i^* be the invariant measure of households that receive shock $i \in \mathcal{E}$, and let ϕ_1 and ϕ_2 be the two parameters that shift the probability mass towards the diagonal and towards the first column of submatrix $\Gamma_{\mathcal{E}\mathcal{E}}$ ⁴⁴, then the recursive procedure that we use to compute the p_{ij} is the following:

- *Step 1:* First, we use parameter ϕ_1 to shift the probability mass from a matrix with vector $\gamma_{\mathcal{E}}^* = (\gamma_1^*, \gamma_2^*, \gamma_3^*, \gamma_4^*)$ in every row towards its diagonal, as follows:

$$\begin{aligned}
p_{51} &= \gamma_1^* + \phi_1 \gamma_2^* + \phi_1^2 \gamma_3^* + \phi_1^3 \gamma_4^* \\
p_{52} &= (1 - \phi_1)[\gamma_2^* + \phi_1 \gamma_3^* + \phi_1^2 \gamma_4^*] \\
p_{53} &= (1 - \phi_1)[\gamma_3^* + \phi_1 \gamma_4^*] \\
p_{54} &= (1 - \phi_1)\gamma_4^* \\
p_{61} &= (1 - \phi_1)\gamma_1^* \\
p_{62} &= \phi_1 \gamma_1^* + \gamma_2^* + \phi_1 \gamma_3^* + \phi_1^2 \gamma_4^* \\
p_{63} &= (1 - \phi_1)[\gamma_3^* + \phi_1 \gamma_4^*] \\
p_{64} &= (1 - \phi_1)\gamma_4^* \\
p_{71} &= (1 - \phi_1)\gamma_1^* \\
p_{72} &= (1 - \phi_1)[\phi_1 \gamma_1^* + \gamma_2^*] \\
p_{73} &= \phi_1^2 \gamma_1^* + \phi_1 \gamma_2^* + \gamma_3^* + \phi_1 \gamma_4^* \\
p_{74} &= (1 - \phi_1)\gamma_4^* \\
p_{81} &= (1 - \phi_1)\gamma_1^* \\
p_{82} &= (1 - \phi_1)[\phi_1 \gamma_1^* + \gamma_2^*] \\
p_{83} &= (1 - \phi_1)[\phi_1^2 \gamma_1^* + \phi_1 \gamma_2^* + \gamma_3^*] \\
p_{84} &= \phi_1^3 \gamma_1^* + \phi_1^2 \gamma_2^* + \phi_1 \gamma_3^* + \gamma_4^*
\end{aligned}$$

⁴⁴A detailed description of this probability mass shifting procedure can be found in Castaneda, Díaz-Giménez, and Ríos-Rull (2003).

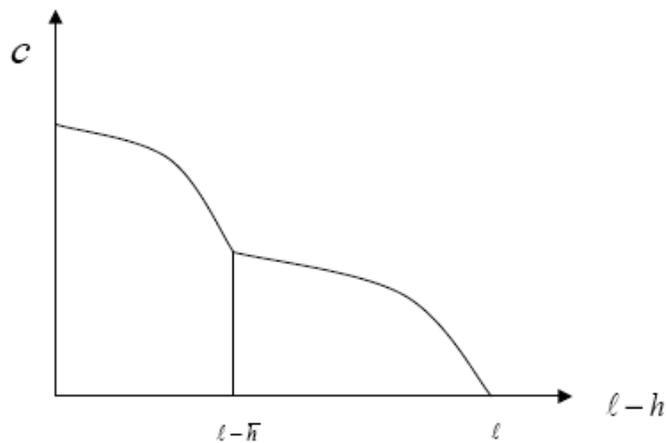
- *Step 2:* Then for $i = 5, 6, 7, 8$ we use parameter ϕ_2 to shift the resulting probability mass towards the first column as follows:

$$\begin{aligned}
p_{i1} &= p_{i1} + \phi_2 p_{i2} + \phi_2^2 p_{i3} + \phi_2^3 p_{i4} \\
p_{i2} &= (1 - \phi_2)[p_{i2} + \phi_2 p_{i3} + \phi_2^2 p_{i4}] \\
p_{i3} &= (1 - \phi_2)[p_{i3} + \phi_2 p_{i4}] \\
p_{i4} &= (1 - \phi_2)p_{i4}
\end{aligned}$$

C Non-convexities

Due to the upper cap in payroll taxes, the marginal tax on labor income has a discontinuity at the income level where the cap is reached. This creates a serious problem when we try to find the optimal household policy. Specifically, for a given value of the choice of end-of-period period assets, z , the budget set of the contemporaneous labor decision becomes non-convex. In Figure 1 we illustrate this point. Consider pair of individual state variables (a, s) and a choice of end-of-period assets, z . Then, equations (7), (8) and (9) and the boundary constraints on c and h define the consumption possibilities set for c and $\ell - h$. In Figure 1 we plot an example of this set for $a = 0$. When the household chooses not to work and to enjoy ℓ units of leisure, its consumption is zero. As the household starts to work, its consumption increases albeit at a decreasing rate. This is because of the progressivity of the personal income tax, τ_y , which reduces the after-tax wage of every extra hour of work. Let \bar{h} be the hours of work such that $e(s)\bar{h}w = a_3$. For $h > \bar{h}$ the marginal payroll tax is zero. Therefore the slope of the consumption possibilities set increases discretely at $h = \bar{h}$ and from that point onwards it decreases monotonically as we increase h , again because of the progressivity of τ_y .

Figure 1: Non-convex constraints



This lack of convexity is twice unfortunate. First, because it implies that the first order necessary conditions are no longer sufficient for optimality and, therefore, they do not identify the optimal solution uniquely. In fact there are two points that potentially satisfy the first order conditions,

one above and one below the threshold \bar{h} , and only one of these points is the optimal solution. Second, as we change the choice of end-of-period assets, z , the optimal choice of hours becomes discontinuous exactly when we move from a solution on one side of \bar{h} to a solution on the other side of \bar{h} . This is much more troublesome for our computational procedure. And it forces us to solve the household decision problem using discrete value function iterations which are much more computationally intensive, than the Euler equation iterations which can only be used when the choice sets are convex.

D Calibration

The model has 43 parameters (actually, there are 46 if we include the three parameters that are needed to perform the tax-reform experiments, but these 'extra' three parameters are not relevant to the calibration). A full description of the parameters and how they are calibrated/estimated is contained in the body of the paper. For convenience we here provide a complete list of parameters and then list which ones are normalizations/calibrated/estimated.

A full list of the parameters:

5 parameters to describe the preferences of the household

$$\beta, \sigma_1, \sigma_2, \chi, \ell$$

2 parameters for production technology

$$\theta, \delta$$

11 parameters for the government policy

$$G, \omega$$

$$a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9$$

Following 3 are "spare" parameters on gov. policy (spare in the sense they are only relevant for the tax reform policy experiment and so not part of the calibration)

$$a_{10}, a_{11}, \text{investmentexpenditures_taxexempt}$$

25 parameters the joint process on age & efficiency labour units

J

$$p_{eg}, p_{gg}$$

$$\phi_1, \phi_2$$

5 so far, 20 more

$$e=[e1, e2, e3, e4, 0,0,0,0];$$

$$(4 \text{ here: } e1, e2, e3, e4)$$

$$\Gamma_{ee} = [\gamma_{11}, \gamma_{12}, \gamma_{13}, \gamma_{14}; \gamma_{21}, \gamma_{22}, \gamma_{23}, \gamma_{24}; \gamma_{31}, \gamma_{32}, \gamma_{33}, \gamma_{34}; \gamma_{41}, \gamma_{42}, \gamma_{43}, \gamma_{44}];$$

$$(\text{and } 16 \text{ here})$$

Normalizations, 6 Parameters: $J=4$, $\ell=1$, and four normalizations on the rows of Γ_{ee} (so that each row adds to one).

Directly Identified, 8 Parameters: $\sigma_1, a_2, a_4, a_5, p_{eg}, p_{gg}, \theta, \delta$.

Estimated by Simulated Method of Moments, 29 Parameters:

Those for preferences, government, and taxation:

$$\beta, \chi, \sigma_2$$

$$G, \omega$$

$$a_1, a_3, a_6, a_7, a_8, a_9$$

And 17 parameters for the joint process on age & efficiency labour units:

ϕ_1, ϕ_2
 $e=[e1, e2, e3, e4, 0,0,0,0]$ (4 here)
 $\Gamma_{ee} = [\gamma_{11}, \gamma_{12}, \gamma_{13}, \gamma_{14}; \gamma_{21}, \gamma_{22}, \gamma_{23}, \gamma_{24}; \gamma_{31}, \gamma_{32}, \gamma_{33}, \gamma_{34}; \gamma_{41}, \gamma_{42}, \gamma_{43}, \gamma_{44}]$;
 (and 12 here, as we have to do the four normalizations, one for each row, in codes the twelve are all the non-diagonal elements)

Remark: The choice to use the diagonals of Γ_{ee} as the elements to be normalized has an important advantage in the computation. Since the non-diagonals are smaller by having the diagonals given by whatever was leftover to make the row sum up to one we avoided the problem that they may end up being negative — something that occurred in an earlier version of the codes where we used the last element of each row for the normalization.

E Computation

This appendix describes the computation, first of the calibration, and then how we calculate the transition paths.

All simulation exercises involved a burn-in of 1000 points (typically starting from the 'mid-point' of the relevant distribution).

E.1 Value Functions and Stationary Distribution

To calculate the decision rules, we discretize the state space and perform Value function iteration using Howard's improvement algorithm.

The size of our state space is $n_k \times n_s = 681 \times 8 = 5,448$ points. The size of our control space is $n_k \times n_n = 681 \times 51 = 34,731$ points. Since the numbers of working-age and retirement states are $n_w = n_r = 4$, the total number of search points is $[(n_k \times (n_w + n_r)) \times (n_k \times n_n)] = 189,214,488$ points.

We approximate the stationary distribution with a discretization of the associated distribution function. The grid for this approximation is the same as that used for the to solve the value function. The stationary distribution is calculated by iterating on the whole distribution, using the optimal policy functions and the transition matrix of the idiosyncratic shocks. This is done until it converges, as measured by a distance criterion based directly on the monotone mixing condition underlying the theory that ensures that a stationary distribution exists (see Hopenhayn and Prescott (1992)).⁴⁵ This process is more demanding computationally than those typically used for calculating the stationary distribution, but that is important here since the model moments relating to the top percentiles, eg. the asset share of the top 1% of asset holders, otherwise varied substantially between different simulations (much smaller simulations were fine for giving stable results for first moments, such as the capital stock, but the top percentile moments can be quite volatile).

⁴⁵To speed up the convergence we start this process of iterating on the stationary distribution from an initial distribution created by a 1,000,000 point simulation. Actually, to take advantage of parallelization this was implemented as *ncore* simulations of 1,000,000/*ncore* points each, where *ncore* = 12 was the number of cores in the computer we had access to.

E.2 Model Moments/Statistics

The model economy’s distributional and aggregate statistics can almost all be computed directly as integrals with respect to the stationary distribution. Since the distribution is approximated as a weight for each point on a grid this just involves taking weighted sums. The exceptions are those that measure the earnings life cycle and the intergenerational correlation of earnings the computations of which we now describe.

Life-cycle profile of earnings: The life-cycle profile of earnings is measured as the ratio of ‘average earnings of households aged 46 to 50’ to ‘average earnings of households aged 26 to 30’. To calculate this statistic we first draw a random ‘newborn’ from the distribution of newborns.⁴⁶ We then simulate this individual for 30 periods (ie. until age 50) recording their productivity in each year and recording both their ‘average earnings of households aged 46 to 50’ to ‘average earnings of households aged 26 to 30’. We do this for 30,000 individuals, drop all of those households which retired before reaching age 50, and then calculate the average ratio across the remaining individuals. We use 30,000 as this ensured that we always ended up with well in excess of 10,000 individuals after dropping all of those individuals who retired; this was enough to ensure that the statistic was stable from one sample to the next.⁴⁷

Intergenerational Correlation of Earnings: The Intergenerational Correlation of Earnings is measured as the correlation between the average annual earnings of two consecutive generations of the same dynasty/household. To calculate this statistic we first draw a random ‘newborn’ from the distribution of newborns (see footnote 46). We then simulate this household, recording it’s annual earnings, until it ‘dies’ twice. From this we calculate the average annual earnings for the first and second generations of this household. This is done for 10,000 households and we then calculate the correlation between average annual earnings of the first and second generations. 10,000 households was enough to ensure stability of this statistic.⁴⁸

E.3 General Equilibrium

The calculation of general equilibrium in this class of models involves finding an interest rate which induces individual behaviour which generates aggregate variables (eg. output and capital) than in turn leads back to the original interest rate; see eg. Aiyagari (1994). The only noteworthy difference in our algorithms is that rather than use a search algorithm on K/Y to find the general equilibrium — the standard approach — we instead discretize the state space for r and use this to find the equilibrium value — the one for which r induces individual behaviour, which generates aggregate variables, that in turn imply the original r . Using a grid allows us to be certain of convergence, and to know if the model were to have multiple solutions — both theoretically uncertain issues with this class of models.⁴⁹

⁴⁶In practice this is implemented as drawing a random retired household, forcibly killing them, and then determining where they would end up as a newborn. Since the probability of death is equal for all retired households this is equivalent to drawing randomly from the distribution of newborns, but saves having to actually calculate the distribution of newborns.

⁴⁷In implementing the code we parallelized across the 30,000 individuals.

⁴⁸In implementing the code we parallelized across the 10,000 individuals.

⁴⁹Many of the optimization algorithms normally applied for this step rely on differentiability, and concavity for convergence, neither of which is known to hold. They also assume that the solution found is a global, rather than simply local, solution.

E.4 Calibration

As we have mentioned in Section 3 the model has 43 parameters. Of these 6 are normalizations, and another 8 are directly identified. This leaves 29 parameters. We estimate these remaining 29 parameters using the Simulated Method of Moments; we find values for the 29 parameters that minimize the distance between 29 moments of the model and the same 29 moments for the US economy. We now describe the implementation of the Simulated Method of Moments for the 29 parameters.

The model is a general equilibrium set-up. Note that since aggregate production is given by a Cobb-Douglas production function, and because the assumption of perfect competition implies that the interest rate equals the marginal product of capital, it is possible to identify the interest rate in terms of K/Y , θ , and δ .⁵⁰ Thus, given our target for K/Y , and since θ and δ are directly identified, we can calculate what the value of the interest rate must be in equilibrium. We exploit this in our calibration process: taking the interest rate as an input, the target value of K/Y becomes in effect the general equilibrium condition. By putting a large weight on the K/Y moment we thus, in effect, insist on the general equilibrium condition. This avoids the need to loop over the calculation of the general equilibrium condition during the calibration process. After the calibration process is completed we then calculate the general equilibrium given the calibrated parameters; this is important as our calculation of the transition paths is about the movement from one general equilibrium to another.

- *Step 1:* We choose a vector of weights, one for each of the 29 moments. These weights measure the relative importance that we attach to each one of our targets. These weights are reported in Table 15
- *Step 2:* We guess an initial value for the 29 unknowns (in implementing this step most of our initial values were based on the calibration results of Castaneda, Díaz-Giménez, and Ríos-Rull (2003)). These initial values are reported in Tables 13 & 14.
- *Step 3:* We compute the optimal policy function and the stationary distribution of households (given the interest rate and parameter values).
- *Step 4:* We compute the values of the 29 moments of the model (given the interest rate and parameter values).
- *Step 5:* Roughly speaking, if the weighted distance of the moments of the model from the moments of the data is small enough our calibration is complete. If not then we choose new values for the parameters and return to Step 3. (More precisely, we use the CMA-ES algorithm, see below.)
- *Step 6:* Having estimated the parameters, we calculate the general equilibrium of the model.

The loop to calibrate the parameters by matching the moments of the model to the moments of the data is implemented using the CMA-ES algorithm (Covariance-Matrix Adaptation – Evolutionary Strategy; see Andreasen (2010) who also provides a Matlab implementation of the algorithm). The use of this algorithm was key; many inbuilt Matlab optimization functions (`fgoalattain`, `fmin-`

⁵⁰Specifically, $r = \theta K^{\theta-1} L^{1-\theta} - \delta = \theta \frac{1}{K/Y} - \delta$.

search, `fminunc`, & `fmincon`; based on, eg., Nelder-Mead simplex, quasi-newton, and trust-region algorithms) simply failed to converge.⁵¹

We now provide a brief description of the CMA-ES algorithm, see Andreasen (2010) for details: the CMA-ES algorithm works by starting out considering the entire parameter space. Parameter vectors are drawn at random (based on the covariance-matrix and a focal-point) and evaluated, based on these evaluations the covariance-matrix and focal-point are updated. As the algorithm progresses the average distance between the parameter vectors drawn and the focal-point is progressively reduced. Once certain convergence criterion are met the focal-point is returned as the estimated value of the true parameter vector.

E.4.1 Calibration Weights

A brief discussion of our choices for the calibration weights is in order. As a default we put a weight of one on each moment, the exceptions to this are as follows. A large weight is put on the capital-output ratio, this is important as this represents our 'general equilibrium target'. We put smaller weights on life-cycle profile of earnings and the intergenerational correlation of earnings as there are not such clean mappings from model to data in terms of this moments (the first due to our use of stochastic aging, the second since in the data it is just measured as father-son correlation). We put a smaller weight on the ratio of government spending to output since it is the 'leftover' difference between the transfer to output ratio and the tax revenue to output ratio, both of which are closely related to other targets. We also put reduced weights on the moments relating to the earnings and wealth distributions on the grounds that these already account for around half of the moments.

E.5 General Equilibrium for a Revenue Neutral Reform

As when trying to find the general equilibrium for the baseline model we discretize the state space for r and use this to find the equilibrium value — the one for which r induces individual behaviour, which generates aggregate variables, that in turn imply the original r . The only added complication is that we now require the reform to be revenue neutral, thus for each value of r we find the tax rate (or tax exemption level) that makes the reform revenue neutral. This is done using Nelder-Mead simplex methods to minimize the square of the difference between actual tax revenue and target tax revenue (the later being the tax revenue in the baseline model); this is implemented using Matlab's `fminsearch`.

E.6 Transition Paths

We now describe the computation of the transition paths. Note that these transition paths are fully general equilibrium. There are two main aspects to this process. The first is to ensure that they are general equilibrium; that the prices are causing individual behaviour, which in turn determines

⁵¹In unpublished work we compared the performance of all these algorithms in performing simulated moment estimation of the model of Pijoan-Mas (2006) (a general equilibrium heterogeneous agent model with idiosyncratic but no aggregate uncertainty; 6 parameters and 6 moments, simple enough that calibration can be used to get the exact values for all the parameters). All of the algorithms (inbuilt Matlab and the CMA-ES) performed fine when the initial guesses for the parameters were close to the true values. But only the CMA-ES was able to reliably converge to the true parameters when the initial guesses were some distance from the true values.

Table 13: Parameter values for the benchmark model economy and their initial values

		Calibrated Value	Initial Value
<i>Preferences</i>			
Time discount factor	β	0.95	0.94
Curvature of consumption	σ_1	1.50	n.a.
Curvature of leisure	σ_2	1.83	1.50
Relative share of consumption and leisure	χ	0.81	1.00
Endowment of discretionary time	ℓ	1.00	n.a.
<i>Technology</i>			
Capital income share	θ	0.38	n.a.
Capital depreciation rate	δ	0.05	n.a.
<i>Age and endowment process</i>			
Probability of retiring	$p_{e\ell}$	0.02	n.a.
Probability of dying	$1 - p_{e\ell}$	0.06	n.a.
Life cycle earnings profile	ϕ_1	0.85	0.90
Intergenerational persistence of earnings	ϕ_2	0.86	0.90
<i>Fiscal policy</i>			
Government consumption	G	0.48	0.40
Retirement pensions	ω	0.17	0.20
Capital income tax function	a_1	0.20	0.25
Payroll tax function	a_2	0.08	n.a.
	a_3	0.71	0.80
	a_4	0.26	n.a.
Household income tax function	a_5	0.77	n.a.
	a_6	1.00	0.80
	a_7	18.28	20.00
Estate tax function	a_8	0.12	0.10
Consumption tax function	a_9	0.09	0.10

Note: Those parameters whose value was determined by normalization or direct parameterization obviously do not have initial values, we thus report their initial value as n.a.

the aggregates, which lead back to the *same* prices. The second is to ensure that the government balances revenue and spending over the transition.

Table 14: The stochastic process for the endowment of efficiency labor units

<i>Estimated Values</i>						
	$e(s)$	γ_s^* (%)	$\Gamma_{\mathcal{E}\mathcal{E}}$ (%) From s To s'			
			$s' = 1$	$s' = 2$	$s' = 3$	$s' = 4$
$s = 1$	0.40	54.82	98.02	1.01	0.96	0.01
$s = 2$	1.36	42.04	0.01	90.42	9.55	0.01
$s = 3$	6.77	3.09	7.67	8.89	83.43	0.01
$s = 4$	198.88	0.05	9.93	7.04	0.01	83.02
<i>Initial Values</i>						
	$e(s)$	γ_s^* (%)	$\Gamma_{\mathcal{E}\mathcal{E}}$ (%) From s To s'			
			$s' = 1$	$s' = 2$	$s' = 3$	$s' = 4$
$s = 1$	0.40	38.62	93.87	2.04	2.04	2.04
$s = 2$	1.20	20.89	2.04	93.87	2.04	2.04
$s = 3$	4.00	20.47	2.04	2.04	93.87	2.04
$s = 4$	200.00	20.02	2.04	2.04	2.04	93.87

Note: $e(s)$ denotes the relative endowments of efficiency labor units; γ_s^* denotes the stationary distribution of working-age households (note that this is not the stationary distribution of $\Gamma_{\mathcal{E}\mathcal{E}}$, it is taken from the stationary distribution of Γ and renormalized); $\Gamma_{\mathcal{E}\mathcal{E}}$ denotes the transition probabilities of the process on the endowment of efficiency labor units for working-age households that are still workers one period later.

Table 15: The Target Moments used in the Simulated Method of Moments estimation and their values in the U.S. data, after calibration/estimation, in general equilibrium, and the weights assigned to the moments estimation.

Target		U.S.	Calib.	Gen. Eqm.	Weight
<i>Macroeconomic & Demographic Trends</i>					
Capital/Output Ratio	K/Y	4.67	4.51	4.55	20.0
Avg. share time allocated to work	H/ ℓ (%)	33.00	32.66	33.10	1.0
Life-cycle profile of earnings		1.30	1.18	1.21	0.2
Intergen. Trans. of Earnings Ability		1.30	0.65	0.15	0.2
<i>Government Policy</i>					
<i>Government Expenditures</i>					
Gov. Expenditure/GDP	G/Y (%)	24.75	20.72	20.81	0.5
Gov. Transfers/GDP	Z/Y (%)	5.53	5.13	5.08	1.0
<i>Government Revenue</i>					
Capital Income Tax Revenue/GDP	τ_k/Y	3.08	2.76	2.96	1.0
Estate Tax Revenue/GDP	τ_e/Y	0.19	0.21	0.21	1.0
Payroll Tax	a_3/\bar{y}	0.82	0.78	0.77	1.0
Personal Income Tax Effective Rate	a_6/\bar{y}	10.95	9.42	9.38	1.0
Estate Tax Exemption	a_7/\bar{y}	20.00	20.04	19.86	1.0
<i>The Distribution of Earnings</i>					
Gini for Earnings		0.64	0.65	0.65	0.8
Earnings Lorenz Curve (%): Quintiles					
1-20: 1st Quintile		0.00	0.00	0.00	0.8
21-40: 2nd Quintile		4.20	3.09	3.01	0.8
41-60: 3rd Quintile		11.70	11.20	11.55	0.8
61-80: 4th Quintile		20.80	21.37	21.34	0.8
81-100: 5th Quintile		63.50	61.46	61.23	0.8
Earnings Lorenz Curve (%): Top Percentiles					
90-95		11.70	9.08	9.03	0.8
96-99		16.60	16.81	16.74	0.8
99-100		18.70	20.26	20.23	0.8
<i>The Distribution of Wealth</i>					
Gini for Wealth		0.82	0.83	0.82	0.6
Wealth Lorenz Curve (%): Quintiles					
1-20: 1st Quintile		0.00	0.00	0.00	0.6
21-40: 2nd Quintile		1.10	0.00	0.00	0.6
41-60: 3rd Quintile		4.50	3.29	3.43	0.6
61-80: 4th Quintile		11.20	12.27	12.26	0.6
81-100: 5th Quintile		83.40	83.11	82.95	0.6
Wealth Lorenz Curve (%): Top Percentiles					
90-95		11.10	14.20	14.44	0.6
96-99		26.70	23.89	23.71	0.6
99-100		33.60	34.65	34.15	0.6

Estimation of Bewley-Huggett-Aiyagari Models: Theory and Implementation

General Equilibrium Heterogeneous Agent Models with Idiosyncratic but no Aggregate Uncertainty

Robert Kirkby
Universidad Carlos III de Madrid

May 5, 2014

Abstract

I provide some theoretical results relating to the estimation of Bewley-Huggett-Aiyagari models; general equilibrium heterogeneous agent models with idiosyncratic but no aggregate uncertainty. Estimation methods include Simulated Moments Estimation and Simulated Likelihood Estimation. The Simulated Method of Moments Estimator presented here is based on targeting moments of the steady-state of the model. The Simulated Likelihood Estimator is based on a two-stage process: in the first stage the micro-parameters are estimated from panel data by simulated maximum likelihood, in the second stage the macro-parameters are then chosen to be consistent both with macro aggregates, and with the prices estimated in the first step. I show that numerical errors in the computation and simulation of the model will disappear asymptotically. I also provide evidence, based on simulations, about which algorithms work best in implementing such estimators.

Contents

1	Introduction	1
2	The Model	4
2.1	Bewley-Huggett-Aiygari Models	4
3	Computational Solution and Simulation of the Model	6
3.1	Computing the Value Function and Optimal Policy Function	6
3.2	Simulating the Steady-State	7
3.2.1	Computing the Moments of the Steady-State	9
4	The Simulated Moments Estimator	9
5	The Two-Stage Simulated Likelihood Estimator	11
5.1	Some Notation	11
5.2	The Estimator	12
5.3	Remarks	13
6	Implementing the Estimators	14
6.1	Model of Pijoan-Mas (2006)	14
6.2	Implementation of the Estimators	15
6.3	Results	17
7	Conclusion	18
A	Numerical Errors in the Value Function and Optimal Policy Function	24
A.1	The Results	26
A.2	Numerical Error Bounds for the Value Function Iteration	28
A.2.1	Uniform Convergence of Value Function Iteration	29
A.2.2	Discrete State Space Approximation	30
A.2.3	The Discretization Procedure	30
A.2.4	Some Constants	33
A.2.5	Intermediate Results	34
A.2.6	Putting it all Together	37
A.2.7	A Remark on the Bounds	40
A.3	Numerical Error Bounds for the Optimal Policy Function	40
A.3.1	Convergence of the Optimal Policy Function	40
A.3.2	Bounding Errors in Optimal Policy Function without Interiority	41
A.3.3	Some More Constants	41
A.3.4	Errors from having an approximation of the true value function	42
A.3.5	Errors from the discretization	43
A.3.6	An Error Bound for the distance between the approximate and true policy functions	45
A.4	SubAppendix: Some Results on Numerical Integration	45
A.5	SubAppendix: Uniform Convergence of Value Function Iteration	48
A.5.1	Howards Improvement Algorithm	50
A.5.2	Bounding the Errors of Value Function Iteration	50
A.6	SubAppendix: Some Results Used for Convergence of Value Function Iteration	51
A.7	SubAppendix: Discretizing the Choice Variable y in Value Function	52
B	Numerical Errors in the Steady-State Distribution	54
B.1	Existence, Uniqueness of, and Convergence to, an Invariant Distribution	55
B.2	The Agents Distribution with the Approximation of the Optimal Policy	56
B.3	Upper Semicontinuity of the Correspondence of Invariant Distributions	56
B.4	Bounding the distance to the true invariant distribution	59
B.5	Combining our results on the Agents Distribution	62

C	Monotone Mixing Condition in Pijoan-Mas (2006) Model	62
D	Monotone Mixing Condition in Other Models	65
E	Evaluating the Likelihood in Model of Pijoan-Mas (2006)	66

1 Introduction

I propose two estimators for the structural estimation of models of the Bewley-Huggett-Aiyagari class; models with heterogeneous agents, incomplete markets, and a competitive general equilibrium, in which there is idiosyncratic but no aggregate uncertainty. In such models analytical solutions are not available, and so simulated estimation methods must be used. The first estimator, a simulated moments estimator (SME), is based on the steady-state of the model. The second estimator, a two-stage simulated likelihood estimator (SLE), is based on estimating the micro-parameters from panel data in the first-stage, and then the macro-parameters in the second-stage. Theoretical results on the consistency of these estimators are provided; with those for the SME estimator being based on assumptions that it can be proven apply to the Bewley-Huggett-Aiyagari class of models. Further theory is provided showing that numerical errors occurring during the computation of the optimal policy and the steady-state moments will not affect, in the limit, the consistency of the SME. An example implementation of these estimators is then given, based on the model of Pijoan-Mas (2006). I provide evidence on which algorithms work in implementing the estimators, and on their reliability in correctly estimating the true parameters of the model.

The Bewley-Huggett-Aiyagari class of models has its beginnings in Bewley (1983) with early quantitative explorations being Huggett (1993) and Aiyagari (1994). Heterogeneous agent models of this class have been used to give quantitative answers to questions on topics as varied as: progressive taxation (Conesa and Krueger, 2006), capital taxation (Domeij and Heathcote, 2004; Conesa, Kitao, and Krueger, 2009), inequality (Castaneda, Díaz-Giménez, and Ríos-Rull, 2003), entrepreneurship (Quadrini, 2000), and working longer hours (Pijoan-Mas, 2006). For more on the applications of heterogeneous agent models, both of the Bewley-Huggett-Aiyagari class and other more-complicated classes, see Ríos-Rull (1995, 2001); Heathcote, Storesletten, and Violante (2009).

Structural estimation by simulation methods is well-understood for certain dynamic models. Early contributions on simulated estimation in dynamic environments such as Lee and Ingram (1991) and Duffie and Singleton (1993) were based on assumptions about the data-generating Markov process — to be understood here as the solution of a dynamic optimization (eg. value function) problem. These results suffered from the issue that it was not clear that the assumptions made about the data-generating Markov-process would necessarily be satisfied by the solution to the value function problems of models like the standard neoclassical growth model. Fernandez-Villaverde, Rubio-Ramirez, and Santos (2006) overcome this issue providing conditions on the data-generating Markov process that can be proven to hold from the fundamentals of the value function problem to be solved. Their approach is based on the Feller property, and the related concept of stochastic contractions, which it can be proven are satisfied by a number of standard dynamic models including the neoclassical growth model; this theory underpins much of the growing field of likelihood-based DSGE estimation, whether by classical or Bayesian likelihood methods.¹

¹Although Akerberg, Geweke, and Hahn (2009) provide a counterexample to Proposition 2 of Fernandez-

The other issue afflicting the earlier work on structural estimation is that the simulations themselves are typically made using a computational approximation to the solution to the value function problem. So numerical errors that occur while computing the solution to the value function may cause problems with the convergence of the simulations themselves. This issue was addressed in Santos and Peralta-Alva (2005). A partial summary of these literatures, concentrating mainly on the later on numerical errors, can be found in Peralta-Alva and Santos (2014).

Heterogeneous agent models of the Bewley-Huggett-Aiyagari class are not based on the Feller property, but instead on the monotone mixing condition. This paper provides results for simulation estimation based on the monotone mixing conditions, and so it can be proven from fundamentals that they apply to models of the Bewley-Huggett-Aiyagari class. Existing results on bounding numerical errors arising from the value function problem and simulations are also not applicable to these models — due, eg., to non-interior optimal policies, non-convex choice sets, and the dependence on the monotone mixing conditions. Another contribution of this paper is thus to extend and adapt existing results to cover these cases.²

Aguirregabiria and Mira (2010) provide a comprehensive recent summary on the structural estimation of dynamic models. They briefly cover competitive-equilibrium models — including the Bewley-Huggett-Aiyagari class of models addressed here — but their main concentration is on the estimation of single-agent models and dynamic games. They focus on the methods used to estimate these various kinds of models, as well as giving examples of their empirical application.

Many standard issues relating to panel-data estimation, such as unobserved heterogeneity, initial observations, self-selection bias, censoring, and attrition bias, may occur when applying the Two-State SLE estimator. For a discussion of these and other related issues in the context of structural estimation see Aguirregabiria and Mira (2010) and Keane, Todd, and Wolpin (2011).

Note that the choice of using moment estimation in the SME likelihood estimation in the SLE is itself arbitrary, the difference between the estimators comes what kind of data they use. One could estimate based on the steady-state using likelihood methods, or estimate the microfoundations from panel data using moment-based methods. The real difference between the two estimators is about the choice between targeting the steady-state (moments) versus a two-stage approach using microeconomic panel data. The decision to use moments-based methods for the first case, and likelihood for the second case simply reflects the (arbitrary) judgement of the author on which estimator is the most suitable in each case.

Both estimators could be trivially extended to allow some of the parameters to be pre-calibrated directly, and then the estimation performed on the remaining parameters; see Peralta-Alva and Santos (2014) for further discussion of this concept in the context Simulated Moment Estimation of a different class of model. In practice this is very often done.

Villaverde, Rubio-Ramirez, and Santos (2006).

²This paper contains two estimators, the SME and the SLE, the results I have proven so far only cover the SME.

Section 2 provides a general description of the model. Section 3 provides theory addressing numerical errors that occur during the computational solution and simulation of the model. Section 4 describes the Simulated Moments Estimation of this class of models — based on moments of the steady-state. Section 5 describes the Two-Stage Simulated Likelihood Estimation of this class of models — based on first estimating the microfoundations.

Macroeconometrics, Heterogeneity and Estimation: Before we get into the paper itself, a comment on the importance of structure, heterogeneity, and estimation in Macroeconomics is in order.

This is a paper about Macroeconometric estimation. Perhaps not in the modern sense of the word, but certainly in the sense in which the term econometrics was coined by Ragnar Frisch, co-recipient of the first Nobel Prize in Economics.

[Frisch] was the first to propose the use of the term *econometrics* to describe a research program that consisted in (1) mathematical formulations of economic theories and (2) systematic tests of the theories using the methods of mathematical statistics...Statistical analysis of economic relationships was in his opinion meaningless unless it was based on rigorous theoretical reasoning, that is, on a mathematically formulated theoretical model (Sandmo, 2011; Chpt 16, pg 376).^{3,4}

Paragraph on importance of structure/mathematical models/microfoundations.

That heterogeneity is important to macroeconomics seems almost self-evident, none-the-less I now make the case. For questions relating to inequality and distributional issues heterogeneity is a prerequisite for creating relevant models — if everyone is the same the concept of inequality is meaningless. But the usefulness of heterogeneity and microfoundations extends well beyond its necessity in addressing such questions. It is also useful for the understanding of aggregate macroeconomic variables such as GDP and the employment rate. The observation is often made that evaluating macroeconomic models is difficult as one often has only fifty years of good quality quarterly data, not nearly enough to differentiate between the many alternative explanations. Adding heterogeneity tackles this problem head on, opening up a vast seam of cross-sectional and panel data that can be used to compare different models. One criticism commonly made is that by modeling individuals we lose sight of the macroeconomy: can we really understand the wealth of nations

³Frisch considered the pairing of statistics with mathematical models to be especially important to the problem of identification, an issue which remains controversial today. See Keane (2010) for a recent statement of this view, to be understood in contrast to the view that identification can be done solely using an experimentalist approach, as embodied in the approach of Angrist and Pischke (2009).

⁴Or as expressed at the beginning of the abstract of a conversation between Frisch, Schumpeter, and Haberler in 1928 about the possibility of forming what is now the Econometric Society (Frisch 1969; pg 13),

The terms econometric and econometrics are interpreted as including both pure economics and the statistical verification of the laws of pure economics. In essential distinction to the purely empirical manipulation of statistical data on economic phenomena.

by looking at the actions of butchers, brewers, and bakers? Fortunately for macroeconomics this criticism is of little relevance: it is a tautology that if we can model the incomes of each individual then we can model GDP, and that if we can model the work choices of individuals then we can model the employment rate — these are the very definitions of GDP and the employment rate! None of this is to suggest that doing so is easy, the modelling of individual choices is difficult, not to mention how they interact. But explicit modeling of heterogeneity does provide us with a path to follow, and access to enough data to decide which turns to take.

So heterogeneity is important, but why estimate the models? One major advantage of estimating heterogeneous agent models is that it solves the problems of synthesis — arbitrarily taking different parameters from different sources. Browning, Hansen, and Heckman (1999) describe the problems of synthesizing as follows:

Synthesizing evidence across micro studies is not a straightforward task. Different microeconomic studies make different assumptions, often implicit, about the economic environments in which agents make their decisions. They condition on different variables and produce parameters with different economic interpretations. A parameter that is valid for a model in one economic environment cannot be uncritically applied to a model embedded in a different economic environment. Different general equilibrium models make different assumptions and require different parameters, many of which have never been estimated in the micro literature.⁵

2 The Model

In this section I give a formal description of the class of models being considered here; namely general equilibrium heterogeneous agent models with incomplete markets and idiosyncratic, but no aggregate, uncertainty.

2.1 Bewley-Huggett-Aiygari Models

In this subsection I provide a formal description of the class of heterogeneous agent models being considered. Notation is loosely based on Stokey, Lucas, and Prescott (1989, henceforth SLP); loosely as SLP do not treat heterogeneous agent models of this type.

The models are those which can be expressed as follows: Let $X \subseteq \mathbb{R}^l$ be the endogenous state variable, $Y = Y_1 \times Y_2 \subseteq X \times \mathbb{R}^c$ be the choice variable, and $Z \subseteq \mathbb{R}^k$ be the exogenous state variable. Let $\Theta = \Theta_1 \times \Theta_2 \subseteq \mathbb{R}^{q_1} \times \mathbb{R}^{q_2}$ be a parameter space; $\theta_1 \in \Theta_1$ is a vector of parameters

⁵For an interesting take on how the results of calibration and estimation methods compare in practice see Ríos-Rull, Schorfheide, Fuentes-Albero, Kryshko, and Santaella-Llopis (2012).

that enter into the value function problem, while $\theta_2 \in \Theta_2$ is a vector of parameters that only affect macroeconomic aggregates.⁶ The state of an agent is then a pair (x, z) . A value function maps $V_{\theta_1, p} : X \times Z \rightarrow \mathbb{R}$. A policy function maps $g_{\theta_1} : X \times Z \rightarrow Y$. Let $S = X \times Z$, and let \mathcal{S} be its Borel σ -field. The measure of agents $\mu_{\theta_1, p}$ is a probability distribution over (S, \mathcal{S}) . The return function maps $F_{\theta_1, p} : X \times Y \times Z \rightarrow \mathbb{R}$, and the discount factor is $0 < \beta < 1$.

Aggregate variables are $A \in \mathbb{A} \subseteq \mathbb{R}^a$. A price vector is $p \in \mathbb{P} \subseteq \mathbb{R}^p$. The exogenous shock follows a Markov-chain with transition function Q_{θ_1} mapping from Z to Z . The aggregation function maps $\mathcal{A} : \mathcal{M}(S, \mathcal{S}) \rightarrow \mathbb{R}^a$, where $\mathcal{M}(S, \mathcal{S})$ is the space of probability measures on (S, \mathcal{S}) . The market clearance function maps $\lambda_{\theta_2} : \mathbb{R}^a \times \mathbb{R}^p \rightarrow \mathbb{R}$.

Definition 1. *A Competitive Equilibrium is an agents value function $V_{(\theta_1, p)}$; agents policy function $g_{(\theta_1, p)}$; vector of prices p ; measure of agents $\mu_{(\theta_1, p)}$; such that*

1. *Given prices p , the agents value function $V_{(\theta_1, p)}$ and policy function $g_{(\theta_1, p)}$ solve the agents problem*

$$V_{(\theta_1, p)}(x, z) = \max_{y=(y_1, y_2) \in Y} \left\{ F_{(\theta_1, p)}(x, y, z) + \beta \int V_{(\theta_1, p)}(y'_1, z') Q_{\theta_1}(z, dz') \right\} \quad (1)$$

2. *Aggregates are determined by individual actions: $A_{(\theta_1, p)} = \mathcal{A}(\mu_{(\theta_1, p)})$.*
3. *Markets clear (in terms of prices): $\lambda_{\theta_2}(A_{(\theta_1, p)}, p) = 0$.*
4. *The measure of agents is invariant:*

$$\mu_{(\theta_1, p)}(x, z) = \int \int \left[\int 1_{x=g_{(\theta_1, p)}(\hat{x}, z)} \mu_{(\theta_1, p)}(\hat{x}, z) Q_{\theta_1}(z, dz') \right] d\hat{x} dz \quad (2)$$

Models fitting this definition include Huggett (1993) and Aiyagari (1994), as well as numerous extensions endogenizing labour supply, introducing taxation, and modeling dynasties. The aggregates in point two generally correspond to the household variables (such as aggregate capital) but in some models may also be aggregates of functions thereof (such as tax revenue). The third point, that prices clear markets, involves rewriting market clearance equations in terms of prices, rather than quantities.

So for example in Aiyagari (1994) the requirement that aggregates are determined by individual actions is that aggregate capital is the sum of individuals capital holdings, $K = \int k'(x, z) d\mu$. While the market clearance conditions is that the interest rate is equal to the marginal product of capital, $\lambda_{(\alpha, \delta)}(K, r) = r - \alpha K^{1-\alpha} - \delta = 0$, where K depends on individual behaviour, and thus on (θ_1, p) .

⁶This differentiation between θ_1 and θ_2 is important for the Two-Stage estimation procedure.

3 Computational Solution and Simulation of the Model

In this section we address the issue of numerical errors that might arise during the solution and simulation of the model. Given the prominent role played by simulation in both the SME and SLE estimators it is good to know that the simulation will not cause problems for the convergence of the estimators. Results are provided, firstly to show that numerical errors in the value function and optimal policy function are bounded, and that they will go to zero as the distance between points on the grids used to approximate the value function and optimal policy function go to zero. It is then shown that if, as was shown, the numerical errors in the optimal policy function are bounded, then numerical errors in the steady-state distribution will also be bounded. It follows almost trivially that numerical errors in the moments of the steady-state distribution are bounded. Furthermore, as numerical errors in the optimal policy function go to zero, so will those in the steady-state distribution and its moments. Together these results ensure that the any role played by numerical errors in the estimation of the SME will go to zero as the distance between points on the grid used in the approximation go to zero. Proving this for the SLE would require results on the numerical errors in finite-length time-series simulations and is not done here.

In what follows we consider in turn: computation of the value function and optimal policy function; simulation of the steady state distribution; computation of moments of the steady state distribution. At every stage emphasis is placed on deriving numerical error bounds that are based on the algorithms that are actually used to implement the estimators.

3.1 Computing the Value Function and Optimal Policy Function

The first step in solving these models is to compute the (value function and the) optimal policy fn. The existence of periodically binding constraints, and non-concave feasible choice sets, are common in this class of models, and so methods based on Euler Equations, and other first order condition based methods, are not applicable. The standard approach is thus value function iteration. Theory on bounding the numerical errors in value function iteration does exist: for example Santos and Vigo-Aguiar (1998) provide results based on partial discretization, while Stachurski (2008) provides results for a variety of fitted value function iteration methods. Here I provide a new result on bounding numerical errors in the value function and more importantly in the optimal policy function based on discretized value function iteration.⁷ The proofs of this result, as well as a discussion of how it differs from existing results, can be found in Appendix A.

Our results are based on the Case 1 value function problem,

$$V(x, z) = \sup_{y=(y_1, y_2) \in \Gamma(x, z)} \left\{ F(x, y, z) + \beta \int V(y_1, z') Q(z, dz') \right\} \quad (3)$$

⁷Previous results depend on assumptions like the interiority of solutions, or differentiability of the value function, that are often not applicable to the kinds of models considered here.

Our intention is to bound the difference between the solution to this problem, V , and the numerical solution to the discretized problem, V_N^G , the latter is the solution after N iterations (once a standard convergence criterion is met) of value function iteration on the discrete grid (hence the G).

We assume that the spaces for the endogenous state X , the control variables Y , and the exogenous state Z are all compact, that the return function F is continuous and bounded⁸, the discount factor β is less than one, and that the transition function Q has the Feller property.

Then we have the following result,

Corollary 1. *For the value function defined in 3. Let X, Y, Z be compact, the return function F be continuous, monotone, and bounded, the discount factor β be less than one, and the transition function Q have the Feller property. Let V_N^G be the numerical solution to the discretized value function problem (discretizing X, Y and Z , using Tauchen method to discretize Q). Then, the numerical errors in the value function, $\|V - V_N^G\|$, and in the optimal policy function, $\|g - g_N^G\|$, converge to zero as the distance between grid points in the dimensions being discretized go to zero.*

Proof: See Appendix A.

This result is based on discretized value function iteration; both a common solution method for models of this kind and the exactly the method that is used to implement the estimators when generating the simulation results in this paper.

3.2 Simulating the Steady-State

Our error bounds are based on discretizing the state space and iterating on the entire agent distribution until it converges. Iterating on the entire agent distribution is needed to get exact bounds on the numerical errors.

In the results closest to those presented here Santos and Peralta-Alva (2005) provide bounds on the numerical errors in invariant distributions and the simulation methods commonly used to generate them. However none of these results is applicable to incomplete market heterogeneous agent models. The reason is two-fold; firstly the bounds on value and policy function errors are all dependent on the interiority of the optimal policy, but in heterogeneous agent models we often have binding borrowing constraints. Secondly, the bounds on invariant distribution errors are dependent on the Feller property, while heterogeneous agent models generally depend on the monotone mixing condition (Hopenhayn and Prescott, 1992).

The following Theorem 1 is taken from Hopenhayn and Prescott (1992) and introduces the monotone mixing condition (MMC). Based on the MMC it proves that for any given interest rate

⁸Since F is a continuous function defined on a compact space, it will therefore also be bounded.

there is a unique invariant distribution.

Theorem 1. *Suppose P is increasing, S contains a lower bound (which we will denote by a) and an upper bound (which we will denote by b) and the following condition is satisfied: Monotone Mixing Condition (MMC): there exists $s^* \in S, m \in \mathbb{Z}, \beta_P > 0$ such that $P^m(b, [a, s^*]) \geq 1 - \beta_P$ and $P^m(a, [s^*, b]) \geq 1 - \beta_P$. Then there is a unique stationary distribution μ^* for process P , and for any initial measure $\mu_0, \mu_n \equiv T^{*n} \mu_0 = \int P^n(s, \cdot) \mu_0(ds)$ converges to μ^* .*

The main result that is needed for the results of this section is,

Theorem 2. *Let $\{g_j\}$ be a sequence of policy functions that converge to g . Let $\{\mu_j^*\}$ be a sequence of probabilities on \mathbb{S} such that $\mu_j^* = T_j^* \mu_j^*$ for each j . Under assumptions 9 and 10, if μ^* is a weak limit point of μ_j^* , then $\mu^* = T^* \mu^*$.*

it appears, with proof, as Theorem Theorem 15 in Appendix B. Have discovered that proof contains an error; it mixes metrics)

Based on the monotone mixing condition we are able to derive the following bound on the numerical errors in the steady-state distribution.

Corollary 2. *Let f be a Lipschitz function with constant L . Let $\|g - \hat{g}\| \leq \delta_g$ for some $\delta_g > 0$. Assume that T^{*m} is a contraction mapping of modulus β_P . Then, under assumptions 9 & 10,*

$$\left| \int f(s) \mu^*(ds) - \int f(s) \hat{\mu}^*(ds) \right| \leq \frac{Lm\delta_g}{1 - \beta_P}$$

and furthermore

$$\|\mu^* - \hat{\mu}^*\| \leq \frac{m\delta_g}{1 - \beta_P}$$

Proof: See Appendix B.

The interpretation of Corollary 2 is that due to the nature of the discretization of the optimal policy and the transition matrix for the exogenous shocks, and in particular due to considering the optimal policy function as a piecewise constant extension on the discretization grid, the discretization of the steady state distribution introduces no new errors. Thus the only errors in the steady state distribution are those which we have already bounded in terms of errors in the optimal policy function, and those coming from stopping after a finite number of iterations. The intuition for why the discretization of the steady state distribution does not create any further errors comes from noting that the approximate transition function $P^G = \hat{g} \cdot Q^G$ is piecewise constant on the partition imposed by discretization; a property it inherits via \hat{g} and Q^G .

3.2.1 Computing the Moments of the Steady-State

Since the moments of the steady-state can be expressed as the integral of a function with respect to the steady-state, as long as the functions defining the moments are Lipschitz we have in fact already addressed this issue in Corollary 2

4 The Simulated Moments Estimator

This section introduces a Simulated Moments Estimator for Bewley-Huggett-Aiyagari models. In implementing the estimator a necessary condition to be avoided having to calculate the general equilibrium during the estimation is that the observed macroeconomic aggregates, together with the parameters, must be sufficient to directly identify the price vector. This necessary condition is satisfied in almost all applications.⁹ However, it is not necessary for the properties of the estimator — the theory developed here relating to the estimator depends on the monotone mixing condition, not on the general equilibrium itself. Based on the theory given in Section 3 it is established that the SME is a consistent estimator (as the distance between the grid points used in the simulation go to zero, the simulated moments estimator converges to a standard moments estimator, which is in turn consistent).

Suppose that we have a set A^D of n data observations that correspond to moments of the steady-state distribution of the model.¹⁰ For a given parameter vector θ , the model produces a set of n corresponding moments $A^M(\theta)$. Since we cannot solve the model exactly, we will have to use computational solution and simulation of the model to generate the set of simulated moments $A^{Ms}(\theta)$. Our aim is to find the true parameter vector θ^* for which the moments of the model $A^M(\theta)$ are equal to those of the data A^D .¹¹ Thus, θ^* is defined as the solution to,

$$\theta^* \equiv \arg \min_{\theta} d(A^D, A^M(\theta)) \quad (4)$$

where $d : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ is a distance function, and is assumed to be continuous. It is assumed that θ^* is the unique solution to this problem. We will use the distance function given by

$$d(A^D, A^M(\theta)) = \sum_i w_i (A_i^D - A_i^M)^2 / \text{var}(A_i^D) \quad (5)$$

This distance function is chosen for a number of reasons: (i) it uses the square of the distance which is standard, simple, and intuitive, (ii) by dividing by variance of the data moments we put more weight on the better 'measured' moments and brings us closer to a standard GMM-estimation

⁹It is satisfied as long as the n observable macroeconomic aggregates, A^D , include as a subset the a moments of the steady-state, A , that enter into the market clearing equation $\mathcal{M}_{\theta_2}(A, p)$.

¹⁰These n moments need to include the a moments used in the definition of the general equilibrium of the model. In practice they will almost always also include further moments.

¹¹By asking the moments to be equal I am implicitly assuming here that the model is exactly identified, and not overidentified. If the model is overidentified the moments would simply need to be 'as close as possible'.

approach, (iii) via the w_i , the researcher can put more weight on moments that are more important or better identified¹², and (iv) it appears to perform well in practice. Note that from the point of view of the model theoretically the aggregate data moments have zero variance and so variance in those moments is to be understood as due to measurement error, which is assumed to be *iid* Normally distributed.

To implement the estimator we will of course have to use empirical estimates of $\text{var}(A_i^D)$, so define $\tilde{\theta}$ as

$$\tilde{\theta} \equiv \arg \min_{\theta} \sum_i w_i (A_i^D - A_i^M)^2 / \widehat{\text{var}}(A_i^D) \quad (6)$$

where $\widehat{\text{var}}(A_i^D)$ is just the standard consistent estimator of sample variance for an *iid* normally distributed variable, namely $\frac{1}{T-1} \sum_{t=1}^T (A_{i,t}^D - \bar{A}_i^D)^2$ (the notation here is based on the idea that, eg., the moment is the captial/output ratio and is observed annually).

One alternative would be to use a standard GMM-style distance measure in which we take the squares of the distances between moments (which we do with $(A_i^D - A_i^M)^2$) and then sum them using as a weighting matrix the inverse of the variance-covariance matrix (instead of a diagonal matrix of w_i/A_i^D terms). I avoid this since in practice, estimation of the aggregate moments is likely to be based on different data sources, and thus trying to estimate their covariances is not an easy problem.¹³

Note that there is, in principle, no uncertainty in the moments of the steady-state distribution of the model. Thus the statistical concept of uncertainty does not make sense unless, as mentioned previously, we assume there is measurement error in the data observations of these moments. For the estimator described in 6 to be consistent for θ^* we then require that:

1. (i) the estimators of the weights, $\widehat{\text{var}}(A_i^D)$ are consistent,
2. (ii) θ^* is the unique solution to $d(A^D, A^M(\theta)) = 0$ (so it will also solve equation 4),
3. (iii) the distance function $d(A^D, A^M(\theta))$ is continuous in θ (this follows immediately from the result of Theorem 2 and continuity of \mathcal{A} function),
4. (iv) the set of possible parameters Θ is compact,
5. (v) $\sup_{\theta \in \Theta} |d(A^D, A^M(\theta))| < \infty$.

These are the standard assumptions needed for consistency of GMM estimators, adapted to the current estimator. Implicitly, we also assume that the parameter space Θ is such that the assumptions on the microfoundations, that we used to prove that the monotone mixing condition, hold.

¹²In practice, this might include putting more weight on those moments related to the general equilibrium, market clearing, requirements

¹³In practice, models of this class often target data moments, some of which come from aggregate national accounts data, while others come from cross-sectional data sources.

Eg., that the parameters relating to the utility function are such that it is increasing in the state variables.

We cannot solve the model exactly, as we do not have the theoretical model moments $A^M(\theta)$. Instead we must use the corresponding simulated moments $A^{M_s}(\theta)$. Let $\hat{\theta}$ be the SME estimate of θ^* defined by,

$$\theta^* \equiv \arg \min_{\theta} d(A^D, A^{M_s}(\theta)) \tag{7}$$

From the results in Section 3, specifically by the trivial combination of Corollary 2, which shows that the numerical errors in moments of the steady-state distribution are uniformly bounded in terms of the errors in the optimal policy function, and Corollary 1, showing that the errors in the optimal policy function converge uniformly to zero as the distance between grid points goes to zero, we have that A^{M_s} converges uniformly to A^M as the distance between grid points goes to zero. It therefore follows that the SME estimate, $\hat{\theta}$, converges uniformly to $\tilde{\theta}$ as the distance between grid points goes to zero. So under the conditions described above for the consistency of $\tilde{\theta}$ it follows that the SME estimator is consistent.

In practice, or when the model is overidentified, the SME estimate of θ^* will often not exactly satisfy $d(A^D, A^M(\theta^*)) = 0$. For this reason it is important, using the estimated parameter vector $\hat{\theta}$, to compute the general equilibrium of the model before using it for any other purpose. In this case it can be a good idea to put large weights on the moments that correspond to the general equilibrium conditions (the a moments in A), so as to ensure that the final general equilibrium still closely replicates the target moments.

5 The Two-Stage Simulated Likelihood Estimator

I first show how reformulate the model into the setup used to derive the qualities of the estimator. The estimator itself is then presented and some of its properties are discussed. Consistency of the SLE is not shown; it would require a result like that of Corollary 2, but for finite-sample time-series.

14

5.1 Some Notation

Define the stochastic process $P_{(\theta_1, p)} : S \rightarrow S$ by $P \equiv g_{(\theta_1, p)} \cdot Q_{\theta_1}$. $P_{(\theta_1, p)}$ is a (first-order) Markov process, and is fully parameterized by (θ_1, p) (SLP, Theorem 9.13). So $P_{(\theta_1, p)}$ takes values in the state-space of the model, and its motion is governed by the solution to the agents problem.

¹⁴Results of this style do exist, but not based on the monotone mixing condition. Eg. Theorem 7 of Santos and Peralta-Alva (2005) provide such a result based on stochastic contractions (stochastic contractions are loosely related to the Feller property).

Models of the Bewley-Huggett-Aiyagari class satisfy the monotone mixing condition (MMC) of Hopenhayn and Prescott (1992). Huggett (1993) proves that his model satisfies this condition; Appendices D and C provide proofs that the models of Aiyagari (1994) and Pijoan-Mas (2006) also satisfy this condition. It follows that $P_{(\theta_1, p)}$ has a stable asymptotic distribution, μ , and that this distribution is parameterized by (θ_1, p) . [Appendix B contains a proof that μ is (upper-semi-) continuous in (θ_1, p) .]

The market clearance equation, $\mathcal{M}_{\theta_2}(A, p)$, needs no reformulation. The key point for our estimator is that A can be observed directly from macroeconomic data, and that we will get estimates of p from our first stage. Thus we do not have to solve the usual fixed-point problem of finding p so that the simulated A satisfies the market clearance conditions. We can instead directly observe A , we can estimate p by SLE from the observed panel data, $\{d_{i,t}\}_{t=1, i=1}^{t=T, i=N}$, and then as a second-stage we can use the two of these together to get θ_2 from the market clearing conditions.

5.2 The Estimator

I now describe the SLE introduced by this paper. At time t an individual's state current state (both the endogenous and exogenous states) is determined by the vector s_t . This state vector s_t evolves according to the stochastic process $P_{(\theta_1, p)}$ an (first-order) Markov-process which is determined by the vector (θ_1, p) , where θ_1 are (individual-level) micro-parameters and p are prices affecting behaviour; we refer to (θ_1, p) as the augmented-micro-parameter vector.

Let θ^* be the true parameter values, and (θ_1^*, p^*) be the true values of the augmented-micro-parameters, where p^* , the true prices, are understood to be the general equilibrium prices that correspond to the true parameter values. The observable panel data $\{d_{i,t}\}_{t=1, i=1}^{t=T, i=N}$ from which we wish to perform the estimation is generated by the Markov-process, $f(P_{(\theta_1^*, p^*)}, (\theta_1^*, p^*))$; note that all of our agents are ex-ante identical, hence we do not differentiate different i subscripts for the process generating each different individual's data. Since $P_{(\theta_1^*, p^*)}$ is itself fully determined by (θ_1^*, p^*) we will denote this data-generating Markov process as $f_{(\theta_1^*, p^*)}$. We also have observations of some macroeconomic aggregates, $A_{(\theta_1^*, p^*)}$.¹⁵ Prices and macroeconomic aggregates are known to be related to each other by markets (eg. interest rates are the marginal product of aggregate capital). This can be thought of as a restriction that $\mathcal{M}_{\theta_2^*}(A_{(\theta_1^*, p^*)}, p^*) = 0$; that aggregates $A_{(\theta_1^*, p^*)}$ and prices p^* satisfy a market clearance condition $\mathcal{M}_{\theta_2^*}(\cdot)$ that depends on parameter vector θ_2^* .

For any augmented-micro-parameter vector (θ_1, p) we can solve the model for the optimal policy function and, combining this with the transition function on the exogenous shocks, get a Markov-process $P_{(\theta_1, p)} : S \rightarrow S$. We also have $f(\cdot, (\theta_1, p))$; note that f may also depend on the solution to the model as it might be related to some of the control variables and therefore depend on the

¹⁵In theory these macroeconomic aggregates are generated as the integrals of functions on the steady-state distribution of this same Markov-process. But importantly we do not need to simulate them, we can just observe them directly from the data.

optimal policy function. Using $P_{(\theta_1, p)}$ we can generate panel data, which combined with $f(\cdot, (\theta_1, p))$ gives us the simulated panel data $\{d_{i,t}(\theta_1, p)\}_{t=1, i=1}^{t=T, i=N}$.

The Simulated Likelihood Estimator is now defined. In the first stage of the nested estimator the augmented-parameter vector (θ_1, p) is chosen so as to maximize the simulated likelihood of the observed panel data, $\{d_{i,t}\}_{t=1, i=1}^{t=T, i=N}$. For individual i , create a simulated time-series of length T , $\{d_{i,t}^j(\theta_1, p)\}$, evaluate the (individual- i -simulation- j -specific) likelihood of individual i 's observed data under this simulation, j . Repeat for $j = 1, \dots, J$ simulations, and then take the sum across these J simulations to get the (individual- i -specific) likelihood of individual i 's observed data under the J simulations. Doing this for each individual, $i = 1, \dots, N$ and then summing across them we get the simulated likelihood of the observed panel $\{d_{i,t}\}_{t=1, i=1}^{t=T, i=N}$. In effect we are using simulated likelihood to measure the (inverse of) distance between the observed panel data $\{d_{i,t}\}_{t=1, i=1}^{t=T, i=N}$, assumed to have been generated by (θ_1^*, p^*) , and the simulated panel data $\{d_{i,t}^j(\theta_1, p)\}_{t=1, i=1, j=1}^{t=T, i=N, j=J}$ generated by (θ_1, p) . So by maximizing the simulated likelihood we will minimize the distance between (θ_1, p) and the true augmented-micro-parameter vector (θ_1^*, p^*) .

In the second stage the macroeconomic parameters relating to market clearance, θ_2 , are then estimated from the market clearance condition using observed macroeconomic aggregate data and our first-stage estimates of p .¹⁶

Under our theoretical models the macroeconomic aggregates are constant, which would in turn lead the likelihood of our model to be minus infinity (due to stochastic singularities). To avoid this we assume measurement error in the observation of macroeconomic aggregates. This way of dealing with stochastic singularities follows the standard approach in the literature on structural estimation (Keane, Todd, and Wolpin, 2011), and is easily rationalized on empirical grounds; eg. that we do not measure gross domestic product or aggregate capital stock with perfect accuracy.

While discussing the SME we considered the issue that the simulations would not be created using the exact solution to the value function iteration problem, but instead would use an approximate computational solution. I will ignore this issue for the SLE, simply because I do not have any results relating to it.

5.3 Remarks

One thing not pursued here is to add observable individual fixed-effects (observable agent types).

There is nothing in the way the SLE estimator is set-up that would prevent the macroeconomic aggregates, A , from also depending on θ_2 (say, letting θ_2 parameterize the functions which we

¹⁶Note that during the first-stage we do not need to simulate the macroeconomic aggregates and ensure that when combined with our estimated prices they will satisfy market clearance. This is very important as it avoids solving a fixed-point problem, common to the computation of Bewley-Huggett-Aiyagari models, that otherwise substantially increases the computation involved in estimating these models.

integrate w.r.t. the steady-state distribution to calculate the macroeconomic aggregates, A)

Using SLE, an interesting idea would be to use further macro aggregates, eg. the investment/output ratio, as over-identifying restrictions for θ_2 . This would allow us to test different market clearance set-ups (say, Cobb-Douglas+perf. competition versus CES+perf. competition) based on their performance in stage 2 of the estimator. Obviously these tests would be conditional on the 'truth' of the structural model used in our stage 1 estimates.

6 Implementing the Estimators

In this section I discuss the implementation of these estimators. Our interests here are two-fold. First, what kinds of numerical optimization algorithms work best in implementing these estimators. Second, how reliable and accurate are these estimators — do they have robust convergence properties — and in the case of the SLE, how much panel data is needed get accurate estimates.

I describe simulation results about which algorithms work in implementing the estimators. These results are based on the model of Pijoan-Mas (2006); essentially Aiyagari (1994) with endogenous labour. The model contains nine parameters which Pijoan-Mas (2006) is able to directly calibrate to exactly reproduce certain data moments. This is advantageous as it means we know the 'solution' to a standard and 'realistic' estimation problem.

6.1 Model of Pijoan-Mas (2006)

A description of the model of Pijoan-Mas (2006) is now given. The model contains a continuum of infinitely-lived that make consumption-savings and consumption-leisure choices. Aggregate capital stock and labour supply are determined by aggregating across the individual households. Output is produced by constant-returns-to-scale firms in perfectly competitive markets. That this model satisfies the monotone mixing condition is proved in Appendix C.

Households: Households are infinitely-lived and make consumption-savings and consumption-leisure choices. They face a borrowing constraint, namely that assets must be greater than zero. Their labour productivity follows an AR(1) process. The households value function problem is given by

$$V(a, z) = \max_{c, l, a'} \left\{ \frac{c^{1-\sigma_1}}{1-\sigma_1} + \chi \frac{(\ell-l)^{1-\sigma_2}}{1-\sigma_2} + \beta E[V(a', z')] \right\}$$

$$\text{s.t.} \quad c + a' = wz(1-l) + (1+r)a$$

$$c \geq 0, \quad 0 \leq l \leq 1, \quad \text{and } a' \geq 0$$

where c is consumption, a assets, and l leisure; w is the wage per efficiency unit of labour, and r the interest rate on assets; z is the households labour efficiency units. Notice that the problem can

be simplified to only choosing l and a' . The idiosyncratic AR(1) process on labour productivity (labour efficiency units) is given by,

$$z_t = \rho z_{t-1} + \epsilon_t$$

where $\epsilon_t \sim N(0, \sigma_\epsilon^2)$.

Production: The aggregate production function is Cobb-Douglas, $Y = F(K, H) = K^{1-\alpha} H^\alpha$.

Markets: The markets for labour and capital are perfectly competitive.

Definition 2. A *Competitive Equilibrium* is a household value function V ; a set of household policy functions $\{g^{a'}, g^l\}$; a pair prices $\{w, r\}$; and a measure of households $\mu(a, z)$; such that

1. Given prices $\{w, r\}$, the household's value fn V and policy functions $\{g^{a'}, g^l\}$ solve the household's optimization problem.
2. Aggregates are determined by aggregating over households actions, $H = \int z(1-l)d\mu$, and $K = \int ad\mu$.
3. Markets clear, so prices are given by marginal productivities; $r = F_K(K, H) - \delta$, $w = F_H(K, H)$.
4. The measure of households is invariant:

$$\mu(a, z) = \int \int \left[\int 1_{a=g(\hat{a}, z)} \mu(\hat{a}, z) Q(z, dz') \right] d\hat{a} dz$$

By Walras' Law, the aggregate resource constraint will be automatically satisfied, $C + K' = F(K, H) + (1 - \delta)K$.

'True' parameter values: We take the true parameter values of the model to be those found by Pijoan-Mas (2006) who calibrated them based on various data targets. These are given in Table 1, with exception of ℓ which is normalized to one, and so plays no further role in the calibration or estimation.¹⁷

6.2 Implementation of the Estimators

Let's start with a summary of how the model maps into our estimation frameworks.

The Parameters to be Estimated: The model has nine parameters, $\{\beta, \sigma_1, \sigma_2, \chi, \delta, \alpha, \ell, \rho, \sigma_\epsilon\}$. The time endowment, ℓ , has to be normalized: I set it to $\ell=1$. There are eight remaining eight parameters. The 'micro' parameters are $\{\beta, \sigma_1, \sigma_2, \chi, \rho, \sigma_\epsilon\}$ — θ_1 in the notation of our estimators.

¹⁷For robustness purposes, Pijoan-Mas (2006) compares two different parameterizations of ρ and σ_ϵ , we follow the first of his two parameterizations.

Table 1: The true parameters and moments for the model of Pijoan-Mas (2006)

True Parameters							
β	σ_1	σ_2	χ	δ	α	ρ	σ_ϵ
0.945	1.458	2.833	0.856	0.083	0.64	0.92	0.21
Target Moments							
K/Y	I/Y	Labour Share	Avg. Hrs Worked	$c.v.(l)$	$corr(l, \epsilon)$		
3	0.25	0.64	1/3	0.22	0.02		

'Labour share' is the labour share of total income, wL/Y . 'Avg. Hrs Worked' is the average hours worked as a fraction of time available $E(l)/\ell$. $c.v.(l)$ is the coefficient of variance of hours worked.

While the 'macro' parameters are $\{\alpha, \delta\} - \theta_2$ in the notation of our estimators. In implementing the SME the parameters ρ and σ_ϵ that determine the AR(1) process on labour productivity will also be pre-calibrated, in imitation of the approach of Pijoan-Mas (2006) who calibrates them based on estimates of wage processes from panel data, rather than from aggregate moments.

The Target Moments: To be able to apply the SME to this model we must define which data moments we are targeting. We have six parameters to estimate. In choosing which data moments it is important to think about how these moments will help to identify the parameters of the model. In this example the moments to be targeted are the capital-output ratio, investment-output ratio, labour share of income, average hours worked, co-efficient of variation of hours, and the correlation between hours worked and the hourly wage. Since our focus here is methodological, for the important discussion of how this particular choice of moments helps to identify the model parameters the reader is referred to Pijoan-Mas (2006). These target moments are given in Table 1.

The Likelihood Function: To be able to apply the SLE to this model we first need to define the likelihood function of the model. In the first step of the SLE we will be estimating the micro parameters from panel data on labour income (wages), labour supply (hours worked), and asset holdings. Following the standard approach to creating the likelihood function in situations of this kind we assume that all three variables (hours, wages and assets) are observed with measurement error.¹⁸

Suppose that for period t the true wage $w_t = wz_t$, labour supply h_t , and assets a_t , are all observed with measurement error. Denote by $(\xi_{W,t}, \xi_{h,t}, \xi_{a,t})$ the vector of measurement errors in observed labour income, labour supply, and assets, respectively. Assume that labour income measurement error is log-normally distributed with mean 1. That is,

$$W_t^D = W_t \xi_{W,t}, \quad \ln(\xi_{W,t}) \sim N\left(-\frac{1}{2}\sigma_{\xi,W}^2, \sigma_{\xi,W}\right)$$

¹⁸Often, to better approach the data, one would also want to include preference shocks in the value function problem itself. Our model unrealistically predicts that everyone in the economy with the same hourly wage and same assets would choose to work exactly the same number of hours, preference shocks help break this direct-link. Since we use simulated data this concern is not relevant here. See, eg., Imai and Keane (2004) for a structural estimation from panel data on hours, wages, and assets in a life-cycle model.

where W_t is the true labour income at period t (which equals $w_t l_t = w z_t l_t$) and W_t^D is the observed labour income in the data.

Assume that the labour supply measurement error is normally distributed. That is,

$$l_t^D = l_t + \xi_{l,t}, \quad \xi_{l,t} \sim N(0, \sigma_{\xi,l})$$

where l_t is the true labour supply at period t and l_t^D is the observed labour supply in the data.

Assume that the asset holdings measurement error is normally distributed. That is,

$$a_t^D = a_t + \xi_{a,t}, \quad \xi_{a,t} \sim N(0, \sigma_{\xi,a})$$

where a_t is the true asset holdings at period t and a_t^D is the observed asset holdings in the data.

In the model we are not really interested in labour income itself, so much as the wage (per unit of time). Since labour income is just the wage times the labour supply we can get the wage by dividing labour income by labour supply, that is

$$w_t = \frac{W_t}{l_t}$$

and analogously for the observed variables. We thus get the following relationship between the observed wage and the true wage,

$$w_t^D = w_t \frac{l_t}{l_t^D} \xi_{W,t}$$

Finally, for the initial period wage we assume the following measurement error,

$$w_{t_0}^D = w_{t_0} \xi_{w,t_0}, \quad \ln(\xi_{w,t_0}) \sim N\left(-\frac{1}{2}\sigma_{\xi,w0}^2, \sigma_{\xi,w0}\right)$$

Simulated Likelihood is employed to evaluate the likelihood of the model. Denote by $\{z_t^m, l_t^m, a_t^m\}$ the sequence of true wage, true labour supply, and true asset holdings at the m th simulation draw. For each simulation we evaluate the likelihood. We repeat the simulation M times (for each individual in the panel) and evaluate the likelihood. The exact steps involved in this simulation and the likelihood are given in Appendix E.

6.3 Results

In implementing the estimators the value function and optimal policy function are calculated by discrete value function iteration. This involves discretizing not just this periods state, but also the control variables, and next periods state (using the Tauchen method to do the numerical integration). To compute the steady-state distribution and for the SME, the agent's distribution is itself discretized and the iterated upon; this is not standard practice but is important to being able to apply our results on bounding numerical errors to the estimator. To evaluate the likelihood in the SLE, simulations are created using the discretized optimal policy function and the discretized

transition matrix of the exogenous state (both created when solving the discretized value function iteration), with linear interpolation used to accommodate off-grid points caused by both the measurement errors and the fact that empirical data does not fit exactly on our grids (the later is not an issue in the simulations used here, but the first is).

In implementing the SME I have not used the weights given when it was defined earlier. I have instead used $d(A_i^D, A_i^{Ms}) \equiv ((A_i^D - A_i^{Ms})^2)/A_i^D$. That is, I have divided by the actual values of the data moments rather than their variances (and with $w_i = 1$ for all i).¹⁹ This will not effect the consistency of the estimator, only it's efficiency. Given that in this model the estimator appears to converge perfectly to the true values anyway the loss of efficiency is not an issue.

The second issue is what algorithm to use to minimize the objective function. I first discuss this issue for the SME. With the SME I tried using a large number of different algorithms, namely active-set, interior-point, SQP, Nelder-Mead simplex direct search, and CMA-ES. When the estimation began using the true parameter vector as the initial parameter vector, all the algorithms were able to recognize this as the solution. However, when the estimation began from the initial parameter vector ($\beta = 0.9, \sigma_1 = 1.2, \sigma_2 = 2, \chi = 0.7, \delta = 0.05, \alpha = 0.5$) only the CMA-ES algorithm succeeded in converging to the true solution.²⁰ Thus, the Covariance-Matrix Adaptation–Evolutionary Strategy (CMA-ES) algorithm appears to be the only contender for implementing the SME estimator. Going forward I intend to test the ability of the CMA-ES algorithm based on starting from a variety of different initial parameter vectors so as to see just how reliable it is, but I have not yet done this.

I have not yet had the opportunity to test which algorithms work best to minimize the objective function (maximize the likelihood) for the SLE.²¹

7 Conclusion

How about actually applying these estimators? In “Flat-tax reform paper” (Diaz-Gimenez, Pijoan-Mas, & Kirkby; see my website) we implement a version of the Simulated Moments Estimator described in this article. The model does not exactly fit the theoretical framework used here — the return function has a non-concavity (due to the ‘ceiling’ on payroll taxes) and you cannot directly choose next periods assets, due to the estate tax — but it gives good idea of how the estimator can be used in practice. Further, two of the target moments are based on finite time-series simulations and not on the steady-state distribution. It includes a discussion of how to approach choosing the weights for the SME for an example application.

¹⁹This was done since Pijoan-Mas (2006) does not report the sample variances of the target moments, since he has no use for them.

²⁰The active-set, interior-point, and SQP algorithms were implemented using Matlab’s inbuilt *fmincon* function for constrained function minimization. The Nelder-Mead simplex direct search algorithm was implemented using Matlab’s inbuilt *fminsearch* function for unconstrained function minimization. The CMA-ES algorithm was performed using the Matlab implementation provided by Andreasen (2010).

²¹A test using the CMA-ES algorithm is currently running.

Estimations of the general 'style' (heterogeneous agents in a competitive general equilibrium) described here are also implemented in Heckman, Lochner, and Taber (1998), Lee (2005), and Lee and Wolpin (2006) (their estimations are based on models that also include aggregate uncertainty).

There are two main alternative ways in which to estimate these models.

The first alternative are constrained (aka. restricted) estimators. Instead of putting a large weight on the general equilibrium condition (as in the SME), or simply relying on asymptotic properties being a good approximation (as in 2-Stage SLE), we could impose the general equilibrium condition as a constraint. We would then run the same kind of estimation, only performing constrained optimization, rather than the unconstrained optimization performed here. This line of approach is not considered, mainly because attempts to implement such estimators run into problems when trying to use any of matlab's inbuilt constrained optimization algorithms. A constrained version of the CMA-ES algorithm which worked well for the unconstrained models does exist, called (1+1)-CMA-ES, but there does not yet appear to be any existing matlab implementation of this algorithm. One advantage of this approach would be that the general equilibrium conditions could then be tested using standard Likelihood-Ratio and Lagrange-Multiplier tests for testing restrictions.

The second alternative estimator not discussed here is to consider the general equilibrium condition as an extra loop. Applied to the estimators the general equilibrium condition would involve an inner-loop to find the general equilibrium for any given vector of parameters, with the problem of minimizing the objective function (the distance between the model and data moments) being an outer-loop. This approach involves a substantial increase in the computational difficulty of the estimators, and for this reason we do not consider it here. Asymptotically both estimators are consistent. One would like to compare simulation results for these estimators to those presented here. Does the substantial increased computational burden of adding an extra loop provide a substantial improvement in the small sample properties of the estimators?

Note that both of these alternatives are simply different ways of dealing with the general equilibrium condition. It is a question for future research how these alternatives would compete/compare with the methods described in this article.

References

- Daniel Akerberg, John Geweke, and Jinyong Hahn. Comments on "convergence properties of the likelihood of computed dynamic models". *Econometrica*, 77(6):2009–2017, 2009.
- Victor Aguirregabiria and Pedro Mira. Dynamic discrete choice structural models: A survey. *Journal of Econometrics*, 156(1):38–67, 2010.
- S. Rao Aiyagari. Uninsured idiosyncratic risk and aggregate saving, 1993. Working Paper 502.

- S. Rao Aiyagari. Uninsured idiosyncratic risk and aggregate saving. Quarterly Journal of Economics, 109(3):659–684, 1994.
- Martin Andreasen. How to maximize the likelihood function for a dsge model. Computational Economics, 35(2):127–154, 2010.
- Joshua Angrist and Jörn-Steffen Pischke. Mostly Harmless Econometrics! Princeton University Press, 2009.
- Dimitri Bertsekas. Dynamic Programming and Stochastic Control. Academic Press, 1976.
- Truman Bewley. A difficulty with the optimum quantity of money. Econometrica, 51:1485–1504, 1983.
- Truman Bewley. Notes on stationary equilibrium with a continuum of independently fluctuating consumers, 1984.
- Rabi Bhattacharya and Oesook Lee. Asymptotics of a class of markov processes which are not in general irreducible. The Annals of Probability, 16(3):1333–1347, 1988. A correction to this article was published in the same journal in 1997, Vol. 25(3), pg. 1541-1543.
- Rabi Bhattacharya and Mukal Majumdar. On a class of stable random dynamical systems: Theory and applications. Journal of Economic Theory, 96:208–229, 2001.
- Martin Browning, Lars Peter Hansen, and James Heckman. Micro data and general equilibrium models. In John B. Taylor and Mark Woodford, editors, Handbook of Macroeconomics, volume 1. 1999.
- Yongyang Cai and Kenneth L. Judd. Advances in numerical dynamic programming and new applications. In Karl Schmedders and Kenneth L. Judd, editors, Handbook of Computational Economics, volume 3, chapter 8. Elsevier, 2014.
- Ana Castaneda, Javier Díaz-Giménez, and Jose Victor Ríos-Rull. Accounting for the u.s. earnings and wealth inequality. Journal of Political Economy, 111(4):818–857, 2003.
- Juan Carlos Conesa and Dirk Krueger. On the optimal progressivity of the income tax code. Journal of Monetary Economics, 53(7):1425–1450, 2006.
- Juan Carlos Conesa, Sagiri Kitao, and Dirk Krueger. Taxing capital? not a bad idea after all! American Economic Review, 99(1):25–48, 2009.
- David Domeij and Jonathan Heathcote. On the distributional effects of decreasing capital taxes. International Economic Review, 45(2):523–544, 2004.
- Darrell Duffie and Kenneth J. Singleton. Simulated moments estimation of markov models of asset prices. Econometrica, 61(4):929–952, 1993.

- Jesus Fernandez-Villaverde, Juan Rubio-Ramirez, and Manuel Santos. Convergence properties of the likelihood of computed dynamic models. Econometrica, 74(1):93–119, 2006.
- Ragnar Frisch. From utopian theory to practical applications: The case of econometrics (nobel prize acceptance speech), 1969.
- Jonathan Heathcote, Kjetil Storesletten, and Giovanni Violante. Quantitative macroeconomics with heterogeneous households. Annual Review of Economics, 1(5):319–354, 2009.
- James Heckman, Lance Lochner, and Christopher Taber. Explaining rising wage inequality: Explorations with a dynamic general equilibrium model of labor earnings with heterogeneous agents. Review of Economic Dynamics, 1(1):1–58, 1998.
- Hugo Hopenhayn and Edward C. Prescott. Stochastic monotonicity and stationary distributions for dynamic economies. Econometrica, 60(6):1387–1406, 1992.
- Mark Huggett. The risk-free rate in heterogenous-agent incomplete-insurance economies. Journal of Economic Dynamics and Control, 17:953–969, 1993.
- Susumu Imai and Michael Keane. Intertemporal labor supply and human capital accumulation. International Economic Review, 45(2):601–641, 2004.
- Kamihigashi and John Stachurski. An order-theoretic mixing condition for monotone markov chains. Statistics and Probability Letters, 82:262–267, 2011.
- Michael Keane. Structural vs. atheoretic approaches to econometrics. Journal of Econometrics, 156(1):3–20, 2010.
- Michael Keane, Petra Todd, and Kenneth Wolpin. The structural estimation of behavioral models: Discrete choice dynamic programming methods and applications. In D. Card and O. Ashenfelter, editors, Handbook of Labor Economics, chapter 4. Elsevier, 2011.
- C. Le Van and John Stachurski. Parametric continuity of stationary distributions. Economic Theory, 33:333–348, 2007.
- Bong-Soo Lee and Beth Fisher Ingram. Simulation estimation of time-series models. Journal of Econometrics, 47:197–205, 1991.
- Donghoon Lee. An estimable dynamic general equilibrium model of work, schooling, and occupational choice. International Economic Review, 46(1):1–34, 2005.
- Donghoon Lee and Kenneth I. Wolpin. Intersectoral labor mobility and the growth of the service sector. Econometrica, 74(1):1–46, 2006.
- Albert Marcet, Francesc Obiols-Homs, and Philippe Weil. Incomplete markets, labor supply and capital accumulation. Journal of Monetary Economics, 54(8):2621–2635, 2007.

- Jiaojun Miao. Competitive equilibria of economies with a continuum of consumers and aggregate shocks. Journal of Economic Theory, 128(1):274–298, 2006.
- Jeno Pál and John Stachurski. Fitted value function iteration with probability one contractions. Journal of Economic Dynamics and Control, 37:251–264, 2013.
- Adrian Peralta-Alva and Manuel Santos. Analysis of numerical errors. In Karl Schmedders and Kenneth L. Judd, editors, Handbook of Computational Economics, volume 3, chapter 9. Elsevier, 2014.
- Josep Pijoan-Mas. Precautionary savings or working longer hours? Review of Economic Dynamics, 9(2):326–352, 2006.
- Vicenzo Quadrini. Entrepreneurship, saving and social mobility. Review of Economic Dynamics, 3(1):1–40, 2000.
- Juan Pablo Rincón-Zapatero and Manuel Santos. Differentiability of the value function without interiority assumptions. Journal of Economic Theory, 144(5):1948–1964, 2009.
- José Víctor Ríos-Rull. Models with heterogenous agents. In T. Cooley, editor, Frontiers in Business Cycle Research, chapter 4. Princeton University Press, 1995.
- José Víctor Ríos-Rull. Computation of equilibria in heterogenous agent models. In R. Marimon and A. Scott, editors, Computational Methods for the Study of Dynamic Economies, chapter 11. Oxford University Press, 2 edition, 2001.
- José-Víctor Ríos-Rull, Frank Schorfheide, Cristina Fuentes-Albero, Maxym Kryshko, and Raul Santaeulalia-Llopis. Methods versus substance: Measuring the effects of technology shocks. Journal of Monetary Economics, 59(8):826–846, 2012.
- Agnar Sandmo. Economics Evolving: A History of Economic Thought. Princeton University Press, 2011.
- Manuel Santos. Accuracy of numerical solutions using the euler equation residuals. Econometrica, 68(6):1377–1402, 2000.
- Manuel Santos and Adrian Peralta-Alva. Accuracy of simulations for stochastic dynamic models. Econometrica, 66:409–426, 2005.
- Manuel Santos and Jesus Vigo-Aguiar. Analysis of a numerical dynamic programming algorithm applied to economic models. Econometrica, 66:409–426, 1998.
- John Stachurski. Continuous state dynamic programming via nonexpansive approximations. Computational Economics, 31(2):141–160, 2008.

O. Stenflo. Ergodic theorems for markov chains represented by iterated function systems. Bulletin of the Polish Academy of Sciences: Mathematics, 49:27–43, 2001.

Nancy Stokey, Robert E. Lucas, and Edward C. Prescott. Recursive Methods in Economic Dynamics. Harvard University Press, 1989.

George Tauchen. Finite state markov-chain approximations to univariate and vector autoregressions. Economics Letters, 20:177–181, 1986.

Whitt. Approximations of dynamic programs, i. Mathematics of Operations Research, 3:231–243, 1978. There is also a II (1979), but I do not refer to it.

A Numerical Errors in the Value Function and Optimal Policy Function

This appendix derives numerical error bounds, both for the value function and the optimal policy function that depend only on monotonicity of the return function. These bounds are applicable to the kinds of models in which other numerical methods cannot be used and value function iteration becomes the method of choice. For example, models of incomplete markets often involve periodically binding borrowing constraints; this leads to a situation in which existing bounds for numerical errors cannot be applied, but those derived here can. The numerical error bounds derived are too loose to be useful in practice.²² However, since the bounds go to zero as the distance between the grid points goes to zero these bounds can be used to prove that numerical errors will go to zero asymptotically. It is to exactly this purpose of proving that numerical errors go to zero asymptotically that these bounds are used in the body of this article. The first few pages of this appendix provide a discussion of the numerical error bounds derived, the conditions they are valid under, a comparison to other results in the literature, and a summary of the main results. The rest of the appendix consists of the derivation of these results.

While slower than other numerical methods value function iteration has certain advantages. One is widespread applicability: many other numerical methods are based on the first-order conditions (FOCs) and so are valid only under certain conditions (eg. the FOCs are a sufficient condition if the return function is continuous, differentiable, and concave; and the choice set is a convex set). Existing theory bounding the numerical errors from value function iteration assume properties — such as interiority, differentiability, strongly concave return functions, and convex choice sets — that make the FOCs are necessary and sufficient so other numerical methods are likely to be used. The bounds for numerical errors arising from value function iteration derived here allow for situations in which the FOCs are not necessary and sufficient conditions, and so most other solution methods are invalid and value function iteration becomes a standard choice.

The numerical bounds here are derived based on discretized value function iteration. They exactly mimic the discretized value function iteration algorithm as it is commonly implemented. Namely by discretizing the state variables, discretizing the control variables (and hence the maximization step), and discretizing the numerical integration (eg. by quadrature methods such as the Tauchen method).²³ Discretized value function iteration is chosen as the basis for the numerical error bounds on the grounds that it is both commonly used and is robust to the kind of situations

²²The numerical error bounds are loose in the sense that the bounds tend to be orders of magnitude larger than the actual numerical errors. A tight bound would mean that the bounds were of roughly the same order of magnitude as the errors that occur.

²³Alternatives to all of these approaches exist. Rather than 'pure' discretization of the state variables the value function can be approximated by fitted value function methods, such as approximating the value function by splines or polynomials. Instead of discretizing the control variables the maximization step can be solved using optimization algorithms, such as binary search, or using the endogenous grid method or envelope condition method. Instead of discretizing the integral with a quadrature method one can use Monte-Carlo integration methods.

— non-differentiability, non-interior choices, non-convex choice sets — which are of interest to us in the application of these numerical error bounds.

Useful references for this appendix include Bertsekas (1976) who takes a similar approach to discrete state space approximations in value function iteration, but only for the finite-horizon case with a discrete iid exogenous shock process, also derives bounds for the optimal policy function, but which depend on strong concavity of the return function. Closest to the results provided here, Whitt (1978) gives numerical error bounds for the infinite horizon value function problem with general shock processes (in fact he allows for what SLP call Case 2 value function problems, and further for state dependent shock processes). However he does not go on to provide numerical error bounds for the optimal policy; essential for our purposes. The results presented here are largely a combination of these two earlier results, but with two further additions — dropping the requirement that the return function is strongly concave²⁴, and explicitly showing how the numerical errors in the value function vary between different parts of the state space.

Other existing results tend to be based on 'partial discretization': considering the errors from approximating the state space, but not the choice variables, nor the the numerical integration. This has the advantage that they are able to derive tighter bounds on numerical errors, and also results relating to the speed at which numerical errors go to zero asymptotically, and on using fitted value function methods. The disadvantage is that they also mostly require a degree of differentiability of the value function, interiority of optimal policies, convexity of choice sets, and concavity of the return function. Santos and Vigo-Aguiar (1998) work with the infinite horizon case looking at partial discretization and using a fitted value function (specifically the value function is modeled using finite elements methods, rather than being a single number for each point in the discretized state space). They also provide results on the speed of convergence to the true value function; namely that it is quadratic in the grid size; and that convergence to the policy function is linear in the grid size. Their numerical error bounds are tighter than those derived here, tight enough to be useful in practice. Stachurski (2008) provides further results on numerical error bounds for partial discretization using a variety of more sophisticated fitted value function methods such as approximating the value function by certain types of splines or polynomials; he shows that the numerical errors resulting from a number of fitted value function methods popular in the literature go to zero asymptotically as the grids get finer. Of general interest are Stachurski's results on shape-preserving fitted approximation methods, a methodology advocated for smooth problems by Cai and Judd (2014). Pál and Stachurski (2013) show how these results can allow for periodically-binding constraints, non-differentiable return function and value function, and non-convex choice set, but do not consider the errors from approximating the choice variables; their results are based on fitted value function iteration, with Monte-Carlo integration.

²⁴We still require that the return function is strictly concave as this is required to prove that the value function will converge and the optimal policy is unique. Strongly concave is a stronger assumption that is by Bertsekas (1976), among others, to derive bounds for the numerical error in the optimal policy function.

In allowing for periodically-binding constraints, non-differentiable return function and value function, and non-convex choice set these errors are more general than those previously derived, and are able to be applied to the value function problems found in many models of the Bewley-Huggett-Aiyagari class. Three possible situations remain for which the error bounds are not applicable: if the slope of the return or value function goes to infinity (say due to Inada conditions), if the return function contains a non-concavity, and if next periods state cannot be chosen directly. These are addressed in turn: (i) the slope of a return or value function going to infinity would occur when Inada conditions are present. At first glance this seems problematic given the prevalence of Inada conditions, but it is actually not likely to be any problem at all. The presence of Inada conditions is generally used to prove theoretically that there exists, eg., a minimum level of consumption, and thus we can work with a bounded space defined using that minimum level of consumption, this ensures all the functions are bounded and thus that we can apply the standard theorems for bounded value functions, having redefined the problem on this new space, the problem of the slope of a return or value function going to infinity would no longer occur, and we could apply the error bounds derived here as usual. (ii) when the return function contains a non-concavity the error bounds are not valid. If one assumes that the non-concavities are of a limited size and are localized some modified results could likely be derived. No attempt to do so is made here. (iii) if next periods state cannot be chosen directly the error bounds derived here are invalid, but extensions to allow for this could be made, Whitt (1978) provides some results of this nature.

A.1 The Results

We provide a quick overview of our two main theorem bounding numerical errors between the solution to the discretized value function iteration problem and corresponding discretized optimal policy function, respectively. We then derive as a Proposition that the numerical errors in the value function and the optimal policy function will go to zero as the distance between the grid points of the discretization goes to zero (ie. as the grid get ever larger and finer). It is this Proposition that is used in the body of this paper. The rest of this appendix then provides a more formal statement, and proofs, of these two theorems.

Our results are based on the Case 1 value function problem,

$$V(x, z) = \sup_{y=(y_1, y_2) \in \Gamma(x, z)} \left\{ F(x, y, z) + \beta \int V(y_1, z') Q(z, dz') \right\} \quad (8)$$

Our intention is to bound the difference between the solution to the this problem, V , and the numerical solution to the discretized problem, V_N^G , the later is the solution after N iterations (once a standard convergence criterion is met) of value function iteration on the discrete grid (hence the G).

We assume that the spaces for the endogenous state X , the control variables Y , and the exoge-

nous state Z are all compact, that the return function F is continuous and bounded²⁵, the discount factor β is less than one, and that the transition function Q has the Feller property.

The bound on the numerical errors in the value function is given by

Theorem 3. *Let V be the value function defined in equation (8). Let $\{V_n^G\}$ be the sequence of functions generated by iteratively applying T^G starting from a function V_0 (ie. $V_1 = T^G V_0$, $V_n = T^G V_{n-1}$). Let V_N^G be the value function at which the algorithm stops; given the stopping criterion $\|V_n - V_{n-1}\| \leq \epsilon_V$. Then*

$$|V(x, z) - V_N^G(x, z)| \leq \frac{1}{1 - \beta} [K_{Fy_1}(i, j) + K_{Fy_2}(i, j) + K_{Vx}(i, j) + K_{Vz}(i, j) + \beta(K_{EVx}(i, j) + K_{EVz}(i, j) + K_{EV}(i, j)) + \beta\epsilon_V]$$

where $K_{Fy_1}(i, j)$, $K_{Fy_2}(i, j)$, $K_{Vx}(i, j)$, $K_{Vz}(i, j)$, $K_{EVx}(i, j)$, $K_{EVz}(i, j)$, and $K_{EV}(i, j)$ are constants that depend on the 'nearest' grid points (i, j) . The constants capture: the numerical errors that occur due to discretizing the control variable ($K_{Fy_1}(i, j)$, $K_{Fy_2}(i, j)$); the errors that occur from discretizing this periods state in this periods value function ($K_{Vx}(i, j)$, $K_{Vz}(i, j)$) and the expectation of next periods value function ($K_{EVx}(i, j)$, $K_{EVz}(i, j)$); and the errors that come from the numerical integration being discretized ($K_{EV}(i, j)$). All the constants are larger when the grid points are further apart, and when the slopes of the return function and value function are 'steeper'.

A bound for those in the optimal policy function is

Theorem 4. *Under Assumptions 1, 2, and 3. For any (x, z) in the partition $X_i \times Z_j$, $i = 1, \dots, n_x, j = 1, \dots, n_z$. y_a and y_c as defined in (28) and (30) satisfy*

$$|y_a(x, z) - y_c(x, z)| \leq dy_l(x, z) \tag{9}$$

where l is given by

$$dy_l(x, z) = \max\left\{\sum_{a=0}^{l_-} dy^{g(-a)}(x, z), \sum_{a=0}^{l_+} dy^{g(+a)}(x, z)\right\}$$

$$l_- = \operatorname{argmin}\left\{\sum_{a=0}^l K_{Fy}^{g(-a)} \geq K_{Fx}^g(i, j) + \beta(K_{Vx}^g(i, j) + K_{Vz}^g(i, j) + K_V^g(i, j)) + \delta_V(i, j)\right\}$$

$$l_+ = \operatorname{argmin}\left\{\sum_{a=0}^l K_{Fy}^{g(+a)} \geq K_{Fx}^g(i, j) + \beta(K_{Vx}^g(i, j) + K_{Vz}^g(i, j) + K_V^g(i, j)) + \delta_V(i, j)\right\}$$

Proofs of these theorems, as well as precise definitions of the constants that make up these numerical error bounds occupies much of this appendix.

²⁵Since F is a continuous function defined on a compact space, it will therefore also be bounded.

Proposition 1. *Under the conditions of Theorems 3 and 4. The numerical errors in the value function, $|V(x, z) - V_N^G(x, z)|$, and the numerical errors in the optimal policy function, $|y_a(x, z) - y_c(x, z)|$, go to zero as the maximum distance between the grid points in the dimensions being discretized go to zero.*

Proof. Follows trivially from Theorems 3 and 4, and from the definitions of the constants (all of which go to zero, since they are simply combinations of the distances between grid points). *Q.E.D.*

In summary, we have derived numerical error bounds for both the value function and the optimal policy function, and shown that they go to zero asymptotically as the distance between grid points goes to zero. They are based on the use of value function iteration with a discretized state space; exactly as it is implemented in the computer.²⁶ We introduced two innovations to the standard approach, the calculation of pointwise bounds, and allowing for non-strongly-concave return functions. The numerical error bounds derived are applicable more widely than existing bounds allowing, among other things, for borrowing constraints, and decisions to work zero hours.

The rest of the appendix is concerned with defining the relevant constants and proving the results given in Theorems 3 & 3. We begin by bounding the numerical errors in the value function, and then turn to the optimal policy function.

A.2 Numerical Error Bounds for the Value Function Iteration

We first bound the distance between the solution to the value function iteration and the true value function, temporarily ignoring the issue of discretization. The convergence of the iterated value function to the true value function is well known. But in practice our value function iteration must stop after a finite number of steps, so it will never reach the true value function. Bounding the distance between the solution given by value function iteration and the true value function is our first step, and involves a well-known property of contraction mappings. We give a brief treatment of this issue drawing on Stokey, Lucas, and Prescott (1989) (henceforth SLP), which doubles as an introduction to the notation used in this paper. This forms Section A.2.1.

But we never get to actually solve the value function iteration — we solve a discrete state space approximation to the value function iteration problem. Our second step is to bound the size of the further errors introduced in by the discretization. This forms Section A.2.2.

Having bounded both of these distances we then simply apply the triangle inequality to get

²⁶A possible source of errors not considered here are rounding errors. Rounding errors arise in all numerical solutions since the accuracy of the numbers which computers can handle is limited so rounding errors will occur (with double floating point numbers the computer is limited to an accuracy of about 15 to 17 digits). Bounds on these could be derived following almost exactly the same methods as used here. Results in Santos and Vigo-Aguiar (1998) (see section 'Stability of the Numerical Method') suggest they will anyhow be orders of magnitude smaller than those we deal with.

a bound on the distance between the solution to the discrete state space dynamic programming problem and the true value function. Section A.2.6 puts this all together.

A.2.1 Uniform Convergence of Value Function Iteration

Results for the uniform convergence of value function iteration are well-known, so we cover them only briefly (a detailed treatment of these same results forms Appendix A.5). Those based on bounded returns are relevant to the models treated here. Our proofs of the validity of discrete state space approximation all require that the spaces for the exogenous and endogenous variables are compact²⁷ so the requirement that the return function be bounded on these spaces is generally trivial. We study value function problems (aka. stochastic dynamic programming, aka. functional equations) of what SLP refer to as type 1. That is ones of the form

$$V(x, z) = \sup_{y \in \Gamma(x, z)} \left\{ F(x, y, z) + \beta \int_{\mathcal{Z}} V(y, z') Q(z, dz') \right\} \quad (10)$$

under the assumption that the return function F is bounded and continuous, the discount factor β is strictly less than one, and the transition function Q has the Feller property.

Preliminaries: Let (X, \mathcal{X}) and (Z, \mathcal{Z}) be measurable spaces of possible values for the endogenous and exogenous state variables, respectively; let $(S, \mathcal{S}) = (X \times Z, \mathcal{X} \times \mathcal{Z})$ be the product space; let Q be a transition function on (Z, \mathcal{Z}) ; let $\Gamma : S \rightarrow X$ be a correspondence describing the feasibility constraints; let A be the graph of Γ ; let $F : A \rightarrow \mathbb{R}$ be the one-period return function; and let $\beta \geq 0$ be the discount factor.

Our metric for the space $C(S)$ is the sup norm, $\|f\| = \sup_{s \in S} |f(s)|$. It is stressed that many of the results below apply much more broadly, and the arguments used here can easily be adapted to other situations.

It is well known that under some general assumptions (see SLP Theorem 9.6) that defining the operator T on $C(S)$ by

$$(Tf)(x, z) = \sup_{y \in \Gamma(x, z)} \left\{ F(x, y, z) + \beta \int_{\mathcal{Z}} f(y, z') Q(z, dz') \right\} \quad (11)$$

Then $T : C(S) \rightarrow C(S)$; T has a unique fixed point V in $C(S)$ and for any $V_0 \in C(S)$,

$$\|T^n V_0 - V\| \leq \beta^n \|V_0 - V\|, \quad n = 1, 2, \dots \quad (12)$$

Moreover, the correspondence $G : S \rightarrow X$ defined by

$$G(x, z) = \left\{ y \in \Gamma(x, z) : V(x, z) = F(x, y, z) + \beta \int_{\mathcal{Z}} V(y, z') Q(z, dz') \right\} \quad (13)$$

is nonempty, compact-valued, and u.h.c.

²⁷Compactness is needed to guarantee that we can limit the distance between any two points on the finite grid.

This result suggests the approach to calculating the true value function V known as value function iteration. Namely, starting from an initial function $V_0 \in C(S)$ we can apply the mapping T defined in (11) to generate a new function V_1 . Iterating on this procedure, ie. $V_n = TV_{n-1}$ we get a sequence $V_0, V_1, \dots, V_n, \dots$ of functions. By (12) we know that $V_n \rightarrow V$ as $n \rightarrow \infty$, and in fact it also tells us the speed of this convergence. Thus the results above prove that value function iteration is globally convergent to the true value function, and gives us a rate of convergence.

So we know that using the value function iteration algorithm our solution will converge to the true value function. However in practice we have to stop after a finite number of iterations. Can we know how close we have ended up? The standard way to decide when to stop is based on a convergence criterion of the form $\|V_n - V_{n-1}\| \leq \epsilon_V$. A well known result for contraction mappings is that the distance of the function V_N at which the algorithm terminates from the true value function based on the convergence criterion satisfies

$$\|V_N - V\| \leq \frac{\beta}{1 - \beta} \epsilon_V \tag{14}$$

We now turn to accounting for approximation errors introduced by the fact that with the computer we cannot solve this problem exactly but only a discrete approximation of it.

A.2.2 Discrete State Space Approximation

We now derive numerical error bounds for the distance between the solution to the value function iteration and the solution to the discretized value function iteration. Our bounds are dependent on two assumptions, monotonicity and concavity, although both are required only in the variables to be discretized. Thus we do not require differentiability, interiority of solutions, or convex choice sets, as in many existing error bounds. The other difference of the approach followed here from others in the literature has two aspects. Firstly, rather than find uniform bounds directly, we find bounds at each point and then take the maximum across the grid points, when this approach is combined with those for bounding errors in optimal policies the improvement is quite substantial.

A.2.3 The Discretization Procedure

Our results are based on the Case 1 value function problem,

$$V(x, z) = \sup_{y=(y_1, y_2) \in \Gamma(x, z)} \left\{ F(x, y, z) + \beta \int V(y_1, z') Q(z, dz') \right\} \tag{15}$$

Our intention is to bound the difference between the solution to the this problem, V , and the numerical solution to the discretized problem, V_N^G . The approach taken is to bound the distances of each of these to solution to the discretized problem, V^G . The bounding of the distance between V and V^G follows closely the approach in Santos and Vigo-Aguiar (1998), albeit without their

assumptions of differentiability. The bounding of the distance between V_N^G and V^G is a well-known property of convergence mappings.

We will also make use of the definition,

$$\mathcal{V} = \{V : X \times Z \rightarrow \mathbb{R} \mid V \text{ is bounded and continuous}\}$$

In what follows the spaces are always assumed to be bounded. This is done as it implies that the return function will be bounded (if it is continuous) and thus that the value function is bounded. It is an appropriate assumption since the discretized state spaces will in any case always be bounded (since they are by definition finite sets). It is not however essential (see Santos and Vigo-Aguiar (1998)), and often does not require the theoretical state space to be bounded or the return function to actually be bounded (see the example in Section 5.1 of SLP).

We discretize the current states, x & z , the choice variable y , and the next period exogenous state z' . The same grid is used for both this and next periods exogenous state. Those elements of the choice variable which correspond to next periods endogenous state use the grid of the latter.

Assumption 1. *Assume,*

- $X \subseteq \mathbb{R}^l$ is compact.
- Control space $Y = Y_1 \times Y_2$, $Y_1 = X$, $Y_2 \subseteq \mathbb{R}^q$ is compact. $[\Gamma(x, z) \subseteq Y, \forall x \in X, z \in Z]$.
- $Z \subseteq \mathbb{R}^k$ is compact. Q is a probability measure.
- Return fn $F(x, y, z)$ is bounded in $x, y, \& z$, increasing in x & z , decreasing in y .

Note that F is bounded implies that V is also bounded.

We partition X into n_x mutually disjoint sets X_1, \dots, X_{n_x} such that $X = \cup_{i=1}^{n_x} X_i$, and select arbitrary points $x_i \in X_i$, $i = 1, \dots, n_x$. Thus, our grid is $X_G = \{x_1, \dots, x_{n_x}\}$. Define the X -grid size

$$d_X = \max_{x \in X} \min_{\hat{x} \in X_G} \|x - \hat{x}\|$$

We partition Z into n_z mutually disjoint sets Z_1, \dots, Z_{n_z} such that $Z = \cup_{j=1}^{n_z} Z_j$, and select arbitrary points $z_j \in Z_j$, $j = 1, \dots, n_z$. Thus, our grid is $Z_G = \{z_1, \dots, z_{n_z}\}$. Define the Z -grid size

$$d_Z = \max_{z \in Z} \min_{\hat{z} \in Z_G} \|z - \hat{z}\|$$

Our partition for $S = X \times Z$ is simply the product of the X and Z partitions. So the grid is given by $S_G = X_G \times Z_G$ and consists of $n_x n_z$ points. We partition next periods exogenous state, z' , in exactly the same way as this periods exogenous state z . We partition Y_1 implicitly via as the intersection of Y_1 and the aforementioned partitioning of X . So we define $Y_{1G} = Y_1 \cap X_G$.

We partition Y_2 into n_{y_2} mutually disjoint sets $Y_{21}, \dots, Y_{2n_{y_2}}$ such that $Y_2 = \cup_{j=1}^{n_{y_2}} Y_{2j}$, and select arbitrary points $y_{2j} \in Y_{2j}$, $j = 1, \dots, n_{y_2}$. Thus, our grid is $Y_{2G} = \{y_{21}, \dots, y_{2n_{y_2}}\}$. Define the Y_2 -grid size

$$d_{Y_2} = \max_{y_2 \in Y_2} \min_{\hat{y}_2 \in Y_{2G}} \|y_2 - \hat{y}_2\|$$

Our partition for $Y = (Y_1, Y_2)$ is simply the combination of the Y_1 and Y_2 partitions. So the grid is given by $Y_G = (Y_1 \cap X_G, Y_{2G})$ and consists of $n_x + n_{y_2}$ points.

We assume that the discretization process for the transition function satisfies

Assumption 2. Q^G is defined as $Q^G(z, z_i) = Q(z, Z_i)$, $i = 1, \dots, n_z$

This assumption is not necessary to derive bounds, but does make them much tighter. Notice that, in the case where z has been previously discretized this assumption will be satisfied by the Tauchen Method (see Tauchen (1986)).

Consider the space of piecewise constant functions (aka step fns),

$$\begin{aligned} \mathcal{V}^G = \{ & V^G : X \times Z \rightarrow \mathbb{R} | V^G \text{ is bounded, continuous, and} \\ & V^G \text{ is constant in } (X_i, Z_j), i = 1, \dots, n_x, j = 1, \dots, n_z \} \end{aligned}$$

Observe that \mathcal{V}^G is a closed subspace of \mathcal{V} equipped with the norm $\|V^G\| = \sup_{(x,z) \in X \times Z} |V^G(x, z)|$ for $V^G \in \mathcal{V}^G$.

Define the mapping $T^G : \mathcal{V} \rightarrow \mathcal{V}^G$ given by

$$T^G(V)(x_i, z_j) = \sup_{y \in \Gamma^G(x_i, z_j)} F(x_i, y, z_j) + \beta \sum_{k=1}^{n_z} V(y, z_k) Q^G(z_j, z_k) \quad (16)$$

for each point (x_i, z_j) & $V \in \mathcal{V}$; where Q^G is defined by $Q^G(z, z_i) = Q(z, Z_i)$ for $i = 1, \dots, n_z$, and zero everywhere else; and $\Gamma^G(x, z) \equiv \Gamma(x, z) \cap Y_G$. Here the maximization & integration are both discretized. Also, nodal values $T^G(V)(x_i, z_j)$ for all the vertex points (x_i, z_j) yield a unique functional extension to the domain $X \times Z$ over the space of piecewise constant functions compatible with a given partition $\{X_i \times Z_j\}$.

The following functional equation will play a central role in the analysis

$$V^G(x_i, z_j) = \sup_{y \in \Gamma^G(x_i, z_j)} F(x_i, y, z_j) + \beta \sum_{k=1}^{n_z} V^G(y, z_k) Q^G(z_j, z_k) \quad (17)$$

for each vertex point (x_i, z_j) . This is the relevant discretized version of the Bellmans equation.

Lemma 1. *Under Assumption 1, equation (17) has a unique solution V^G in \mathcal{V}^G*

Proof. The proof is a standard one. One immediately sees that T^G is a contraction mapping with modulus $0 < \beta < 1$. By a well known fixed point theorem, equation 17 has a unique fixed point V^G in \mathcal{V}^G . Q.E.D.

A.2.4 Some Constants

We now define a variety of constants which will form part of our error bounds. First a comment on notation. Since x (and y, z) is potentially of more than one dimension x_i and x_{i-1} have no obvious relation. For this purpose we introduce the notation $x_i(-1)$ to be one grid point less than x_i in every (continuous) dimension (which is being discretized); likewise $x_i(+1)$ is one grid point more. We define a constants that are different for each (i, j) grid point. This has the advantage that we will later derive bounds on the numerical errors in the value function from which it is clear that the numerical errors in the value function tend to be larger where the grid points are further apart, and where the slopes of the value function and the return function are 'steeper'.

- Bound on the value function in the x -dimension.

$$K_{Vx}(i, j) = \max\{|V(x_i(+1), z_j) - V(x_i, z_j)|, |V(x_i, z_j) - V(x_i(-1), z_j)|\} \quad (18)$$

- Bound on the value function in the z -dimension.

$$K_{Vz}(i, j) = \max\{|V(x_i, z_j(+1)) - V(x_i, z_j)|, |V(x_i, z_j) - V(x_i, z_j(-1))|\} \quad (19)$$

- Bound on the expectation of the value function in the x -dimension.

$$K_{EVx}(i, j) = \max_{x_l \in X^G} \sum_{k=1}^{n_z} K_{Vx}(x_l, z_k) Q^G(z_j, z_k) \quad (20)$$

- Bound on the expectation of the value function in the z -dimension.

$$K_{EVz}(i, j) = \max_{x_l \in X^G} \sum_{k=1}^{n_z} K_{Vz}(x_l, z_k) Q^G(z_j, z_k) \quad (21)$$

- Bound on the return function in the y_1 -dimension.

$$K_{Fy_1}(i, j) = \max_{(y_{1k}, y_{2l}) \in Y^G} \max\{|F(x_i, y_{1k}, y_{2l}, z_j) - F(x_i, y_{1k}(-1), y_{2l}, z_j)|, \\ |F(x_i, y_{1k}(+1), y_{2l}, z_j) - F(x_i, y_{1k}, y_{2l}, z_j)|\} \quad (22)$$

- Bound on the return function in the y_2 -dimension.

$$K_{Fy_2}(i, j) = \max_{(y_{1k}, y_{2l}) \in Y^G} \max\{|F(x_i, y_{1k}, y_{2l}, z_j) - F(x_i, y_{1k}, y_{2l}(-1), z_j)|, \\ |F(x_i, y_{1k}, y_{2l}(+1), z_j) - F(x_i, y_{1k}, y_{2l}, z_j)|\} \quad (23)$$

- Bound on the errors resulting from numerical integration.

$$K_{EV}(i, j) = \max_{x_l \in X^G} \sum_{k=1}^{n_z} |V(x_l, z_k(+1)) - V(x_l, z_k(-1))| Q^G(z_j, z_k) \quad (24)$$

Remark: These constants will all exist and be finite since F , and therefore V are assumed to be bounded.

Remark: I use two different constants here, $K_{EV_x}(i, j)$ and $K_{EV_z}(i, j)$ to illustrate this possibility. One could simply use one constant to bound of V in both the x and z dimensions as it makes things simpler. But in trying to get tighter bounds it is useful to note that the only gradients of the function being discretized that matter are those occurring in dimensions along which discretization is occurring: what I do here to separate x and z can of course equally be used to separate two different dimensions (variables) in x , the obvious extension along these lines provides the proof that we can ignore the differences in the discrete dimensions (which do not need to be discretized) as the constant associated with an already discrete dimension will be zero.

A.2.5 Intermediate Results

We begin by bounding the difference between the true discretized value function and that we get from our iteration, the result follows directly from equation (14).

Corollary 3. *Let V^G be the value function defined in equation (17). Let $\{V_n^G\}$ be the sequence of functions generated by iteratively applying T^G starting from a function V_0 (ie. $V_1 = T^G V_0$, $V_n = T^G V_{n-1}$). Let V_N^G be the value function at which the algorithm stops; given the stopping criterion $\|V_n - V_{n-1}\| \leq \epsilon$. Then*

$$\|V^G - V_N^G\| \leq \frac{\beta}{1 - \beta} \epsilon$$

We now turn to bounding the difference between the true value function and the true discretized value function. To do this we break the current problem of the errors that come from discretizing (x, z, y, z') up into a series of problems of discretizing first (x, z) , then y and finally z' , these are then combined later in Lemma 5.

In the same way that $T^{G(x,y,z,z')} = T^G$ has been defined above as the discretization of all the variables, we will now define operators for each step of the discretization. Thus, Let $T^{G(x,z)}$ be the contraction mapping associated with the discretization of x and z , thus $T^{G(x,z)} : \mathcal{V} \rightarrow \mathcal{V}^G$ is given by

$$T^{G(x,z)}(V)(x_i, z_j) = \sup_{y \in \Gamma(x_i, z_j)} F(x_i, y, z_j) + \beta \int_Z V(y, z') Q(z_j, dz') \quad (25)$$

for each point (x_i, z_j) & $V \in \mathcal{V}$. Here the maximization & integration procedures are performed exactly (as we are not discretizing y or z').

Let $T^{G(y|x,z)}$ be the contraction mapping associated with the discretization of y , when x & z are already discrete, thus $T^{G(y|x,z)} : \mathcal{V}^G \rightarrow \mathcal{V}^G$ is given by

$$T^{G(y|x,z)}(V)(x, z) = \sup_{y \in \Gamma^G(x,z)} F(x, y, z) + \beta \int_Z V(y, z') Q(z, dz') \quad (26)$$

for each $V \in \mathcal{V}$. Where $\Gamma^G(x, z) \equiv \Gamma(x, z) \cap Y_G$. Here the maximization is no longer performed exactly (integration is still performed exactly as we are not discretizing z').

Let $T^{G(z'|x,y,z)}$ be the contraction mapping associated with the discretization of z' when $x, y, & z$ are already discrete, thus $T^{G(z'|x,y,z)} : \mathcal{V}^G \rightarrow \mathcal{V}^G$ is given by

$$T^{G(z'|x,y,z)}(V)(x, z) = \sup_{y \in \Gamma(x,z)} F(x, y, z) + \beta \sum_{i=1}^{n_z} V(y, z_i) Q^G(z, z_i) \quad (27)$$

for each $V \in \mathcal{V}$. Where Q^G is defined by $Q^G(z, z_i) = Q(z, Z_i)$ for $i = 1, \dots, n_z$, and zero everywhere else. Here the integration is no longer performed exactly.

Our result for the discretization of (x, z) is

Lemma 2. *Let V be the value fn defined in equation (15). Let $K_{V_x}(i, j)$, $K_{V_x}(i, j)$, $K_{EV_x}(i, j)$ & $K_{EV_z}(i, j)$ be as defined in equations (18), (18), (20), & (20). Then under Assumption 1 for (x, z) in a given partion $X_i \times Z_j$, it holds that*

$$|V(x, z) - T^{G(x,z)}V(x, z)| \leq (K_{V_x}(i, j) + K_{V_z}(i, j)) + \beta(K_{EV_x}(i, j) + K_{EV_z}(i, j))$$

Proof. The proof is standard, and follows closely that in Santos and Vigo-Aguiar (1998).

Note that by defns of T & $T^{G(x,z)}$ we have that $TV(x_i, z_j) = V(x_i, z_j) = T^{G(x,z)}V(x_i, z_j)$ for every vertex point (x_i, z_j) , and that the function $T^{G(x,z)}V$ is piecewise constant.

Consider now an arbitrary point (x, z) is a given partion $X_i \times Z_j$. Then

$$\begin{aligned} |V(x, z) - T^{G(x,z)}V(x, z)| &\leq |V(x, z) - V(x_i, z_j)| + |T^{G(x,z)}V(x, z) - T^{G(x,z)}V(x_i, z_j)| \\ &\quad + |V(x_i, z_j) - T^{G(x,z)}V(x_i, z_j)| \\ &\leq (K_{V_x}(i, j) + K_{V_z}(i, j)) + \beta(K_{EV_x}(i, j) + K_{EV_z}(i, j)) + 0 \end{aligned}$$

by the triangle inequality, boundedness of V , and as $T^{G(x,z)}$ is a contraction mapping of modulus β . Q.E.D.

The lemma only relies on points 1,3, and 4 of Assumption 1, point 2 is not needed here.

We turn now to the (further) discretization of y , our result is

Lemma 3. *Let V be the value fn defined in equation (15). Let $K_{F_{y_1}}(i, j)$ & $K_{F_{y_2}}(i, j)$ be as defined in equations (22) and (23). Then under Assumption 1 it must hold that*

$$|T^{G(x,z)}V(x_i, z_j) - T^{G(y|x,z)}V(x_i, z_j)| \leq K_{F_{y_1}}(i, j) + K_{F_{y_2}}(i, j)$$

Proof. The proof is an refinement of the standard approach which can be found in Bertsekas (1976), Chpt 5.2.

Consider an arbitrary point (x, z) . Then

$$\begin{aligned}
& |T^{G(x,z)}V(x_i, z_j) - T^{G(y|x,z)}V(x_i, z_j)| \\
&= | \sup_{y \in \Gamma(x_i, z_j)} \{F(x_i, y, z_j) + \beta \int_Z V(y_1, z')Q(z_j, dz')\} \\
&\quad - \sup_{y \in \Gamma^G(x_i, z_j)} \{F(x_i, y, z_j) + \beta \int_Z V(y_1, z')Q(z_j, dz')\} | \\
&\leq | \sup_{y \in \Gamma(x_i, z_j)} F(x_i, y, z_j) - \sup_{y \in \Gamma^G(x_i, z_j)} F(x_i, y, z_j) | \\
&\quad + | \sup_{y \in \Gamma(x_i, z_j)} \beta \int_Z V(y_1, z')Q(z_j, dz') \\
&\quad \quad - \sup_{y \in \Gamma^G(x_i, z_j)} \beta \int_Z V(y_1, z')Q(z_j, dz') | \\
&\leq K_{Fy_1}(i, j) + K_{Fy_2}(i, j) \\
&\quad + | \sup_{y \in \Gamma(x_i, z_j)} \beta \int_Z V(y_1, z')Q(z_j, dz') - \sup_{y \in \Gamma^G(x_i, z_j)} \beta \int_Z V(y_1, z')Q(z_j, dz') | \\
&\leq K_{Fy_1}(i, j) + K_{Fy_2}(i, j) + \beta 0 \\
&= K_{Fy_1}(i, j) + K_{Fy_2}(i, j)
\end{aligned}$$

where $\hat{V}(y_1, z') \equiv V(y_{1j}, z)$ for $y_1 \in Y_{j}$, by the triangle inequality, boundedness of V , and as T^G is a contraction mapping of modulus β . Note that the reason the difference between the two supremums of the integrals is zero is because the function V is only defined on the already discretized (x, z) anyway. Q.E.D.

We now present our result for the discretization of z' given that (x, y, z) have all been previously discretized (actually the result is exactly the same whether the others have been discretized beforehand or not).

Lemma 4. *Let V be the value fn defined in equation (15). Let $K_{EV}(i, j)$ be as defined in equation (24). Then under Assumptions 1 & 2 it must hold that*

$$|T^{G(x,y,z)}V(x_i, z_j) - T^{G(z'|x,y,z)}V(x_i, z_j)| \leq \beta K_{EV}(i, j)$$

Proof. Consider an arbitrary point (x_i, z_j) . Then

$$\begin{aligned}
& |T^{G(x,y,z)}V(x_i, z_j) - T^{G(z'|x,y,z)}V(x_i, z_j)| \\
&= \left| \sup_{y \in \Gamma^G(x_i, z_j)} \left\{ F(x_i, y, z_j) + \beta \int_Z V(y_1, z') Q(z_j, dz') \right\} \right. \\
&\quad \left. - \sup_{y \in \Gamma^G(x_i, z_j)} \left\{ F(x_i, y, z_j) + \beta \sum_{k=1}^{n_z} V(y, z_k) Q^G(z_j, z_k) \right\} \right| \\
&\leq \beta \sup_{y \in \Gamma^G(x_i, z_j)} \left| \int_Z V(y, z') Q(z_j, dz') - \sum_{k=1}^{n_z} V(y, z_k) Q^G(z_j, z_k) \right| \\
&\leq \beta K_{EV}(i, j)
\end{aligned}$$

where the first step comes from the definitions of $T^{G(x,y,z)}$ and $T^{G(z'|x,y,z)}$, the third is Lemma 13 on bounding errors in numerical integration. *Q.E.D.*

Note that while the result itself does not depend on (x, y, z) having been previously discretized. That Assumption 2 is satisfied by the Tauchen method is only true when z has been previously discretized.

A.2.6 Putting it all Together

The following result combines all of the individual lemmas we had for discretizing (x, z) , y , and z' respectively. Because we have used constants that depend on (i, j) it is clear that the size of the numerical errors will be different in different areas (different (i, j)). Numerical errors will be higher where the points are further apart, and where the first derivatives of the return and value functions are 'steeper' — this can be seen by looking at the definitions of the 'local' constants.

Lemma 5. *Let V be the value function defined in equation (15). Let V^G be the value function defined in equation (17). Then for $(x, z) \in X_i \times Z_j$ we have*

$$\begin{aligned}
|V(x, z) - V^G(x, z)| &\leq \frac{1}{1 - \beta} [K_{Fy_1}(i, j) + K_{Fy_2}(i, j) + K_{Vx}(i, j) + K_{Vz}(i, j) \\
&\quad + \beta(K_{EVx}(i, j) + K_{EVz}(i, j) + K_{EV}(i, j))]
\end{aligned}$$

Proof.

$$\begin{aligned}
|V(x, z) - V^G(x, z)| &= |TV(x, z) - T^G V^G(x, z)| \\
&= |TV(x, z) - T^G(x, y, z, z') V^G(x, y, z, z')(x, z)| \\
&\leq |V(x, z) - T^G(x, z) V(x, z)| + |T^G(y|x, z) V(x, z) - T^G(x, z) V(x, z)| \\
&\quad + |T^G(z'|x, y, z) V(x, z) - T^G(x, y, z, z') V(x, z)| \\
&\quad + |T^G(z'|x, y, z) V(x, z) - T^G(x, y, z, z') V(x, z)| \\
&\quad + |T^G(x, y, z, z') V(x, z) - T^G(x, y, z, z') V^G(x, z)| \\
&\leq |V(x, z) - T^G(x, z) V(x, z)| + |T^G(x, z) V(x, z) - T^G(x, y, z) V(x, z)| \\
&\quad + |T^G(x, y, z) V(x, z) - T^G(x, y, z, z') V(x, z)| \\
&\quad + 0 + \beta |V(x, z) - V^G(x, z)|
\end{aligned}$$

thus

$$\begin{aligned}
|V(x, z) - V^G(x, z)| &\leq \frac{1}{1-\beta} (|V(x, z) - T^G(x, z) V(x, z)| \\
&\quad + |T^G(x, z) V(x, z) - T^G(y|x, z) V(x, z)| \\
&\quad + |T^G(x, y, z) V(x, z) - T^G(z'|x, y, z, z') V(x, z)|)
\end{aligned}$$

so by Corollaries 2, 3, & 4,

$$\begin{aligned}
|V(x, z) - V^G(x, z)| &\leq \frac{1}{1-\beta} ([K_{V_x}(i, j) + K_{V_z}(i, j)] + \beta(K_{EV_x}(i, j) + K_{EV_z}(i, j))) \\
&\quad + [K_{F_{y_1}}(i, j) + K_{F_{y_2}}(i, j)] + [\beta K_{EV}(i, j)]
\end{aligned}$$

rearranging

$$\begin{aligned}
|V(x, z) - V^G(x, z)| &\leq \frac{1}{1-\beta} [K_{F_{y_1}}(i, j) + K_{F_{y_2}}(i, j) + K_{EV_x}(i, j) + K_{EV_z}(i, j) \\
&\quad + \beta(K_{EV_x}(i, j) + K_{EV_z}(i, j) + K_{EV}(i, j))]
\end{aligned}$$

Q.E.D.

Combining these two previous results we get our desired result; a bound on the distance between the true value function and the discretized value function after we stop iterating, ie. the one we will actually have.

Theorem 5. *Let V be the value function defined in equation (15). Let $\{V_n^G\}$ be the sequence of functions generated by iteratively applying T^G starting from a function V_0 (ie. $V_1 = T^G V_0$, $V_n = T^G V_{n-1}$). Let V_N^G be the value function at which the algorithm stops; given the stopping criterion $\|V_n - V_{n-1}\| \leq \epsilon_V$. Then*

$$\begin{aligned}
|V(x, z) - V_N^G(x, z)| &\leq \frac{1}{1-\beta} [K_{F_{y_1}}(i, j) + K_{F_{y_2}}(i, j) + K_{V_x}(i, j) + K_{V_z}(i, j) \\
&\quad + \beta(K_{EV_x}(i, j) + K_{EV_z}(i, j) + K_{EV}(i, j)) + \beta\epsilon_V]
\end{aligned}$$

Proof.

$$\begin{aligned}
|V(x, z) - V_N^G(x, z)| &\leq |V(x, z) - V^G(x, z)| + |V^G(x, z) - V_N^G(x, z)| \\
&\leq \frac{1}{1-\beta} [K_{Fy_1}(i, j) + K_{Fy_2}(i, j) + K_{Vx}(i, j) + K_{Vz}(i, j) \\
&\quad + \beta(K_{EVx}(i, j) + K_{EVz}(i, j) + K_{EV}(i, j)) + \beta\epsilon_V]
\end{aligned}$$

by the triangle inequality and then applying the results of Lemma 5 & Corollary 3. *Q.E.D.*

Since the number of grid points is finite, a uniform bound can be trivially follows by simply maximizing across all of the constants that depend on (i, j) . Thus,

Corollary 4. *Let V be the value function defined in equation (15). Let $\{V_n^G\}$ be the sequence of functions generated by iteratively applying T^G starting from a function V_0 (ie. $V_1 = T^G V_0$, $V_n = T^G V_{n-1}$). Let V_N^G be the value function at which the algorithm stops; given the stopping criterion $\|V_n - V_{n-1}\| \leq \epsilon_V$. Then*

$$\|V - V_N^G\| \leq \frac{1}{1-\beta} [K_{Fy_1} + K_{Fy_2} + (1+\beta)(K_{Vx} + K_{Vz}) + \beta K_{EV} + \beta\epsilon_V]$$

where

$$\begin{aligned}
K_{Fy_1} &= \max_{\substack{i=1, \dots, n_x, \\ j=1, \dots, n_z}} K_{Fy_1}(i, j) \\
K_{Fy_2} &= \max_{\substack{i=1, \dots, n_x, \\ j=1, \dots, n_z}} K_{Fy_2}(i, j) \\
K_{Vx} &= \max_{\substack{i=1, \dots, n_x, \\ j=1, \dots, n_z}} K_{Vx}(i, j) \\
K_{Vz} &= \max_{\substack{i=1, \dots, n_x, \\ j=1, \dots, n_z}} K_{Vz}(i, j) \\
K_V &= \max_{\substack{i=1, \dots, n_x, \\ j=1, \dots, n_z}} K_V(i, j)
\end{aligned}$$

Or alternatively, letting $\delta_V(i, j)$ be defined as the (right-hand side) bounding constant in Theorem 5, we have

$$\|V - V_N^G\| \leq \delta_V := \max_{\substack{i=1, \dots, n_x, \\ j=1, \dots, n_z}} \delta_V(i, j)$$

Proof. Follows trivially from the observation that $K_{Vx} \geq \max_{\substack{i=1, \dots, n_x, \\ j=1, \dots, n_z}} K_{EVx}(i, j)$ due to definitions of K_{Vx} , $K_{Vx}(i, j)$, and $K_{EVx}(i, j)$ and fact that $Q^G(\cdot, \cdot) \leq 1$; as well as analagous result for K_{Vz} . *Q.E.D.*

A.2.7 A Remark on the Bounds

The constants used to bound the errors are based on the differences in various functions between adjacent points on the grid. In the case of models where some dimensions are continuous (and we want to discretize them) while others are already discrete these differences only need to be taken in the dimensions that are being discretized (those that in theory are continuous). The differences in the already discrete variables should be considered to be zero. Care must be taken in the constants that are evaluated in the neighbourhood of a certain point as that neighbourhood should also be taken along the already discrete dimensions, as well as the dimensions being discretized.

A.3 Numerical Error Bounds for the Optimal Policy Function

In economic models what interests us is often not the value function itself but the optimal policy function. It is this later that determines the actions of the agent, and thus the behaviour of variables in equilibrium. For this reason we now turn to bounding errors in the optimal policy function. There are two existing approaches. First, one can bound errors in the policy function in terms of those in the value function based on differentiability of value fn, interiority of optimal policy, and the curvature of the return function (Santos and Vigo-Aguiar, 1998)²⁸. The second approach is that of Euler Equation Errors (Santos, 2000) which has the strength of being applicable to any numerical method for calculating the optimal policy, not just value function iteration. Both approaches depend on differentiability of the value function and interiority of the optimal policy. The approach developed here takes advantage of the fact that in many applications the value function is monotone (generally increasing) in the state variables²⁹.

A.3.1 Convergence of the Optimal Policy Function

See SLP Theorems 3.7, 3.8, 4.9 and 9.9. These tell us that if X and Z are both compact (and F strictly concave, continuous, etc.) then the optimal policies converge uniformly as the value function converges. However they do not give us fixed bounds on this convergence (it does not appear to be a contraction mapping). Thus in what follows we concentrate on getting bounds on the degree of convergence. The standard approach in the literature is to use a strong concavity on the return function, we use a different approach here based on monotonicity. We do this as it is not clear how strong concavity of the return function would relate to the concavity of the utility function under things such as adjustment costs.

²⁸ Rincón-Zapatero and Santos (2009) provide results on the differentiability of the value function in the presence of borrowing constraints.

²⁹I avoid the assumption of strong concavity of the return function, common in the literature, as it is not obvious how this relates to, eg., strong concavity of the utility function in the presence of adjustment costs and/or periodically binding constraints.

A.3.2 Bounding Errors in Optimal Policy Function without Interiority

Similarly to the value function we proceed in three steps. First we bound errors coming from the fact that we have only an approximation of the true value function. Secondly we bound errors coming from the discretization. Lastly we combine these two results to get our numerical error bounds.

With this in mind we define the following: 1) The true policy function

$$y_a(x, z) = \arg \max_{y \in \Gamma(x, z)} F(x, y, z) + \beta \int_Z V(y, z') Q(z, dz') \quad (28)$$

for all $(x, z) \in X \times Z$. 2) The policy function on the grid that would results from the true value function

$$y_b(x, z) = \arg \max_{y \in \Gamma(x, z)} F(x, y, z) + \beta \int_Z \tilde{V}(y, z') Q(z, dz') \quad (29)$$

for all $(x, z) \in X \times Z$. And, 3) the approximate policy function, which is that we will actually have

$$y_c(x, z) = \arg \max_{y \in \Gamma_G(x, z)} F(x, y, z) + \beta \int_Z \tilde{V}(y, z') Q(z, dz') \quad (30)$$

for all $(x, z) \in X_G \times Z_G$, and extended linearly from this to the rest of $X \times Z$. Where V is the true value function, and \tilde{V} is an approximation of the true value function satisfying $\|V - \tilde{V}\| \leq \delta_V$. Thus the first step is to bound $\|y_a - y_b\|$ and the second to bound $\|y_b - y_c\|$; putting these together will yield a bound on $\|y_a - y_c\|$.

A new assumption will be required for both steps, which did not need to be made previously when our interest was just in the value function. Namely,

Assumption 3. *F is decreasing, bounded, and strictly concave in y.*

A.3.3 Some More Constants

Some of the constants we will need when bounding the errors in the optimal policy function are simply the same as those we already used to bind errors in the value function. Some additional ones will however also be required and these are now given.

- Coarseness of y -grid (in neighbourhood of the optimal policy)

$$dy^g(i, j) = \max\{|g(x_i, z_j)(+1) - g(x_i, z_j)|, |g(x_i, z_j) - g(x_i, z_j)(-1)|\} \quad (31)$$

- Bound in the neighbourhood of the optimal policy of the return function in the y -dimension.

$$K_{Fy}^g(i, j) = \max\{|F(x_i, g(x_i, z_j)(+1), z_j) - F(x_i, g(x_i, z_j), z_j)|, \\ |F(x_i, g(x_i, z_j), z_j) - F(x_i, g(x_i, z_j)(-1), z_j)|\} \quad (32)$$

- Bound in the neighbourhood of the optimal policy of the return function in the x -dimension.

$$K_{Fx}(i, j) = \max_{y_k \in Y^G} \max\{|F(x_i(+1), y_k, z_j) - F(x_i, y_k, z_j)|, |F(x_i, y_k, z_j) - F(x_i(-1), y_k, z_j)|\} \quad (33)$$

A.3.4 Errors from having an approximation of the true value function

We begin by bounding the $\|y_a - y_b\|$ errors. This is where our derivation of errors in the value function that are dependent on the grid points, (x_i, z_j) , become useful. Since the largest ones sometimes occur at the points which are optimal policies corresponding to where the return function $F(x, y, z)$ is sensitive to the value of y they will not cause such large errors in the optimal policy function. This point-by-point approach to building the bounds is often substantially better, and cannot possibly be worse, than not exploiting this margin. Note that $dy_l(x, z)$ evolves in each lemma (it remains the same concept, but its precise definition evolves).

The further assumption that will be required to derive the bounds here is the existence of the kind of error bounds that we have derived previously for the value function.

Assumption 4. For any $i = 1, \dots, n_x, j = 1, \dots, n_z$, for all $(x, z) \in X_i \times Z_j$
 $|V(x, z) - \tilde{V}(x, z)| \leq \delta_V(x_i, z_j)$

After this subsection, this assumption will not be needed as it will be implied by assumptions that are anyway required for the discretization procedure.

Lemma 6. Under Assumptions 3 and 4, and that V is bounded, y_a and y_b as defined in (28) and (29) satisfy

$$|y_a(x, z) - y_b(x, z)| \leq dy_l(x, z) \quad (34)$$

where l is given by

$$dy_l(x, z) = \max\left\{\sum_{a=0}^{l_-} dy^{g(-a)}(x, z), \sum_{a=0}^{l_+} dy^{g(+a)}(x, z)\right\}$$

$$l_- = \operatorname{argmin}\left\{\sum_{a=0}^l K_{Fy}^{g(-a)} \geq \delta_V(x, z)\right\}$$

$$l_+ = \operatorname{argmin}\left\{\sum_{a=0}^l K_{Fy}^{g(+a)} \geq \delta_V(x, z)\right\}$$

Proof. Let y_a be as defined above. First, let us assume $y_b > y_a$. We want to find a limit to how much more y_b can be than y_a . So let's take the most extreme situation by assuming that $V \geq \tilde{V}$, then the value of

$$F(x, y, z) + \beta \int_Z \tilde{V}(y, z') Q(z, dz')$$

the right-hand side of the equation for y_b , increases by at most $\delta_V(x, z)$. This gives us the maximum gain that results. Meanwhile the minimum loss from choosing $\hat{y} > y_a$ (within l grid points distance of y_a) is

$$F(x, \hat{y}, z) - F(x, y_a, z) \leq - \sum_{a=0}^l K_{Fy}^{g(+a)}(x, z)$$

Now, \hat{y} can only be the argmax (ie. equal to y_b) if it does equal or better than y_a , that is, if the maximum gains minus the minimum losses are greater than zero,

$$\delta_V(x, z) - \sum_{a=0}^l K_{Fy}^{g(+a)}(x, z) \geq 0 \tag{35}$$

The maximum distance we will have to go is thus that which corresponds to the minimum value of l for which (35). The distance corresponding to l is given by $\sum_{a=0}^l dy^{g(+a)}(x, z)$.

This gives us the bound captured in the lemma by l_+ . The analogous argument for $\hat{y} < y_a$ leads to the bound given in the lemma by l_- . The maximum of the two thus gives us the overall bound $dy_l(x, z)$. *Q.E.D.*

A.3.5 Errors from the discretization

We now turn to the $\|y_b - y_c\|$ errors. Again, we proceed first to derive grid-point specific bounds. The discretization procedure is exactly the same as for the value function, as are many of the necessary assumptions; they will thus not be repeated.

Again we proceed first with the discretization of (x, z) , then y , and finally z' .

Discretizing (x, z) we get

Lemma 7. *Under assumptions 1 & 3, discretizing x and z gives a maximum error of*

$$|y_b(x_i, z_j) - y_c(x_i, z_j)| \leq dy_l(x, z) \tag{36}$$

where $dy_l(x, z)$ is given by

$$dy_l(x, z) = \max\left\{ \sum_{a=0}^{l_-} dy^{g(-a)}(x, z), \sum_{a=0}^{l_+} dy^{g(+a)}(x, z) \right\}$$

$$l_- = \operatorname{argmin} \left\{ \sum_{a=0}^l K_{Fy}^{g(-a)} \geq K_{Fx}(i, j) + \beta(K_{Vx}(i, j) + K_{EVz}(i, j)) \right\}$$

$$l_+ = \operatorname{argmin} \left\{ \sum_{a=0}^l K_{Fy}^{g(+a)} \geq K_{Fx}(i, j) + \beta(K_{Vx}(i, j) + K_{EVz}(i, j)) \right\}$$

Proof. Following the same logic as proof A.3.4. The max gain in $\beta \int \tilde{V}Q(z, dz')$ from getting x and z onto the grid is $\beta(K_{EVx}(i, j) + K_{EVz}(i, j))$. The max gain in F from getting x onto grid is

$K_{F_x}(i, j)$ (note that F is increasing in x and decreasing in y). Thus net effect on value of $f(x, y, z) = F(x, y, z) + \beta \int \tilde{V}Q(z, dz')$ of getting onto the grid is at most $\beta(K_{EV_x}(i, j) + K_{EV_z}(i, j)) + K_{F_x}(i, j)$. Thus the maximum shift is that corresponding to the minimum l that makes this net effect negative. This gives the bound relating to l_+ , with the analogous argument for the min loss of getting x and z onto the grid giving l_- . Q.E.D.

The discretization of y , given that (x, z) have already been discretized is trivial. We have simply that

Lemma 8. $|y_b(x, z) - y_c(x, z)| \leq dy$

(result follows trivially from SLP Theorem 9.9; we want, in getting onto the grid, to go as little distance from y_b as possible, dy provides a bound on this distance). Or using separate $y = (y_1, y_2)$; $\|y_b - y_c\| \leq d_x + d_{y_2}$

Discretizing z' we have

Lemma 9. Under assumptions 1,2, and 3, discretizing z' gives a maximum error of

$$|y_b(x_i, z_j) - y_c(x_i, z_j)| \leq dy_l(x, z) \tag{37}$$

where $dy_l(x, z)$ is given by

$$dy_l(x, z) = \max\left\{\sum_{a=0}^{l_-} dy^{g(-a)}(x, z), \sum_{a=0}^{l_+} dy^{g(+a)}(x, z)\right\}$$

$$l_- = \operatorname{argmin}\left\{\sum_{a=0}^l K_{F_y}^{g(-a)} \geq \beta K_{EV}(i, j)\right\}$$

$$l_+ = \operatorname{argmin}\left\{\sum_{a=0}^l K_{F_y}^{g(+a)} \geq \beta K_{EV}(i, j)\right\}$$

Proof. Following the same logic as proof A.3.4. The max gain in $\beta \int \tilde{V}Q(z, dz')$ from getting z' onto the grid is $\beta K_V(i, j)$ (as shown in lemma). Discretizing z' has no effect on F . Thus net effect on value of $f(x, y, z) = F(x, y, z) + \beta \int \tilde{V}Q(z, dz')$ of getting onto the grid is at most $\beta K_{EV}(i, j)$. Thus the maximum shift is that corresponding to the minimum l that makes this net effect negative. Q.E.D.

Combining these three lemmas we get our bounds for the numerical errors in the optimal policy directly introduced by the discretization procedure

Theorem 6. Under Assumptions 1 2, and 3, for (x, z) in partition $X_i \times Z_j$ discretizing x, y and z gives a maximum error of

$$|y_b(x_i, z_j) - y_c(x_i, z_j)| \leq dy_l(x, z) \tag{38}$$

where $dy_l(x, z)$ is given by

$$dy_l(x, z) = \max\left\{\sum_{a=0}^{l_-+1} dy^{g(-a)}(x, z), \sum_{a=0}^{l_++1} dy^{g(+a)}(x, z)\right\}$$

$$l_- = \operatorname{argmin}\left\{\sum_{a=0}^l K_{Fy}^{g(-a)} \geq K_{Fx}(i, j) + \beta(K_{EVx}(i, j) + K_{EVz}(i, j) + K_{EV}(i, j))\right\}$$

$$l_+ = \operatorname{argmin}\left\{\sum_{a=0}^l K_{Fy}^{g(+a)} \geq K_{Fx}(i, j) + \beta(K_{EVx}(i, j) + K_{EVz}(i, j) + K_{EV}(i, j))\right\}$$

Proof. A simple combination of Theorems 7 & 8 and section 9 (the discretization of y here appears as making it the distance corresponding to $l + 1$). *Q.E.D.*

A.3.6 An Error Bound for the distance between the approximate and true policy functions

. This is simply a matter of combining the previous two errors.

Theorem 7. *Under Assumptions 1, 2, and 3. For any (x, z) in the partition $X_i \times Z_j$, $i = 1, \dots, n_x, j = 1, \dots, n_z$. y_a and y_c as defined in (28) and (30) satisfy*

$$|y_a(x, z) - y_c(x, z)| \leq dy_l(x, z) \tag{39}$$

where $dy_l(x, z)$ is given by

$$dy_l(x, z) = \max\left\{\sum_{a=0}^{l_-+1} dy^{g(-a)}(x, z), \sum_{a=0}^{l_++1} dy^{g(+a)}(x, z)\right\}$$

$$l_- = \operatorname{argmin}\left\{\sum_{a=0}^l K_{Fy}^{g(-a)} \geq K_{Fx}(i, j) + \beta(K_{EVx}(i, j) + K_{EVz}(i, j) + K_{EV}(i, j)) + \delta_V(i, j)\right\}$$

$$l_+ = \operatorname{argmin}\left\{\sum_{a=0}^l K_{Fy}^{g(+a)} \geq K_{Fx}(i, j) + \beta(K_{EVx}(i, j) + K_{EVz}(i, j) + K_{EV}(i, j)) + \delta_V(i, j)\right\}$$

Proof. The result follows directly from the obvious combination of Theorem 6 and Lemma 6. *Q.E.D.*

A.4 SubAppendix: Some Results on Numerical Integration

The first part of the following is copied from Whitt (1978) where it constitutes Lemmas 6.1 & 6.2. The material after the second lemma is an adaptation of these results to be of use here.

Let (S, \mathcal{S}) be a measurable set. Let $\|\cdot\|$ be the sup-norm on the set of bounded real-valued functions on S . Let $\gamma(f) = \sup_{s \in S} f(s) - \inf_{s \in S} f(s)$.

We first state a lemma which does not exploit partitions.

Lemma 10. *If μ_1 and μ_2 are two finite measures on S , then*

$$\left| \int_S f d\mu_1 - \int_S f d\mu_2 \right| \leq \gamma(f) \min\{\mu_1(S), \mu_2(S)\} + \|f\| (|\mu_1(S) - \mu_2(S)|)$$

Proof. Observe that the upper integrals satisfy

$$\begin{aligned} \int_S f d\mu_1 - \int_S f d\mu_2 &\leq \mu_1(S) \sup_{s \in S} f(s) - \mu_2(S) \inf_{s \in S} f(s) \\ &\leq \min\{\mu_1(S), \mu_2(S)\} \left(\sup_{s \in S} f(s) - \inf_{s \in S} f(s) \right) \\ &\quad + (\mu_1(S) - \min\{\mu_1(S), \mu_2(S)\}) \sup_{s \in S} f(s) \\ &\quad - (\mu_2(S) - \min\{\mu_1(S), \mu_2(S)\}) \inf_{s \in S} f(s) \\ &\leq \gamma(f) \min\{\mu_1(S), \mu_2(S)\} + \|f\| (|\mu_1(S) - \mu_2(S)|) \end{aligned}$$

To obtain the inequality in the other direction, change the subscripts of μ_1 and μ_2 . *Q.E.D.*

(the thing about upper integrals is just about avoiding assuming measurability of f)

We now exploit the partitions. For this purpose, let

$$\begin{aligned} \gamma_n(f) &= \sup_{s \in S_n} f(s) - \inf_{s \in S_n} f(s), \quad \text{and} \\ \|f\|_n &= \sup_{s \in S_n} |f(s)| \end{aligned} \tag{40}$$

Let

$$K_\mu = \sum_{n=1}^{\infty} |\mu_1(S_n) - \mu_2(S_n)|$$

where μ_1 and μ_2 are finite measures on S .

Lemma 11. (a) *If μ_1 and μ_2 are two finite measures on S , then*

$$\begin{aligned} \left| \int_S f d\mu_1 - \int_S f d\mu_2 \right| &\leq \sum_n [\gamma_n(f) \min\{\mu_1(S_n), \mu_2(S_n)\} + \|f\|_n (|\mu_1(S_n) - \mu_2(S_n)|)] \\ &\leq (\sup_n \gamma_n(f)) [(\mu_1(S) + \mu_2(S) - K_\mu)/2] + \|f\| K_\mu \end{aligned}$$

(b) *If also $\mu_1(S) = \mu_2(S)$, then*

$$\begin{aligned} \left| \int_S f d\mu_1 - \int_S f d\mu_2 \right| &\leq \sum_n [\gamma_n(f) \min\{\mu_1(S_n), \mu_2(S_n)\}] + \gamma(f) K_\mu / 2 \\ &\leq (\sup_n \gamma_n(f)) [\mu_1(S) - K_\mu / 2] + \gamma(f) K_\mu / 2 \end{aligned}$$

Proof. (a) Apply the triangle inequality with Lemma 10, using the fact that $\min\{x, y\} = (x + y - |x - y|)/2$ in the last step.

(b) Apply the proof of Lemma 10 on the partition subsets to obtain

$$\begin{aligned}
\int_S f d\mu_1 - \int_S f d\mu_2 &= \sum_n \left(\int_{S_n} f d\mu_1 - \int_{S_n} f d\mu_2 \right) \\
&\leq \sum_n \min\{\mu_1(S_n), \mu_2(S_n)\} \left(\sup_{s \in S_n} f(s) - \inf_{s \in S_n} f(s) \right) \\
&\quad + \sum_n (\mu_1(S_n) - \min\{\mu_1(S_n), \mu_2(S_n)\}) \sup_{s \in S_n} f(s) \\
&\quad - \sum_n (\mu_2(S_n) - \min\{\mu_1(S_n), \mu_2(S_n)\}) \inf_{s \in S_n} f(s) \\
&\leq \sum_n [\gamma_n(f) \min\{\mu_1(S_n), \mu_2(S_n)\}] + \gamma(f) K_\mu / 2
\end{aligned}$$

Since $\gamma_n(f) \leq \sup_n \gamma_n(f)$ and $\min\{x, y\} = (x + y - |x - y|)/2$, the second inequality in (b) holds too. Q.E.D.

Lemma 12. *If μ_1 & μ_2 are probability measures on S , then*

$$\left| \int_S f d\mu_1 - \int_S f d\mu_2 \right| \leq \sup_{s \in S} f(s) - \inf_{s \in S} f(s)$$

Proof.

$$\begin{aligned}
\left| \int_S f d\mu_1 - \int_S f d\mu_2 \right| &\leq \mu_1(S) \sup_{s \in S} f(s) - \mu_2(S) \inf_{s \in S} f(s) \\
&= \sup_{s \in S} f(s) - \inf_{s \in S} f(s)
\end{aligned}$$

since as they are probability measures $\mu_1(S) = \mu_2(S) = 1$. Q.E.D.

Definition 3. *Let $\{S_n\}$ be a partition of S . We say that two probability measures, μ_1 & μ_2 , on S coincide on the partition $\{S_n\}$ if $\mu_1(S_n) = \mu_2(S_n)$ for all n .*

Lemma 13. *If μ_1 & μ_2 are probability measures on S , and they coincide on the partition $\{S_n\}$ then*

$$\left| \int_S f d\mu_1 - \int_S f d\mu_2 \right| \leq \sum_n ([\sup_{s \in S_n} f(s) - \inf_{s \in S_n} f(s)] \mu_1(S_n))$$

Proof.

$$\begin{aligned}
\left| \int_S f d\mu_1 - \int_S f d\mu_2 \right| &\leq \sum_n (\mu_1(S_n) \sup_{s \in S_n} f(s) - \mu_2(S_n) \inf_{s \in S_n} f(s)) \\
&= \sum_n ([\sup_{s \in S_n} f(s) - \inf_{s \in S_n} f(s)] \mu_1(S_n))
\end{aligned}$$

Q.E.D.

A.5 SubAppendix: Uniform Convergence of Value Function Iteration

This section provides the detailed results for dynamic programming under bounded returns that we summarized in Section A.2.1. It is based heavily on SLP Section 9.2. We study value function problems (aka. stochastic dynamic programming, aka. functional equations) of what SLP refer to as type 1. That is ones of the form

$$V(x, z) = \sup_{y \in \Gamma(x, z)} \{F(x, y, z) + \beta \int_Z V(y, z') Q(z, dz')\} \quad (41)$$

under the assumption that the return function F is bounded and continuous, the discount factor β is strictly less than one, and the transition function Q has the Feller property.

Preliminaries: Let (X, \mathcal{X}) and (Z, \mathcal{Z}) be measurable spaces of possible values for the endogenous and exogenous state variables, respectively; let $(S, \mathcal{S}) = (X \times Z, \mathcal{X} \times \mathcal{Z})$ be the product space; let Q be a transition function on (Z, \mathcal{Z}) ; let $\Gamma : S \rightarrow X$ be a correspondence describing the feasibility constraints; let A be the graph of Γ ; let $F : A \rightarrow \mathbb{R}$ be the one-period return function; and let $\beta \geq 0$ be the discount factor. Thus, the givens for the problem we will study are (X, \mathcal{X}) , (Z, \mathcal{Z}) , Q , Γ , F , & β . We will use A_z , A_{yz} , and so on to denote the sections of A .

Definition 4. A transition function Q on (Z, \mathcal{Z}) has the **Feller property** if the associated operator T maps the space of bounded continuous functions on Z into itself; that is, if $T : C(Z) \rightarrow C(Z)$.

(Markov operators that have the Feller property are also said to be *stable*.)

Definition 5. A transition function Q on (Z, \mathcal{Z}) is **monotone** if the associated operator T has the property that for every nondecreasing function $f : Z \rightarrow \mathbb{R}$, the function Tf is also nondecreasing.

We are now ready to make some assumptions that will be necessary for our results.

Assumption 5. X is a convex Borel set in \mathbb{R}^l , with its Borel subsets \mathcal{X} .

Assumption 6. Z is a compact (Borel) set in \mathbb{R}^k , with its Borel subsets \mathcal{Z} , and the transition function Q on (Z, \mathcal{Z}) has the Feller property.

Notice that if Z is countable, then all functions on Z are continuous, so in this case the requirement that Q satisfy the Feller property would be vacuous.

Our metric for the space $C(S)$ is the sup norm, $\|f\| = \sup_{s \in S} |f(s)|$. It is stressed that many of the results below apply much more broadly, and the arguments used here can easily be adapted to other situations.

The following lemmas shows that, under these two assumptions, integration preserves the required properties of the integrand in (41) — boundedness, continuity, monotonicity, and concavity.

Lemma 14. *Let (X, \mathcal{X}) , (Z, \mathcal{Z}) , and Q satisfy Assumptions A.5 & A.6. If $f : X \times Z \rightarrow \mathbb{R}$ is bounded and continuous, then Tf defined by*

$$(Tf)(y, z) = \int f(y, z')Q(z, dz'), \quad \text{all } (y, z) \in X \times Z$$

is also; that is $T : C(S) \rightarrow C(S)$. If f is (strictly) increasing in each of its first l arguments, then so is Tf ; and if f is (strictly) concave jointly in its first l arguments, then so is Tf .

Proof: See SLP, pg 261

In some situations the requirement that the set $Z \subseteq \mathbb{R}^k$ be compact is very unattractive. In fact, it can be dispensed with; but the proof becomes more complicated. See SLP Section 12.6 for this extension.

We now make two more assumptions

Assumption 7. *The correspondence $\Gamma : X \times Z \rightarrow X$ is nonempty, compact-valued, and continuous.*

Assumption 8. *The function $F : A \rightarrow \mathbb{R}$ is bounded and continuous, and $\beta \in (0, 1)$.*

If Z is a countable set, we interpret Assumption A.7 to mean that for each fixed $z \in Z$, the correspondence $\Gamma(\cdot, z) : X \rightarrow X$ is nonempty, compact-valued, and continuous. Similarly, in this case Assumption A.8 means that for each fixed $z \in Z$, the function $F(\cdot, \cdot, z) : A_z \rightarrow \mathbb{R}$ (the z -section of F) is continuous.

We note in passing the under Assumptions A.5-8 the Theorems 9.2 & 9.4 of SLP, tying together the the sequence problem and the functional equation, hold.

Under these same assumptions, we have the following basic result

Theorem 8. *Let (X, \mathcal{X}) , (Z, \mathcal{Z}) , Q , Γ , F , and β satisfy Assumptions A.5-8, and define the operator T on $C(S)$ by*

$$(Tf)(x, z) = \sup_{y \in \Gamma(x, z)} \left\{ F(x, y, z) + \beta \int f(y, z')Q(z, dz') \right\} \quad (42)$$

Then $T : C(S) \rightarrow C(S)$; T has a unique fixed point V in $C(S)$ and for any $V_0 \in C(S)$,

$$\|T^n V_0 - V\| \leq \beta^n \|V_0 - V\|, \quad n = 1, 2, \dots \quad (43)$$

Moreover, the correspondence $G : S \rightarrow X$ defined by

$$G(x, z) = \left\{ y \in \Gamma(x, z) : V(x, z) = F(x, y, z) + \beta \int V(y, z')Q(z, dz') \right\} \quad (44)$$

is nonempty, compact-valued, and u.h.c.

Proof. Fix $f \in C(S)$. Then it follows from Lemma 14 that

$$(Tf)(y, z) = \int f(y, z')Q(z, dz')$$

is a bounded continuous function of (y, z) , that is $T : C(S) \rightarrow C(S)$. Moreover, since $Q(z, \cdot)$ is a probability measure, $T(f + c) = Tf + c$, for any constant function c . Thus we have that T satisfies Blackwell's sufficient conditions for a contraction (SLP Thm 3.3; Appendix A.6, Thm 12) and so is a contraction mapping. Since $C(S)$ is a Banach space (SLP Thm 3.1; Appendix A.6, Thm 10), it follows from the contraction mapping theorem (SLP Thm 3.2; Appendix A.6, Thm 11) that T has a unique fixed point $V \in C(S)$, and (43) holds. The stated properties of G then follow from the Theorem of the Maximum (SLP Thm 3.6; Appendix A.6, Thm 13), applied to (41). *Q.E.D.*

Theorem 8 suggests the approach to calculating the true value function V known as value function iteration. Namely, starting from an initial function $V_0 \in C(S)$ we can apply the mapping T defined in (42) to generate a new function V_1 . Iterating on this procedure, ie. $V_n = TV_{n-1}$ we get a sequence $V_0, V_1, \dots, V_n, \dots$ of functions. By (43) we know that $V_n \rightarrow V$ as $n \rightarrow \infty$, and in fact it also tells us the speed of this convergence. Thus the results of Thm 8 prove that value function iteration is globally convergent to the true value function, and gives us a rate of convergence.

A.5.1 Howards Improvement Algorithm

When performing infinite-horizon value function iteration in practice we often Howards improvement algorithm to speed it up. A description of this algorithm and a proof (for the non-stochastic case) that its use does not in any way effect the convergence results just derived forms Exercise 4.4 of SLP. Intuitively, the increase in speed comes about because it requires us to make less use of the maximization operation, which is computationally costly.

A.5.2 Bounding the Errors of Value Function Iteration

While we know from Theorem 8 that using the value function iteration algorithm our solution will converge to the true value function, in practice we have to stop after a finite number of iterations. So how can we know how close we have ended up? The standard way to decide when to stop is based on a convergence criterion of the form $\|V_n - V_{n-1}\| \leq \epsilon$. This section gives a theorem that bounds the distance of the function V_N at which the algorithm terminates from the true value function based on the convergence criterion.

Theorem 9. *Let (X, ρ) be a complete metric space. Let $\{x_n\}$ be a sequence of elements of X converging to x and satisfying $\rho(x_n, x) \leq \beta\rho(x_{n-1}, x)$. Let $\epsilon > 0$ satisfy $\rho(x_n, x_{n-1}) < \epsilon$. Then*

$$\rho(x_n, x) < \frac{\beta}{1 - \beta}\epsilon \tag{45}$$

Proof.

$$\begin{aligned}
\rho(x_n, x) &\leq \rho(x_n, x_{n+1}) + \rho(x_{n+1}, x) \\
&\leq \rho(x_{n+1}, x_n) + \rho(x_{n+2}, x_{n+1}) + \rho(x_{n+2}, x) \\
&\vdots \\
&\leq \sum_{i=1}^j \rho(x_{n+i}, x_{n+i-1}) + \rho(x_{n+1+j}, x) \\
&\text{taking limit as } j \rightarrow \infty \\
&= \sum_{i=1}^{\infty} \rho(x_{n+i}, x_{n+i-1}) + 0 \\
&\leq \sum_{i=1}^{\infty} \beta^i \rho(x_n, x_{n-1}) \\
&= \frac{\beta}{1-\beta} \rho(x_n, x_{n-1}) \\
&\leq \frac{\beta}{1-\beta} \epsilon
\end{aligned}$$

Q.E.D.

Corollary 5. *Let $\{f_n\}$ be a sequence of functions converging to f and satisfying $\|f_n - f\| \leq \beta \|f_{n-1} - f\|$. Let $\epsilon > 0$ satisfy $\|f_n - f_{n-1}\| < \epsilon$. Let $\|\cdot\|$ be the sup norm. Then*

$$\|f_n - f\| < \frac{\beta}{1-\beta} \epsilon \tag{46}$$

Applying this Corollary to Value Function Iteration we get that $\|V_N - V\| \leq \frac{\beta}{1-\beta} \epsilon$. Thus we now have a bound on the errors of value function iteration which is expressed solely in terms of known parameters, β & ϵ . How tight are these error bounds? For many economic models appropriate values of β are 0.9, 0.95, & 0.99 implying $\beta/(1-\beta) = 9, 19, \& 99$, respectively.

A.6 SubAppendix: Some Results Used for Convergence of Value Function Iteration

Contains Theorems 3.1, 3.2, 3.3, & 3.6 of SLP, which should be consulted for proofs, rewritten into the notation in which they apply to the stochastic case

Theorem 10 ($C(S)$ is a Banach space). *Let $S \subseteq \mathbb{R}^{l+k}$, and let $C(S)$ be the set of bounded continuous functions $f : S \rightarrow \mathbb{R}$ with the sup norm $\|f\| = \sup_{x \in X} |f(x)|$. Then $C(S)$ is a complete normed vector space, a.k.a. a **Banach space**. (Note that if S is compact then every continuous function is bounded. Otherwise the restriction to bounded functions must be added.)*

Theorem 11 (Contraction Mapping Theorem). *If (M, ρ) is a complete metric space and $T : M \rightarrow M$ is a contraction mapping with modulus β , then*

1. T has exactly one fixed point V in M , and
2. for any $V_0 \in M$, $\rho(T^n V_0, V) \leq \beta^n \rho(V_0, V)$, $n = 0, 1, 2, \dots$

Theorem 12 (Blackwell's sufficient conditions for a contraction). *Let $S \subseteq \mathbb{R}^{l+k}$, and let $B(S)$ be the space of bounded functions $f : S \rightarrow \mathbb{R}$, with the sup norm. Let $T : B(S) \rightarrow B(S)$ be an operator satisfying*

1. (monotonicity) $f, g \in B(S)$ and $f(s) \leq g(s)$, for all $s \in S$, implies $(Tf)(s) \leq (Tg)(s)$, for all $s \in S$
2. (discounting) there exists some $\beta \in (0, 1)$ such that $[T(f + a)(s)](x) \leq (Tf)(x) + \beta a$, all $f \in B(S)$, $a \geq 0$, $s \in S$.

[Here $(f + a)(x)$ is the function defined by $(f + a)(x) = f(x) + a$. Then T is a contraction with modulus β .

Theorem 13 (Theorem of the Maximum). *Let $S \subseteq \mathbb{R}^{l+k}$ and $Y \subseteq \mathbb{R}^m$, let $f : S \times Y \rightarrow \mathbb{R}$ be a continuous function, and let $\Gamma : S \rightarrow Y$ be a compact-valued and continuous correspondence. Then the function $h : S \rightarrow \mathbb{R}$ defined by*

$$h(s) = \max_{y \in \Gamma(s)} f(s, y)$$

is continuous, and the correspondence $G : X \rightarrow Y$ defined by

$$G(x) = \{y \in \Gamma(s) : f(s, y) = h(s)\}$$

is nonempty, compact-valued, and u.h.c.

A.7 SubAppendix: Discretizing the Choice Variable y in Value Function

The following is a version of Lemma 3 in which it is not previously assumed that (x, z) has already been discretized.

Lemma 15. *Let V be the value fn defined in equation (15). Let K_{Vx} , $K_{Fy_1}^g$ & $K_{Fy_2}^g$ be the maximums across i and j of $K_{Vx}(i, j)$, $K_{Fy_1}^g(i, j)$ & $K_{Fy_2}^g(i, j)$ respectively, as defined in equations (18), (22) and (23). Then under points 1, 2 and 4 of Assumption 1 it must hold that $\|TV - T^G V\| \leq K_{Fy_1}^g + K_{Fy_2}^g + \beta K_{Vx}$*

Proof. The proof is an improvement on the standard approach which can be found in Bertsekas (1976), Chpt 5.2.

(The improvement being the definition of $K_{Fy_1}^g$ & $K_{Fy_2}^g$ to be in the neighbourhood of the optimal policy. This modification is very important for many economic models for reasons are explained in

the text above).

Define $N(y^*)$ (which depends on (x, z)) to be a grid-neighbourhood of $y^* = g(x, z)$, that is to be the subspace of Y containing just y^* and it's nearest grid points in every direction.

Consider an arbitrary point (x, z) . Then

$$\begin{aligned}
& |TV(x, z) - T^G(z')V(x, z)| \\
&= \left| \sup_{y \in \Gamma(x, z)} \left\{ F(x, y, z) + \beta \int_Z V(y_1, z') Q(z, dz') \right\} \right. \\
&\quad \left. - \sup_{y \in \Gamma^G(x, z)} \left\{ F(x, y, z) + \beta \int_Z V(y_1, z') Q(z, dz') \right\} \right| \\
&= \sup_{y \in N(y^*)} \left\{ \left| \sup_{y \in \Gamma(x, z)} \left\{ F(x, y, z) + \beta \int_Z V(y_1, z') Q(z, dz') \right\} \right. \right. \\
&\quad \left. \left. - \sup_{y \in \Gamma^G(x, z)} \left\{ F(x, y, z) + \beta \int_Z V(y_1, z') Q(z, dz') \right\} \right| \right\} \\
&\leq \sup_{y \in N(y^*)} \left\{ \left| \sup_{y \in \Gamma(x, z)} F(x, y, z) - \sup_{y \in \Gamma^G(x, z)} F(x, y, z) \right| \right. \\
&\quad \left. + \left| \sup_{y \in \Gamma(x, z)} \beta \int_Z V(y_1, z') Q(z, dz') - \sup_{y \in \Gamma^G(x, z)} \beta \int_Z V(y_1, z') Q(z, dz') \right| \right\} \\
&\leq \sup_{y \in N(y^*)} \left\{ \left| \sup_{y \in \Gamma(x, z)} F(x, y, z) - \sup_{y \in \Gamma^G(x, z)} F(x, y, z) \right| \right. \\
&\quad \left. + \sup_{y \in N(y^*)} \left\{ \left| \sup_{y \in \Gamma(x, z)} \beta \int_Z V(y_1, z') Q(z, dz') - \right. \right. \right. \\
&\quad \quad \left. \left. \sup_{y \in \Gamma^G(x, z)} \beta \int_Z V(y_1, z') Q(z, dz') \right| \right\} \\
&\leq K_{Fy_1}^g + K_{Fy_2}^g \\
&\quad + \sup_{y \in N(y^*)} \left\{ \left| \sup_{y \in \Gamma(x, z)} \beta \int_Z V(y_1, z') Q(z, dz') - \right. \right. \\
&\quad \quad \left. \left. \sup_{y \in \Gamma^G(x, z)} \beta \int_Z V(y_1, z') Q(z, dz') \right| \right\} \\
&\leq K_{Fy_1}^g + K_{Fy_2}^g \\
&\quad + \left| \sup_{y \in \Gamma(x, z)} \beta \int_Z V(y_1, z') Q(z, dz') - \sup_{y \in \Gamma^G(x, z)} \beta \int_Z V(y_1, z') Q(z, dz') \right| \\
&\leq K_{Fy_1}^g + K_{Fy_2}^g + \beta \sup_{y \in \Gamma(x, z)} \int_Z |V(y_1, z') - \hat{V}(y_1, z')| Q(z, dz') \\
&\leq K_{Fy_1}^g + K_{Fy_2}^g + \beta \sup_{y \in \Gamma(x, z)} \int_Z K_{Vx} Q(z, dz') \\
&= K_{Fy_1}^g + K_{Fy_2}^g + \beta K_{Vx}
\end{aligned}$$

where $\hat{V}(y_1, z') \equiv V(y_{1j}, z')$ for $y_1 \in Y_{j, z'}$, by the triangle inequality, boundedness of V , and as T^G is a contraction mapping of modulus β . Since (x, z) was arbitrary, we have

$$\|TV - T^G V\| \leq K_{Fy_1}^g + K_{Fy_2}^g + \beta K_{Vx}$$

B Numerical Errors in the Steady-State Distribution

NOTE: This section is currently incomplete; The Proof of Theorem 15 is incorrect as it mixes metrics. Am currently working on fixing this.

Appendix A established bounds on the errors occouring in the value function V and optimal policy function g due to discretization. In computing Bewley-Huggett-Aiyagari models the optimal policy function is used to compute the steady-state distribution of agents. Do the (bounded) errors in the optimal policy function cause errors in the steady-state distribution to explode? It is shown here that they do not; that the errors in the steady-state agents distribution can be bounded in terms of the errors in the optimal policy function. Key to this result is contraction property of the monotone-mixing condition. The monotone-mixing condition can be shown to hold for models of the Bewley-Huggett-Aiyagari class and is part of the foundation for existing theory on the existence of equilibria in these models (Hopenhayn and Prescott, 1992).

Define $P = g \circ Q$, where g is the optimal policy function and Q is the transition mapping for the exogenous state. A common approach to the convergence of the distributions for economic models is to use the assumption that P satisfies the Feller property, which is the approach taken in SLP (Chpt 12) and Santos and Peralta-Alva (2005). However as we wish to apply our results to heterogeneous agents models this assumption is not applicable (in particular the difficulties arise when the borrowing constraints are binding³⁰). In proving the existence of equilibria for heterogeneous agent models we turn instead to Markov chain convergence theorems based on monotonicity and the monotone mixing condition, which is developed in Hopenhayn and Prescott (1992). Huggett (1993) for example demonstrates the use of this condition, which is also described as being more useful for heterogeneous agents in Ríos-Rull (2001). The results do require compactness of the state space, something which can be avoided if instead of a monotone mixing condition we use 'splitting conditions' (see Bhattacharya and Lee (1988), Bhattacharya and Majumdar (2001)), or even more general 'order mixing conditions' (see Kamihigashi and Stachurski (2011)). Our use of the monotone mixing condition is based on two things; its usefulness for heterogeneous agent models, and the fact that our main use for them here is not in proving the existence of equilibria for economic models, but in bounding the errors of discrete state space approximations, and with a discrete state space the assumptions of compact spaces are requisite. The results presented here largely follow the presentation of Hopenhayn and Prescott (1992), but with some of the most important for our use being adaptated from Bhattacharya and Lee (1988) and Bhattacharya and Majumdar (2001).

Following SLP we use P to denote the transition function on $(S, \mathcal{S}) = (X \times Z, \mathcal{X} \times \mathcal{Z})$ resulting

³⁰It has not been proved that it is not applicable, but nor has it been proved that it is applicable.

from the combination of the (optimal) policy function³¹ $g : X \times Z \rightarrow X$, and the transition function Q on (Z, \mathcal{Z}) . Our interest is then in the existence and uniqueness of a probability distribution μ over $S = \mathcal{X} \times \mathcal{Z}$ which is a stationary distribution for P , ie. $P\mu = \mu$, and on convergence to this distribution. We begin with some results from Hopenhayn and Prescott (1992) which can be used for proving the existence of a unique equilibria. Let $\mathcal{M} = \mathcal{P}(S)$, the set of probability measures on S .

B.1 Existence, Uniqueness of, and Convergence to, an Invariant Distribution

The monotone mixing condition is introduced, which, when combined with a requirement trivially satisfied by compact subspaces of \mathcal{R}^n needed for the existence of invariant distributions, gives us results for uniqueness and global convergence. This result is Theorem 2 of Hopenhayn and Prescott (1992)

Theorem 14. *Suppose P is increasing, S contains a lower bound (which we will denote by a) and an upper bound (which we will denote by b) and the following condition is satisfied: Monotone Mixing Condition (MMC): there exists $s^* \in S, m \in \mathbb{Z}, \beta_P > 0$ such that $P^m(b, [a, s^*]) \geq 1 - \beta_P$ and $P^m(a, [s^*, b]) \geq 1 - \beta_P$. Then there is a unique stationary distribution μ^* for process P , and for any initial measure $\mu_0, \mu_n \equiv T^{*n}\mu_0 = \int P^n(s, \cdot)\mu_0(ds)$ converges to μ^* .*

The following result, adapted from Bhattacharya and Lee (1988; Lemma 2.3) and Bhattacharya and Majumdar (2001) shows how fast the sequence $\mu_n = T^{*n}\mu_0$ converges to it's invariant distribution μ^* . First we define a distance d_1 stronger than that we have been using till now³². For $a \geq 0$, let \mathcal{G}_a denote the class of all real-valued Borel measurable nondecreasing functions f on S satisfying $0 \leq f(s) \leq a, \forall s \in S$. Define

$$d_a(\mu, \nu) = \sup\left\{ \left| \int f d\mu - \int f d\nu \right| : f \in \mathcal{G}_a \right\}, \quad \mu, \nu \in \mathcal{P}(S)$$

Proposition 2. *Under the assumptions of Theorem 14 we have further that*

$$d_1(T^{*n}\mu, T^{*n}\nu) \leq \beta_P^{[n/m]} d_1(\mu, \nu), \quad \forall \mu, \nu \in \mathcal{P}(S)$$

and

$$d(T^{*n}\mu_0, \mu^*) \leq (1 - \beta_P)^{[n/m]} \quad \forall \mu_0 \in \mathcal{P}(S) \tag{47}$$

Bounding the distance between the iterated agents distribution and the discretized agents distribution is then just a well-known property of contraction mappings,

³¹Or more accurately the restriction g_{y_1} of the optimal policy $g : X \times Z \rightarrow Y = (Y_1, Y_2)$ onto $Y_1 = X; g_{y_1} : X \times Z \rightarrow Y_1 = X$.

³²Till now we used the measure $d(\mu, \nu) = \sup\{|\mu(A) - \nu(A)| : A \in \mathcal{S}\}, \mu, \nu \in \mathcal{P}(S)$, where F_μ is the distribution function of μ .

Corollary 6. *Under the assumptions of Theorem 14. Let $\{\mu_n\}$ be the sequence defined by repeated application of P , ie. $\mu_n = P\mu_{n-1}$, $\mu_1 = P\mu_0$, from an arbitrary starting probability measure $\mu_0 \in \mathcal{S}$. Let $d(\mu_N, \mu_{N+m}) < \epsilon$, then*

$$d((\mu_N, \mu^*) \equiv \leq \frac{1}{1 - \beta_P} \epsilon$$

Remark: Convergence in the d_1 metric is stronger than convergence in moments. Since in heterogeneous agents what we often care about is the aggregate value (which is just the first moment) we could use analogues of this result for convergence in moments.

B.2 The Agents Distribution with the Approximation of the Optimal Policy

So far we have looked at value function iteration and how close our approximate solution is to the true solution (and analogously for the optimal policy function), and also at the agents distribution and how close our approximate invariant distribution is to the true invariant distribution. One aspect we have however ignored in looking at the agents distribution is the fact that when we calculate it we are using not the true optimal policy function. It is this point that we address now: how close is the invariant distribution generated by the approximation of the optimal policy function to that generated by the approximation of the true optimal function.

This issue is also addressed in Santos and Peralta-Alva (2005) using the Feller property rather than the monotone mixing condition used here (in both cases compact state spaces are assumed): loosely speaking, their Theorem 2 shows that the invariant distribution is continuous in the policy function (technically, that the invariant distribution correspondence is upper semicontinuous). They also provide a Generalized Law of Large Numbers (their Theorem 3) showing convergence of markov chain simulations; something we will not attempt to do here (these would what be required to provide the results bounding numerical errors relevant to the SLE). They provide a condition which will give them the Feller property (their Condition C; that P is a stochastic contraction and stochastically bounded (the later following trivially from S being compact), see Stenflo (2001)) and thus allows them to bound the errors (their Theorem 6); in contrast it will follow for us directly from the monotone mixing condition and does not require further assumptions on the optimal policy function.

B.3 Upper Semicontinuity of the Correspondence of Invariant Distributions

We first look at the issue of whether the sequence of invariant distributions associated with the sequence of approximations of the optimal policy function converges to the true invariant distribution as the approximations of the optimal policy function converge to the true optimal policy function, that is, whether the invariant distribution correspondence is upper semicontinuous. Note that, like Santos and Peralta-Alva (2005), we do not assume uniqueness of the invariant distribution

generated by the approximation of the optimal policy function³³. Rather than explicitly assuming compactness of the state space and of the monotone mixing condition we will simply assume here that the adjoint operator, T^* , associated with P is a contraction mapping (a property implied by the former two conditions).

Following Santos and Peralta-Alva (2005) and Stenflo (2001) the theory of this section is developed using the iterated function systems notation. Up till now the solution to the individuals problem has been given by the optimal policy function³⁴ $g(x, z) : X \times Z \rightarrow X$, which is then combined with the transition function Q on (Z, \mathcal{Z}) to give $P = g \cdot Q$, with $P((x, z), A \times B) = 1_{\{g(x, z) \in A\}} Q(z, B)$, where $(x, z) \in X \times Z$ and $A \times B \in \mathcal{X} \times \mathcal{Z}$. In iterated function systems notation we instead consider a function from the state space $S = (X, Z)$ into itself whose value depends on shocks coming from Ω , in this way we can make the shock process an iid variable without losing any of the richness of the environment. Let $\varphi : X \times Z \times W \rightarrow X \times Z$ and let $\Omega : (W, \mathcal{W}) \rightarrow \mathbb{R}$ be an iid random variable. Then, for any process which can be represented by $P = g \cdot Q$ can also be represented as $P = \varphi \cdot \Omega$, with $P(s, C) = \Omega(\{w | \varphi(s, w) \in C\})$, where $s \in S = (X, Z)$ and $C \in \mathcal{S} = \mathcal{X} \times \mathcal{Z}$ (cf Stenflo (2001)).

Let us start by laying out the necessary assumptions

Assumption 9. S is a compact set.

Assumption 10. Let T^* be the adjoint-operator of $P = g \cdot Q = \varphi \cdot \Omega$ (in our standard, and the iterated function systems notations respectively). The m -times application of T^* , T^{*m} is a contraction mapping of modulus β_P .

Theorem 15. Let $\{g_j\}$ be a sequence of policy functions that converge to g . Let $\{\mu_j^*\}$ be a sequence of probabilities on \mathbb{S} such that $\mu_j^* = T_j^* \mu_j^*$ for each j . Under assumptions 9 and 10, if μ^* is a weak limit point of μ_j^* , then $\mu^* = T^* \mu^*$.

Proof. (Proof is similar to that of Santos and Peralta-Alva (2005) Theorem 2)

Define $P = \varphi \cdot \Omega$ and $\hat{P}_j = \varphi_j \cdot \Omega$ (rewriting into iterated function systems notation).

For an associated pair (P, T^*) and a probability μ , let $P \cdot \mu$ stand for $T^* \mu$, and likewise $P^m \cdot \mu$ stand for $T^{*m} \mu$

Then, the theorem will be established if we can show the continuity of the evaluation map $ev(P, \mu) = P \cdot \mu$.

Recall that the space of probability measures is endowed with the topology of weak convergence, and the distance function in the space of mappings is given by

$$d(\varphi, \hat{\varphi}) = \max_{s \in S} \left[\int \|\varphi(s, w) - \hat{\varphi}(s, w)\| |\Omega(dw)| \right]$$

where $\varphi, \hat{\varphi}$ are as above, and $\|\cdot\|$ is the sup-norm on \mathbb{R}^{l+k} .

³³When we come to apply these results later on we will know that it is unique.

³⁴Again, more accurately, the restriction of the optimal policy function onto Y_1 , see footnote 31.

As is well known, the topology of weak convergence can be defined by the metric

$$d(\mu, \nu) = \sup_{f \in \mathcal{A}} \left\{ \left| \int f(s) \mu(ds) - \int f(s) \nu(ds) \right| \right\} \quad (48)$$

where \mathcal{A} is the space of Lipschitz functions on S with constant $L \leq 1$ and such that $-1 \leq f \leq 1$.

Let $f \in \mathcal{A}$. Then for any two mappings P^m and \hat{P}^m , and any two measures μ & ν , we have

$$\begin{aligned} & \left| \int f(s) [P^m \cdot \mu(ds)] - \int f(s) [\hat{P}^m \cdot \nu(ds)] \right| \\ & \leq \left| \int \left[\int f(s) P^m(ds) \right] \mu(ds) - \int \left[\int f(s) P^m(ds) \right] \nu(ds) \right| \\ & \quad + \left| \int \left[\int f(s) P^m(ds) \right] \nu(ds) - \int \left[\int f(s) \hat{P}^m(ds) \right] \nu(ds) \right| \\ & \leq \left| \int \left[\int f(s) P^m(ds) \right] [\mu(ds) - \nu(ds)] \right| + d(P^m, \hat{P}^m) \end{aligned}$$

where the first step is the triangle inequality and the second from the definition of $d(P^m, \hat{P}^m)$.

Then, by (48) the theorem will be established if we can show that for every arbitrary $\eta > 0$ there exists a weak neighbourhood $V(\mu)$ of μ such that for all $\nu \in V(\mu)$ and for all $f \in \mathcal{A}$,

$$\left| \int \left[\int f(s) P^m(ds) \right] [\mu(ds) - \nu(ds)] \right| < \eta \quad (49)$$

Under assumption 9 the Arzela-Ascoli theorem implies that the set \mathcal{A} is compact. Hence, we can find a finite set of elements $\{f^j\}$ such that for every f in \mathcal{A} there exists an element f^j such that in the sup norm, $\|f - f^j\| < \eta/3$. Since f is continuous, it follows that for every f^j there exists a weak neighbourhood $V_j(\mu)$ such that for all $\nu \in V_j(\mu)$,

$$\left| \int f^j [\mu(ds) - \nu(ds)] \right| < \eta/(3\beta_P)$$

Therefore, for all f with $\|f - f^j\| < \eta/3$, we have that

$$\left| \int f [\mu(ds) - \nu(ds)] \right| < \eta/\beta_P$$

Letting $V(\mu) = \bigcap_j V_j(\mu)$. Thus $\left| \int f [\mu(ds) - \nu(ds)] \right| < \eta/\beta_P$ holds for all $\nu \in V(\mu)$ and all $f \in \mathcal{A}$.

Thus, for all $\nu \in V(\mu)$,

$$\begin{aligned} \left| \int \left[\int f(s) P^m(ds) \right] [\mu(ds) - \nu(ds)] \right| & \leq d(T^{*m} \mu, T^{*m} \nu) \\ & \leq \beta_P d(\mu, \nu) \\ & \leq \eta \end{aligned}$$

where the first line follows from the definition of d since $f \in \mathcal{A}$, the second as T^* is a contraction mapping of modulus β_P , the third by the definition of $V(\mu)$. Q.E.D.

This results establishes that as the approximation of the optimal policy function converges to the optimal policy function, the invariant distribution it implies (which need not be unique) will converge to the invariant distribution of the true optimal policy function. It establishes this using the assumption that the adjoint-operator of P^m , namely T^{*m} , is a contraction mapping; analogously to the result of Theorem 2 of Santos and Peralta-Alva (2005) which is based on an assumption that P has the Feller property.

B.4 Bounding the distance to the true invariant distribution

In this section we present a theorem providing a bound on the distance between the invariant distribution associated with the true optimal policy function and that associated with the approximation of the optimal policy function. An analagous result is given by Santos and Peralta-Alva (2005) based on an assumption that P has the Feller property, S is compact, and P is a Lipschitz function.

We start with the definition of a Lipschitz function: f is a Lipschitz function with constant L if, $|f(s) - f(s')| \leq L||s - s'||$, $\forall s, s' \in S$.

Theorem 16. *Let f be a Lipschitz function with constant L . Let $d(P^m, \hat{P}^m) \leq \delta$ for some $\delta > 0$. Assume that T^{*m} is a contraction mapping of modulus β_P . Then, under assumptions 9 & 10,*

$$\left| \int f(s)\mu^*(ds) - \int f(s)\hat{\mu}^*(ds) \right| \leq \frac{L\delta}{1 - \beta_P}$$

Proof.

$$\begin{aligned} \left| \int f(s)\mu^*(ds) - \int f(s)\hat{\mu}^*(ds) \right| &= \left| \int f(s)P^m \cdot \mu^*(ds) - \int f(s)\hat{P}^m \cdot \hat{\mu}^*(ds) \right| \\ &\leq \left| \int f(s)P^m \cdot \mu^*(ds) - \int f(s)\hat{P}^m \cdot \hat{\mu}^*(ds) \right| \\ &\leq \left| \int f(s)P^m \cdot \mu^*(ds) - \int f(s)P^m \cdot \hat{\mu}^*(ds) \right| \\ &\quad + \left| \int f(s)P^m \cdot \hat{\mu}^*(ds) - \int f(s)\hat{P}^m \cdot \hat{\mu}^*(ds) \right| \\ &\leq \left| \int f(s)P^m \cdot \mu^*(ds) - \int f(s)P^m \cdot \hat{\mu}^*(ds) \right| + Ld(P^m, \hat{P}^m) \\ &\leq \beta_P \left| \int f(s)\mu^*(ds) - \int f(s)\hat{\mu}^*(ds) \right| + Ld(P^m, \hat{P}^m) \\ &\leq \beta_P \left| \int f(s)\mu^*(ds) - \int f(s)\hat{\mu}^*(ds) \right| + L\delta \end{aligned}$$

first line as invariant distributions, second by triangle inequality, third by definition of d and since f is Lipschitz, fourth as T^* is a contraction mapping of modulus β_P , fifth as $d(P^m, \hat{P}^m) \leq \delta$. The theorem follows from simply rearranging the terms. Q.E.D.

Corollary 7. Let $d(P^m, \hat{P}^m) \leq \delta$ for some $\delta > 0$. Assume that T^{*m} is a contraction mapping of modulus β_P . Then, under assumptions 9 & 10,

$$\|\mu^* - \hat{\mu}^*\| \leq \frac{\delta}{1 - \beta_P}$$

Proof. Let $f : S \rightarrow S$ be defined by $f(s) = s$. Then f is a Lipschitz function with constant $L = 1$. Result is then just an application of Theorem 16. Q.E.D.

To make this results useful it remains for us to get it out of the iterated function systems notation in which it is currently expressed (recall the definition of $d(P^m, \hat{P}^m)$). For this we use the following result

Lemma 16. Let $g, \hat{g} : (X, Z) \rightarrow X$ be two policy functions satisfying $\|g - \hat{g}\| \leq \delta_g$. Define $P = g \cdot Q$ and $P^G = \hat{g} \cdot Q^G$, where Q, Q^G are stochastic transition matrices on (Z, \mathcal{Z}) , and Q^G satisfies assumption 2. Then $d(P^m, P^{Gm}) \leq m\delta_g$.

Proof. Define $\hat{P} = \hat{g} \cdot Q$.

First, observe that

$$\begin{aligned} d(P^m, \hat{P}^m) &= \max_{s \in S} \left[\int \|P^m(s, w) - \hat{P}^m(s, w)\| \Omega(dw) \right] \\ &\leq m \max_{s \in S} \left[\int \|P(s, w) - \hat{P}(s, w)\| \Omega(dw) \right] \\ &= m \max_{s \in S} \left[\int \|(g(s), Q(z, dz(dw))) - (\hat{g}(s), Q(z, dz(dw)))\| \Omega(dw) \right] \\ &\leq m \max_{s \in S} \left[\int \|g(s) - \hat{g}(s)\| \Omega(dw) \right] \\ &\leq m \max_{s \in S} \left[\int \delta_g \Omega(dw) \right] \\ &= m \max_{s \in S} \delta_g \\ &= m\delta_g \end{aligned}$$

where the first step is by the definition of $d(P^m, \hat{P}^m)$. The third from the our ability to use the standard notation in place of the iterated function systems representation (cf Stenflo (2001)). The remainder since the Q s are the same, the assumption that $\|g - \hat{g}\| \leq \delta_g$, and since δ_g is a constant and Ω a distribution function.

Similarly, observe that

$$\begin{aligned}
d(\hat{P}^m, P^{Gm}) &= \max_{s \in S} \left[\int \|\hat{P}^m(s, w) - P^{Gm}(s, w)\| \Omega(dw) \right] \\
&\leq m \max_{s \in S} \left[\int \|\hat{P}(s, w) - P^G(s, w)\| \Omega(dw) \right] \\
&= m \max_{s \in S} \left[\int \|(\hat{g}(s), Q(z, dz(dw))) - (\hat{g}(s), Q^G(z, dz(dw)))\| \Omega(dw) \right] \\
&= m \max_{s \in S} \left[\sum_{\substack{i=1, \dots, n_x, \\ j=1, \dots, n_j}} \int_{X_i \times Z_j} \|(\hat{g}(s), Q(z, dz(dw))) - (\hat{g}(s), Q^G(z, dz(dw)))\| \Omega(dw) \right] \\
&= m \max_{s \in S} \left[\sum_{\substack{i=1, \dots, n_x, \\ j=1, \dots, n_j}} \int_{X_i \times Z_j} 0 \Omega(dw) \right] \\
&= 0
\end{aligned}$$

where the fourth step is because the $X_i \times Z_j$, $i = 1, \dots, n_x$, $j = 1, \dots, n_j$ form a partition of $X \times Z$, and the fifth as \hat{g} is piecewise constant on the partition and by the assumption 2 on Q^G .

Combining these two by the triangle inequality we get,

$$d(P^m, P^{Gm}) \leq d(P^m, \hat{P}^m) + d(\hat{P}^m, P^{Gm}) \leq m\delta_g + 0 = m\delta_g$$

Q.E.D.

The interpretation of the just presented Lemma 16 is that: due to the nature of the discretization of the optimal policy and the transition matrix for the exogenous shocks, and in particular due to considering the optimal policy function as a piecewise constant extension on the discretization grid, the discretization of the steady state distribution introduces no new errors. Thus the only errors in the steady state distribution are those which we have already bounded in terms of errors in the optimal policy function, and those coming from stopping after a finite number of iterations. The intuition for why the discretization of the steady state distribution does not create any further errors comes from noting that the approximate transition function $P^G = \hat{g} \cdot Q^G$ is piecewise constant on the partition imposed by discretization; a property it inherits via \hat{g} and Q^G .

Thus, our result becomes

Corollary 8. *Let f be a Lipschitz function with constant L . Let $\|g - \hat{g}\| \leq \delta_g$ for some $\delta_g > 0$. Assume that T^{*m} is a contraction mapping of modulus β_P . Then, under assumptions 9 & 10,*

$$\left| \int f(s) \mu^*(ds) - \int f(s) \hat{\mu}^*(ds) \right| \leq \frac{Lm\delta_g}{1 - \beta_P}$$

and furthermore

$$\|\mu^* - \hat{\mu}^*\| \leq \frac{m\delta_g}{1 - \beta_P}$$

B.5 Combining our results on the Agents Distribution

Combining our results on the discretization of the agents distribution, and on the errors caused by only having an approximation of the optimal policy we get

Proposition 3. *Let μ_g^* be the true distribution. Let $\mu_{\hat{g},G}^N$ be the distribution obtained by iterating on the discretization using the approximate optimal policy until the convergence criterion $\|\mu_{\hat{g},G}^{N+m} - \mu_{\hat{g},G}^N\| \leq \epsilon_\mu$ is reached. Let the approximation of the optimal policy function be sufficiently accurate, in the sense that $\|g - \hat{g}\| \leq \delta_g$. Then*

$$\|\mu_g^* - \mu_{\hat{g},G}^N\| \leq \frac{1}{1 - \beta_P^G} (m\delta_g + \beta_P^G \epsilon_\mu) \quad (50)$$

Proof. By the triangle inequality,

$$\|\mu_g^* - \mu_{\hat{g},G}^N\| \leq \|\mu_g^* - \mu_{\hat{g}}^*\| + \|\mu_{\hat{g}}^* - \mu_{\hat{g},G}^N\| \quad (51)$$

and then simply apply Corollary 8 to the first term, and Corollary 6 to the second. *Q.E.D.*

C Monotone Mixing Condition in Pijoan-Mas (2006) Model

Pijoan-Mas (2006) investigates an extension of Aiyagari (1994) which incorporates endogenous labour choice in a framework of efficiency wages. The general eqm components of the model are the same (ie. a Cobb-Douglas firm, etc.) so we will look here just at the households maximization problem and proving that the model satisfies the monotone mixing condition. Perhaps the most important observation of the endogenizing of labour in heterogeneous agent models is that variable labour supply provides an alternative form of partial self insurance to precautionary saving (for more on this, see Pijoan-Mas (2006), and Marcet, Obiols-Homs, and Weil (2007)).

The household's dynamic programming problem is,

$$\begin{aligned} V(a, z) &= \max_{c, l, a'} \{u(c) + n(l) + \beta \sum_{z'} V(a', z') Q(z, z')\} \\ \text{s.t. } & c + a' = wz(1 - l) + (1 + r)a \\ & c \geq 0, 1 \geq l \geq 0, a' \geq \underline{a} \end{aligned}$$

where r & w are the return on assets and the wage per efficiency unit of labour. u , n are the utilities of consumption and leisure, respectively, both are strictly increasing and strictly concave (note that we assume separability of utility, this is standard in most models in the literature which use separable CES; it is used in the proofs below). c is consumption, l is leisure, a is assets, and \underline{a} is the borrowing constraint. z is the labour productivity shock which takes values in $E = \{z_1, \dots, z_{n_z}\}$ and evolves according to transition matrix Q .

We denote the optimal policies that solve this problem by $g^a(a, z)$, $g^c(a, z)$, & $g^l(a, z)$. To simplify notation we define $S = E \times A$, where $A = [\underline{a}, \infty)$ is the set of possible asset choices (we will later show that in fact we can bound A from above). That the solution to this problem will display the following properties; $V(a, z)$ is strictly increasing in a and increasing in z , $V(a, z)$ is strictly concave in a . V is continuously differentiable, follows from some standard results.³⁵ For the purpose of all the following theory we simply assume that the value function to be solved is of the more general form

$$V(a, z') = \max_l, a' F(a, a', l, z) + \beta E[V(a', z')|z]$$

where z takes values in a compact set with minimum z_1 and maximum z_{n_z} ; F is str. increasing and str. concave in a & l , inc. and concave in z , str. decreasing and str. concave in a' ; F_l is independent of a , a' and z . The model of Pijoan-Mas (2006) fits this general form.

We begin our proofs by showing that, for all points where the optimal policies are interior, the optimal polices for both assets and leisure are strictly increasing (we cannot just invoke the standard theorems because of the presence of the leisure choice, which is not a state).

Lemma 17. *Let (a, z) be such that $g^a(a, z) > \underline{a}$ and $g^l(a, z) < 1$, then $g^a(a, z)$ and $g^l(a, z)$ are both strictly increasing in a .*

Proof. For optimality, the first-order conditions imply that

$$-F_{a'}(a, g^a(a, z), g^l(a, z), z) = \beta E[V_a(g^a(a, z), z')|z] \quad (52)$$

and

$$-F_{a'}(a, g^a(a, z), g^l(a, z), z) = F_l(a, g^a(a, z), g^l(a, z), z) \quad (53)$$

(we know that V can be derived even when the borrowing constraint binds from Rincón-Zapatero and Santos (2009)).

Let $a^2 > a^1$, and $g^a(a^1, z) > \underline{a}$.

Assume $g^a(a^2, z) \leq g^a(a^1, z)$ and $F_{a'}$ is independent of l (a sufficient condition for this would be that utility is seperable in consumption and leisure, as is the case here).

³⁵Marcet, Obiols-Homs, and Weil (2007) cite Thm 9.6 of SLP for existence of a solution. This is in fact incorrect. The required Thm's is the extensions of Thm 9.6 to the Case 2 value function (SLP's categorization) in Exercise 9.7 of SLP. For the rest of our results we use the extensions of Thm 9.7 and 9.11 contained in Exercise 9.7 of SLP. The 'strictly' in Thm 9.11 (V strictly inc. in z) must be dropped as the return function is only increasing in z (because of possibility that labour supply equals zero/leisure equals one). That V is differentiable, even in presence of the borrowing constraint, is proved in Rincón-Zapatero and Santos (2009)

Then,

$$\begin{aligned}
-F_{a'}(a^1, g^a(a^1, z), g^l(a^1, z), z) &= \beta E[V_a(g^a(a^1, z), z')|z] \\
&\leq \beta E[V_a(g^a(a^2, z), z')|z] \\
&= -F_{a'}(a^2, g^a(a^2, z), g^l(a^2, z), z) \\
&< -F_{a'}(a^1, g^a(a^2, z), g^l(a^2, z), z) \\
&\leq -F_{a'}(a^1, g^a(a^1, z), g^l(a^2, z), z) \\
&= -F_{a'}(a^1, g^a(a^1, z), g^l(a^1, z), z)
\end{aligned}$$

where, steps are by FOC; by strict concavity of V in a ; by FOC; as $a^2 > a^1$ & F is increasing and strictly concave in a ; as $g^a(a^2, z) \leq g^a(a^1, z)$, F is dec. and concave in a' ; as utility is seperable in c & l .

– Contradiction.

Thus, $g^a(a^2, z) \geq g^a(a^1, z)$.

So g^a is strictly increasing in a for (a, z) s.t. $g^a(a, z) > \underline{a}$.

That $g^l(a, z)$ must then also be str. inc. in l for (a, z) s.t. $g^a(a, z) > \underline{a}$ & $g^l(a, z) < 1$, then follows immediately from one of the FOCs together with the envelope condition. The envelope condition, together with that V is str. inc. in a and that $g^a(a, z)$ is str. inc. in a implies that the LHS of 53 is str. dec. in a ; which implies that the RHS of 53 is str. dec. in a ; which implies that g^l is str. inc. in a (observe again that this argument uses the seperability of the utility fn in c & l).

Trivially, these results for strictly increasing g^a & g^l on the 'interior' (choice-wise) could be extended to a result of increasing for all (a, z) . *Q.E.D.*

We now turn to two lemmas that will be used (following the approach of Huggett (1993)) to show the mixing condition.

Lemma 18. $g^a(a, z_1) < a$, $\forall(a, z_1)$ s.t. $g^l(a, z_1) < 1$, $a > \underline{a}$.

Proof. Observe that, $V_a(a, z_1) > V_a(a, z)$, $\forall z > z_1$, $\forall(a, z_1)$ s.t. $g^l(a, z_1) < 1$.

$\implies V_a(a, z_1) > \beta E[V_a(a, z')|z_1]$, $\forall(a, z_1)$ s.t. $g^l(a, z_1) < 1$.

$\implies g^a(a, z_1) < a$, $\forall(a, z_1)$ s.t. $g^l(a, z_1) < 1$, $a > \underline{a}$.

where this last step follows from the envelope condition, FOC, and that V is str. inc. and str. concave in a by the following reasoning,

Env. Condn $\implies V_a(a, z_1) = -F_{a'}(a, g^a(a, z_1), g^l(a, z_1), z_1)$

FOC $\implies = \beta E[V_a(g^a(a, z_1), z')|z_1]$

Therefore, $E[V_a(g^a(a, z_1), z')|z_1] > E[V_a(a, z')|z_1]$ which in turn implies $g^a(a, z) < a$. *Q.E.D.*

Notice that if $g^l(a, z_1) = 1$, we would get $g^a(a, z_1) \leq a$.

Lemma 19. *There exists and a s.t. $g^a(a, z_{n_z}) = a$.*

Proof. Suppose not. Then $g^a(a, z_{n_z}) > a, \forall a$.

Since $l \in [0, 1]$ a compact set, & F is inc. and concave in l it follows that there exists a $\min_{l \in [0, 1]} F_l$ (because of seperable utility assumption; observe also that it is given by $F_l(\cdot, \cdot, 1, \cdot)$). Now, by assumption $F_{a'} \rightarrow 0$ as $a' \rightarrow \infty$, and we have seen that g^a is str inc. for $g^l < 1$, and inc. everywhere. Thus, there exists some a^* s.t $-F_{a'}(a^*, g^a(a^*, z), g^l(a^*, z), z) < \min_{l \in [0, 1]} F_l$. Now, since for interior solution the FOCs require that $-F_{a'} = F_l$, it follows that the optimal choice at (a^*, z_{n_z}) cannot be interior, so $g^l(a^*, z_{n_z}) = 1$.

Rest of proof follows as g^a is inc. in z^{36} , so then for all $a > a^*$, the choice of leisure equals ones means the the budget constraint becomes $c + a' \leq (1 + r)a$ and $u(c)$ is str. inc. and str. concave, so it is well known that there is some maximum optimal choice of a (the maximum a may be below a^* , but this doesn't matter, as we still get a bound). This is a contradiction. Q.E.D.

Now we use the proceeding three lemmas to show that the Monotone Mixing Condition holds. Denote \bar{a} as $\min_a g^a(a, z_{n_z}) = a$.

Proposition 4. *The model of Pijoan-Mas (2006) satisfies the Monotone Mixing Condition.*

Proof. That P is monotone follows immediately from monotonicity of Q (which was assumed by model), and monotonicity of $g^a(\cdot, z), \forall z$ (Lemma 17).

The mixing condition: Choose $s^* = ((a(\underline{a}, z_{n_z}) + \bar{a})/2, z_{n_z})$. Define a sequence $x_1 = \underline{a}, x_2 = g^a(x_1, z_{n_z}), x_3 = g^a(x_2, z_{n_z}), \dots$ and a sequence $y_1 = \bar{a}, y_2 = g^a(y_1, z_1), y_3 = g^a(y_2, z_1), \dots$. Note that $\{x_n\} \rightarrow \bar{a}$ monotonically and $\{y_n\} \rightarrow \underline{a}$ monotonically. Therefore, there is an N_1 such that an agent goes from (\underline{a}, z_1) to $\{s \in S : s \geq s^*\}$ with positive probability in N_1 or greater steps and there is an N_2 such that an agent goes from (\bar{a}, z_{n_z}) to $\{s \in S : s \leq s^*\}$ with positive probability in N_2 or greater steps. Let $N = \{N_1, N_2\}$ in the mixing condition. Q.E.D.

D Monotone Mixing Condition in Other Models

Huggett (1993) proves that his model satisfies the monotone mixing condition. Marcet, Obiols-Homs, and Weil (2007) claim in a footnote that their model also satisfies the MMC but provide no proof of this result. In the rest of this appendix I show that the monotone mixing condition, and many of the other conditions used in the paper, apply to the model of Aiyagari (1994). Given that the model of Aiyagari (1994) is used as a standard example of this class of models it seems useful to have this result which is otherwise absent from the literature (Aiyagari does not provide it).

³⁶A simple implication of observation $z^2 > z^1$ implies $V_a(a, z^1) \geq V_a(a, z^2)$, which by envelope condn implies $-F_{a'}(a, g^a(a, z^1), g^l(a, z^1), z^1) = -F_{a'}(a, g^a(a, z^2), g^l(a, z^2), z^2)$. From there it is just a matter of following the later steps of the proof of Lemma 17.

Aiyagari (1993) contains proofs of the following results (albeit for the iid case, but they are extended to a general Markov process case in Miao (2006)): (i) $g(x, z)$ is increasing in x and z ; (ii) If Z is compact, then so is X , and thus so is $S = X \times Z$. (iii) That the bounds on $X = [\underline{x}, \bar{x}]$ are given by $\underline{x} = \lim_{t \rightarrow \infty} x_t^{low}$ and $\bar{x} = \lim_{t \rightarrow \infty} x_t^{high}$, where $x_t^{low} = g(x_{t-1}^{low}, \underline{z})$ and $x_t^{high} = g(x_{t-1}^{high}, \bar{z})$.

Proposition 5. *The model of Aiyagari (1994) satisfies that P is increasing.*

Proof. Optimal policies are increasing, and Q is increasing. Thus $P = g \cdot Q$ is increasing. *Q.E.D.*

Proposition 6. *The model of Aiyagari (1994) satisfies the monotone mixing condition.*

Proof. Let $s^* = (\underline{x} + \bar{x})/2$. Define a sequence $a_1 = \underline{x}$, $a_2 = g(a_1, \bar{z})$, $a_3 = g(a_2, \bar{z}), \dots$ and a sequence $b_1 = \bar{x}$, $b_2 = g(b_1, \underline{z})$, $b_3 = g(b_2, \underline{z}), \dots$. Note that $\{a_i\} \rightarrow \bar{x}$ monotonically and $\{b_i\} \rightarrow \underline{x}$ monotonically. Therefore, there exists an N_1 such that an agent goes from \underline{x} to $\{x \in S : x \geq s^*\}$ with positive probability in N_1 or less steps, and likewise an N_2 such that an agent goes from \bar{x} to $\{x \in S : x \leq s^*\}$ with positive probability in N_2 or less steps. Choose $N = \max\{N_1, N_2\}$ in the mixing condition. *Q.E.D.*

Thus we have that S is compact, P is increasing, and the MMC is satisfied. So we can apply all of our above results on bounding errors.

\tilde{r} is continuous in r (and w). But monotonicity need not hold. Thus I have to go and make some assumption that \tilde{r} is not going crazy between the grid points. This continuity of \tilde{r} in r is apparently proved in Bewley (1984)³⁷. In any case a simple combination of the results presented here suffices to prove this. Namely the combination of (i) r is a continuous function of the invariant distribution, (ii) the invariant distribution is upper-semicontinuous in the policy function (by Thm 15, as we have MMC), and (iii) that the the policy function is continuous in the parameters of the maximization problem (side-effect of Berge's max thm (see SLP Thm 3.8) and that return function F is continuous in parameters)³⁸. Note that this approach also proves continuity of \tilde{r} in w and β .

Remark: The price function, $r(K) = \alpha K^{\alpha-1} - \delta$, is monotone and can easily be used to evaluate $|\tilde{r} - \hat{r}| = |r(K) - r(\hat{K})| \leq \max\{r(A + \delta_A) - r(A), r(A) - r(A - \delta_A)\}$.

E Evaluating the Likelihood in Model of Pijoan-Mas (2006)

I here describe the steps involved in the simulated likelihood evaluation of the model of Pijoan-Mas (2006). The idea is that for each individual in the panel M simulations are performed and

³⁷I say apparently as I have never seen this document. I assume given that google is unable to find it that it does not exist in digital form.

³⁸Putting these together is easy enough here since the spaces are all compact, and so the continuities can all be considered as based in uniform convergences meaning that interchanging their orders/combining them is no problem; for more see Le Van and Stachurski (2007).

their likelihoods are evaluated. We then simply sum the likelihoods across simulations and across individuals to evaluate the likelihood of the entire panel. This process is described in three steps. The first explains how an single individual-simulation is created. The second how that individual-simulation likelihood, L_i^m is evaluated. The third and final step explains how to sum across the individual-simulation likelihoods to create the likelihood of the entire panel.

Step 1 (Create an individual-simulation): Simulate $\{a_t^m, z_t^m, l_t^m\}_{t=t_0}^T$ starting from the initial period as follows,

- i) Draw the true initial asset $a_{t_0}^m$.

First, draw the initial period asset measurement error $\tilde{\xi}_{a,t_0}$ and derive

$$a_{t_0}^m = a_{t_0}^D - \tilde{\xi}_{a,t_0}$$

- ii) Draw the true initial wage $w_{t_0}^m$; use to calculate $z_{t_0}^m$.

First, draw the initial period wage measurement error $\tilde{\xi}_{w,t_0}$ and derive

$$w_{t_0}^m = \frac{w_{t_0}^D}{\tilde{\xi}_{w,t_0}}$$

then calculate

$$z_{t_0}^m = \frac{w_{t_0}^m}{w}$$

- iii) We now have initial state $\{a_{t_0}^m, z_{t_0}^m\}$. We can use the optimal policy functions to calculate $l_{t_0}^m = g^l(a_{t_0}^m, z_{t_0}^m)$, and next period assets $a_{t_0+1}^m = g^a(a_{t_0}^m, z_{t_0}^m)$. Calculate next periods exogenous state $z_{t_0+1}^m$ by drawing randomly from the transition function $pi(\cdot, z_{t_0}^m)$.

- iv) Repeating iii until the end of period T we construct the sequence of variables $\{a_t^m, z_t^m, l_t^m\}_{t=t_0}^T$.

Step 2 (Evaluate L_i^m): Given the simulated sequence of variables $\{a_t^m, z_t^m, l_t^m\}_{t=t_0}^T$, we then derive the measurement error. Then, we calculate the log-likelihood increment to person i at the m th simulation as follows.

Let us denote,

$$\begin{aligned} \tilde{\xi}_{a,t} &= a_t^D - a_t^m \\ \tilde{\xi}_{l,t} &= l_t^D - l_t^m \\ \tilde{\xi}_{w,t} &= \ln(w_t^D) + \ln(l_t^D) - \ln(w_t^m) - \ln(l_t^m), \quad w_t^m \equiv w z_t^m \end{aligned}$$

for $t = t_0 + 1, \dots, t_0 + T$ (for l_t also including the case $t = t_0$).

Then, log-likelihood increment of simulation m for person i is,

$$L_i^m = \sum_{t=t_0}^T \left[\frac{\tilde{\xi}_{a,t}^2}{-2\sigma_{\xi,a}^2} - \ln(\sigma_{\xi,a}) \right] \\
\sum_{t=t_0}^T \left[\frac{\tilde{\xi}_{l,t}^2}{-2\sigma_{\xi,l}^2} - \ln(\sigma_{\xi,l}) \right] \\
\sum_{t=t_0+1}^T \left[\frac{(\tilde{\xi}_{w,t} + \frac{1}{2}\sigma_{\xi,W})^2}{-2\sigma_{\xi,W}^2} - \ln(\sigma_{\xi,W}) \right] + \left[\frac{\tilde{\xi}_{w,t_0}^2}{-2\sigma_{\xi,w_0}^2} - \ln(\sigma_{\xi,w_0}) \right]$$

Step 3 (Likelihood L): We repeat the simulation and likelihood increment calculation for $m = 1, \dots, M$ and derive the simulated log-likelihood increment for individual i as follows,

$$L_i = \ln \left[\sum_{m=1}^M \exp(L_i^m) \right]$$

The total log-likelihood is then just the sum across individuals

$$L = \sum_{i=1}^T L_i$$