Inequality, Business Cycles, and Monetary-Fiscal Policy*

Anmol Bhandari  David Evans  Mikhail Golosov
U of Minnesota     U of Oregon   U of Chicago

Thomas J. Sargent
NYU

February 13, 2020

Abstract

We study optimal monetary and fiscal policy in a model with heterogeneous agents, incomplete markets, and nominal rigidities. We show that functional derivative techniques can be applied to approximate equilibria in such economies quickly and efficiently. Our solution method does not require approximating policy functions around some fixed point in the state-space and is not limited to first-order approximations. We apply our method to study Ramsey policies in a textbook New Keynesian economy augmented with incomplete markets and heterogeneous agents. Responses differ qualitatively from those in a representative agent economy and are an order of magnitude larger. Conventional price stabilization motives are swamped by an across person insurance motive that arises from heterogeneity and incomplete markets.

Key words: Sticky prices, heterogeneity, business cycles, monetary policy, fiscal policy

*We thank Adrien Auclert, Benjamin Moll, and seminar participants at numerous seminars and conferences for helpful comments
1 Introduction

We compute optimal monetary and fiscal policies in a New Keynesian economy populated by agents who face aggregate and idiosyncratic risks. Agents differ in their wage, exposure to aggregate shocks, holdings of financial assets, and ability to trade assets. Incomplete financial markets prevent agents from fully insuring risks. Firms are monopolistically competitive. Price adjustments are costly. We examine how a Ramsey planner’s choices of nominal interest rates, transfers, and flat-rate taxes on labor earnings, dividends, and interest income respond to aggregate shocks.

Analysis of Ramsey policies in settings like ours faces substantial computational challenges. The aggregate state in a recursive formulation of the Ramsey problem includes the joint distribution of individual asset holdings and auxiliary promise-keeping variables that had been previously chosen by the planner. The law of motion for that high-dimensional object must be determined together with the optimal policies, and the distributions along the transition path differ substantially from the invariant distribution without aggregate shocks. These aspects render inapplicable common computational strategies that either approximate policy functions after summarizing cross-sectional distributions with a small number of moments or linearize around some time-invariant distribution.

To overcome this challenge, we develop a new computational approach that can be applied to economies with substantial heterogeneity and does not require knowing their long-run properties in advance. Our approach builds on a perturbation theory that uses small noise expansions with respect to a one-dimensional parameterization of uncertainty. Each period, along a sample path, we approximate policy functions by applying a perturbation algorithm evaluated at the current cross-sectional distribution. We use approximate decision rules for the current period to determine outcomes, including the cross-sectional distribution next period. Then we obtain approximations of next period’s decision rules by perturbing around that new distribution. In this way, we sequentially update points around which policy functions are approximated along the equilibrium path.

Our perturbation approach requires repeatedly computing derivatives of policy functions with respect to all state variables. One state variable is a distribution over a multi-dimensional space of agents’ characteristics. Except for very simple models of heterogeneity, it is impractical to compute the derivative with respect to this distribution (i.e., the Frechet derivative). We prove that in a large class of heterogeneous agent competitive equilibrium models this computationally-intensive step can be avoided and the problem of finding parameters of the approximated policy functions can be written as a collection of low-dimensional systems of linear equations that are independent of each other. This property emerges because in standard competitive environments, conditional on prices and aggregate quantities,
each agent’s optimal choices can be solved separately from those of others. These choices can then be aggregated into feasibility constraints to solve a low-dimensional fixed point problem that determines prices and aggregate quantities. Because these systems of linear equations are independent across agents, they are easily parallelizable. This lets us handle the ample heterogeneity present in our model. We go on to show that an analogous, computationally convenient, linear structure prevails for second- and higher-order expansions, making our approach applicable to many optimal policy problems in which aggregate risks have important effects on equilibrium dynamics.

We apply our approach to a textbook New Keynesian sticky price model (see, e.g., Galí (2015)) augmented with heterogeneous agents in the spirit of Bewley-Hugget-Aiyagari. Financial markets are incomplete and agents can trade only non-state-contingent nominal debt. Agents’ wages are subject to idiosyncratic and aggregate shocks that we calibrate to match the U.S. business cycle and cross-sectional properties of labor earnings. We choose the initial joint distribution of nominal claims, real claims, and wages to match cross-sectional moments in the Survey of Consumer Finances. We posit two types of aggregate shocks: a productivity shock and a shock to the elasticity of substitution between differentiated intermediate goods that affects firms’ optimal markups. We study two types of Ramsey policies. First, a “purely monetary policy” planner is required to keep all tax rates fixed, an assumption commonly used with New Keynesian models. In this case, the planner can adjust only nominal interest rates and a uniform lump-sum transfer. Second, a more powerful “monetary-fiscal” Ramsey planner can adjust tax rates on all sources of income in addition to interest rates and transfers.

In conventional representative agent New Keynesian models, the role for policy is inherited from a motive to stabilize inflation and the “output gap,” i.e., the difference between actual output and the first-best level of output. In our setting, households differ in their shares of labor, dividend, and bond income and, therefore, in their exposures to aggregate shocks. Without a complete set of Arrow securities agents cannot hedge risk, so aggregate shocks affect agents differently. While the planner uses the expected path of taxes and transfers to redistribute and provide insurance with respect to idiosyncratic risk, how optimal policy deviates from this path is governed by a necessity to provide insurance against aggregate shocks. In the calibrated economy, the welfare gains from providing insurance are much larger than the benefits of price stability, and the optimal monetary responses in our heterogeneous agent economy are both quantitatively larger and qualitatively different from those in its representative agent counterpart.

To understand how insurance concerns shape policy responses, consider the optimal monetary response to a one-time positive markup shock. One effect of this shock, traditionally emphasized by the New Keynesian literature, is that firms want to increase their prices.
To maintain price stability, the planner can increase nominal interest rates which lowers aggregate demand and, hence, firm’s marginal costs. This response forms the basis for the classical prescription of “leaning against the wind” and raising interest rates when firms’ desired markups are high (Galí (2015)). The markup shock also changes the relative shares of payments that go to labor and equity. Such movements in factor shares have no welfare consequences when firm owners and wage earners are the same people, but has important insurance implications when agents are heterogeneous. When agents cannot trade Arrow securities, the planner can use monetary and fiscal policy to substitute for missing insurance markets. Since a positive markup shock creates an unexpected drop in wage income (and a windfall in profits), the planner provides insurance by lowering nominal rates to boost wages. A negative markup shock reverses insurance needs, and therefore the optimal response is a mirror image of the response to a positive markup shock.

Quantitatively, the strength of the insurance motive depends on the correlation of labor and capital income, with smaller correlations calling for higher insurance needs. In the data the distribution of stock ownership is much more skewed than the distribution of labor earnings, which implies large welfare gains from insurance. As a result, in our calibrated economy optimal monetary responses to markup shocks are an order of magnitude larger and go in the opposite direction from those in the representative agent economy.

The insurance motive also shapes Ramsey responses to TFP shocks. While TFP shocks affect profits and wages in the same way, the effects of TFP shocks are not shared equally across borrowers and lenders. When borrowing is not state contingent, a TFP shock changes total output while keeping nominal obligations unchanged. A negative TFP shock hurts borrowers while a positive TFP shock hurts lenders. The planner can provide insurance and improve welfare by lowering (raising) the real return on debt in response to a negative (positive) TFP shock. Such a policy contrasts with the standard New Keynesian prescription of adjusting the nominal rate one-for-one with the “natural” rate of interest in the flexible price economy. Since there is a large dispersion of nominal claims in the data, the insurance motive is strong in the calibrated economy. We find that the optimal response to a mean-reverting negative TFP shock is to lower nominal as well as real rates of interest despite the fact that the “natural” rate increases.

We also show that, in providing insurance, there exists a trade-off between the on-impact magnitude and the overall duration of the monetary response to a shock. The New Keynesian Phillips curve implies that current inflation is equal to the present value of all future wages. When all agents can freely borrow and lend, their utilities depend only on the present value of their income. Therefore, an efficient monetary response minimizes costs of future inflation by front-loading all insurance: monetary policy responds aggressively to the shock at the time of the impact but this response is short lived. When some agents face trading frictions,
the planner also needs to smooth the insurance provision over time. Monetary responses become smaller in magnitude but longer lasting.

The role of monetary policy is diminished when the planner can also use fiscal instruments to respond to aggregate shocks. In the representative agent economy, the planner can fully offset the effect of markup shocks by changing labor taxes, financed through lump sum transfers, making monetary policy redundant. This response is no longer optimal when agents are heterogeneous, because it increases the volatility of the after-tax income of low wage earners relative to high wage earners. We show that it is more efficient to use dividend or corporate profit taxes in response to markup shocks when agents are heterogeneous. For similar reasons, adjusting taxes on interest income is an efficient fiscal response to TFP shocks.

1.1 Related literature

Our paper contributes to two literatures, one that approximates equilibria of incomplete markets economies with heterogeneous agents, and another that computes Ramsey plans for fiscal and monetary policies.

We use small noise expansions around transition paths like those deployed by Fleming (1971), Fleming and Souganidis (1986), and Anderson et al. (2012). Those authors apply this approach only to problems with low-dimensional states. The state in our model includes a joint distribution over a multi-dimensional domain, a feature that gives us a much larger state space and makes direct application of the approaches taken in those earlier papers computationally infeasible. To meet these challenges, we introduce functional derivatives techniques\footnote{Childers (2018) combines related functional derivative techniques with a Reiter (2009) method, but unlike our approach, his still requires that distributions remain close to the invariant distribution of a no-aggregate-shock economy. We build on and extend Evans (2015).} to cope with the most computationally intensive step and reformulate the problem of computing approximate policy functions as a manageable collection of small-dimensional linear equations. We also show how our techniques can be used to construct second- and higher-order approximations via a convenient set of recursions.

There are trade-offs in applying our method as compared to other popular approaches such as Krusell and Smith (1998) or Reiter (2009). Relative to Krusell and Smith (1998), our method is not restricted to problems in which a small number of moments are enough to summarize how policy functions depend on the distribution of agents’ characteristics. In a Reiter-type approach, policy functions are accurate with respect to idiosyncratic shocks, but are approximated only to the first order with respect to aggregate shocks around the invariant distribution of the no-aggregate-shock economy. In comparison, our method is not restricted to linear approximations, and we repeatedly update the coefficients of the
approximation as the state of the economy moves along the equilibrium path. Our method is particularly well suited when the economy features non-trivial or even non-stationary transition dynamics; when impulse responses depend on past realizations of shocks; or when higher-order moments of aggregate variables, such as risk premium, play an important role. It comes at a cost that we approximate less well the dependence of policy functions on idiosyncratic shocks.

We study the numerical accuracy of our method by computing the competitive equilibrium given fixed government policies for a special case of our model, which corresponds to the economy studied by Acharya and Dogra (2018). The advantage of this economy is that it is possible to compute the equilibrium analytically without requiring approximations. This allows us to compare both our approximation and Reiter’s to the true solution. Under Acharya and Dogra (2018) calibration, the maximum numerical errors in our approximated policy functions are less than 0.05%. Moreover, the approximation errors in policy functions for aggregate variables are two orders of magnitude smaller than the ones obtained under Reiter’s method. While our method is less accurate with respect to idiosyncratic shocks, those errors wash out in the aggregate; on the other hand, the second-order errors from the aggregate shocks under Reiter’s method do not. Moreover, we also show how these errors from aggregate shocks compound over time under Reiter’s approach, affecting the distribution in the long run. Finally, we also illustrate how the drift of the distribution of assets away from the point of approximation leads to errors in impulse responses under the Reiter-based method which do not exist in our method.

A substantial literature on optimal monetary and fiscal policies in the Ramsey tradition has mostly studied economies with limited or no sources of heterogeneity. For a textbook level treatment of optimal monetary policies in representative agent New Keynesian models see Galí (2015) and Woodford (2003). Papers such as Bilbiie and Ragot (2017), Challe (2017), Bilbiie (2019), and Debortoli and Gali (2017) study optimal monetary policy in economies with very limited heterogeneity and in which a cross-sectional distribution disappears from the formulation of the Ramsey problem and the analysis can be done using traditional techniques. Like us, those papers emphasize that uninsurable aggregate shocks create reasons for the planner to compromise price stability. Recent papers by Legrand and Ragot (2017) and Nuno and Thomas (2016) develop alternative methods different from ours to approximate Ramsey allocations in incomplete market economies with heterogeneity. Legrand and Ragot (2017) apply their method to a neoclassical economy and Nuno and Thomas (2016) study transition dynamics of a Ramsey allocation in an small open economy setting.

There are several papers that study optimal monetary and fiscal policies in calibrated representative agent settings. For instance, see Chari and Kehoe (1999) for a neoclassical setup, Schmitt-Grohe and Uribe (2004a) and Siu (2004) for optimal responses to government spending shocks in setups with nominal rigidities.
2 Environment

Our economy is inhabited by a continuum of infinitely lived households. Individual $i$’s preferences over final consumption good $c_{i,t}$ and hours $n_{i,t}$ are ordered by

$$E_t \sum_{t=0}^{\infty} \beta^t u(c_{i,t}, n_{i,t}),$$

(1)

where $E_t$ is a mathematical expectation operator conditioned on time $t$ information, $\beta \in (0, 1)$ is a time discount factor, and $u$ is an infinitely differentiable utility function that is concave in $c$ and $-n$ and satisfies the Inada conditions. We denote its derivatives by $u_c(c_{i,t}, n_{i,t})$, $u_n(c_{i,t}, n_{i,t})$, and so on. A random variable with subscript $t$ is measurable with time $t$ information.

Agent $i$ who works $n_{i,t}$ hours supplies $\epsilon_{i,t}n_{i,t}$ units of effective labor, where $\epsilon_{i,t}$ is an exogenous productivity process. Effective labor receives nominal wage $P_tW_t$, where $P_t$ is the nominal price of the final consumption good at time $t$. Agents trade a one-period risk-free nominal bond with price $Q_t$ in units of the final consumption good. We use $P_tb_{i,t}$ to denote face value of nominal bonds owned by agent $i$ and $P_td_{i,t}$ to denote nominal dividends from intermediate goods producers. Let $\Pi_t = \frac{P_t}{P_{t-1}} - 1$ denote the net inflation rate. The budget constraint of household $i$ at date $t$ in units of final goods is

$$c_{i,t} + Q_tb_{i,t} = (1 - \Upsilon^a_t) W_t \epsilon_{i,t} n_{i,t} + T_t + (1 - \Upsilon^d_t) d_{i,t} + (1 - \Upsilon^b_t) \frac{b_{i,t-1}}{1 + \Pi_t}.$$  

(2)

All agents receive the same uniform lump-sum transfer $T_t$, and face a linear tax $\Upsilon^a_t$ on their labor earnings, a tax $\Upsilon^d_t$ on their dividends, and a tax $\Upsilon^b_t$ on their bond income.$^3$ The government’s budget constraint at time $t$ is

$$\bar{G} + T_t + \frac{B_{t-1}}{1 + \Pi_t} = \int_i \left[ \Upsilon^a_t W_t \epsilon_{i,t} n_{i,t} + \Upsilon^d_t d_{i,t} + \Upsilon^b_t \frac{b_{i,t-1}}{1 + \Pi_t} \right] di + Q_tB_t,$$

where $\bar{G}$ is a time-invariant level of non-transfer government expenditures.

A final good $Y_t$ is produced by competitive firms that use a continuum of intermediate goods $\{y_t(j)\}_{j \in [0,1]}$ in a production function

$$Y_t = \left[ \int_0^1 y_t(j) \frac{\Phi_t^{-1}}{\Phi_t^{-1}} dj \right]^{\frac{\Phi_t^{-1}}{\Phi_t^{-1}}},$$

$^3$Although $\Upsilon^b_t$ multiplies $b_{i,t-1}$, we refer to it as a tax on the bond income because it is equivalent to a tax on the return on a one-period bond. To see this, rewrite the budget constraint using the market value of nominal debt $b_{i,t} = Q_tb_{i,t}$, and notice that the term $(1 - \Upsilon^b_t) \frac{b_{i,t-1}}{1 + \Pi_t} = (1 - \Upsilon^b_t) (R_{t-1,t}) b_{i,t-1}$, where $R_{t-1,t} = \left( \frac{1}{Q_{t-1}} \right) \left( \frac{1}{1 + \Pi_t} \right)$ is the real return from holding a nominal bond from $t-1$ to $t$. 

7
where the elasticity of substitution $\Phi_t$ is stochastic. Final good producers take the final good price $P_t$ and the intermediate goods prices $\{p_t(j)\}_j$ as given and solve

$$\max_{\{y_t(j)\}_{j \in [0,1]}} P_t \left[ \int_0^1 y_t(j) \frac{\Phi_t - 1}{\Phi_t} \, dj \right] - \int_0^1 p_t(j) y_t(j) \, dj. \tag{3}$$

Outcomes of optimization problem (3) are a demand function for intermediate goods

$$y_t(j) = \left( \frac{p_t(j)}{P_t} \right)^{-\Phi_t} Y_t, \tag{4}$$

and a final goods price satisfying

$$P_t = \left( \int_0^1 p_t(j)^{1-\Phi_t} \right)^{1/(1-\Phi_t)}.$$

Intermediate goods $y_t(j)$ are produced by monopolists with production functions

$$y_t(j) = [n_t^D(j)]^\alpha \, [h_t(j)]^{1-\alpha}, \tag{5}$$

where $n_t^D(j)$ is effective labor hired by firm $j$ and $h_t(j)$ is an intermediate input measured in units of the final good. Intermediate goods monopolists face downward sloping demand curves $\left( \frac{p_t(j)}{P_t} \right)^{-\Phi_t} Y_t$ and choose prices $p_t(j)$ while bearing quadratic Rotemberg (1982) price adjustment costs $\frac{\psi}{2} \left( \frac{p_t(j)}{p_{t-1}(j)} - 1 \right)^2$ measured in units of the final consumption good. Intermediate goods producing firm $j$ chooses prices $\{p_t(j)\}_t$ and factor inputs $\{h_t(j), n_t^D(j)\}_t$ that solve

$$\max_{\{p_t(j), h_t(j), n_t^D(j)\}_t} \mathbb{E}_0 \sum_t S_t (1 - \Upsilon_t) \left\{ \frac{p_t(j)}{P_t} y_t(j) - W_t n_t^D(j) - h_t(j) - \frac{\psi}{2} \left( \frac{p_t(j)}{p_{t-1}(j)} - 1 \right)^2 \right\} \tag{6}$$

subject to (4) and (5), where $W_t$ is the real wage per unit of effective labor and $S_t$ is the stochastic discount factor (SDF) defined recursively via

$$S_t = S_{t-1} Q_{t-1} (1 + \Pi_t) / \left( 1 - \Upsilon_t \right),$$

with $S_{-1} = 1.4$ In a symmetric equilibrium, $p_t(j) = P_t$, $y_t(j) = Y_t$, $h_t(j) = H_t$, $n_t^D(j) = N_t$

---

4In economies with heterogeneous agents and incomplete markets, a stand must be taken on how firms are valued. To explain our numerical methods most transparently we chose a simple specification of the SDF that discounts future profits at the after-tax real risk-free rate. Our quantitative results are virtually identical when we use other popular choices of SDFs.
for all $j$. Market clearing conditions in labor, goods, and bond markets are

\[
C_t = \int c_{i,t} di, \quad N_t = \int \epsilon_{i,t} n_{i,t} di, \quad D_t = Y_t - H_t - W_t N_t - \frac{\psi}{2} \Pi^2_t, \tag{8}
\]

\[
Y_t = N_t^\alpha H_t^{1-\alpha}, \quad \Pi_t = P_t / P_{t-1} - 1
\]

\[
C_t + G = Y_t - H_t - \frac{\psi}{2} \Pi^2_t, \tag{9}
\]

\[
\int b_{i,t} di = B_t. \tag{10}
\]

There are aggregate and idiosyncratic shocks. Aggregate shocks are a “markup” shock $\Phi_t$ and an aggregate productivity $\Theta_t$ that follow AR(1) processes

\[
\ln \Phi_t = \rho_\Phi \ln \Phi_{t-1} + (1 - \rho_\Phi) \ln \bar{\Phi} + \mathcal{E}_{\Phi,t},
\]

\[
\ln \Theta_t = \rho_\Theta \ln \Theta_{t-1} + (1 - \rho_\Theta) \ln \bar{\Theta} + \mathcal{E}_{\Theta,t},
\]

where $\mathcal{E}_{\Phi,t}$ and $\mathcal{E}_{\Theta,t}$ are mean-zero random variables that are i.i.d. over time.

Individual productivity $\epsilon_{i,t}$ follows a stochastic process described by

\[
\ln \epsilon_{i,t} = \ln \Theta_t + \ln \theta_{i,t} + \epsilon_{\epsilon,i,t}, \tag{12}
\]

\[
\ln \theta_{i,t} = \rho_\theta \ln \theta_{i,t-1} + \epsilon_{\theta,i,t}, \tag{13}
\]

where innovations $\epsilon_{\epsilon,i,t}$ and $\epsilon_{\theta,i,t}$ are mean-zero, uncorrelated with each other, and i.i.d. over time.

We set the initial price level $P_{-1} = 1$. Agent $i$ in period 0 is characterized by a triple $(\theta_{i,-1}, b_{i,-1}, s_i)$, where $\theta_{i,-1}$ is agent $i$’s persistent component of productivity, $b_{i,-1}$ denotes the bonds that agent $i$ initially owns, and $s_{i}$ denotes agent $i$’s initial ownership of equity. In what will serve as our baseline specification, we assume that agent $i$’s dividends in period $t$ are $d_{i,t} = s_i D_t$. This imposes that agents do not trade equity and that $s_i$ is a permanent characteristic. We denote the collection $\{\theta_{i,-1}, b_{i,-1}, s_i\}_i$ as an initial condition.

**Definition 1.** Given an initial condition and a monetary-fiscal policy $\{Q_t, \Upsilon_t, T_t\}_t$, a competitive equilibrium is a sequence $\{\{c_{i,t}, n_{i,t}, b_{i,t}\}_i, C_t, N_t, B_t, W_t, P_t, Y_t, H_t, D_t, \Pi_t, S_t\}_t$ such that: (i) $\{c_{i,t}, n_{i,t}, b_{i,t}\}_i$ maximize (1) subject to (2) and natural debt limits;\(^5\) (ii) final goods firms choose $\{y_t(j)\}_j$ to maximize (3); (iii) intermediate goods producers’ prices and factor

\(^5\)Natural debt limits are enforced by requiring a transversality condition

\[
\lim_{s \to \infty} \mathbb{E}_t \left( \prod_{k=1}^{s} Q_{t+k} \right) b_{i,t+s} = 0.
\]
inputs solve (6) and satisfy \( p_t(j) = P_t, \ y_t(j) = Y_t, \ h_t(j) = H_t, \ n_t^d(j) = N_t \) for all \( j \); and (iv) market clearing conditions (8)-(11) are satisfied.

A Ramsey planner orders allocations by

\[
\mathbb{E}_0 \int \sum_{t=0}^{\infty} \beta^t \vartheta_i u(c_{i,t}, n_{i,t}) \, di, \tag{14}
\]

where \( \vartheta_i \geq 0 \) is a Pareto weight attached to agent \( i \) with \( \int \vartheta_i di = 1 \).

**Definition 2.** Given an initial condition and a time-invariant tax policy satisfying \( \Upsilon_t = \bar{\Upsilon} \) for some \( \bar{\Upsilon} \), an optimal monetary policy is a sequence \( \{Q_t, T_t\}_t \) that implements a competitive equilibrium allocation that maximizes (14). Given an initial condition, an optimal monetary-fiscal policy is a sequence \( \{Q_t, \Upsilon_t, T_t\}_t \) that implements a competitive equilibrium allocation that maximizes (14). A maximizing monetary or monetary-fiscal policy is called a Ramsey plan; an associated allocation is called a Ramsey allocation.

We characterize competitive equilibria by the feasibility constraints (7), (8), (9), and (10); the consumer and firm optimality conditions

\[
(1 - \Upsilon_t^b) W_t \epsilon_{i,t} u_{c,i,t} = -u_{n,i,t}, \tag{15}
\]

\[
Q_t u_{c,i,t} = \beta \mathbb{E}_t u_{c,i,t+1} \left( 1 - \Upsilon_{t+1}^b \right) / (1 + \Pi_{t+1})^{-1}, \tag{16}
\]

\[
\frac{1}{\psi} Y_t \left[ 1 - \Phi_t \left( 1 - \frac{1}{1 - \alpha} \left( \frac{1 - \alpha}{\alpha} W_t \right)^{\alpha} \right) \right] - \Pi_t(1 + \Pi_t) + \mathbb{E}_t \frac{S_t + 1}{S_t} \left( \frac{1 - \Upsilon_{t+1}^d}{1 - \Upsilon_t^d} \right) \Pi_{t+1}(1 + \Pi_{t+1})^2 = 0, \tag{17}
\]

\[
\frac{1 - \alpha}{\alpha} W_t = \frac{H_t}{N_t}, \tag{18}
\]

and agents’ budget constraints that, by using equation (15) to eliminate \( W_t \) and \( \Upsilon_{t+1}^d \), we can represent as

\[
c_{i,t} - T_t - (1 - \Upsilon_t^d) s_i D_t - \frac{(1 - \Upsilon_t^b) b_{i,t-1}}{1 + \Pi_t} = \left( \frac{u_{n,i,t}}{u_{c,i,t}} \right) n_{i,t} + \mathbb{E}_t \left( \frac{u_{i,t+1}}{u_{c,i,t}} \right) \left( 1 - \Upsilon_{t+1}^b \right) b_{i,t} / (1 + \Pi_{t+1}). \tag{19}
\]

We can construct an optimal monetary-fiscal policy and an associated competitive equilibrium to maximize the welfare criterion (14) subject to these constraints. Choice of an optimal monetary policy is subject to the additional constraints \( \Upsilon_t = \bar{\Upsilon} \) for all \( t \geq 0 \).
2.1 Discussion of the environment

We want to understand how optimal policies respond to aggregate shocks and what economic forces motivate those responses. To help us do so, we use two baselines that differ in whether or not the Ramsey can adjust tax rates. In what we call our optimal monetary-fiscal policy baseline, in response to aggregate shocks, the Ramsey planner can freely adjust both the nominal interest rate and all tax rates. In what we call our optimal monetary policy baseline, the Ramsey planner can adjust only the nominal interest rate. Our use of a pure monetary policy baseline follows a New Keynesian tradition that, in our notation, assumes time-invariant tax rates \( \Upsilon_t = \bar{\Upsilon} \) and that only the nominal interest rate \( Q_t^{-1} \) can respond to shocks with \( T_t \) adjusting to satisfy the government’s budget constraint. A popular justification for this restriction is that central banks adjust interest rates fast enough to react to shocks at business cycle frequencies, while institutional constraints prevent adjusting tax rates quickly and often. In principle, one can study optimal monetary policy for any arbitrary \( \bar{\Upsilon} \), but in the spirit of the New Keynesian literature, in our section 4 quantitative application, we focus on the level of \( \bar{\Upsilon} \) that maximizes welfare (14) under the optimal monetary policy associated with that \( \bar{\Upsilon} \).

Our environment extends a textbook New Keynesian model along the lines of Galí (2015, ch. 3) to allow for incomplete markets and heterogeneous agents in the tradition of Bewley-Hugget-Aiyagari. An advantage of using this canonical New Keynesian setup is that normative prescriptions for the representative agent economy are widely understood. That allows us to isolate modifications of those prescriptions that heterogeneity and incomplete markets bring. In our baseline environment, we model heterogeneity using a standard process for wage dynamics from the macro labor literature, for instance as in Low et al. (2010) or Storesletten et al. (2001). In section 6.3, we enrich the baseline process for wage dynamics to allow for diverse responses of labor earnings to recessions that are documented by Guvenen et al. (2014).

In our two baseline models, we assume that all agents can freely trade bonds subject to natural debt limits. This means Ricardian equivalence holds and timing of transfers is undetermined.\(^6\) This is a natural baseline as economies with ad hoc debt limits often prescribe ad hoc non-stationary optimal fiscal policy by front-loading transfers to undo those constraints.\(^7\) We relax this assumption in section 6.1 when we prevent a subset of agents

\(^6\) In the general formulation of the Ramsey problem, we do not restrict lump-sum transfers \( T_t \) to be positive. However, in our section 4 quantitative application, transfers are always positive since households are unequal and planner cares about redistribution.

\(^7\) In Bhandari et al. (2017) we provide a comprehensive treatment of a Ramsey problem with ad hoc debt limits. In the economy with ad hoc debt limits the planner can simply choose the timing of transfers to undo such ad hoc debt limits. If the planner enforces debt and tax liabilities equally then welfare in the economies with ad hoc and natural debt limits coincides. Welfare can sometimes be improved in the economy with ad hoc debt limits if the planner commits not to enforce private debt contracts (see also Yared (2013) for a
from trading the risk-free bond or other assets.

In our baseline models, we also assume that agents can trade debt but not equity. This specification yields several insightful special cases that we find useful for explaining our numerical methods and economic forces in the model. We relax the nontradability of claims to dividends in section 6.2 when we introduce mutual funds that hold corporate equity and government debt and that issue mutual fund shares to households who trade them in a competitive market.

3 Approximation method

We approximate Ramsey plans for heterogeneous agents (HA) economies that by their very nature have state vectors that include joint distributions of agents’ characteristics. For reasons anticipated in section 1.1, this feature impelled us to depart from approximation methods used by earlier authors such as the projection method of Krusell and Smith (1998) and the perturbation-around-a-no-aggregate-shock-economy method used by Reiter (2009).

The Krusell-Smith approach works well when the dimension of the state is low (e.g., a univariate distribution), and policy functions admit approximate aggregation. Even in the simplest versions of our problem, the state variable is a multivariate distribution, and application of a Krusell-Smith strategy would require tracking higher-order moments and be computationally very expensive. The Reiter approach was designed for situations in which a no-aggregate-shock invariant distribution is easy to compute and when it is known that the state in an economy with aggregate shocks always remains close to the no-aggregate shock invariant distribution. This condition might prevail in competitive equilibria under arbitrarily fixed government policies, but it is not in our setting.8

In this section, we propose an alternative method of solving HA economies. The key step in our analysis is to demonstrate how functional derivative techniques can be used to characterize the dependence of policy functions on a high-dimensional state that changes over time in response to aggregate shocks. These computational techniques are fast and work at any order of approximation.9

---

8The long-run behavior of the state variables even in the simplest Ramsey problems can differ dramatically with and without aggregate shocks in otherwise identical economies. One can easily see this from the classic tax-smoothing paper by Barro (1979). In his economy, government debt is the only state variable. It stays always at its initial level in the economy without aggregate shocks and follows a random walk with aggregate shocks. This law of motion has very different implications for the long-run distribution of debt in these two cases. Similarly, Aiyagari et al. (2002), Farhi (2010), Bhandari et al. (2017) all study Ramsey policies and find that the invariant distribution of the state variables, while being well defined in all cases they consider, is discontinuous with respect to the size of aggregate shocks around the no-aggregate-shock level.

9We extensively discuss the comparison of our techniques to alternatives in section 3.3, but a reader may find the following informal summary helpful at this point. Perturbation methods of Reiter (2009) and Kaplan et al. (2018) are exact with respect to the dependence of policy functions on idiosyncratic...
Section 3.1 considers a special case of our environment by setting parameters so that the sole state variable in a recursive formulation of a continuation Ramsey problem is a joint distribution of agents’ characteristics. This case allows us to describe our techniques in a fairly transparent way. Section 3.2 then tells how to extend our approach to the general environment. Section 3.3 discusses numerical accuracy and computational speed and makes comparisons with earlier methods for approximating equilibria of HA economies.

3.1 An informative special case

To obtain a case that explains our methods in a transparent manner, we focus on optimal monetary-fiscal policy for a utilitarian planner and impose: (i) equity holdings $s_i$ are uniform across households; (ii) $\alpha = 1$ so that no intermediate goods are used as inputs; (iii) all shocks are i.i.d. These three restrictions are designed to make the state space as small as possible while keeping it large enough to preserve the essence of our technique. The first two assumptions imply that the Phillips curve constraint (17) is slack in all periods and so can be omitted from the Ramsey planner’s maximization problem. The third assumption implies that past shocks do not appear as arguments in optimal policy functions. Therefore, the state becomes just a distribution of individual endogenous characteristics that summarize individual optimality conditions (15), (16), and (19).

Our focus in this section will be on characterizing how policy functions depend on that distributional state. There are several mathematically equivalent ways to choose the state space to formulate our problem recursively. Our approach works best when a state space satisfies an independence property that we formally define below. In the simple economy that we consider in this section, most commonly used choices for state variables satisfy this property. For clarity of exposition and consistency, we adopt a recursive formulation of a Ramsey problem that preserves the independence property in more general settings.

Let $M_t \equiv \int u_{c,i,t}di$ be the average marginal utility of consumption at time $t$, and let $m_{i,t} \equiv u_{c,i,t}/M_t$ be scaled marginal utility of consumption of agent $i$ at time $t$. We can interpret $m_{i,t}$ as (an inverse of) a Pareto-Negishi weight that the planner attaches to agent shocks but only first-order approximate with respect to aggregate shocks, all around a fixed distribution $\Omega$. The approximation errors in this approach are thus on the order of $O(\sigma_{agg}^2, \|\Omega - \Omega\|^2)$, where $\sigma_{agg}$ is a measure of the size of aggregate shocks. Childers (2018) provides a formal treatment. Our approach uses expansions with respect to both aggregate and idiosyncratic shocks around the time $t$ distribution $\Omega_t$ in that period, and can be done to an arbitrary order of approximation. Approximation errors are of the order $O\left(\sigma_{agg}^{n+1}, \sigma_{idiosync}^{n+1}\right)$ for arbitrary $n$. The two approaches are therefore complementary and have advantages and disadvantages that depend on specific applications.
policy functions, \( \tilde{x} \) 

The law of motion for the aggregate state variable is denoted by \( \tilde{x} \) as well as \( \tilde{z} \). Policy functions for individual states \( \tilde{x} \) determine all lower-case time \( t \) values of all upper-case choice variables \( \tilde{z} \). We use tildes to denote policy functions for the \( t > 0 \) continuation Ramsey plan. The aggregate policy functions determine the time \( t \) values of all upper-case choice variables in problem (21). We denote the \( n_X \) dimensional vector of these functions by \( \tilde{X}(\Omega, \mathcal{E}) \), where \( \mathcal{E} \) are aggregate shocks. Individual policy functions determine all lower-case time \( t \) choice variables for the planner in problem (21). We denote individual policy functions by \( \tilde{a}(z, \Omega, \varepsilon, \mathcal{E}) \), where \( \varepsilon \) are idiosyncratic shocks. Policy functions for individual states \( \tilde{z} \) are components of \( \tilde{x} \). We define \( \Pi \) to be a selection matrix that returns \( \tilde{z} \) from \( \tilde{x} \), i.e., \( \tilde{z} = \Pi \tilde{x} \). The law of motion for the aggregate state variable is denoted by \( \tilde{\Omega}(\Omega, \mathcal{E}) \).

Consider the full set of \( t \geq 1 \) first-order optimality conditions for problem (21). These conditions can be split into two groups. The first group consists of optimality conditions for individual choices that connect current period individual and aggregate policy functions \( \tilde{a}, \tilde{X}, \) current period realizations of shocks \( \varepsilon, \mathcal{E}, \) and expectations of current and next period policy functions, \( \mathbb{E}[\tilde{a}(\tilde{\Omega}(\Omega, \mathcal{E}), \tilde{\Omega}(\Omega, \mathcal{E}), \tilde{a}(\tilde{\Omega}(\Omega, \mathcal{E}), \tilde{\Omega}(\Omega, \mathcal{E}), \varepsilon, \mathcal{E})] \). To economize on
notation, we denote these two mathematical expectations by $\mathbb{E}_{-}\tilde{x}$ and $\mathbb{E}_{+}\tilde{x}$, respectively. The first group of conditions can be written as

$$F\left(\mathbb{E}_{-}\tilde{x}, \tilde{x}, \mathbb{E}_{+}\tilde{x}, \bar{X}, \varepsilon, \mathcal{E}, z\right) = 0$$

(22)

for a collection of functions.\textsuperscript{10} The second group of optimality conditions for a continuation Ramsey problem are various aggregate feasibility constraints and first-order conditions with respect to $\bar{X}$ that connect aggregate functions and integrals of individual policy functions. These conditions can be written as

$$R\left(\int \tilde{x}d\Omega, \bar{X}, \mathcal{E}\right) = 0$$

(23)

for some mapping $R$. The law of motion for measure $\Omega$ is

$$\tilde{\Omega}(\Omega, \mathcal{E})(z) = \int \ell(\tilde{z}(y, \Omega, \varepsilon, \mathcal{E}) \leq z) d\text{Pr}(\varepsilon) d\Omega(y) \quad \forall z$$

(24)

where $\ell(\tilde{z} \leq z)$ is 1 if all elements of $\tilde{z}$ are less than or equal to all elements of $z$, and zero otherwise.

We use perturbation methods to approximate the dependence of continuation Ramsey policy functions on $\varepsilon, \mathcal{E}$ shocks around the cross-section distribution $\Omega_t$ of individual characteristics at each time $t$ along a simulated history. From these approximations, we can deduce how the aggregate shock $\mathcal{E}_{t+1}$ affects the time $t + 1$ measure $\Omega_{t+1}$.

To construct our small noise approximations of policy functions, we consider a family of economies parameterized by a positive scalar $\sigma$ that scales \textit{all} shocks $\varepsilon, \mathcal{E}$, so that policy functions are $\bar{X}(\Omega, \sigma \mathcal{E}; \sigma)$ and $\bar{x}(z, \Omega, \sigma \varepsilon, \sigma \mathcal{E}; \sigma)$. Let $\bar{X}(\Omega)$ and $\bar{x}(z, \Omega)$ denote these functions evaluated at $\sigma = 0$. We will often suppress dependence on $\Omega$ when it is clear from the context.\textsuperscript{11} We assume that policy functions are smooth enough to justify the derivatives that we compute and let $\bar{X}_\varepsilon, \bar{x}_\varepsilon(z), \bar{x}_\varepsilon(z)$ denote gradients of policy functions with respect to aggregate and idiosyncratic shocks, and $\bar{X}_\sigma$ and $\bar{x}_\sigma(z)$ denote their derivatives with respect to $\sigma$, all evaluated at $\sigma = 0$. Similarly, $\tilde{\Omega}_\mathcal{E}$ refers to the gradient of $\tilde{\Omega}(\Omega, \sigma \mathcal{E}; \sigma)$ with respect to aggregate shocks at $\sigma = 0$. First-order small noise expansions of policy functions are

$$\bar{X}(\Omega, \sigma \mathcal{E}; \sigma) = \bar{X} + \sigma (\bar{X}_\varepsilon \mathcal{E} + \bar{X}_\sigma) + \mathcal{O}(\sigma^2)$$

(25)

\textsuperscript{10}Strictly speaking, if $\tilde{x}$ consists of all lower-case choice variables and multipliers in problem (21) then the relevant objects are $\mathbb{E}_{-}f(\tilde{x})$ and $\mathbb{E}_{+}g(\tilde{x})$ for some transformations $f$ and $g$. Our exposition is without loss of generality after we extend the definition of $\tilde{x}$ to include variables $f(\tilde{x})$ and $g(\tilde{x})$, for example, by including variable $\tilde{u}_c$ in vector $\tilde{x}$ and its definition $\tilde{u}_c = u_c(\tilde{c}, \tilde{n})$ in mapping $F$.

\textsuperscript{11}For instance, $\bar{X}$ would refer to the function $\bar{X}(\Omega) \equiv \bar{X}(\Omega, 0, 0)$, $\bar{x}(z)$ would refer to the function $\bar{x}(z, \Omega) \equiv \bar{x}(z, \Omega, 0, 0, 0)$, and so on.
and
\[ \tilde{x}(z, \Omega, \sigma \varepsilon, \sigma \mathcal{E}; \sigma) = \tilde{x}(z) + \sigma (\tilde{x}_\varepsilon(z) \varepsilon + \tilde{x}_\mathcal{E}(z) \mathcal{E} + \tilde{x}_\sigma(z)) + \mathcal{O}(\sigma^2). \]  

(26)

3.1.1 Zeroth-order expansions

Because higher-order approximations of policy functions use inputs from lower-order approximations, we start with zeroth-order approximations. To do this, we must study the economy without shocks.

**Lemma 1.** For any \( \Omega \), policy functions satisfy \( \tilde{z}(z, \Omega) = z \) for any \( z \) and therefore \( \tilde{\Omega}(\Omega) = \Omega \).

**Proof.** The first-order condition with respect to \( b_{i,t-1} \) in (21) is
\[ \mathbb{E} \left[ \frac{\tilde{u}_c(z, \Omega, \cdot, \cdot)}{1 + \tilde{\Pi}(\Omega, \cdot, \cdot)} (\mu - \bar{\mu}(z, \Omega, \cdot, \cdot)) \right] = 0, \]
which implies \( \bar{\mu}(z, \Omega) = \mu \) for all \( z, \Omega \). Equation (20) to the zeroth-order is
\[ \bar{Q}(\Omega) \bar{M}(\Omega) m = \beta \bar{m}(z, \Omega) \bar{M}(\bar{\Omega}(\Omega)) (1 + \bar{\Pi}(\bar{\Omega}(\Omega)))^{-1}. \]
Since Pareto-Negishi weights \( m(z, \Omega) \) integrate to one, this equation implies
\[ \bar{Q}(\Omega) \bar{M}(\Omega) = \beta \bar{M}(\bar{\Omega}(\Omega)) (1 + \bar{\Pi}(\bar{\Omega}(\Omega)))^{-1}, \]
and therefore that \( \bar{m}(z, \Omega) = m \) for all \( z, \Omega \).

That the cross-sectional distribution of characteristics \( \Omega \) stays constant reflects the fact that in the \( \sigma = 0 \) economy the planner would want to keep agents' Pareto-Negishi weights and associated multipliers on agents' budget constraints constant over time. This makes sense because the \( \sigma = 0 \) economy is deterministic and in effect has complete markets.

Lemma 1 implies that \( \tilde{x}(\tilde{z}(z, \Omega), \tilde{\Omega}(\Omega)) = \tilde{x}(z, \Omega) \). Therefore, we can compute \( \bar{X} \) and \( \tilde{x}(z) \) by solving a system of non-linear equations
\[ \bar{F}(z) \equiv F(\bar{x}(z), \bar{x}(z), \bar{x}(z), \bar{X}, 0, 0, z) = 0, \bar{R} \equiv \bar{F} \left( \int \bar{x}(z) d\Omega(z), \bar{X}, 0 \right) = 0. \]  

(27)

From zeroth-order terms \( \bar{X} \) and \( \tilde{x}(z) \), we construct several objects to be used in computing higher-order terms. Let \( \bar{R}_x \) be the derivative of the mapping \( R \) with respect to its first argument, that is, the value of \( \int \bar{x} d\Omega \), \( \bar{R}_X \) and \( \bar{R}_\mathcal{E} \) be the derivatives of \( R \) with respect to its second and third arguments, respectively, all evaluated at \( \sigma = 0 \). Similarly, let subscripts \( x-, x, x+, X, \varepsilon, \mathcal{E} \) and \( z \) of \( \bar{F} \) denote the corresponding derivatives of \( F \) with respect
to each of its arguments evaluated at $\sigma = 0$. From the implicit function theorem we have

$$\bar{x}_z(z) = \left[ F_{x-}(z) + \bar{F}_z(z) + \bar{F}_{x+}(z) \right]^{-1} \bar{F}_z(z).$$

All of these objects can be constructed from $\bar{X}, \bar{x}(z)$.

Finally, we use $\partial \bar{x}(z, \Omega), \partial \bar{X}(\Omega)$ to denote Frechet derivatives of $\bar{x}(z, \Omega)$ and $\bar{X}(\Omega)$ with respect to the measure $\Omega$. Frechet derivatives generalize the notion of gradients to infinite-dimensional variables. They capture how changes in the aggregate distribution $\Omega$ affect policy functions. In principle, these Frechet derivatives could be calculated from (27), but unfortunately that approach is impractical except for very simple cases because the number of unknowns in the operators $\partial \bar{x}(\cdot, \Omega)$ and $\partial \bar{X}(\Omega)$ grows exponentially with the size of $\Omega$. An important part of our contributions is a way to overcome this problem by imposing what we will call the independence property.

**Corollary 1.** *(The independence property)* $\partial \bar{z}(z, \Omega) = 0$ for all $z, \Omega$.

Corollary 1 asserts that at $\sigma = 0$ the Frechet derivative of policy functions for individual states equals zero. The benefit of this property is that it provides tractability in calculating $\partial \bar{\Omega}$, which is a key intermediate term for the constructing our expansions. In the case studied in this section, corollary 1 and lemma 1 imply $\partial \bar{\Omega} = I$, but more generally, we show that as long as the independence property is satisfied, $\partial \bar{\Omega}$ can be expressed in terms of $\bar{z}_z(z)$ which are easy to compute.

### 3.1.2 First-order expansions

We can now use the first-order Taylor expansion of equations (22)-(24). As a preliminary step, observe that expansions of $E \pm \bar{x}$ and $E \pm \bar{x}$ are, using lemma 1,

$$E_+ \bar{x} = \bar{x}(z) + \left[ \bar{x}_z(z) \bar{\bar{p}} \bar{E}(z) + \partial \bar{x}(z) \cdot \bar{\Omega} \bar{E} \right] \sigma \bar{E} + \left[ \bar{x}_\sigma(z) \bar{p} \bar{x}_z(z) \right] \sigma \bar{E} + \bar{x}_\sigma(z) \sigma + O(\sigma^2),$$

$$E_- \bar{x} = \bar{x}(z) + \bar{x}_\sigma(z) \sigma + O(\sigma^2).$$

---

12 A Frechet derivative of some variable $\bar{X}(\Omega)$ is a linear operator from the space of distributions $\Omega$ to $\mathbb{R}$ with a property that $\lim_{\Delta \to 0} \frac{\|\bar{X}(\Omega + \Delta) - \bar{X}(\Omega) - \partial \bar{X}(\Omega) \Delta\|}{\|\Delta\|} = 0$. It can be found by fixing a feasible direction $\Delta$ and calculating a directional (Gateaux) derivative, since, when both derivatives exist, they coincide, $\partial \bar{X}(\Omega) \cdot \Delta = \lim_{\Delta \to 0} \frac{\bar{X}(\Omega + \Delta) - \bar{X}(\Omega)}{\|\Delta\|}$. Following Luenberger (1997), we refer to $\partial \bar{X}(\Omega) \cdot \Delta$ as a Frechet derivative of $\bar{X}$ at a point $\Omega$ with increment $\Delta$. Roughly speaking, $\partial \bar{X}(\Omega)$ is a measure and $\partial \bar{X}(\Omega) \cdot \Delta$ is the value of the integral of function $\Delta$ with respect to $\partial \bar{X}(\Omega)$. 

---

17
This implies that the Ramsey planner’s optimality conditions equations (22) and (23) satisfy, up to $O(\sigma^2)$,

$$
\bar{F}(z) + [(\bar{F}_x(z) + \bar{F}_x^+(z) \bar{x}_x(z)p) \bar{x}_\varepsilon(z) + \bar{F}_x^+(z) \bar{c}_\Delta(z) \cdot \bar{\Omega}_\varepsilon + \bar{F}_X(z) \bar{X}_\varepsilon + \bar{F}_\varepsilon(z)] \sigma \varepsilon
$$

and

$$
\bar{R} + \left[ \bar{R}_x \int \bar{x}_\varepsilon(z) d\Omega + \bar{R}_X \bar{X}_\varepsilon + \bar{R}_\varepsilon \right] \sigma \varepsilon + \left[ \bar{R}_x \int \bar{x}_\sigma(z) d\Omega + \bar{R}_X \bar{X}_\sigma \right] \sigma = 0. \quad (28)
$$

The system of equations (28) and (29) must hold for all $\varepsilon$, $\varepsilon$ and $\sigma$ and characterizes \{\bar{x}_\varepsilon(z), \bar{x}_\sigma(z), \bar{X}_\varepsilon, \bar{X}_\sigma, \bar{X}_\varepsilon(z), \bar{X}_\varepsilon\}. Let’s consider each of these objects in turn. From (28), we immediately get

$$
\bar{x}_\varepsilon(z) = - (\bar{F}_x(z) + \bar{F}_x^+(z) \bar{x}_x(z)p)^{-1} \bar{F}_\varepsilon(z). \quad (30a)
$$

All the terms on the right-hand side are known from the zeroth-order expansion, so we can compute $\bar{x}_\varepsilon(z)$ via matrix inversion. This step is easily parallelizable because the computation is to be done independently for all $z$. Terms $\bar{x}_\sigma(z)$ and $\bar{X}_\sigma$ can be computed in a similar way but it is straightforward to verify that they are equal to zero.

Calculating $\bar{x}_\varepsilon(z)$ and $\bar{X}_\varepsilon$ is more difficult. The aggregate shock $\varepsilon$ changes the next period state by $\bar{\Omega}_\varepsilon$ and that alters expectations of next period policies by $\partial \bar{x}(z) \cdot \bar{\Omega}_\varepsilon$, as can be seen from the first square bracket in (28). Neither $\partial \bar{x}(z)$ nor $\bar{\Omega}_\varepsilon$ are known at this stage. The next theorem and its proof show how to use functional derivative techniques to construct $\partial \bar{x}(z) \cdot \bar{\Omega}_\varepsilon$.

Theorem 1. *From the zeroth-order expansion we can construct matrices $A(z)$ and $C(z)$ that satisfy*

$$
\partial \bar{x}(z) = C(z) \partial \bar{X}, \quad (30a)
$$

$$
\bar{x}(z) \cdot \bar{\Omega}_\varepsilon = C(z) \partial \bar{X} \cdot \bar{\Omega}_\varepsilon = C(z) \int A(y) \bar{x}_\varepsilon(y) d\Omega(y). \quad (30b)
$$

Proof. Lemma 1 implies $\partial \bar{\Omega} = 1$. The Frechet derivatives of (22) and (23) with arbitrary increment $\Delta$ satisfy

$$
(\bar{F}_x(z) + \bar{F}_x^+(z) + \bar{F}_x^+(z)p) \partial \bar{x}(z) \cdot \Delta + \bar{F}_X(z) \partial \bar{X} \cdot \Delta = 0, \quad (31a)
$$

$$
\bar{R}_x \partial \left( \int \bar{x}(y) d\Omega(y) \right) \cdot \Delta + \bar{R}_X \partial \bar{X} \cdot \Delta = 0. \quad (31b)
$$
The first equation yields (30a) with $C(z) = - \left( \bar{F}_x(z) + \bar{F}_x(z) + \bar{F}_x(z) + \bar{F}_x(z) \bar{z}(z)p \right)^{-1} \bar{F}_X(z)$.

Since directional and Frechet derivatives coincide, by fixing any direction $\Delta$ and computing the directional derivative (see footnote 12) we obtain

$$\partial \left( \int \bar{x}(y) d\Omega(y) \right) \cdot \Delta = \int (\partial \bar{x}(y) \cdot \Delta) d\Omega(y) + \int \bar{x}(y) d\Delta(y). \quad (32)$$

We want to evaluate the integral on the right side at $\Delta = \bar{\Omega}_E$. Differentiating (24) at any $z = (m, \mu)$ and applying lemma 1 we get

$$\bar{\Omega}_E(m, \mu) = - \int_{y_2 \leq \mu} \bar{m}_E(m, y_2) \omega(m, y_2) dy_2 - \int_{y_1 \leq m} \bar{\mu}_E(y_1, \mu) \omega(y_1, \mu) dy_1,$$

where $\omega$ is the density of $\Omega$. The density of $\bar{\Omega}_E(m, \mu)$, which we denote with $\bar{\omega}_E(m, \mu)$, is then

$$\bar{\omega}_E(m, \mu) = - \frac{d}{dm} [\bar{m}_E(m, \mu) \omega(m, \mu)] - \frac{d}{d\mu} [\bar{\mu}_E(m, \mu) \omega(m, \mu)].$$

Substitute this equation and (30a) into (32) to get

$$\partial \left( \int \bar{x}(y) d\Omega(y) \right) \cdot \bar{\Omega}_E = \int C(y) \partial \bar{X} \cdot \bar{\Omega}_E d\Omega(y) - \int \bar{x}(y) \frac{d}{dm} [\bar{m}_E(y) \omega(y)] dy - \int \bar{x}(y) \frac{d}{d\mu} [\bar{\mu}_E(y) \omega(y)] dy = (\partial \bar{X} \cdot \bar{\Omega}_E) \int C(y) d\Omega(y) + \int \bar{x}(y) p \bar{x}_E(y) d\Omega(y),$$

where the second equality is obtained via integration by parts. Substitute this expression into (31b) and solve for $\partial \bar{X} \cdot \bar{\Omega}_E$ to obtain

$$\bar{X}_E \equiv \partial \bar{X} \cdot \bar{\Omega}_E = \int A(y) \bar{x}_E(y) d\Omega(y), \quad (33)$$

where $A(z) = - \left( \bar{R}_x \int C(y) d\Omega(y) + \bar{R}_X \right)^{-1} \bar{R}_x \bar{x}_E(z)$. Together with (30a) we get (30b).

The second equality is obtained via integration by parts. Substitute this expression into (31b) and solve for $\partial \bar{X} \cdot \bar{\Omega}_E$ to obtain

$$\bar{X}_E \equiv \partial \bar{X} \cdot \bar{\Omega}_E = \int A(y) \bar{x}_E(y) d\Omega(y), \quad (33)$$

where $A(z) = - \left( \bar{R}_x \int C(y) d\Omega(y) + \bar{R}_X \right)^{-1} \bar{R}_x \bar{x}_E(z)p$. Together with (30a) we get (30b).

Economic forces underlie the proof of theorem 1. In a competitive equilibrium, agents care about the distribution $\Omega$ only because it helps them predict aggregate prices and income. That means that effects from a perturbation of distribution $\Omega$ on individual variables, $\partial \bar{x}$ can be factored into its effects on aggregate variables, $\partial \bar{X}$, and a known loading matrix $C(z)$ that captures how individual variables respond to changes in the aggregates. Equation (30a) captures this.

Feasibility and market clearing impose a tight relationship between how individual policy
functions respond to aggregate shocks in the current period, $\bar{x}_E(z)$, and how aggregates can be expected to change next period, $\bar{X}_E'$. This relationship sets up a fixed point problem that we represent in equation (33). Together with (30a), this equation allows us to express the Frechet derivative $\partial \bar{x} \cdot \bar{\Omega}_E$ as a linear function of $\bar{x}_E(z)$.

These calculations put us in a position to compute the coefficients $\bar{x}_E$ and $\bar{X}_E'$. Setting the first square brackets in (28) and (29) to zero and using the definition of $\bar{X}_E'$ from (33) we obtain the following system of linear equations in the unknowns $\bar{X}_E$, $\bar{x}_E(z)$ for all $z$.

$$\begin{align*}
(F_x(z) + F_x(z)\bar{x}_E(z))\bar{x}_E(z) + F_x(z)C(z)\bar{X}_E' + F_X(z)\bar{X}_E + \bar{F}_E(z) &= 0, \quad (34a) \\
\bar{R}_x \int \bar{x}_E(y)d\Omega(y) + \bar{R}_X\bar{X}_E + \bar{R}_E &= 0. \quad (34b)
\end{align*}$$

This linear system has a computationally convenient structure because it allows us to split one large problem of simultaneously finding $\bar{x}_E(z)$ for all $z$ into a large number of small problems that independently characterize $\bar{x}_E(z)$ for each $z$. Thus, we use equation (34a) to calculate matrices $D_0(z)$ and $D_1(z)$ that define the affine function

$$\bar{x}_E(z) = D_0(z) + D_1(z) \cdot \begin{bmatrix} \bar{X}_E & \bar{X}_E' \end{bmatrix}^T.$$

Substitute this function into equations (33) and (34b) to compute $\bar{X}_E$ and $\bar{X}_E'$. Values of $\bar{x}_E(z)$ can be found either by substituting back into the previous equation or from (30a). This completes the calculations necessary for the first-order terms.

3.1.3 Second- and higher-order expansions

Our approach extends to second- and higher-order expansions while preserving computationally convenient linear and parallelizable structure. The key observation is that theorem 1 generalizes to higher-order expansions. The independence property, $\partial \bar{z}(z, \Omega) = 0$, allows a counterpart to equation (30a) to hold for any order of perturbation of $\Omega$. This allows us to solve for higher order analogs of $\partial \bar{x} \cdot \bar{\Omega}_E$ explicitly as weighted sums of higher-order coefficients $\bar{x}_{EE}, \bar{x}_{E\sigma}, \bar{x}_{\sigma\sigma} \ldots$, with weights known from lower-order expansions. We then can form higher-order analogs of the system of equations (34). As before, the mathematical structure of these equations allows us to split one large system of equations into a large number of low-dimensional linear problems that can be solved fast and simultaneously. Formal proofs and constructions are notation intensive, but the steps mirror those in section 3.1.2. We provide details in the online appendix.
3.2 Approximations in the general case

We now apply our small noise approximation method to the economy described in section 2. Our recursive formulation of the problem for the section 2 model adds two features to computations described in section 3.1. First, the optimality condition (17) generally binds and cannot be omitted. We add this constraint to our Lagrangian formulation (21), so its multiplier, \( \Lambda \), now becomes an additional aggregate state variable. Second, since shocks are persistent, policy functions also depend on previous period values of aggregate shocks \( \Theta = (\Theta, \Phi) \) as well as idiosyncratic shocks \( \theta \). Thus, in the general case, \( z = (m, \mu, s, \theta, \bar{\theta}) \) is the individual state, \( \Omega \) is a measure over \( z \), and the aggregate and individual policy functions are functions \( \bar{X} (\Omega, \Lambda, \Theta, \varepsilon) \) and \( \bar{x} (z, \Omega, \Lambda, \Theta, \varepsilon, \varepsilon) \), respectively. The zeroth-order terms have non-trivial (deterministic) transition paths which can be computed using a shooting algorithm.

With persistent shocks, there are two ways to perturb policy functions that result in approximation errors of the same order of magnitude. One is to scale \( \{\sigma E, \sigma \varepsilon\} \), and expand with respect to \( \sigma \) around current values of \( (\Theta, \theta) \), and \( \Omega \). Since to the zeroth-order \( \bar{\theta} (z, \Omega) \neq \theta \), it is no longer the case that \( \bar{\Omega} (\Omega) = \Omega \), and so lemma 1 does not hold.\(^{13}\) However, functional derivative techniques used in the proof of theorem 1 still apply and we can construct the relevant Frechet derivatives along the transition path. Tractability is preserved because policy functions still satisfy the independence property, i.e., \( \partial \bar{X} (z, \Omega) = 0 \) for all \( z, \Omega \). The law of motion for exogenous variables does not depend on distribution \( \Omega \) and thus adding those variables to vector \( z \) leaves the independence property unaffected.

An alternative approach is to scale \( \{\sigma E, \sigma \varepsilon, \sigma \Theta, \sigma \theta\} \) and then expand around \( \sigma = 0 \). Since \( \theta \) is a part of \( z \), this means that \( z \) and therefore \( \Omega \) are now also functions of the scaling parameter \( \sigma \). The zeroth-order approximation satisfies lemma 1, but now expansions of policy functions require computing additional Frechet derivatives such as \( \partial \bar{X} \cdot \Omega \sigma \) and \( \partial \bar{x} (z) \cdot \Omega \sigma \). These derivatives are easy to compute using the same techniques as in the proof of theorem 1.

Although the two approaches imply errors of the same order of approximation, one can be better than the other depending on circumstances. For example, in some cases, the second approach may not require computing a transition path and therefore can be faster to implement. In the online appendix, we provide explicit formulas and extensions of theorem 1 for both approaches.

\(^{13}\)When persistence of the idiosyncratic shocks \( \rho_\theta \) is close to one, we can recover lemma 1 if we approximate \( \rho_\theta \) by \( \rho_\theta (\sigma) = 1 - \sigma \rho \) for some \( \rho \geq 0 \) and expand \( \rho_\theta (\sigma) \) with respect to \( \sigma \).
3.3 Accuracy and comparison to other methods

Our approach builds on the perturbation techniques widely used in computational economics (see, for example, Judd and Guu (1993), Judd and Guu (1997), Schmitt-Grohe and Uribe (2004b)). Our particular implementation of perturbations “around the current state” is closely related to earlier work by Fleming (1971), Fleming and Souganidis (1986), Anderson et al. (2012), Bhandari et al. (2017), and Phillips (2017). In all of those applications, the state space is simple, and approximations do not require computing high-dimensional Frechet derivatives. In contrast, our approach applies even when the underlying state is a complicated, high-dimensional object, as is frequently encountered in HA economies. To the best of our knowledge, ours is the first method that fully incorporates effects of the complete current state on policy rules and thereby can approximate equilibrium dynamics of such economies well.\footnote{Methods used here were first explored in the Ph.D. thesis Evans (2015).}

Our approach can be applied to a large class of HA economies for which equilibrium dynamics can be written in the form given by equations (22)-(24).\footnote{Note that HA economies with inequality constraints, such as ones with additional ad-hoc debt limits, can also be written in this form by including appropriate complementary slackness conditions. Inequality constraints often imply that policy functions have kinks. These kinks violate the smoothness assumption that we imposed on equations (25) and (26). However, we foresee no impediments to extending our method to such cases. We hope to explore such possibilities in future work.}

To verify the accuracy of our method, we study a simpler problem of computing a competitive equilibrium for a given monetary-fiscal policy. We proceed by simplifying our general environment in a way that allows us to compute an equilibrium analytically and exactly without using approximations. We then compare that analytical solution to our approximation by varying key parameters that would a priori affect approximation errors.

Specifically, we follow Acharya and Dogra (2018), and assume that labor is supplied inelastically at $n_{i,t} = 1$; preferences are given by $U(c_t,n_t) = -\exp(-\gamma c_t)$; equity holdings are uniform across consumers; there are no aggregate shocks; idiosyncratic shocks $\epsilon_{i,t}$ are i.i.d. normally distributed; government spending and all tax rates equal zero; and interest rates are set according to a Taylor rule $Q_t^{-1} - 1 = a_0 (1 + \Pi_t)^{a_1}$ for coefficients $a_0$ and $a_1$ chosen so that steady state inflation is zero.

Under these assumptions, household income $W_t\epsilon_{i,t} + T_t + D_t$ is normally distributed. This property, together with the CARA assumption on the utility function, means that the consumption-saving problem has an analytically tractable solution. One can then derive explicit expressions for both the steady-state aggregate quantities and deterministic transition paths from given initial conditions.

Acharya and Dogra (2018) call this a Pseudo Representative Agent New Keynesian economy (abbreviated as PRANK) and used it to illustrate many insights that emerge in more complicated HANK models. They showed that their PRANK economy has a unique steady
state in which all the aggregate variables, such as output, inflation, and real interest rates, are constant. In a steady state, individual assets follow a random walk, so the dispersion of asset holdings across agents grows without limit. Since explicit expressions are available for policy functions along the transition path, there are explicit expressions for how this PRANK economy is affected by one-time, fully unanticipated aggregate shock.

We list all the equilibrium conditions and calibrated values for the parameters in the online appendix. We start at the steady state, and study equilibrium responses to a one-time, unanticipated, 1.23% shock to aggregate productivity in period $t$ which then decays deterministically.\textsuperscript{16} We compare our second-order approximation to the exact solution. We report two versions of this comparison: one in which the shock occurs in period $t = 1$ and another one in which it occurs in $t = 250$. In both versions, shocks arrive when all aggregate variables are at the same steady state values; the two cases differ only in the degree of asset inequality at the time of the shock.

Blue and black solid lines in figure I show the exact and approximate impulse responses of output, inflation, and asset inequality measured by the standard deviation of individual wealth. They are almost identical in both experiments. The PRANK economy is engineered in such a way that impulse responses of output and inflation are independent of the asset distribution, so dynamics of output and inflation are the same in the top and bottom rows of figure I. This is not the case for other potential variables of interest, such as the dynamics of asset inequality, as can be seen in the rightmost panels in figure I.

By comparing the two experiments, we can evaluate the precision of our approximation and how the errors deteriorate with the time horizon. In the PRANK economy, we can calculate distribution $\Omega_t$ exactly for all $t$ and the corresponding impulse responses. Our approach computes sequential approximations of policy functions and distributions for $t = 1, 2, ...$. Thus, computational errors potentially accumulate over time as we compute responses far into the future. Because individual wealth follows a random walk process, making any approximation errors accrued by our method permanent, this environment might be a worst case for testing our approximation method. Despite that, we find that our approximate distribution is very close to the exact distribution at $t = 250$ (see online appendix), and we capture responses of asset inequality to aggregate shock in that period very well.

To document the accuracy and computational speed of our approach we compute approximation errors of our policy functions relative to the true solution and report the results in Table I.\textsuperscript{17} All the errors are reported for a quadratic approximation. The percent errors relative to the true solution for the individual consumption policies are less than 0.05% and vary between 0.004%-0.033% as we double volatility of idiosyncratic shocks, while the per-

\textsuperscript{16} This corresponds to a one standard deviation shock to productivity in our baseline economy.

\textsuperscript{17} See the online appendix for details of how the approximation errors are computed. Following Den Haan (2010), we also provide Euler Equation and Dynamic Euler Equation errors.
Figure I: Comparisons of impulse responses to a 1% TFP shock at $t = 1$ in the top panel and $t = 250$ in the bottom panel across approximation methods. The bold lines are are the exact solution (black) and our method applied to second-order (blue). The dashed black line are responses under the Reiter method.

Percentage errors for aggregate output, inflation and the interest rate range from $4.3 \times 10^{-5}\%$ to $0.0007\%$. In terms of percentage errors, increasing the risk aversion to 3 is roughly on par with doubling the volatility of the idiosyncratic shocks. Since our approach is easily parallelizable and sidesteps the computationally intensive step of computing Frechet derivatives, it works fast and allows us to simulate transition dynamics of a path of 100 periods in 1.5 seconds on a dual AMD EPYC 7351 processor with 32 cores.

### 3.4 Comparison to the Reiter method

We can also compare our method to the widely-used approach of Reiter (2009). Our method with order $n$ yields approximation errors that scale with the size of both the idiosyncratic and aggregate shocks, that is, $O \left( \sigma_{agg}^{n+1}, \sigma_{idiosync}^{n+1} \right)$. A Reiter-type approximation delivers policy functions that are linear in aggregate shocks and globally accurate with respect to idiosyncratic shocks around a fixed distribution $\Omega_t = \bar{\Omega}$. Therefore errors in such an approximation scale with both the size of the aggregate shock and the distance of the current distribution from the point of approximation $O \left( \sigma_{agg}^2, \| \Omega - \bar{\Omega} \|^2 \right)$. We use the PRANK setting to show the trade-offs involved in these two types of errors. All our comparisons will be for $n = 2$. 
<table>
<thead>
<tr>
<th>Maximum Errors (%)</th>
<th>Ind. Consumption</th>
<th>Agg. Output</th>
<th>Inflation</th>
<th>Interest Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2\textsuperscript{nd} Order</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\gamma = 1, \sigma_\epsilon = 0.50$</td>
<td>0.0039</td>
<td>4.2e-6</td>
<td>3.1e-5</td>
<td>4.3e-5</td>
</tr>
<tr>
<td>$\gamma = 1, \sigma_\epsilon = 0.75$</td>
<td>0.0134</td>
<td>2.6e-5</td>
<td>1.5e-4</td>
<td>2.2e-4</td>
</tr>
<tr>
<td>$\gamma = 1, \sigma_\epsilon = 1.00$</td>
<td>0.0328</td>
<td>8.2e-5</td>
<td>4.9e-4</td>
<td>6.9e-4</td>
</tr>
<tr>
<td>$\gamma = 3, \sigma_\epsilon = 0.5$</td>
<td>0.0453</td>
<td>0.0011</td>
<td>0.0024</td>
<td>0.0034</td>
</tr>
<tr>
<td>Reiter-based</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\gamma = 1, \sigma_\epsilon = 0.50$</td>
<td>0.0374</td>
<td>0.0616</td>
<td>0.0337</td>
<td>0.0505</td>
</tr>
<tr>
<td>$\gamma = 1, \sigma_\epsilon = 0.75$</td>
<td>0.0466</td>
<td>0.0610</td>
<td>0.0335</td>
<td>0.0501</td>
</tr>
<tr>
<td>$\gamma = 1, \sigma_\epsilon = 1.00$</td>
<td>0.0492</td>
<td>0.0602</td>
<td>0.0329</td>
<td>0.0493</td>
</tr>
<tr>
<td>$\gamma = 3, \sigma_\epsilon = 0.5$</td>
<td>0.0896</td>
<td>0.2252</td>
<td>0.1327</td>
<td>0.1991</td>
</tr>
</tbody>
</table>

TABLE I: Percentage errors in policy functions in response to an one standard deviation unanticipated shock to aggregate TFP. The values reported are the maximum errors across states $(b, \epsilon)$ and time $t$ relative to the true solution.

The bottom four rows of Table I present the percentage errors of the Reiter-based method\textsuperscript{18} relative to the true solution allowing us to compare our method directly to Reiter’s approach. Both approaches yield accurate approximations of the true solution with small percentage errors on the order of 0.1%. Despite that, we observe that our method yields errors for the individual consumption policy rules are consistently smaller than those of the Reiter-based approach while the errors for the aggregate variables are two orders of magnitude smaller than the Reiter-based approach. Our approach is less accurate with respect to the idiosyncratic risk, but those errors partially wash out in the aggregate. Meanwhile, second-order errors in aggregates variables under the Reiter-based approach propagate down to the individual policy rules.

To observe the long run consequences of these approximation errors, we augment the PRANK economy with a non-degenerate stochastic process for aggregate TFP. In line with standard calibrations, aggregate shocks are much less volatile than idiosyncratic shocks (standard deviation of innovations about 1% for aggregate TFP vs. 50% for individual productivities). We then simulate a long sequence of aggregate shocks and compare the distribution of assets at $t = 250$ using both our method and Reiter’s. We find that the obtained asset distributions are visibly different under the two methods, with the standard deviation of assets being more than 1 percent larger under Reiter’s approach. To understand the source of these differences, we drop the second-order terms with respect to aggregate risk

\textsuperscript{18}Conventional applications of Reiter’s method requires expanding policy functions around the invariant distribution in the economy without aggregate shocks. Since individual assets follow a random walk, such distribution does not exist in PRANK economy. In our application of Reiter’s method, we used initial distribution $\Omega_0$ as the point of expansion.
from our expansions and re-compute the long run distribution using this inferior approximation. This distribution (which has error of the order $O\left(\sigma_{agg}^{2}, \sigma_{agg}^{3}\right)$) is almost identical to the distribution generated by the Reiter method, which allows us to conclude that ignoring higher-order effects of aggregate shocks can inject long-run drifts into approximation errors. We report details about this and other related experiments in the online appendix.

The consequences of ignoring movements of $\Omega_t$ away from $\bar{\Omega}$ can be seen in the responses of the standard deviation of assets in figure I. While the PRANK economy is constructed so that responses of output and inflation do not depend on the asset distribution at all, the responses of other moments, such as asset inequality, do depend on it. This implies that Reiter’s approximation of these impulse responses becomes progressively poorer when the distribution of assets drifts away from the point of approximation. In this example, the movement of the distribution away from the point of approximation is due to idiosyncratic income risk but similar issues would emerge in economies where aggregate shocks follow non-degenerate stochastic processes for which nonlinear impulse responses would generally depend on the past history of aggregate shocks or on the current state.

4 Calibration

Our model builds on two literatures – New Keynesian monetary models and Bewley-Huggert-Aiyagari heterogeneous agent models. To focus on the key trade-offs that our Ramsey planner faces, we keep the baseline economy close to commonly used specifications. Our initial calibration ignores some features such as ample heterogeneity in marginal propensities to consume and heterogeneous labor earning responses in recessions. We incorporate them as extensions in section 6.

Preferences and technology parameters

We assume that preferences are isoelastic $u(c, l) = \frac{c^{1-\nu}}{1-\nu} - \frac{n^{1+\gamma}}{1+\gamma}$, and set $\nu = 3$, $\gamma = 2$. We calibrate to annual data and set discount factor $\beta = 0.96$. We set $\bar{\Theta} = 1$ and $\bar{\Phi} = 6$ to attain average markups of 20%. We abstract from the use of intermediates in production and set $\alpha = 1$. We choose the cost of nominal price changes $\psi$ to match the slope of the aggregate Phillips curve. Sbordone (2002) estimates the slope of the U.S. Phillips curve using quarterly data to be about 0.05. To convert that to an annual frequency we multiply her number by 4. To a first order approximation the slope of the Phillips curve in our model is $(\bar{\Phi} - 1) / \psi$, which implies $\psi = 20$. 

26
Idiosyncratic and aggregate uncertainty

We assume that all shocks are Gaussian and set the standard deviations of $\varepsilon_{t,i}$ and $\theta_{t,i}$ to 8.7% and 10.3%, and the autocorrelation $\rho_{\theta} = 0.992$ to match evidence on individual wage dynamics from Low et al. (2010).

The stochastic process for the markup shocks is calibrated so that movements in the labor share of output are consistent with movements in the U.S. corporate sector’s labor share (Table 1.14, NIPA) over the period 1947-2016. Calibrated values for $(\rho_{\Phi}, \sigma_{\Phi})$ are $(0.85, 4.6%)$.\textsuperscript{19}\textsuperscript{19} The stochastic process for aggregate labor productivity, $\log \Theta_t$, is calibrated so that output per hour is consistent with detrended U.S. non-farm real output per hour (BLS) over the period 1947-2016. Calibrated values for $(\rho_{\Theta}, \sigma_{\Theta})$ are $(0.73, 1.23%)$.

Initial conditions

A common approach in the heterogeneous agent macro-labor literature is to specify government policy $(\bar{G}, \Upsilon_t, Q_t)$ and study long-run allocations in an associated competitive equilibrium. A well-known property of several models in this class is that the invariant distribution underpredicts the amount of wealth inequality. As we shall see, asset inequality has important implications for optimal policy responses. We calibrate initial conditions $\{\theta_{i,-1}, b_{i,-1}, s_i\}_i$ to be consistent with empirical distributions of wages, nominal claims, and claims to real firm profits. In section 5.1.1, we show that the drift from this distribution is slow and its presence matters very little for the optimal policy responses.\textsuperscript{20}\textsuperscript{20}

We use the 2007 wave of the Survey of Consumer Finances (SCF) as our benchmark for earnings and asset inequality. We follow the procedure proposed by Doepke and Schneider (2006) to map household-level direct and indirect holdings of financial assets to the joint distribution of claims to nominal debt and claims to equity.\textsuperscript{21}\textsuperscript{21} Table II reports summary statistics for our sample. Of particular relevance to results presented in section 5 is that earnings and assets are positively correlated and that assets inequality is much larger than earnings inequality.

The final parameter is the level of government expenditures $\bar{G}$, which we calibrate to be consistent with the ratio of non-transfer government expenditures to tax revenues. To

\textsuperscript{19}There is large range of choices for how markup shocks are modeled and calibrated. In the DSGE literature, for instance, Smet and Wouters (2007), Justiniano et al. (2010), Gali et al. (2007) use ARMA(1,1) processes and estimate the quarterly persistence to be in the range of 0.90–0.95. In the finance literature, for instance, Greenwald et al. (2014) estimate factor share shocks with a monthly persistence of 0.995. Our calibrated value for $\rho_{\Phi} = 0.85$ lies within the range of these estimates.

\textsuperscript{20}Incorporating proposed “fixes” in the literature to obtain a sufficiently skewed invariant distribution of wealth, for instance allowing persistent shocks to discount factors (Krusell and Smith (1998)), bequests (De Nardi (2004)), entrepreneurial choice (Cagetti and De Nardi (2006)), or persistent idiosyncratic differences in returns to financial assets (Benhabib et al. (2019)), and then computing a Ramsey allocation for such an economy is interesting but beyond the scope of this paper.

\textsuperscript{21}The online appendix contains details about how we apply the Doepke and Schneider (2006) procedure.
TABLE II: FIT OF THE INITIAL DISTRIBUTION

<table>
<thead>
<tr>
<th>Moments</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraction of pop. with zero equities</td>
<td>30%</td>
</tr>
<tr>
<td>Std. share of equities</td>
<td>2.63</td>
</tr>
<tr>
<td>Std. bond</td>
<td>6.03</td>
</tr>
<tr>
<td>Gini of financial wealth</td>
<td>0.82</td>
</tr>
<tr>
<td>Std. ln wages</td>
<td>0.80</td>
</tr>
<tr>
<td>Corr(share of equities, ln wages)</td>
<td>0.40</td>
</tr>
<tr>
<td>Corr(share of equities, bond holdings)</td>
<td>0.62</td>
</tr>
<tr>
<td>Corr(bond, ln wages)</td>
<td>0.33</td>
</tr>
</tbody>
</table>

Notes: The moments correspond to SCF 2007 wave with sample restrictions explained in the text and after scaling wages, equity holdings, and debt holdings by the average yearly wage in our sample. The share of equities refers to the ratio of individual equity holdings to the total in our sample such that the weighted sum of shares equals one. Financial wealth is defined as the sum of nominal and real claims.

obtain tax revenues, we model a stylized U.S. tax system. To be consistent with estimates on consolidated federal and state-level average marginal tax rates calculated in Bhandari and McGrattan (2019), we assume that taxes on labor income, dividends, and interest income are time invariant and set to \((\bar{\Upsilon}^n, \bar{\Upsilon}^d, \bar{\Upsilon}^b) = \bar{\Upsilon}^{US} = (38\%, 34\%, 0\%)\). We set \(\bar{G}\) so that on average the ratio of non-transfer government expenditures to total tax receipts equals 50%, also estimated by Bhandari and McGrattan (2019).

Optimal responses depend on the joint distribution of assets and after-tax income. For our baseline simulations, we choose Pareto weights so that average optimal levels of taxes are similar to the U.S. data. In particular, we assume that Pareto weights are described by \(\vartheta_i \propto \exp(\delta_1 \theta_{i,-1}) + \exp(\delta_2 s_{i,-1}) + \exp(\delta_3 b_{i,-1})\), where \((\theta_{i,-1}, s_{i,-1}, b_{i,-1})\) are the three dimensions of initial heterogeneity and \(\delta = (\delta_1, \delta_2, \delta_3)\) are parameters that are set so that in the non-stochastic economy setting \(\bar{\Upsilon} = \bar{\Upsilon}^{US}\) is optimal. In section 5.1.1, we discuss how results vary with alternative choices of Pareto weights.

5 Optimal monetary and monetary-fiscal policies

We now turn to a discussion of the optimal policies. We start with general observations about taxes, nominal interest rates, and transfers in our model and then contrast them with widely-studied representative agent counterparts.

---

22To arrive at the estimate of the marginal tax rate on capital income, we combine the Bhandari and McGrattan (2019) estimates of the effective marginal tax rate on corporate business income, on distributed dividends adjusted, and the schedule of marginal tax rates on non-corporate business income into a single number using the Barro and Redlick (2011) procedure. The same steps are used to combine the schedule of marginal tax rates on wage income into the flat tax rate used above. The tax on bond income is set to be zero to approximate the feature that most of government bonds are held through tax-deferred accounts.
The Ramsey planner in our economy has several instruments at their disposal: the nominal interest rates, tax rates on labor income, dividends, and bond income as well as lump-sum transfers. Revenues are raised directly through taxes and indirectly through nominal interest rates and inflation. These revenues are spent on servicing government debt, paying exogenous expenditures $\bar{G}$, and financing transfers. The timing of transfers is undetermined since Ricardian equivalence holds.\(^{23}\) Taxes and inflation that raise these revenues generate dead-weight losses. Average levels of taxes and transfers depend on the underlying inequality and on the Pareto weights that appear in the planner’s objective function.

We compare optimal allocations in our baseline setting (to be abbreviated as HANK) to optimal responses in a representative agent version of our model (to be abbreviated as RANK). In RANK, welfare considerations arise from concerns for price stabilization which are summarized by fluctuations in the rate of inflation, and concerns for output stabilization which are summarized by fluctuations in output relative to its first-best level. To achieve the first best, the planner can set nominal rates equal to the natural rate of interest (the real rate that prevails in the absence of nominal rigidities), and then implement a subsidy on labor income, $\Upsilon^t = -1/\Phi_t$ to eliminate time-varying markups, financed by a reduction in lump-sum transfers. Such a policy sets the inflation and output gap to zero at all dates, leaving no role for taxes on dividends or bond income.

With heterogeneous agents, such a policy is not optimal. For example, the burden of transfers falls disproportionately on poor households who finance a large portion of their consumption using these transfers. Our HANK setting is designed to match across-household differences in shares of labor, dividend, and bond income and, therefore, generates differential exposures to aggregate shocks. By setting the average level of taxes optimally, the planner provides redistribution and some insurance against idiosyncratic risks. However, due to incomplete asset markets, there is also a role for the planner to vary policy with shocks in order to ensure relative consumption shares are smooth across states and time. Providing this within-person insurance in response to aggregate shocks is a new force that is absent in RANK.

Adjusting the tax rates on bond and dividend income allows the planner, with minimal distortions, to influence after-tax real returns and helps in replacing the missing Arrow securities. When taxes are fixed, monetary policy has to take up the insurance role. By varying nominal rates in response to shocks, the planner can affect real wages, real returns, and therefore provide imperfect insurance.

These optimal responses are conveniently represented in terms of impulse response func-

\(^{23}\)We explore the optimal timing of transfers in section 6.1 where we extend the model to incorporate heterogeneity in the marginal propensity to consume.
tions, which will be our focus for the rest of the paper. An impulse response of variable $X_t$ to unexpected shock $E_k$ of size $\delta$ in period $k$ is defined as $\mathbb{E}_0 [X_t|E_k = \delta] - \mathbb{E}_0 [X_t|E_k = 0]$. For simulations, we set size $\delta$ to be one standard deviation of $E$.

We begin with the monetary responses to markup shocks and build several insights that carry over to the rest of the experiments.

### 5.1 Monetary policy responses to a markup shock

In this section, we consider the optimal monetary response to a negative innovation in $E_{\Phi,t}$. Because this shock increases the desired markup $1/(\Phi_t - 1)$, we refer to it as a positive markup shock. Figure II plots responses of the nominal interest rate, inflation, the real pre-tax wage per unit of effective labor, and real output to a positive markup in period one.

Figure II shows that optimal responses in HANK and RANK economies differ significantly. While the RANK planner slightly increases nominal interest rates in response to a markup shock, the HANK planner aggressively cuts them. The response of inflation in HANK is an order of magnitude higher, and the path of real wages and output are temporarily above the RANK counterparts.

We first isolate the economic mechanism responsible for those differences. Heterogeneity adds three motivations for the HANK planner that a RANK planner can ignore: redistribution, provision of insurance against idiosyncratic shocks, and provision of insurance against aggregate shocks. To spot which of these motivations influence policy responses, we construct a decomposition by successively studying economies located between HANK and RANK as follows. We (i) start with our calibrated HANK economy (plotted as solid blue line labeled as “HANK”), then (ii) shut down idiosyncratic shocks (plotted as dashed lines with square markers labeled as “HANK No Idio. risk”), then (iii) allow agents to trade Arrow securities contingent on aggregate shocks (plotted as dashed with circle markers labeled as “HANK CM”), and finally (iv) shut down heterogeneity in initial productivities and assets to obtain our RANK economy (plotted as solid red line labeled as “RANK”). This procedure allows us to isolate the contributions of provision of insurance against idiosyncratic shocks by comparing responses of economy (ii) to economy (i); the contribution of insurance provision against aggregate shocks by comparing the responses of economy (iii) to economy (ii); and redistribution by comparing responses in economy (iv) to economy (iii).

The decomposition in figure II shows that nearly all the differences in policy responses between HANK and RANK are driven by the planner’s wish to provide insurance against aggregate shocks. With lump-sum transfers, complete and incomplete market versions of RANK are identical, which implies that the interaction of market incompleteness and heterogeneity creates a powerful concern for the HANK planner. Monetary policy is ineffective.
in providing redistribution or insurance against idiosyncratic shocks beyond what can be achieved through taxes $\bar{Y}$. In contrast, monetary policy can provide insurance against aggregate shocks because it can act as an imperfect substitute for the missing Arrow securities.

It is easiest to understand the provision of insurance and the differences between RANK and HANK economies by considering a one-time, fully transient positive markup shock. The textbook effect of this shock is an increase in inflationary pressure. Monetary policy can offset that force by depressing marginal costs. Since marginal costs are proportional to aggregate demand, a contractionary increase in nominal rates is optimal in the RANK economy. Galí (2015) dubs this mechanism “leaning against the wind.”

This one-time markup shock also changes the mix of factor payments: it increases dividends and lowers wages. When households are homogeneous as in the RANK economy, or can all trade Arrow securities ex-ante, the change in the composition of firms’ payments does not affect welfare: agents who mainly receive wage income will hold the portfolio of Arrow securities that provide a payoff that offsets the drop in wage income. When the appropriate Arrow securities are missing, monetary policy can improve welfare by providing insurance against aggregate shocks. To offset the drop in labor income, the Ramsey planner sets interest rates to stimulate real wages. Since real wages are firms’ marginal costs, the monetary policy action that provides insurance is opposite to one that promotes price stability.

The net effect of a markup shock on optimal monetary policy depends on the relative strengths of planner’s price stability and insurance provision motives. The cost of inflation is set by the price adjustment cost parameter $\psi$ while insurance provision motives depend on inequalities in stock ownership relative to inequalities in wage income. If stock holdings are perfectly aligned with labor earnings, that is, if the share of stocks that each person owns equals his or her share of aggregate labor compensation, then the insurance provision motive vanishes; a positive effect on dividends exactly offsets the negative effect on labor earnings. In such a case, optimal responses are similar to those in RANK.24 The reason that HANK responses in figure II differ so much from what would support price stability is that stock holdings are much more skewed than labor earnings in U.S. data, making insurance concerns strong in our calibration.

Figure II also shows that the most efficient way to provide insurance is to front-load it: virtually all differences in HANK and RANK optimal outcomes are driven by policies in period $t = 1$. Agents can borrow and lend freely, and as a result, their utility depends on the present value of factor payments but not their exact timing. The first-order approximation

---

24 The RANK economy is a special case, where both shares are equal to one, but the same result will hold in the economy with arbitrary heterogeneity so long as shares of dividends and earnings are aligned. See online appendix for an illustration.
of firm's optimality condition (17) can be written as

\[ \pi_t = \beta E_t \pi_{t+1} + \text{const} \cdot \hat{w}_t + \text{const} \cdot (-\ln \Phi_t), \]

(35)

where \( \pi_t \) is log inflation, \( \hat{w}_t \) is the log deviation of the wage per unit of effective labor from its average level. Equation (35) shows that inflation in period \( t \) is proportional to the present discounted value of future wages. As a consequence, an increase in wage in some period \( k \), increases inflation in all \( t \leq k \). Therefore, to minimize costs of inflation it is optimal to deliver all the adjustment in the present discounted value of wages at the time of the shock.

The optimal response to a negative markup shock is virtually a mirror image of an optimal response to a positive markup shock. Averaged over time, the net expected flow of resources to each agent generated by the monetary policy response is approximately zero; an outcome consistent with optimal responses driven mainly by the planner’s insurance rather than the redistribution motives.

Figure II: Optimal monetary response to a markup shock. The bold blue and red lines are the calibrated HANK and RANK responses respectively. The dashed black lines with squares and circles are responses under HANK with idiosyncratic shocks shutdown and with complete markets, respectively.
5.1.1 Role of key parameters

Optimal responses vary with parameters that change the trade-off between price stabilization and insurance motives. The magnitude of the price stabilization motive is driven by the value of the price adjustment cost parameter $\psi$. The strength of the planner’s insurance motive depends on the relative post-tax inequality in dividends and labor earnings. The inequality, in turn, depends on the distribution of assets and labor productivities, and on the levels of taxes $\bar{\Upsilon}$ that are determined by the choice of Pareto weights. In this section, we study how optimal monetary policy responses vary with these objects. We report the main findings here and relegate details to the online appendix.

We set the value for $\psi$ following estimates of Sbordone (2002). In a staggered price adjustment model, such an estimate would correspond to firms changing prices every nine months. Studies using micro evidence on price changes (see Nakamura and Steinsson (2013)) recover estimates that range from around 6.8 to 12 months. In the online appendix, we vary $\psi$ between half to twice of our baseline value to study the sensitivity of our results with respect to $\psi$. We find that that the on-impact change in the nominal rates is fairly similar to the baseline, while the peak response of inflation varies roughly linearly in $\psi$ for this range. In addition, we also report a case where we set $\psi \approx 0$ to study a benchmark with fully flexible prices.

Next we study sensitivity with respect to Pareto weights. For a wide range of weights, we repeat our decomposition in figure II and show that virtually all the difference between HANK and RANK responses are driven by insurance needs. Quantitatively, insurance needs depend on correlation between individual shares of after-tax labor and dividend incomes. Lower correlations generally imply larger differences between HANK and RANK responses. For Pareto weights that correspond to a fairly large range of taxes around the baseline U.S. levels, impulse responses are quantitatively similar to the baseline responses in figure II.\footnote{Eliminating inequality in assets, say by a 100% tax on dividends, does not eliminate the need for insurance. An economy with unequal stock ownership and a 100% tax on dividends is isomorphic to an economy without dividend taxes and equal stock ownership. In such an economy, a positive markup shock benefits agents with low labor earnings (since windfall from profits is greater than their drop in earnings) and hurts high-earners. All households, however, would still prefer to eliminate this risk. In the online appendix we show that optimal responses of a Utilitarian planner look very similar to those in figure II, although the identities of the agents who benefit from a positive mark up shock differs from our baseline calibration.}

Since inequality drifts over time in our economy, optimal responses depend, in principle, on the time $t$. In our calibrated economy, we found the drift to be slow and therefore responses in $t = 25$ and $t = 50$ appear very similar to those at $t = 1$.  

\[25\]
5.2 Monetary-fiscal response to a markup shock

We now turn to studying a Ramsey planner who chooses both monetary and fiscal policy actions. Figure III shows optimal responses to a markup shock when the planner chooses tax rates in addition to nominal interest rates. The planner offsets the impact of the shock using a combination of a labor subsidy and a dividend tax while the nominal interest rate is unchanged.

In the RANK economy, lump-sum taxes are non-distortionary, and taxes on dividend and interest income are redundant. Under the optimal monetary-fiscal policy, the planner achieves the first best by setting $\Upsilon^m_t = -1/\Phi_t$ to offset monopoly distortions, choosing the path of nominal rates that delivers a constant price level and setting lump-sum taxes to satisfy the government’s budget constraint.

In the HANK economy, the burden of lump-sum taxes falls disproportionately on poor households. The RANK prescription of a proportional labor subsidy financed by reducing lump-sum transfers is therefore not optimal. Instead, the planner finances the labor income subsidy using a one-time tax on dividends. This tax exactly offsets the gains to stock owners from the higher markups and thus the policy response provides complete insurance to all agents with respect to the markup shock. Finally, the flat nominal rate implements a constant price level.

5.3 Optimal responses to productivity shocks

Figure IV shows optimal monetary response to a negative TFP shock. The monetary responses in HANK differ significantly from RANK, and the decomposition reveals that the difference is once again driven by the lack of insurance against aggregate shocks.

To understand why the planner wants to provide insurance, first observe that a TFP shock leads to the same reduction in both profits and labor earnings. Thus, firm owners and workers suffer equally from the shock. The impact of a TFP shock has different effects on different agents because agents are heterogeneous in their bond holdings: agents who hold a lot of bonds suffer less from the TFP shock than otherwise identical agents who are in debt. Market incompleteness is the necessary friction that prevents borrowers and savers from structuring contracts to provide an effective hedge against TFP shocks. Monetary policy fills in this gap by lowering returns on debt which transfers resources from the savers to borrowers and keeps the relative consumption shares smooth.

The optimal policy response departs conceptually from RANK, and more generally, from the prescription that monetary policy should aim to minimize fluctuations in inflation and the New Keynesian concept of an “output gap.” In response to TFP shocks, setting the interest rate to the “natural level” (a level that would prevail in a flexible price economy)
the planner can eliminate fluctuations in both inflation and the output gap. The optimal response in the RANK economy and in the complete market version of a HANK economy follow this prescription. See in figure IV solid red line labeled as “RANK” and dashed black line labeled as “HANK CM”

That prescription is not optimal in the incomplete market economy since it does not provide insurance to borrowers and savers. To transfer resources from savers to borrowers, the planner lowers ex-post real returns on debt. Lower ex-post returns are achieved by generating surprise inflation as well as pushing the expected real rate below the natural rate. A lower real rate requires temporarily higher output today relative to the future. A higher output triggers more inflation, as can be seen from equation (35). To offset this extra inflationary pressure, the planner also commits to deflation at $t = 2$.

The magnitude of the planner’s insurance motive depends on heterogeneity in holdings of nominal bonds. The U.S. data indicate considerable heterogeneity in nominal asset holdings, which explains the large difference between the optimal policy in the HANK economy relative to RANK.

---

\[26\] This is a common property of New Keynesian models; Blanchard and Gali (2007) refer to it as the “divine coincidence.”
Figure IV: Optimal monetary response to a TFP shock. The bold blue and red lines are the calibrated HANK and RANK responses respectively. The dashed black lines with squares and circles are responses under HANK with idiosyncratic shocks shutdown and with complete markets respectively.

If fiscal policy can be adjusted in response to the TFP shock, then optimal insurance is provided to borrowers and savers directly by a state-contingent tax on interest income. Such a tax effectively completes markets and brings optimal policy responses for HANK and RANK economies together. See figure V.

6 Extensions

We describe several extensions to the baseline calibration and environment. The first two extensions alter trading arrangements regarding risk-free assets and stocks. The last one incorporates a richer stochastic process for individual skills with heterogeneous labor income exposures to aggregate TFP shocks.

6.1 Liquidity frictions

So far, we assumed that households have unrestricted access to risk-free bond markets. Empirical work documenting large marginal propensities to consume points to presence of liquidity constrained households (see, for instance Jappelli and Pistaferri (2014), Johnson et al. (2006)). In this section, we investigate the implications of liquidity frictions for optimal
monetary policy. We follow Jappelli and Pistaferri (2014) and augment our model with “hand-to-mouth” agents who own equities and hold a fixed amount of nominal assets. They can consume dividends, interest on their nominal holdings, and their labor income but cannot trade the financial assets. Therefore, the budget constraint of a hand-to-mouth agent satisfies equation (2) subject to the restriction that the market value of nominal debt holdings $P_t Q_i h_{i,t}$ must be constant over time. Thus, we modify our baseline model to allow an additional dimension of permanent heterogeneity, namely, an indicator variable $h_i \in \{0, 1\}$, where $h_i = 1$ if the agent is a hand-to-mouth type and $h_i = 0$ if the agent is not a hand-to-mouth type. All other aspects of the baseline model remain the same, including how we set initial conditions.

To calibrate the distribution of hand-to-mouth agents, we use Jappelli and Pistaferri’s computations of average MPCs by cash-in-hand defined as financial wealth plus current period wage minus taxes. We put agents into cash-in-hand quantile bins and for each quantile bin we assign the hand-to-mouth status randomly to match the model-generated MPC gradient with respect to cash-in-hand observed in Jappelli and Pistaferri (2014). Because Jappelli and Pistaferri detect high MPCs throughout the wealth distribution, this calibration results in substantial numbers of both asset rich and asset poor hand-to-mouth agents. This
approach also preserves the distribution of real and nominal claims of the baseline model which results in the same insurance motives and, therefore, allows us to distinguish effects coming from the ‘liquidity frictions’ of the hand-to-mouth agents.27

Figures VI and VII show optimal monetary responses to markup and aggregate TFP shocks for the calibration with hand-to-mouth agents and compares them to optimal responses in the baseline calibration. Evidently, the trading frictions that give rise to the heterogeneities in MPCs make the paths for nominal rates, real wages, and inflation smoother than they are in the baseline model. Since Ricardian equivalence no longer holds, the optimal path of transfers is now uniquely pinned down. We show next that the deviation of the optimal policy responses from those in our baseline model are driven by the inability of the hand-to-mouth agents to borrow and save to smooth their consumption over time as well as the substantial heterogeneity within the set of liquidity constrained households.

We start with optimal monetary policy responses to a markup shock in figure VI. In the baseline calibration (solid blue line), the planner provides insurance against the shock by front-loading higher wages and avoiding additional future inflation from firms anticipating higher future marginal costs. That front-loading policy would be costly for hand-to-mouth agents – they would have too much income in the short run relative to future which they cannot smooth over time. Thus, in addition to providing insurance, the planner must also ensure that the consumption paths of the hand-to-mouth agents are smooth over time.

A natural way to both achieve insurance and to allow for smoother consumption after the shock would be to set the timing of lump-sum transfers appropriately. However, the planner is confronted with the problem that hand-to-mouth agents are not homogeneous. A path of transfers that would smooth the consumption of poor hand-to-mouth agents reliant on wage income would exacerbate the volatility of consumption of rich hand-to-mouth agents who rely mostly on dividend income. Thus, heterogeneity among liquidity constrained agents makes transfers a less effective tool.28

In addition to picking an appropriate path for transfers, the planner distorts allocations to induce a smoother path (dashed black line) for the real wage. A smoother path of real

27Our specification and calibration is also in line with Kaplan et al. (2014) and Kaplan et al. (2018) who emphasize that data indicate substantial heterogeneity among liquidity constrained agents. Their baseline calibration assigns higher MPCs to agents having low liquid wealth. In their calibration, about two thirds of agents who have zero liquid wealth nevertheless have positive illiquid wealth. Our specification captures both features of their calibration because our high-s hand-to-mouth agents behave like the wealthy hand-to-mouth agents in Kaplan et al. (2018).

28To highlight this point and isolate the role of wealthy hand-to-mouth, in the online appendix, we consider an alternative calibration in which the bottom 15% of the cash-in-hand distribution is set to be hand-to-mouth. This example captures the flavor of typical calibrations of a one-asset-Aiyagari model because the constrained hand-to-mouth agents are more homogeneous and depend almost entirely on their labor incomes. We show that the dynamics of interest rates, inflation, output and wages is almost identical to that of the baseline economy with no liquidity frictions. When liquidity-constrained agents are homogeneous, the planner can effectively borrow on their behalf and use transfers to smooth their consumption.
wages means a real wage above the “natural wage” (a wage rate that would prevail with flexible prices) when markups are high, and below the “natural wage” as the economy recovers. Implementing such a path for wages requires expansionary monetary policy and associated inflation on impact with the shock, and a commitment to a contractionary monetary policy in the future, leading to persistent deflation.

Figure VII displays the response of optimal policy to a productivity shock. Overall the policy actions help to approximate the within-agent transfers that would occur if all agents, including the “liquidity constrained” hand-to-mouth agents, were free to trade assets. The TFP shock generates disparities in total income between hand-to-mouth agents who are debtors and those who have savings. As in the baseline, the planner inflates away debt in the short run, thereby transferring resources from savers towards debtors during a recession. In addition, the planner engineers a persistent but small deflation. The reason for this difference is that a higher price level acts like a tax on wealth; it lowers the obligations of hand-to-mouth agents that are in debt and reduces the asset holdings of hand-to-mouth agents with savings. Since the shock itself is transitory, the agents would want to undo this permanent reshuffling of resources as the effect of the shock wears out. Since they cannot trade, the planner uses the price level as a tool, and the subsequent deflation generates a smooth path of repayments from debtors to asset holders as TFP reverts to its steady-state value. Similarly to markup shocks, these paths of prices and real rates require an expansionary monetary policy for a few periods, and then commitment to high interest rates in the future making the path of nominal rates smoother relative to the baseline.

Our discussion above suggests that not only the dispersion of MPCs but also their persistence has important quantitative implications for optimal policy. The calibration in this section makes an implicit assumption that individual MPCs are permanent. The more frequently the identity of liquidity constrained agents switches, the smaller this persistence will be. This will reduce the needs of the planner to smooth insurance benefits over time and bring impulse responses closer to the baseline case. The empirical work on persistence in MPCs is still largely in its infancy, although the work of Auclert et al. (2018) and Kaplan et al. (2018) provides some valuable insights into MPC dynamics over time.

### 6.2 Mutual Fund

In the baseline model, we calibrated households’ portfolios to their empirical counterparts in the SCF but imposed that households could not trade claims to dividends. In this section, we follow Gornemann et al. (2016) by having agents trade shares in a mutual fund. A competitive mutual fund sector invests in corporate equity and government bonds and remits after-tax earnings to households in proportion to their holdings of the mutual fund. Shares in the mutual fund are indirect claims to returns from a common financial portfolio.
Figure VI: Optimal monetary response to a markup shock with liquidity frictions. The bold blue lines are responses under the baseline without hand-to-mouth agents, the dashed black lines with circles are responses with hand-to-mouth agents. Transfer are not plotted for the baseline because Ricardian equivalence holds and timing of transfers is indeterminate.

and are traded by all households in a competitive market.

The mutual fund solves

\[
\max_{B_t} \mathbb{E}_0 \sum_t S_{mf}^t D_{a,t} \\
Q_t B_t + D_{a,t} = \frac{(1 - \Upsilon^b_t) B_{t-1}}{1 + \Pi_t} + (1 - \Upsilon^d_t) D_t,
\]

where we follow Gornemann et al. (2016) and set \( S_{mf}^t \) to be the asset-weighted average of intertemporal marginal rates of substitutions across households. Households’ budget constraint (2) is modified to become

\[
c_{i,t} + P_{a,t} a_{i,t} = (1 - \Upsilon^b_t) W_t \epsilon_{i,t} n_{i,t} + T_t + (D_{a,t} + P_{a,t}) a_{i,t-1},
\]

where \( a_{i,t} \) are household \( i \)'s holdings of the mutual fund and \( \int a_{i,t} di = 1 \). Households freely trade \( a_{i,t} \). The supply side of the model is kept unchanged.

We study optimal responses to markup and TFP shocks and contrast them with those for the section 2 baseline model. To make results comparable, our calibration in this mutual fund setting closely follows the baseline model in section 4. The distribution of mutual fund
Figure VII: Optimal monetary response to a TFP shock with liquidity frictions. The bold blue lines are responses under the baseline without hand-to-mouth agents, the dashed black lines with circles are responses when hand-to-mouth agents. Transfer are not plotted for the baseline because Ricardian equivalence holds and timing of transfers is indeterminate.

holdings is initialized using the distribution of financial wealth from SCF 2007 formed by summing all claims to bonds and stocks, and Pareto weights are again reset to rationalize observed U.S. average tax rates. All other parameters are the same as the baseline model. When we study optimal monetary policy, we impose $B_t = B_0$.\(^{29}\)

Figure VIII plots optimal monetary policy responses to a markup shock. Responses in the baseline model and those for the mutual fund are very close. The reason for this is simple. As discussed extensively in section 5, insurance motives largely drive optimal policy, and magnitudes are determined by cross-sectional heterogeneity in exposures of labor and non-labor incomes to aggregate shocks. Even after we sum the bond and stock claims, total financial wealth remains quite skewed relative to labor earnings. As before, the planner provides insurance in response to markup shocks by boosting the present discounted value of wages. In the online appendix, we show that responses to the TFP shock are also similar

\(^{29}\)This is mainly done for the comparability of results. Different from the baseline, in which we had Ricardian equivalence, the planner with the mutual fund has incentives to vary the level of debt and change the riskiness of the holding period returns of the mutual fund. Setting the debt level to a constant imposes parity between the baseline monetary planner and the mutual fund monetary planner in their ability to affect returns. We relax this restriction when we study optimal monetary-fiscal policies with a mutual fund in which the planner can use taxes on bonds income or dividends to directly affect returns. See the online appendix.
Figure VIII: Optimal monetary responses to a markup shock with mutual fund. The bold blue lines are responses under the baseline and the dashed black lines with circles are responses under the mutual fund setting.

to the baseline.

6.3 Heterogeneous labor income exposures

In our baseline calibration, the percentage fall in labor income during recessions is the same across workers. Using administrative data on W2 forms over the period 1978-2010, Guvenen et al. (2014) document that relative to a typical worker, individuals who have low past incomes face larger drops in earnings in recessions. In this section, we compute optimal monetary and fiscal responses under a richer stochastic process for idiosyncratic risk that captures the Guvenen et al. (2014) patterns.

We modify equation (12) to

$$\ln \epsilon_{i,t} = (1 + f(\theta_{i,t-1})) \ln \Theta_t + \ln \theta_{i,t} + \epsilon_{i,t},$$

and set $f(\theta)$ so that the aggregate productivity shock has different loadings for agents with different earning histories. We assume a quadratic function $f(\theta) = f_0 + f_1 \theta + f_2 \theta^2$ and normalize $f_0$ so that an agent with median productivity faces a drop similar to the drop in aggregate TFP. We then simulate the competitive equilibrium for 30 periods and extract “recessions” as consecutive periods in which the growth rate of output falls one
standard deviation below zero. Following the empirical procedure in Guvenen et al. (2014), we rank workers by percentiles of their average log labor earnings 5 years prior to the shock and compute the percent earnings loss for each percentile relative to the median. We set parameters $f_1$ and $f_2$ to match earnings losses of the 5th, 95th percentiles.

In figure IX, we report the the optimal monetary policy response to a TFP shock with heterogeneous exposures. The additional inequality induced by a recession increases the gains that the planner earns from providing insurance. As compared to our baseline monetary response, the planner further lowers the nominal rate and thereby induces higher inflation in the short run and a lower ex-ante real rate.\footnote{In the online appendix we plot the monetary-fiscal responses. With optimal monetary as well as fiscal policy, the planner responds by increasing labor income taxes in addition to the tax on bond income. For reasons similar to Werning (2007), the marginal costs extracting resources from high-income agents is lower in times of higher inequality. Therefore optimal labor tax rates are higher on impact and then revert, as the effects of the inequality shocks wear out, to a permanently higher level.}
7 Concluding Remarks

Our paper makes two contributions. It develops a way to approximate optimal plans in economies with heterogeneous agents, and reevaluates lessons for monetary and fiscal policy drawn from New Keynesian economies. Relative to price stability motives that typically drive policy prescriptions, heterogeneity adds a quantitatively important insurance motive. We developed these insights using a fairly canonical setting that combined the most basic versions of the new Keynesian model and the incomplete market models. Our methods can be extended to environments that feature a more detailed modeling of household balance sheets, richer labor market dynamics including realistic wage-setting frictions, and richer asset markets formulations that are consistent with the observed returns and monetary transmission channels. We suspect that insurance concerns with respect to aggregate shocks and determinants of heterogeneous income exposures in the cross section will still be key in understanding policy responses. What will change with these features are the details of how this insurance is implemented. Cataloging this is left for future work.
References


Online Appendix

A Additional details for section 3

In this section, we fill in the missing steps for section 3. First we show, in the context of the section 3.1 economy, how our method extends to higher-order approximations. Second, we show how to generalize the expansions so that we can deal with persistent aggregate and idiosyncratic shocks as well as additional state variables, as discussed in section 3.2.

A.1 Higher order approximations for section 3.1

We start with a second-order approximation to the model presented in section 3.1. These are given by

\[ \tilde{X}(\Omega, \sigma\varepsilon; \sigma) = \bar{X} + \sigma(\tilde{X}\varepsilon\varepsilon + \tilde{X}_\sigma) + \frac{1}{2}\sigma^2(\tilde{X}\varepsilon\varepsilon\varepsilon + 2\tilde{X}\varepsilon\sigma\varepsilon + \tilde{X}_\sigma\sigma) + O(\sigma^3), \]

where the symbol \( a \cdot (b, c) \) denotes a bilinear map.\(^{31}\) A similar expansion can be written for \( \tilde{x}(z, \Omega, \sigma\varepsilon, \sigma\varepsilon; \sigma) \).

To obtain the necessary terms, we proceed in two steps: section A.1.1 computes intermediate terms including higher-order Frechet derivatives for individual and aggregate policy functions, and section A.1.2 uses these terms to compute the second-order expansion. Although the second-order expansion requires additional notation, the steps below highlight ways in which the same fundamental insights presented in section 3 maintain the tractability of the problem.

\(^{31}\)Specifically, if \( a \) is a \( n_1 \times n_2 \times n_3 \) tensor, \( b \) is a \( n_2 \times n_4 \) matrix and \( c \) is a \( n_3 \times n_5 \) matrix then \( d = a \cdot (b, c) \) is a \( n_1 \times n_4 \times n_5 \) tensor defined by

\[ d_{ilm} = \sum_{j,k} a_{ijk} b_{jl} c_{km}. \]

This definition generalizes to when \( a, b, \) or \( c \) is infinite dimensional, such as with \( \partial \tilde{x}_z \).
A.1.1 Intermediate terms for second-order expansions

Differentiating equation (22) twice with respect to $z$ we find

$$0 = \bar{F}_x \cdot \bar{x}_{zz} + \bar{F}_x \bar{x}_{zz} + \bar{F}_{x+} \cdot (\bar{x}_{zz} + \bar{x}_z p \bar{x}_{zz})$$

$$+ \bar{F}_{zz} + \bar{F}_{x z} \cdot (I, \bar{x}_z) + \bar{F}_{x x} \cdot (I, \bar{x}_z) + \bar{F}_{x x+} \cdot (I, \bar{x}_z)$$

$$+ \bar{F}_{x z} \cdot (\bar{x}_z, I) + \bar{F}_{x x-} \cdot (\bar{x}_z, \bar{x}_z) + \bar{F}_{x x} \cdot (\bar{x}_z, \bar{x}_z) + \bar{F}_{x x+} \cdot (\bar{x}_z, \bar{x}_z)$$

$$+ \bar{F}_{x z} \cdot (\bar{x}_z, I) + \bar{F}_{x x-} \cdot (\bar{x}_z, \bar{x}_z) + \bar{F}_{x x} \cdot (\bar{x}_z, \bar{x}_z) + \bar{F}_{x x+} \cdot (\bar{x}_z, \bar{x}_z),$$

where $I$ represents the identity matrix and we use $a \cdot (b, c)$ to denote a bilinear map. Lines 2-5 appear complicated but are actually simply combining all of the already known derivatives of $x$ with cross derivatives of $F$. It will prove convenient to combine all of these terms into a single term:

$$\sum_{\alpha, \beta \in \{z, x^-, x^+\}} \bar{F}_{\alpha \beta} \cdot ((\bar{\alpha}_z, \bar{\beta}_z))$$

with the knowledge that $\bar{z}_z \equiv I, \bar{x}_z^- \equiv \bar{x}_z$, and $\bar{x}_z^+ \equiv \bar{x}_z$. In doing this $\bar{x}_{zz}$ can be represented by a simple linear equation

$$\bar{x}_{zz} = -\left[\bar{F}_{x-} + \bar{F}_x + \bar{F}_{x+} \cdot (I + \bar{x}_z p)\right]^{-1} \sum_{\alpha, \beta \in \{z, x^-, x^+\}} \bar{F}_{\alpha \beta} \cdot (\bar{\alpha}_z, \bar{\beta}_z).$$

In a similar manner one can show that

$$\partial x_z \cdot \Delta = -\left[\bar{F}_{x-} + \bar{F}_x + \bar{F}_{x+} \cdot (I + \bar{x}_z p)\right]^{-1} \sum_{\alpha \in \{z, x^-, x^+\}} \bar{F}_{\alpha \beta} \cdot (\bar{\alpha}_z, \partial \beta \cdot \Delta),$$

where we use $\partial x^- \cdot \Delta \equiv \partial x^+ \cdot \Delta \equiv \partial x \cdot \Delta$.

The last of the derivatives with respect to the state variables that are required for the second order expansion is $\partial^2 \bar{x} \cdot (\Delta_1, \Delta_2)$. We will use the pre computed expressions for $\partial \bar{x}$ and $\partial \bar{X}$ evaluating them in the directions $\Delta_1$ and $\Delta_2$. Differentiating (22) we find

$$0 = \bar{F}_x \cdot \partial^2 \bar{x} \cdot (\Delta_1, \Delta_2) + \bar{F}_x \partial^2 \bar{x} \cdot (\Delta_1, \Delta_2) + \bar{F}_{x+} \cdot \partial^2 \bar{x} \cdot (\Delta_1, \Delta_2) + \bar{F}_x \cdot (\partial^2 \bar{x} \cdot (\Delta_1, \Delta_2)) + \bar{F}_x \partial^2 \bar{X} \cdot (\Delta_1, \Delta_2)$$

$$+ \sum_{\alpha, \beta \in \{x^-, x^+, X\}} \bar{F}_{\alpha \beta} \cdot (\partial \alpha \cdot \Delta_1, \partial \beta \cdot \Delta_2).$$

In solving this equation for $\partial^2 \bar{x} \cdot (\Delta_1, \Delta_2)$ we find

$$\partial^2 \bar{x} \cdot (\Delta_1, \Delta_2) = A(z, \Delta_1, \Delta_2) + C(z) \partial^2 \bar{X} \cdot (\Delta_1, \Delta_2).$$
where

\[ A(z, \Delta_1, \Delta_2) = - \left[ \bar{F}_x - \bar{F}_x + \bar{F}_{xx} + (I + \bar{x}_z p) \right]^{-1} \sum_{\alpha, \beta \in \{ x^-, x, x^+, X \}} \bar{F}_{\alpha\beta} \cdot (\partial\bar{\alpha} \cdot \Delta_1, \partial\bar{\beta} \cdot \Delta_2) \]

from terms already known and \( C(z) \) is the same term computed in section 3.1. To find \( \partial^2 \bar{X} \cdot (\Delta_1, \Delta_2) \) we differentiate (23) to find

\[ 0 = \bar{R}_x \int \partial^2 \bar{x}(y) \cdot (\Delta_1, \Delta_2) d\Omega(y) + \bar{R}_x \partial^2 \bar{X} \cdot (\Delta_1, \Delta_2) + \int \sum_{\alpha, \beta \in \{ x(y), X \}} \bar{R}_{\alpha\beta} \cdot (\partial\bar{\alpha} \cdot \Delta_1, \partial\bar{\beta} \cdot \Delta_2) d\Omega(y) \]

\[ + \bar{R}_x \int \partial \bar{x}(y) \cdot \Delta_1 d\Delta_2(y) + \bar{R}_x \int \partial \bar{x}(y) \cdot \Delta_2 d\Delta_1(y). \]

Plugging in for \( \partial^2 \bar{x} \cdot (\Delta_1, \Delta_2) \) yields a linear equation which can be easily solved for \( \partial^2 \bar{X} \cdot (\Delta_1, \Delta_2) \).

### A.1.2 Second-order expansions

We can use these derivatives to compute the second-order terms. To find \( \bar{x}_{\varepsilon\varepsilon} \), differentiate \( F \) twice with respect to \( \varepsilon \) to get the linear equation

\[ 0 = \bar{F}_x \bar{x}_{\varepsilon\varepsilon} + \bar{F}_{xx} \bar{x}_z p \bar{x}_{\varepsilon\varepsilon} + \sum_{\alpha, \beta \in \{ x, x^+ \}} \bar{F}_{\alpha\beta} \cdot (\bar{\alpha}_\varepsilon, \bar{\beta}_\varepsilon), \]

where \( \bar{x}_\varepsilon^+ \equiv \bar{x}_z p \bar{x}_\varepsilon \) and \( \varepsilon \equiv I \). Similarly, \( \bar{x}_{\varepsilon\varepsilon} \) solves the following linear equation

\[ 0 = \bar{F}_x \bar{x}_{\varepsilon\varepsilon} + \bar{F}_{xx} \bar{x}_z p \bar{x}_{\varepsilon\varepsilon} + \sum_{\alpha \in \{ x, x^+ \}} \bar{F}_{\alpha\beta} \cdot (\bar{\alpha}_\varepsilon, \bar{\beta}_\varepsilon), \]

with the understanding that \( \bar{E}_\varepsilon \equiv I \) and \( \bar{x}_\varepsilon^+ \equiv \bar{x}_z p \bar{x}_\varepsilon + \partial \bar{x} \cdot \bar{\Omega}_\varepsilon \).

Differentiating twice with respect to \( \varepsilon \) yields

\[ 0 = \bar{F}_x \bar{x}_{\varepsilon\varepsilon} + \bar{F}_{xx} \bar{x}_z p \bar{x}_{\varepsilon\varepsilon} + \partial \bar{x} \cdot \bar{\Omega}_{\varepsilon\varepsilon} + \bar{F}_x \bar{X}_{\varepsilon\varepsilon} \]

\[ + \bar{F}_{xx} \bar{x}_z \cdot (p \bar{x}_\varepsilon, p \bar{x}_\varepsilon) + \partial \bar{x}_z \cdot (p \bar{x}_\varepsilon, \bar{\Omega}_\varepsilon) + \partial \bar{x}_z \cdot (\bar{\Omega}_\varepsilon, p \bar{x}_\varepsilon) + \partial^2 \bar{x} \cdot (\bar{\Omega}_\varepsilon, \bar{\Omega}_\varepsilon) \]

\[ + \sum_{\alpha, \beta \in \{ x, x^+ \}} \bar{F}_{\alpha\beta} \cdot (\bar{\alpha}_\varepsilon, \bar{\beta}_\varepsilon) \]

\[ \text{for parsimony we have dropped the dependence on } z \text{ when not necessary.} \]
and
\[
\int R_x \tilde{x}_{\epsilon\epsilon}(y) + R_X \bar{X}_{\epsilon\epsilon} + \sum_{\alpha, \beta \in \{\bar{x}(y), \bar{x}\}} R_{\alpha\beta} \cdot (\tilde{x}_{\epsilon\epsilon}, \beta_{\epsilon\epsilon}) d\Omega(y).
\] (39)

All the terms in the second line can be computed from our analysis of the previous section and all the terms in the third line are known. What remains is to find \(\tilde{x}_{\epsilon\epsilon}\) and \(\bar{X}_{\epsilon\epsilon}\). This requires us extend the steps we used in the proof of theorem 1.

Differentiating (24) twice with respect to \(\epsilon\), evaluated at \(\sigma = 0\), yields
\[
\bar{\Omega}_{\epsilon\epsilon}(y) = -\int \sum_i \delta(z^i - y^i) \prod_{j \neq i} \zeta(z^j - y^j, z_{\epsilon\epsilon}(z)) d\Omega(z)
\]
\[
- \int \sum_i \delta'(z^i - y^i) \prod_{j \neq i} \zeta(z^j - y^j, z_{\epsilon\epsilon}(z))^2 d\Omega(z)
\]
\[
+ \int \sum_i \delta(z^i - y^i) \sum_{j \neq i} \delta(z^j - y^j) \prod_{k \neq i, j} \zeta(z^k - y^k, z_{\epsilon\epsilon}(z)) d\Omega(z).
\]

The density is then
\[
\bar{\omega}_{\epsilon\epsilon}(y) = \frac{\partial^{n_{\epsilon\epsilon}}}{\partial y_1 \partial y_2 \cdots \partial y_{n_{\epsilon\epsilon}}} \bar{\Omega}_{\epsilon\epsilon}(y) = -\sum_i \frac{\partial}{\partial y^i} (\tilde{z}_{\epsilon\epsilon}(y) \omega_{\epsilon}(y)) + \sum_i \sum_j \frac{\partial^2}{\partial y^i \partial y^j} (\tilde{z}_{\epsilon\epsilon}(y) \tilde{z}_{\epsilon\epsilon}(y) \omega_{\epsilon}(y)).
\]

The identical steps to (1) then show that
\[
\partial \tilde{x}(z) \cdot \Omega_{\epsilon\epsilon} = C(z) \partial X \cdot \Omega_{\epsilon\epsilon} \equiv C(z) X'_{\epsilon\epsilon}
\]
with
\[
X'_{\epsilon\epsilon} = - \left( R_x \int C(y) d\Omega(y) + R_X \right)^{-1} R_x \left( \int \tilde{x}_{zz}(y) p \tilde{x}_{\epsilon\epsilon}(y) + \tilde{x}_{z}(y) \cdot (p \tilde{x}_{\epsilon\epsilon}, p \tilde{x}_{\epsilon}) d\Omega(y) \right).
\] (40)

As with \(\bar{X}_{\epsilon\epsilon}\), rather than solving for \(\tilde{x}_{\epsilon\epsilon}(z)\) and \(X_{\epsilon\epsilon}\) jointly, we substitute for \(\partial \tilde{x}(z) \cdot \Omega_{\epsilon\epsilon}\) in (38) and solve for \(\tilde{x}_{\epsilon\epsilon}(z)\) yielding the linear relationship
\[
\tilde{x}_{\epsilon\epsilon}(z) = D_1(z) \cdot \left[ \begin{array}{c} \bar{X}_{\epsilon\epsilon} \cr \bar{X}'_{\epsilon\epsilon} \end{array} \right] + D_2(z)
\]
where \(D_1(z)\) is identical to the \(D_1\) in section 3.1. We then use this relationship to substitute into equations (39) and (40) to find \(X_{\epsilon\epsilon}\) and \(X'_{\epsilon\epsilon}\).

A key part of the second-order approximations is capturing the effect of risk via the terms \(\tilde{x}_{\sigma\sigma}(z)\) and \(X_{\sigma\sigma}(z)\).\(^{33}\) Let \(C_{\epsilon} \equiv \bar{E}^T_{\epsilon} E\) and \(C_{\epsilon} \equiv \bar{E}^T_{\epsilon} E\) be the variance-covariance matrix of the idiosyncratic and aggregate shocks respectively. Differentiating (22) and (23)

\(^{33}\)It is easy to verify that the cross derivatives with shocks and \(\sigma\) are zero.
yields

\[ 0 = \bar{F}_x - (\bar{x}_{ee} \cdot \bar{C}_e + \bar{x}_{EE} \cdot \bar{C}_E) + \bar{F}_x \bar{x}_{\sigma\sigma} + \bar{F}_X \bar{X}_{\sigma\sigma} + \bar{R}_x \int \bar{x}_{\sigma\sigma} \cdot \bar{C}_e d\Omega(y) + \bar{R}_X \bar{X}_{\sigma\sigma} \]

(41)

and

\[ 0 = \bar{R}_x \int \bar{x}_{\sigma\sigma}(y) + \bar{x}_{ee}(y) \cdot \bar{C}_e d\Omega(y) + \bar{R}_X \bar{X}_{\sigma\sigma}. \]

(42)

Before this set of equations can be solved for \( \bar{x}_{\sigma\sigma} \), we must evaluate \( \bar{\Omega}_{\sigma\sigma} \). Differentiating (24) and evaluating at \( \sigma = 0 \) yields

\[
\bar{\Omega}_{\sigma\sigma}(y) = - \int \sum_i \delta(z^i - y^i) \prod_{j \neq i} \lambda(z^j - y^j) \left( z_{\sigma\sigma}^i(z) + z_{ee}^i \cdot \bar{C}_e \right) d\Omega(z)
\]

\[ - \int \sum_i \delta'(z^i - y^i) \prod_{j \neq i} \lambda(z^j - y^j) \left[ z_{ee}^i(z) \right]^2 \cdot \bar{C}_e d\Omega(z)
\]

\[ + \int \sum_i \delta(z^i - y^i) \sum_{j \neq i} \delta(z^j - y^j) \prod_{k \neq i} \lambda(z^k - y^k) \left( z_{\sigma\sigma}^i(z) z_{ee}^i(z) \right) \cdot \bar{C}_e d\Omega(z)
\]

which gives

\[
\bar{\omega}_{\sigma\sigma}(y) = - \sum_i \frac{\partial}{\partial y^i} \left( (z_{\sigma\sigma}^i(y) + z_{ee}^i(y) \cdot \bar{C}_e) \omega(y) \right)
\]

\[ + \sum_i \sum_j \frac{\partial^2}{\partial y^i \partial y^j} \left( (z_{ee}^i(y) z_{ee}^j(y)) \cdot \bar{C} \omega(y) \right). \]

Following the identical steps as theorem (1) to show that show that

\[ \partial \bar{x}(z) \cdot \bar{\Omega}_{\sigma\sigma} = C(z) \partial \bar{X} \cdot \bar{\Omega}_{\sigma\sigma} \equiv C(z) \bar{X}'_{\sigma\sigma} \]

with

\[
\bar{X}'_{\sigma\sigma} = - \left( \bar{R}_x \int C(y) d\Omega(y) + \bar{R}_X \right)^{-1} \bar{R}_x \int \left( \bar{x}_z(y) p(\bar{x}_{\sigma\sigma}(y) + \bar{x}_{ee}(y) \cdot \bar{C}_e)
\]

\[ + \bar{x}_{zz}(y) \cdot (p \bar{x}_e, p \bar{x}_e) \cdot \bar{C}_e \right) d\Omega(y). \]

(43)

We then substitute for \( \partial \bar{x} \cdot \bar{\Omega}_{\sigma\sigma} = C(z) \bar{X}'_{\sigma\sigma} \) in (41) to and solve for \( \bar{x}_{\sigma\sigma}(z) \) to find the

---

34 If \( a \) is a \( n_1 \times n_2 \times n_2 \) tensor and \( C \) is a \( n_2 \times n_2 \) matrix then \( d = a \cdot C \) is length \( n_1 \) vector defined by

\[ d_i = \sum_{j,k} a_{ijk} C_{jk}. \]
linear relationship

\[ \bar{x}_{\sigma\sigma}(z) = E_0(z) + E_1(z) \left[ \bar{X}_{\sigma\sigma} \bar{X}'_{\sigma\sigma} \right]^\top. \]

This relationship can then be plugged into (42) and (43) to yield a linear equation for \( \bar{X}_{\sigma\sigma} \) and \( \bar{X}'_{\sigma\sigma} \).

### A.2 Expansions in the general case of section 3.2

We extend our method to handle persistent shocks and other endogenous persistent state variables besides the distributional state \( \Omega \). To do so, we extend the equilibrium conditions in the following manner

\[ F \left( E_{-\tilde{x}}, \tilde{x}, E_{+\tilde{x}}, \tilde{X}, \Lambda, \Theta, \varepsilon, E, z \right) = 0, \quad (44) \]

which must hold for all \( z \) in the support of \( \Omega \),

\[ R \left( \int \tilde{x} d\Omega dPr(\varepsilon), \tilde{X}, E_{+\tilde{X}}, \Lambda, \Theta, E \right) = 0, \quad (45) \]

and a first-order vector autoregression model \( \Theta' = \rho_\Theta \Theta + (1 - \rho_\Theta) \tilde{\Theta} + E \) for the exogenous shocks. The law of motion of the distribution is given by

\[ \tilde{\Omega}(\Omega, \Lambda, \Theta, E)(z) = \int \tilde{\iota}(\tilde{y}, \Omega, \Lambda, \Theta, \varepsilon, E) \leq z) dPr(\varepsilon) d\Omega(y) \quad \forall z. \quad (46) \]

We consider a family of perturbations indexed by a positive scalar \( \sigma \) that scales all shocks \( \varepsilon, E \) so that the policy functions are \( \tilde{X}(\Omega, \Lambda, \Theta, \sigma E; \sigma) \) and \( \tilde{x}(z, \Omega, \Lambda, \Theta, \sigma \varepsilon, \sigma E; \sigma) \). We will use \( \tilde{\cdot} \) to denote these functions evaluated at \( \sigma = 0 \).

Unlike section 3.1, we cannot assume that \( \tilde{\Omega}(\Omega, \Lambda, \Theta) \) is stationary but we recover the independence property

**Lemma 2.** For any \( \Omega, \Lambda, \Theta \), the policy functions \( \tilde{z}(z, \Omega, \Lambda, \Theta) \) satisfy \( \partial \tilde{z}(z, \Omega, \Lambda, \Theta) = 0 \) for all \( z \) and \( \tilde{z}(z, \Omega, \Lambda, \Theta) \) independent of \( z \).

**Proof.** We proceed similar to the proof of Lemma 1 in the main text. The first order condition with respect to \( b_{i,t-1} \) yields

\[ \mathbb{E} \left[ \frac{[\tilde{\iota}(\tilde{y}, \Omega, \Lambda, \Theta, \cdot, \cdot)]^\nu}{1 + \Pi(\Omega, \Lambda, \Theta, \cdot)} (\mu - \bar{\mu}(z, \Omega, \Lambda, \Theta, \cdot, \cdot)) \right] = 0. \]

When \( \sigma = 0 \), this yields \( \bar{\mu}(z, \Omega, \Lambda, \Theta) = \mu \) for all \( z \). While equation (20) to the zeroth order
\[
\bar{Q}(\Omega, \Lambda, \Theta) \bar{M}(\Omega, \Lambda, \Theta) m = \bar{m}(z, \Omega, \Lambda, \Theta) \bar{M}(\bar{\Omega}(\Omega, \Lambda, \Theta)) (1 + \bar{\Pi}(\bar{\Omega}(\Omega, \Lambda, \Theta)))^{-1}.
\]

By construction, the Pareto-Negishi weights integrate to one which implies \(\bar{m}(z, \Omega, \Lambda, \Theta) = m\) for all \(z\). Finally, the law of motion for \(\theta\) implies
\[
\bar{\theta}(z, \Omega, \Lambda, \Theta) = \rho_\theta \theta.
\]

Together they imply \(\partial \bar{z}(z, \Omega, \Lambda, \Theta) = 0\) for all \(z\) and \(\bar{z}_z(z, \Omega, \Lambda, \Theta)\) independent of \(z\). \(\square\)

A by product of 2 is that \(\bar{z}_z\) is diagonal. Although we exploit this property in the next section it is not essential.

We start by showing how our expansion extends to the transition path. We assume for a given \(\Omega, \Lambda, \Theta\) we have solved for the \(\sigma = 0\) transition dynamics \(\{\bar{\Omega}^n, \bar{\Lambda}^n, \bar{\Theta}^n\}_{n=0}^{N}\) with \((\bar{\Omega}^0, \bar{\Lambda}^0, \bar{\Theta}^0) = \Omega, \Lambda, \Theta\) and \((\bar{\Omega}^N, \bar{\Lambda}^N, \bar{\Theta}^N) = (\bar{\Omega}, \bar{\Lambda}, \bar{\Theta})\) at a non-stochastic steady state. Solving the the transition dynamics is eased by the fact that we know, a priori, the transition dynamics of \(\Omega\). For the remainder of this appendix we use \(\bar{\cdot}\) to denote derivatives evaluated at \((\bar{\Omega}^n, \bar{\Lambda}^n, \bar{\Theta}^n)\) and, to save on notation, and use \(\tilde{\cdot}\) to denote derivatives evaluated at the steady state \((\bar{\Omega}^N, \bar{\Lambda}^N, \bar{\Theta}^N)\). We’ll start by showing how to compute derivatives at the steady state and then show how to evaluate derivatives along the path.

The policy rules for \(X\) and \(x\) can then be approximated via Taylor expansion. The first order expansions for these variables are given by
\[
\tilde{X}(\Omega, \Lambda, \Theta, \sigma \mathcal{E}; \sigma) = \bar{X}^0 + \sigma (\bar{X}^0_{\mathcal{E}} \mathcal{E} + \bar{X}^0_\sigma) + \mathcal{O}(\sigma^2)
\]
and
\[
\tilde{x}(z, \Omega, \Lambda, \Theta, \sigma \mathcal{E}; \sigma) = \tilde{x}^0(z) + \sigma (\bar{x}^0_{\mathcal{E}}(z) \mathcal{E} + \bar{x}^0_{\sigma}(z)) + \mathcal{O}(\sigma^2).
\]

For brevity, we present the necessary derivatives for the first order expansions. Higher order terms extend analogously to section A.1.

### A.2.1 Derivatives at the steady state

The derivatives of the policy functions with respect to \(\Lambda\) and \(\Theta\) as well as the Frechet derivative with respect to the distribution \(\Omega\) are used repeatedly in what follows.

Differentiating (44) with respect to \(\Lambda\) yields (lemma 2 implies that \(\bar{\Omega}_\Lambda = 0\)).
\[
\tilde{F}_x(z)\bar{x}_\Lambda(z) + \tilde{F}_x(z)\bar{x}_\Lambda(z) + \tilde{F}_x(z) (\bar{x}_\Lambda(z)\bar{\Lambda}_\Lambda) + \tilde{F}_X(z)\bar{X}_\Lambda = 0
\]
and 
\[ R_x \int \overline{x}_\Lambda(z) d\Omega(z) + \overline{R}_X X_\Lambda + \overline{R}_X + X_\Lambda \Lambda + \overline{R}_\Lambda = 0. \]

The object \( \Lambda_\Lambda \) is unknown. It requires solving a nonlinear equation which we show below can be expressed using operations that involve matrices of small dimension. First note that 
\[ \overline{x}_\Lambda(z) = -(\overline{F}_x(z) + \overline{\Lambda}_x(z) - 1 \overline{F}_X(z) X_\Lambda \]

Let \( A(z) = -(\overline{F}_x(z) + \overline{\Lambda}_x(z) - 1 \overline{F}_X(z) X_\Lambda \), then
\[ \overline{X}_\Lambda = - \left( R_x \int A(z) d\Omega(z) + \overline{R}_X + \overline{\Lambda}_\Lambda R_X + \overline{\Lambda}_\Lambda \right)^{-1} \overline{R}_\Lambda. \]

Let \( P \) be such that \( \Lambda = PX \). Therefore, \( \Lambda_\Lambda \) must solve
\[ \Lambda_\Lambda = -P \left( R_x \int A(z) d\Omega(z) + \overline{R}_X + \overline{\Lambda}_\Lambda R_X + \overline{\Lambda}_\Lambda \right)^{-1} \overline{R}_\Lambda. \]

This can be found easily with a 1-dimensional root solver as all the matrices that need to be inverted are small dimensional.

Next differentiating (44) with respect to \( \Theta \) yields (lemma 2 implies that \( \overline{\Omega}_\Theta = 0 \)).
\[ \overline{F}_x(z) \overline{x}_\Theta(z) + \overline{F}_x(z) \overline{x}_\Theta(z) + \overline{F}_x(z) \overline{x}_\Theta(z) \rho_\Theta + \overline{x}_\Lambda(z) P \overline{X}_\Theta \right) + \overline{F}_X(z) \overline{X}_\Theta + \overline{F}_\Theta(z) = 0. \]

This yields a linear equation in \( \overline{x}_\Theta \) and \( \overline{X}_\Theta \) which we can solve for \( \overline{x}_\Theta \).\(^{35}\) Plugging in for the linear relationship between \( \overline{x}_\Theta \) and \( \overline{X}_\Theta \) in
\[ R_x \int \overline{x}_\Theta(z) d\Omega(z) + \overline{R}_X \overline{X}_\Theta + \overline{R}_X + \overline{X}_\Theta \rho_\Theta + \overline{R}_X + \overline{X}_\Lambda P \overline{X}_\Theta + \overline{R}_\Theta = 0. \]

yields a linear equation for \( \overline{X}_\Theta \).

Finally to determine the Frechet derivative, we differentiate (44) along the direction \( \Delta \).

Doing so yields
\[ (\overline{F}_x(z) + \overline{F}_x(z)) \partial x(z) \cdot \Delta + \overline{F}_x(z) \partial x(z) \cdot \partial \Sigma \cdot \Delta + \overline{F}_x(z) x_\Lambda(z) \rho_\Theta \partial \overline{X} \cdot \Delta + \overline{F}_X(z) \partial \overline{X} \cdot \Delta = 0. \]

We first derive an analogue of the property \( \partial \overline{\Omega} \cdot \Delta = \Delta \). This holds in the simple section 3.1 economy but fails in the more general case. We proceed by showing that we can evaluate \( \partial \overline{\Omega} \) along a direction \( \Delta^j \) that satisfies the property that there exists a function \( a(\cdot) \) such

\(^{35}\)Easiest to exploit \( \rho_\Theta = \begin{pmatrix} \rho_\Theta & 0 \\ 0 & \rho_\Phi \end{pmatrix} \) and solve for each column of \( \overline{x}_\Theta \) separately.
that the density of $\Delta^j$ takes the form

$$\frac{\partial}{\partial y^j} (a(y) \tilde{\omega}(y))$$

Begin by differentiating the law of motion for $\tilde{\Omega}$ at $\sigma = 0$. Since $\partial \tilde{z} = 0$, we get

$$(\partial \tilde{\Omega} \cdot \Delta^j)(y) = \int \prod \delta(\tilde{z}^i(z) - y^i) \frac{\partial}{\partial z^i} (a(z) \tilde{\omega}(z)) \, dz$$

$$= \int \sum \delta(\tilde{z}^i(z) - y^i) \prod \delta(z^k(z) - y^k) \frac{\partial \tilde{z}^i}{\partial z^i}(z) a(z) \tilde{\omega}(z) dz.$$  

$$= \tilde{z}_j \int \delta(\tilde{z}^i - y^i) \prod \delta(z^k - y^k) a(z) \tilde{\omega}(z) dz$$

where the second line was achieved through integration by parts. The third line was achieved by noting that $\tilde{\omega}$ is the density of the steady state so $\tilde{z}(z) = z$ for all $z$ in its support and exploiting that $\frac{\partial \tilde{z}^i}{\partial z^i}(z)$ is both independent of $z$ and diagonal. We can also compute the density of $(\partial \tilde{\Omega} \cdot \Delta^j)(y)$ by applying the derivative $\frac{\partial^n}{\partial y^i \partial y^i \cdots \partial y^i}$ which gives

$$\tilde{z}_j \frac{\partial}{\partial y^i} \int \prod \delta(z^k - y^k) a(z) \tilde{\omega}(z) dz = \tilde{z}_j \frac{\partial}{\partial y^i} (a(y) \tilde{\omega}(y)).$$

We conclude that $\partial \tilde{\Omega} \cdot \Delta^j = \tilde{z}_j \Delta^j$.

Evaluating the Frechet derivative of (44) in this particular direction $\Delta^j$ we find

$$(\bar{F}_x - (z) + \bar{F}_x(z) + \bar{F}_x(z) x z p + z^j \bar{F}_x(z)) \partial \bar{x}(z) \cdot \Delta^j + \bar{F}_x(z) \bar{x}_\Lambda(z) \mathbf{P} \partial \bar{X} \cdot \Delta^j + \bar{F}_x(z) \partial \bar{X} \cdot \Delta^j = 0.$$  

Solving for $\partial \bar{x}(z) \cdot \Delta^j$ we conclude that

$$\partial \bar{x}(z) \cdot \Delta^j = -(\bar{F}_x - (z) + \bar{F}_x(z) + \bar{F}_x(z) x z p + z^j \bar{F}_x(z))^{-1} (\bar{F}_x(z) \bar{x}_\Lambda(z) \mathbf{P} + \bar{F}_x(z)) \partial \bar{X} \cdot \Delta^j$$

$$\equiv C^j(z) \partial \bar{X} \cdot \Delta^j.$$  

Taking the derivative of $R$ along this direction we get

$$\partial \bar{X} \cdot \Delta^j = \left( \tilde{R}_x \int C^j(z) \bar{d} \Omega(z) + \tilde{R}_x + \bar{X} \mathbf{P} + \tilde{X} \right)^{-1} \tilde{R}_x \int \bar{x}(z) d \Delta^j(z)$$

$$\equiv (D^j)^{-1} \tilde{R}_x \int \bar{x}(z) d \Delta^j(z).$$

For generality we have written this as one value of $C$ for each individual state, i.e. $C^j$. In fact, one only needs one value for each level of $\tilde{z}_j$ which in our case is two: 1 and $\rho_\theta$. 

9
From the definition of $\Delta^j$, we can use integration by parts to find that

$$\partial \mathbf{X} \cdot \Delta^j = (D^j)^{-1} \bar{R}_x \int \bar{x}_{z,j}(z)a(z)d\Omega(z).$$

A.2.2 Expansion along the path

We will use the derivatives of the state variables at the end of the transition path to evaluate our expansion along the path using backward induction. This approach is recursive, so we'll compute the derivatives at $(\bar{\Omega}^n, \bar{\Lambda}^n, \bar{\Theta}^n)$ assuming derivatives at period $n+1$ of the transition are known.

Differentiating (44) and (45) with respect to $\Lambda$ we obtain

$$\bar{F}_{x_{-}}^n(z)\bar{x}_{\Lambda}^n(z) + \bar{F}_{x_{+}}^n(z)\bar{x}_{\Lambda}^n(z) + \bar{F}_{x_{+}}^n(z)\bar{x}_{\Lambda}^{n+1}(z)P \bar{X}_{\Lambda}^n + \bar{F}_X(z)\bar{X}_{\Lambda}^n = 0$$

and

$$\bar{R}_{x}^n \int \bar{x}_{\Lambda}^n(z)d\Omega(z) + \bar{R}_{x}^n \bar{X}_{\Lambda}^n + \bar{R}_{x}^n \bar{X}_{\Lambda}^{n+1}P \bar{X}_{\Lambda}^n + \bar{R}_{\Lambda}^n = 0.$$ 

As both $\bar{x}_{\Lambda}^{n+1}(z)$ and $\bar{X}_{\Lambda}^{n+1}$ are already known we can solve for $\bar{x}_{\Lambda}^n(z)$ to find

$$\bar{x}_{\Lambda}^n(z) = \left(\bar{F}_{x_{-}}^n(z) + \bar{F}_{x_{+}}^n(z)\right)^{-1} \left(\bar{F}_{x_{+}}^n(z)\bar{x}_{\Lambda}^{n+1}(z)P + \bar{F}_X(z)\right) \bar{X}_{\Lambda}^n.$$ 

and, therefore, $\bar{X}_{\Lambda}^n$ equals

$$-\left(\bar{R}_{x}^n \int (\bar{F}_{x_{-}}^n(z) + \bar{F}_{x_{+}}^n(z))^{-1} (\bar{F}_{x_{+}}^n(z)\bar{x}_{\Lambda}^{n+1}(z)P + \bar{F}_X(z)) d\Omega(z) + \bar{R}_{x}^n + \bar{R}_{x}^n \bar{X}_{\Lambda}^{n+1}P \right)^{-1} \bar{R}_{\Lambda}^n$$

Differentiating with respect to $\Theta$ we find

$$\bar{F}_{x_{-}}^n(z)\bar{x}_{\Theta}^n(z) + \bar{F}_{x_{+}}^n(z)\bar{x}_{\Theta}^n(z) + \bar{F}_{x_{+}}^n(z) \left(\bar{x}_{\Theta}^{n+1}(z)\rho_{\Theta} + \bar{x}_{\Lambda}^{n+1}(z)P \bar{X}_{\Theta}^n\right) + \bar{F}_X(z)\bar{X}_{\Theta}^n + \bar{F}_{\Theta}(z) = 0.$$ 

This yields a linear equation in $\bar{x}_{\Theta}^n$ and $\bar{X}_{\Theta}^n$ which we can solve for $\bar{x}_{\Theta}^n$ as a linear function of $\bar{X}_{\Theta}^n$. Plugging in for the linear relationship between $\bar{x}_{\Theta}^n$ and $\bar{X}_{\Theta}^n$ in

$$\bar{R}_{x}^n \int \bar{x}_{\Theta}^n(z)d\Omega(z) + \bar{R}_{X}^n \bar{X}_{\Theta}^n + \bar{R}_{X}^n \bar{X}_{\Theta}^{n+1}\rho_{\Theta} + \bar{R}_{X}^n \bar{X}_{\Theta}^{n+1}P \bar{X}_{\Theta}^n + \bar{R}_{\Theta}^n = 0$$

yields a linear equation that can be solved for $\bar{X}_{\Theta}^n$.

To compute the Frechet derivative, we'll evaluate the derivative in direction $\Delta^{j,n}$ with density of the form

$$\frac{\partial}{\partial y_j} (a^n(y)\bar{\omega}^n(y))$$

where $\bar{\omega}^n$ is the density of $\bar{\Omega}^n$ and $a^n(y)$ is some arbitrary function. For this derivative we
Similarly, differentiating \( \bar{R}_x \) generates

\[
R^n_x \int \partial \bar{x}^n(z) \cdot \Delta \cdot \Omega + z^j \bar{R}^n_x \partial \bar{X}^n \cdot \Delta \cdot \Lambda + \bar{X}^n \cdot \partial \bar{x}^n \cdot \Delta = 0
\]
In order to proceed, we need to determine the density of the transition path at time 1, \( \bar{\Omega}(z) \), where \( \bar{\Omega} \) is therefore defined as the measure with density

\[
\bar{\omega}(z) = -\sum_j \frac{\partial}{\partial y} \int \prod_i \delta(\hat{z}^i(z) - y^i) \hat{z}_E^i(\hat{z}(z)) \bar{\omega}(z)dz = -\sum_j \frac{\partial}{\partial y} \left( \hat{z}_E^j(\bar{z}^{-1}(y)) \bar{\omega}(y) \right) \equiv \sum_j \bar{\omega}_x^{0,j,1},
\]

where 1 here represents that the objects are evaluated using the density of the transition path at time 1, \( \bar{\omega}(y) \). If we define \( \bar{\Omega}_x^{0,j,1} \) as the measure with density

\[
\bar{\omega}_x^{0,j,1}(y) = -\frac{\partial}{\partial y^i} \left( \hat{z}_E^j(\bar{z}^{-1}(\cdots \bar{z}^{-1}(y)) \bar{\omega}(y) \right),
\]

then

\[
\partial \hat{z}(z) \cdot \bar{\Omega}_x^{0,j,1} = \sum_{k=0}^{N-1} C_k^j(z) \partial \hat{X}^{1+k} \cdot \bar{\Omega}_x^{0,j,1+k} = \sum_{k=0}^{N-1} C_k^j \hat{X}_x^{1+k}.
\]

Combined with (47) gives a linear system

\[
M(z) \hat{X}_x^0(z) = N(z) \left[ I \hat{X}_x^{1,1} \hat{X}_x^{2,1} \ldots \hat{X}_x^{n,N} \right]^T
\]
and which can be solved for $\bar{x}_0^0(z)$. To find $X_{J,n}^{i,n}$, we note that they satisfy the equation

$$X_{J,n}^{i,n} = - (D^{J,n})^{-1} \left( R^n_{xJ} + R^n_{x} \int \left[ x_{zJ}^n(z) \bar{x}_0^0 \left( \bar{\varepsilon}^{-1}(\ldots \bar{\varepsilon}^{-1}(z)) \right) \right] d\Omega^n(z) + \sum_{k=1}^{N-n} E_{J,n}^{i,n} X_{J,n}^{i,n+k} \right).$$

Combining the previous equation with

$$\bar{R}_x \int \bar{x}_0^0(z) d\Omega^0(z) + \bar{R}_X X_0^0 + \bar{R}_X^0 (X_\Theta + X_{\bar{\varepsilon}}^1 + X_{\bar{\varepsilon}}^1 + X^0 P X^0) + \bar{R}_x^0 = 0$$

yields a linear system

$$O \cdot \begin{bmatrix} X_0^0 & X_{J,n}^{1,1} & X_{J,n}^{2,1} & \cdots & X_{J,n}^{r,n,N} \end{bmatrix}^T = P$$

which can be solved for $X_0^0$.

The term $\bar{x}_0^0(z)$ satisfies

$$\bar{x}_0^0(z) = (F_{x}^0(z) + F_{x+}^0(z) \bar{x}_0^0(z)) p^{-1} F_{x}^0(z).$$

**A.2.3 An Alternative Approximation**

In this section we present the alternative approach highlighted in section 3.2 where we scale \{\sigma, \sigma, \sigma, \sigma, \sigma, \sigma, \sigma\} and expand with respect to \sigma instead of just \{\sigma, \sigma\}.

For this approach, the full policy function for $X$ can be written as $\bar{X}(\Omega(\sigma), \Lambda, \sigma\Theta, \sigma\varepsilon; \sigma)$ where $\Omega(y; \sigma)$ incorporates the fact that we are scaling $\theta$ with $\sigma$ and therefore also scaling $\Omega$. Formally, we have (assuming the simplest case where $m, \mu$ and $\theta$ are the only individual state variables for our problem)

$$\Omega(y; \sigma) = \int \nu(m \leq y_1) \nu(\mu \leq y_2) \nu(\sigma \theta \leq y_3) d\Omega(m, \mu, \theta).$$

The same proof can be used to show that Lemma 2 holds for this approximation as well. There still may be transition dynamics with respect to $\Lambda$, at which point it will be necessary to follow sections A.2.1 and A.2.2 to compute the relevant derivatives for the expansion.\(^{37}\)

$\bar{X}$ and $\bar{x}$ can then be approximated using Taylor expansions with respect to $\sigma$. For brevity we only report the first order expansion of $\bar{X}$ which given by

$$\bar{X}(\Omega(\sigma), \Lambda, \sigma\Theta, \sigma\varepsilon; \sigma) = \bar{X}^0 + \sigma (\partial \bar{X}^0 \cdot \bar{\Omega}_\sigma + \bar{X}_\Theta^0 \cdot \bar{\Theta}_\sigma + \bar{X}_\varepsilon^0 \cdot \bar{\varepsilon}_\sigma) + O(\sigma^2).$$

\(^{37}\)In the case where there is no endogenous aggregate state variable only section A.2.1 is required.
To obtain \( \partial \mathbf{X} \cdot \Omega_{\sigma} \), we differentiate (48) with respect to \( \sigma \) to obtain

\[
\Omega_{\sigma}(y) = - \int \nu(m \leq y_1) \nu(\mu \leq y_2) \delta(0 - y_3) \theta d\Omega(m, \mu, \theta).
\]

The density of this object is constructed by applying the derivative \( \partial^3 / \partial y_1 \partial y_2 \partial y_3 \) to get

\[
\omega_{\sigma}(y) = - \frac{\partial}{\partial y_3} \left( \int \delta(m - y_1) \delta(\mu - y_2) \delta(0 - y_3) \theta d\Omega(m, \mu, \theta) \right)
= - \frac{\partial}{\partial y_3} \left( \delta(0 - y_3) \int \omega(y_1, y_2, \theta) \theta d\theta \right)
= - \frac{\partial}{\partial y_3} \left( E\theta(y_1, y_2) \omega(y) \right)
\]

where in the last equality we defined \( E\theta(y_1, y_2) = \int \omega(y_1, y_2, \theta) \theta d\theta \) as the cross-sectional mean of \( \theta \) conditional on \((m, \mu) = (y_1, y_2)\). From this expression, we know that \( \partial \mathbf{X}_0 \cdot \Omega_{\sigma} \) can be solved for in the same manner as \( \partial \mathbf{X}_0 \cdot \Omega^0_{E} \) using the tools in section (A.2.2).

A.2.4 Simulation and Clustering

To simulate an optimal policy at each date with \( N \) agents, we discretize the distribution across agents with \( K \) grid points that we find each period using a k-means clustering algorithm. Let \( \{z_i\}_{i=1}^N \) represent the current distribution of agents. The k-means algorithm generates \( K \) points \( \{\bar{z}_k\}_{k=1}^K \) with each agent \( i \) assigned to a cluster \( k(i) \) to minimize the squared error \( \sum_i \|z_i - \bar{z}_{k(i)}\|^2 \). We let \( \Omega \) represent the distribution of \( N \) agents and \( \bar{\Omega} \) represent our approximating distribution of clusters.\(^{38}\) At each history, we compute \( \bar{\Omega} \) and then apply our algorithm to approximate the optimal policies around \( \bar{\Omega} \).\(^{39}\) When \( K = N \) we exactly approximate around \( \Omega \), but for \( K < N \) we can speed up the computations by a factor of \( \frac{N}{K} \).

A.2.5 Solving the \( t = 0 \) problem

For the Ramsey problem (21), optimality conditions at \( t = 0 \) are different from \( t \geq 1 \). The full set of optimality conditions are represented by expanding equations (22)–(24). We describe how to apply our procedure for the section 3.1 simple case. The extension to the general problem in section 2 is straightforward.

We start with some notation. Let \( \Omega^B \) be a measure over the claims to risk-free debt. Denote the \( t = 0 \) aggregate policy functions as \( \tilde{X}_0(\Omega^E, \epsilon_0) \) and individual policy functions

---

\(^{38}\)Formally \( \Omega(z) \) has density \( \sum_i \frac{1}{N} \delta(z - z_i) \) while \( \bar{\Omega}(z) \) has density \( \sum_i \frac{1}{K} \delta(z - \bar{z}_{k(i)}) \).

\(^{39}\)Similar to section (A.2.3) this is done by constructing a distribution \( \Omega(\sigma) \) with density \( \sum_i \frac{1}{N} \delta(z - \bar{z}_{k(i)} - \sigma(z_i - \bar{z}_{k(i)})) \) and then computing \( \partial \mathbf{X}^0 \cdot \bar{\Omega}_\sigma \) in the same manner as section (A.2.3).
as $\tilde{x}_0(b, \Omega^B, \varepsilon_0, \mathcal{E}_0)$. Augment the system (22)–(24) with mappings $F_0$ and $R_0$, capturing the time 0 first order conditions, such that

$$F_0\left(\tilde{x}_0, \varepsilon, \dot{X}_0, \varepsilon_0, \mathcal{E}_0, b_0\right) = 0$$  \hspace{1cm} (49)

$$R_0 \left(\int \tilde{x}_0 d\Omega^B, \dot{X}_0, \mathcal{E}_0\right) = 0$$  \hspace{1cm} (50)

Policy functions for $t \geq 1$ individual states $z_0 = (m_0, \mu_0)$ are components of $\tilde{x}_0$. Let function $\Omega_0 (\Omega^B, \mathcal{E}_0)$ map the initial condition $\Omega^B$ and aggregate shock $\mathcal{E}_0$ to a measure $\Omega$ over $z$ using

$$\Omega_0 (\Omega^B, \mathcal{E}_0) (z) = \int \iota (z_0 (y, \Omega^B, \varepsilon, \mathcal{E}_0) \leq z) d \Pr (\varepsilon) d\Omega^B (y) \ \forall z$$  \hspace{1cm} (51)

Section 3.1 characterizes the small-noise approximations of the $t \geq 1$ policy functions around an arbitrary $\Omega$. We update $\Omega$ along the path by iterating between an approximation and a simulation step. At some $t \geq 1$, taking as input $\Omega_{t-1}$, we draw idiosyncratic shocks $\varepsilon$ for each agent as well as aggregate shocks $\mathcal{E}$, and use the policy functions approximated around $\Omega_{t-1}$ to move to the next period $\Omega_t$. All that remains to be specified is how the $t = 1$ state, $\Omega_0$, is obtained. We do that below by constructing small-noise approximations to $t = 0$ policy functions: $\dot{X}_0(\Omega^B, \sigma \mathcal{E}_0; \sigma)$ and $\tilde{x}_0(b, \Omega^B, \sigma \varepsilon_0, \sigma \mathcal{E}_0; \sigma)$. We present a first order expansion. Higher order expansions along the lines of A.1 are analogous.

1. Zeroth-order: For some choice of $\Omega^B$, the $\sigma = 0$ allocation consists of $\{\tilde{x}_0(b), \tilde{x}(b)\}$ for $b$ in support of $\Omega^B$ as well as $\{\dot{X}_0, \dot{X}\}$ such that

$$F_0 \left(\tilde{x}_0, \tilde{x}, \dot{X}_0, 0, 0, b_0\right) = 0, \quad R_0 \left(\int \tilde{x}_0 (b) d\Omega^B (b), \dot{X}_0, \mathcal{E}_0\right) = 0$$

$$F \left(\tilde{x}, \tilde{x}, \dot{X}, 0, 0, \tilde{x}\right) = 0, \quad R \left(\int \tilde{x} (b) d\Omega^B (b), \dot{X}, 0\right) = 0$$

2. To compute derivatives $\{\tilde{x}_{0,b} (b), \tilde{x}_{0,\varepsilon} (b), \tilde{x}_{0,\sigma} (b), \dot{X}_{0,\varepsilon}, \dot{X}_{0,\sigma}\}$, we use the formulas from section (A.2.2). The expressions that appear in section (A.2.2) use superscript $n$ to denote the period of transition path for the $\sigma = 0$ allocation. We can obtain $\{\tilde{x}_{0,b} (b), \tilde{x}_{0,\varepsilon} (b), \tilde{x}_{0,\sigma} (b), \dot{X}_{0,\varepsilon}, \dot{X}_{0,\sigma}\}$ by using those formulas after replacing $F_0^0$ with $F_0^0, F_n^0$ with $F$, for $n \geq 1$ and similarly for $R_0^0$ and $R_n^0$.

3. Simulation: Draw idiosyncratic shocks $\varepsilon_0$ for each agent as well as aggregate shocks $\mathcal{E}_0$ and use the approximations to policy functions

$$\dot{X}_0(\Omega^B, \sigma \mathcal{E}_0; \sigma) = \dot{X}_0 + \sigma (\dot{X}_{0,\varepsilon} \mathcal{E}_0 + \dot{X}_{0,\sigma}) + \mathcal{O}(\sigma^2)$$
and
\[
\tilde{x}(b, \Omega^B, \sigma \varepsilon_0, \sigma \mathcal{E}_0; \sigma) = \bar{x}_0(b) + \sigma \left( \bar{x}_{0,\varepsilon}(b) \varepsilon_0 + \bar{x}_{0,\mathcal{E}}(b) \mathcal{E}_0 + \bar{x}_{0,\sigma}(b) \right) + \mathcal{O}(\sigma^2)
\]
to obtain the $\Omega_0(z_0)$ for $t = 1$.

A.3 Additional details for section 3.3

In this section, we provide more details concerning Acharya and Dogra (2018) “PRANK” economy, which we use as a laboratory to test the accuracy of our algorithm and compare it to alternative methods. We start with equilibrium conditions, next we discuss the calibration, and report the accuracy tests for our method. Finally, we present a simplified version of this economy for which we can solve for all gradients in closed form.

Equilibrium in the PRANK economy

To obtain the PRANK setting we impose the following assumptions: (i) labor is supplied inelastically, and period utility function $U(c_t, n_t) = -\exp(-\gamma c_t)$, (ii) the distribution of shares is uniform, (iii) Idiosyncratic productivity shocks are i.i.d, and (iv) All tax rates are constant and the monetary policy follows a Taylor rule given by
\[
Q_t^{-1} - 1 = a_0 (1 + \Pi_t)^{a_1}
\]  
(52)

In the PRANK economy, a perfect foresight equilibrium is constructed as follows. For a sequence of innovations to TFP $\{\mathcal{E}_{t,s}\}_{s=0}^T$, Agent $i$’s consumption $c_{i,t}$ satisfies
\[
c_{i,t} = C_t + \mu_t \left( \frac{b_{i,t-1}}{1 + \Pi_t} + y_{i,t} \right),
\]  
(53)
where $y_{i,t} = (1 - \Upsilon_t) W_t \epsilon_{i,t} n_{i,t} + T_t + d_{i,t}$ is the households income at date $t$. The two parameters $C_t$ and $\mu_t$ that are common to all agents are given by
\[
\mu_t = \frac{\mu_{t+1} \left( \frac{Q_t}{1 + \Pi_{t+1}} \right)}{1 + \mu_{t+1} \left( \frac{Q_t}{1 + \Pi_{t+1}} \right)}
\]  
(54)
\[
\mathcal{C}_t \left[ 1 + \mu_{t+1} \left( \frac{Q_t}{1 + \Pi_{t+1}} \right) \right] = \frac{1}{\gamma} \ln \beta \left( \frac{Q_t}{1 + \Pi_{t+1}} \right) + C_{t+1} + \mu_{t+1} \bar{y}_{t+1} - \frac{\gamma \mu_{t+1}^2 \sigma_{y,t+1}^2}{2}
\]  
(55)
where $\bar{y}_{t+1} = \int y_{i,t+1} di$ is the average household income, and $\sigma_{y,t+1}^2$ is the variance in household level income. A perfect foresight equilibrium can by solving equations (52)–(55) along with equations (8), (10),(17), and (18).

We check the accuracy of our approximations using an exact solution to the perfect foresight equilibrium. Section (A.3) discusses how we calibrate the PRANK economy, section
(A.3) and (A.3) compare approximation errors using several diagnostics.

**Calibration** We study several cases. For the parameters that are common across these cases we use Acharya and Dogra (2018) targets which are quite standard in the representative agent New Keynesian literature. The discount rate $\beta$ to 0.96 to get a real rate of 4% per year, the elasticity of substitution parameter, $\Phi$, to 6 to target an average markup of 20%. The share of intermediate inputs $\alpha$, is set to 0.6 to target a labor income share of $2/3$, and we set the adjustment cost parameter $\psi$ to 41.6 to target a slope of the Phillips curve of 0.06. Aggregate productivity follows an AR(1) process with a decay parameter 0.73, and the standard deviation of the innovation is set to 1.23% to be consistent with de trended output per hour and we turn off the markup shocks. For the Taylor rule parameters, we set $a_1 = 1.5$ and choose $a_0$ to target 0% inflation rate in absence of aggregate risk. We vary the standard deviation of idiosyncratic risk, $\sigma_\epsilon \in \{0.5, 0.75, 1\}$, and the risk aversion parameter, $\gamma \in \{1, 3\}$. Our calibrations cover a range that includes Acharya and Dogra (2018) as well as what we use in our baseline section 4. Since the distribution of assets is non-stationary, we set $\Omega_0(b)$ to be Gaussian and calibrate the parameters to be consistent with the distribution of wealth in the SCF. For simulation, we approximate the distribution with 150 points and the idiosyncratic shocks with 10 point Gaussian quadrature.

**Diagnostics** In this section, we compare the accuracy of our policy functions in two settings. We start with a stationary environment with no aggregate risk and study the policy function for individual consumption as well as values for the aggregate variables. Then, we study impulse responses of several aggregate variables to a TFP shock. As mentioned before, the advantage of PRANK is that in both cases, the true solution can be solved for exactly.

We report three types of approximation errors for the individual policy functions: (i) % gap in the approximated and true policies (ii) % gap in the approximated and policy rules implied by the Euler equation (or the Euler equation errors) and (iii) dynamic Euler Equation errors from Den Haan (2010). The first two errors are acquired from the formulas:

$$E_{c,t}^{pol}(b, \epsilon) \equiv 100 \times \frac{|c_{t}^{True}(b, \epsilon) - c_{t}^{approx}(b, \epsilon)|}{C_t}$$

$$E_{c,t}^{EE}(b, \epsilon) \equiv 100 \times \frac{|c_{t}^{imp}(b, \epsilon) - c_{t}^{approx}(b, \epsilon)|}{C_t},$$

As CARA preference feature an aversion to absolute risk and, possibly, negative consumption we report the absolute errors scaled by average consumption.
where \( c_t^{\text{True}}(b, \epsilon) \) are constructed using the exact solution and \( c_t^{\text{imp}}(b, \epsilon) \) is defined as
\[
c_t^{\text{imp}}(b, \epsilon) = -\frac{1}{\alpha} \log \left( \frac{\beta}{Q_t} \mathbb{E}_t \left[ \frac{1}{1 + \pi_{t+1}} \exp(-\alpha c_{t+1}^{\text{approx}}(b', \epsilon')) \right] \right).
\]

The dynamic Euler equation errors are constructed by simulating an alternative series \( \tilde{b}_t \) and \( \tilde{c}_t \), for a group of agents, derived using the budget constraint and \( c_t^{\text{imp}} \) evaluated at \( \tilde{b}_{t-1} \). We then compare to \( c_t \) simulated using \( c_t^{\text{approx}} \) and the same sequence of shocks. These errors have the advantage of allowing for the possibility of small errors accumulating into large errors over time. We denote them as \( E_{\text{dynEE}}(c_t, b, \epsilon) \). For aggregate variable \( X_t \) we simply report
\[
E_{X,t} = 100 \times \frac{|X_t^{\text{True}} - X_t^{\text{approx}}|}{X_t^{\text{True}}}.
\]

We will often report maximum errors where the max is taken over state space \((b, \epsilon)\) as well as \(t\).

For all the experiments we use a second-order approximation of our method. As a point of comparison, we report the errors when policy functions are approximated using the Reiter-approach (also used in Acharya and Dogra (2018)) in which the no-aggregate risk economy is solved exactly and then the policy functions are linearized with respect to aggregate shocks. In all our plots, our method will be represented by a bold blue line, the Reiter approximation will be represented by a dashed black line and the exact solution will be a bold black line.

We begin with the approximation errors turning off aggregate risk. By construction, the errors for the Retier-method are zero and so we report the error diagnostics just for our method for several values of \( \{\sigma_{\epsilon}, \gamma\} \) in table 1. The maximum percent errors in the individual policy rules for consumption relative to the exact solution are small starting at 0.0039\% for our baseline calibration and raising to only 0.0328\% when we double the size of the idiosyncratic risk. The Euler equation errors are comparable. We see a similar pattern in the errors for the aggregate variables, though the errors for those are an order of magnitude smaller.

Next we compare the errors in the policy functions in response to an one time one standard deviation unanticipated shock to aggregate productivity. We report these errors in table 2. For the individual consumption policy functions, the maximum errors (across the state space \((b, \epsilon)\) and across time \(t\)) for our second-order approach are comparable to the Reiter-based method. In fact, while the Euler equation errors \( E_{\text{EE}}(c_t, b, \epsilon) \) for the Reiter method are generally smaller than our second-order approximation, the errors relative to the exact solution \( E_{\text{pol}}(c_t, b, \epsilon) \) are an order of magnitude larger (0.0039\% vs 0.0404\%). The diagnostic errors \( E_{\text{pol}}(c_t, b, \epsilon) \) clearly captures errors coming from aggregate shocks that are not reflected in the Euler equation errors. We also see that the dynamic Euler equation errors
remain small and comparable to those of the Reiter-based approach, which indicates that
one should not be too concerned with errors accumulating over time.

Tables 1 and 2 also report errors for the alternative calibrations where we increase risk
aversion, $\gamma$, to 3. Not surprisingly increasing risk aversion leads to the largest policy errors
for both our second-order approximation as well as the Retier-based methods, but the policy
errors remain small and are comparable to those from the Retier-based approach. Figure 1
reproduces the impulse responses in figure I of the main text for $\gamma = 3$. For larger values of
risk aversion, we see a visible deviation of the Retier-based approach from both the exact
solution and our second-order approximations. The visible deviation reflects the errors in
aggregates of the Retier-based approach documented in Table 2.
Figure 1: Comparisons for impulse responses to a 1% TFP shock at $t = 1$ in the top panel and $t = 250$ in the bottom panel.

**Long run errors**  In the PRANK economy, individual assets follow an approximate random walk and, therefore, the distribution of individual savings drifts over time. Since our method approximates with respect to the size of idiosyncratic risk, a diagnostic for whether small errors at a point in time accumulate to large error over time, we check how well the approximated distribution of assets tracks the true distribution with our method as well as with the Reiter method.

In figure 2, we plot the distribution of assets obtained at $t = 250$ after a one-standard deviation shock at $t = 0$. We see the second-order approximation lines up very closely with the Reiter-based method and to the outcomes from the exact solution. This figure also explains the finding in section 3.3, why our method captures the the response of inequality to an unanticipated TFP shock $t = 250$ so well.

We next compare the distribution of assets after a sequence of TFP shocks in a stochastic PRANK economy. The TFP shocks follow an AR(1) and for this exercise we do not have the true solution in an analytic form. However, we can still compare our second-order approximation and the Reiter-based approach. In addition, we include a “hybrid” method where we take a second-order approximation with respect to idiosyncratic shocks and a first-order approximation with respect to aggregate shocks.

In figure 3, we see that the hybrid and Reiter-based approaches produce nearly identical distributions after these shocks, but the second-order approach delivers a tighter distribution.
Figure 2: Distribution of assets at $t = 250$ following a one time unanticipated TFP shock at $t = 0$. 
over time. As the hybrid approach was obtained by dropping the second-order terms with respect to aggregate shocks, we take this as evidence that, in this model, ignoring those second-order terms can lead long run drift away from the true solution.

**A Simplified Example** To illustrate how our approach from section 3 is applied to the PRANK economy, we present a version of the PRANK economy where we can explicitly show how to compute all the gradients that appear that section. We assume aggressive enough monetary policy to ensure $\Pi_t = 0$ for all $t$; that share of intermediate inputs, $1 - \alpha$, is 0 which ensures that output is linearly related to productivity; and finally that $\Phi \to \infty$ to ensure that there are no markups and dividends. The environment is similar to the well known Huggett (1993) model and can be trivially solved with standard methods; we use it to illustrate transparently how to construct all the objects that appear in Section 3.

The economy is populated with a continuum of infinitely lived consumers who receive endowment shocks. Let $e_{i,t}$ be endowment of consumer $i$ in period $t$. Endowments are subject to aggregate shock $\mathcal{E}_t$ and idiosyncratic shock $\varepsilon_{i,t}$ and satisfy

$$e_{i,t} = 1 + \varepsilon_{i,t} + \mathcal{E}_t.$$
Shocks \( \mathcal{E}_t \) and \( \varepsilon_{i,t} \) are mean zero and i.i.d. over time.

Competitive equilibrium in this economy is fully characterized by consumer budget constraint and the Euler equation

\[
\begin{align*}
c_{i,t} + Q_t b_{i,t} - 1 - \varepsilon_{i,t} - \mathcal{E}_t - b_{i,t-1} &= 0 \\
Q_t \exp(-\gamma c_{i,t}) - \beta \mathbb{E}_{i,t} \exp(-\gamma c_{i,t}) &= 0
\end{align*}
\]
as well as the feasibility

\[
\int c_{i,t} di - 1 - \mathcal{E}_t = 0.
\]

We now show how to use our approximation techniques in this simple example to find competitive equilibrium. To make it similar to our notation in Section 3, let \( y = \exp(-\gamma c) \) and re-write this problem as

\[
\begin{align*}
c_{i,t} + Q_t b_{i,t} - 1 - \varepsilon_{i,t} - \mathcal{E}_t - b_{i,t-1} &= 0 \\
Q_t y_{i,t} - \beta \mathbb{E}_{i,t} y_{i,t+1} &= 0 \\
y_{i,t} - \exp(-\gamma c_{i,t}) &= 0
\end{align*}
\]

The first pair of equations correspond to (22) and define mapping \( F \), the last equation corresponds to (23) and define \( R \). This problem is recursive in the distribution of agents’ assets. In our notation of Section 3 we have \( z = b \) and \( \Omega \) is the distribution of \( b \) such that \( \int b d\Omega = 0 \). Vector \( \tilde{x} \) of individual policy functions is given by three policy functions \( \begin{bmatrix} \tilde{b} & \tilde{c} & \tilde{y} \end{bmatrix}^T \), and \( \tilde{Q} \) is the only aggregate policy function in vector \( X \). Selection matrix \( p \) is simply \( \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \).

It is immediate to verify that without shocks consumption smoothing implies that \( \tilde{b}(b) = b \) for all \( b \), so that Lemma 1 holds and equation (27) become

\[
\begin{align*}
\tilde{c}(b) + \tilde{Q} \tilde{b}(b) - 1 - b &= 0 \\
\tilde{Q} \tilde{y}(b) - \beta \tilde{y} \left( \tilde{b}(b) \right) &= 0 \\
\tilde{y}(b) - \exp(-\gamma \tilde{c}(b)) &= 0
\end{align*}
\]

and

\[
\int \tilde{c}(b) d\Omega - 1 = 0,
\]

which immediately gives \( \tilde{Q} = \beta \) and \( \tilde{c}(b) = 1 + (1 - \beta) b \), \( \tilde{y}(b) = \exp(-\gamma \tilde{c}(b)) \). From these
we construct mappings $R_x = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}, R_X = 0, R_\varepsilon = -1$, and $F_{x-}(b) = 0$,

\[
F_x(b) = \begin{bmatrix} \tilde{Q} & 1 & 0 \\ 0 & 0 & \tilde{Q} \\ 0 & -\gamma \exp(-\gamma \tilde{c}(b)) & 1 \end{bmatrix}, \quad F_{x+}(b) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -\beta \\ 0 & 0 & 0 \end{bmatrix}, \quad F_X(b) = \begin{bmatrix} b \\ \tilde{y}(b) \\ 0 \end{bmatrix},
\]

\[
F_\varepsilon(b) = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \quad F_\varepsilon(b) = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \quad F_z(z) = \begin{bmatrix} -1 \\ 0 \end{bmatrix}.
\]

All elements of these matrices are known from the zeroth order expansion. Using them, we construct first order approximations of policy functions as described in the text.

B Additional details for section 4

In this section we provide details of how we calibrate the initial distribution of nominal and real claims using the Doepke and Schneider (2006) procedure. Then we show the dynamics of the calibrated competitive equilibrium using simulations.

**Initial distribution of nominal and real claims** We combine the rich house-level data on financial assets from the Survey of Consumer Finances (SCF) and the aggregated Flow of Funds for intermediate investors to obtain nominal and real exposures. We start with the 2007 Wave of the SCF and restrict our sample to married households who work at least 100 hours. We drop observations where equity or bond holdings are more than 100 times the average yearly wage. These turned out to be about 0.5% of the total sample. We extract household-level data on their financial holdings and categorize them into (i) deposits, government bonds, liquid assets (net of unsecured credit), (ii) direct holdings of claims to corporate equities and corporate bonds, and (iii) indirect holdings of (i) and (ii) through mutual funds and retirement accounts.

We then use Flow of Funds data to obtain balance sheet information for private pensions (Table L.118), for state and local pensions (Table L.119) and mutual funds (Table L.122). Since pension funds have a nontrivial exposure to mutual funds and not vice versa, we start with the aggregated mutual fund balance sheet and map it into broad categories that represent deposits, corporate bonds, government bonds, corporate equities. In the year 2007, mutual funds invested 84 percent of their assets in corporate equities and bonds, 16 percent in government bonds, and other liquid claims.

We next turn pension funds and after aggregating private and public pension funds categorize the combined assets into deposits, government-issued debt, corporate debt, corporate
equities, and mutual funds. For the year 2007, the pension funds assets were invested 22 percent in mutual funds, 63 percent in corporate equities and bonds, and rest 15 percent in government bonds and other liquid claims.

We define nominal claims as money-like assets plus government issued bonds and claims to real profits as corporate bonds plus corporate equities. Using the information above, we first consolidate the mutual funds into these two categories and then reassign the mutual funds to pension funds, and finally the mutual funds and pensions funds to the individuals in the SCF.

To fit initial states, we sample directly from the SCF log wages, nominal claims, and claims to real profits that we constructed. The SCF provides population weights for each observation. Given these weights, we set the initial condition by drawing with replacement a random sample of 100000 agents from a discrete distribution.

Properties of the competitive equilibrium In this section, we report several moments from our calibrated competitive equilibrium along a transition path. We draw a sequence of markup and TFP shocks of length 100 and simulate the competitive equilibrium policies using 100000 agents. When we simulate the competitive equilibria, we keep the tax rates $\gamma_t = \bar{\gamma}$ and use a Taylor rule with $Q_t^{-1} = \frac{1}{\beta} \Pi_t^{2.5}$.

In figure 4, we plot the time series for aggregate output and labor share. We see that the aggregates are quite stationary and exhibit small fluctuations due to productivity and markup shocks. In the table 3 we report cross-sectional moments at dates $t \in \{10, 25, 50, 75\}$. Here we notice a small drift in the distribution of the risk-free assets. The more significant drifts are in the correlations of log wages and dividend shares as well as risk-free assets and dividend shares which steadily declines over time and the correlation between bonds and log wages that increase over time. These patterns are the outcomes of the features in the baseline that claims to equity are not traded and households are subject to natural debt limits.

C Additional details for section 5

In the main body, we focused on the results under the baseline calibration and briefly discuss sensitivity checks and special cases. In this section, we provide all the omitted details.

C.1 Sensitivity with respect to price adjustment costs

In this section, we present the impulse responses under alternative choices for the price adjustment cost parameter $\psi$. As mentioned in section 5.1.1, we vary $\psi$ from twice the baseline calibration to half of the baseline calibration, and also when $\psi$ is approximately
Table 3: Distributional Moments Along the Path

<table>
<thead>
<tr>
<th>Moments</th>
<th>DATA</th>
<th>MODEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Std. share of equities</td>
<td>2.63</td>
<td>2.62</td>
</tr>
<tr>
<td>Std. bond</td>
<td>6.03</td>
<td>6.18</td>
</tr>
<tr>
<td>Std. ln wages</td>
<td>0.80</td>
<td>0.81</td>
</tr>
<tr>
<td>Std. ln hours</td>
<td>0.42</td>
<td>0.42</td>
</tr>
<tr>
<td>Corr(share of equities, ln wages)</td>
<td>0.40</td>
<td>0.37</td>
</tr>
<tr>
<td>Corr(share of equities, bond holdings)</td>
<td>0.62</td>
<td>0.59</td>
</tr>
<tr>
<td>Corr(bond, ln wages)</td>
<td>0.33</td>
<td>0.40</td>
</tr>
</tbody>
</table>

Notes: The data moments correspond to SCF 2007 wave with sample restrictions explained in the text and after scaling wages, equity holdings, and debt holdings by the average yearly wage in our sample. The share of equities refers to the ratio of individual equity holdings to the total in our sample such that the weighted sum of shares equals one. The model columns correspond to simulated sample of 100000 agents using the baseline calibration from section 4.
Figure 5: Optimal monetary response to a markup shock. The bold blue lines are the responses for the baseline calibration. The dashed black lines with squares, circles and triangles are responses under a calibration in which we double the price adjustment costs parameter, half the price adjustment cost parameter, and finally set it near zero, respectively.

As is readily apparent in both figures the effect on inflation is roughly linear for a large range of $\psi$. Doubling $\psi$ leads to a halving of inflation while halving $\psi$ leads to a doubling of inflation. The effect on the nominal rate is quite small. In the limit as $\psi$ approaches zero, the planner can no longer effect real variables through monetary policy and instead relies more on unexpected inflation to provide insurance through the ex-post real return as instead of distorting the allocation by varying the ex-ante real rate.

C.2 Example with perfectly aligned distribution of equity shares

As noted in section 5.1.1, the quantitative driver of the need for insurance concerns against the markup shocks is the misalignment of dividend income from labor income. To illustrate this point, we construct a calibration with non trivial amount of inequality but in which these shares are perfectly aligned. To achieve this, we take the distribution of labor productivities from the benchmark calibration; assume Pareto weights such that optimal tax rates $\bar{\Upsilon}$ equal zero; and then assign dividend shares such that individuals initial share of labor income $\frac{\epsilon_{i,o} a_{i,0}}{J_i \epsilon_{i,o} a_{i,0}}$ equals their share of dividend income $s_i$. Figure 7 plots the optimal response and
Figure 6: Optimal monetary response to a TFP shock. The bold blue lines are the responses for the baseline calibration. The dashed black lines with squares, circles and triangles are responses under a calibration in which we double the price adjustment costs parameter, half the price adjustment cost parameter, and finally set it near zero, respectively.

...its when the shares are aligned. We see that the alignment of shares nearly removes all need for insurance bringing the policy responses in line with those of the representative agent.\textsuperscript{41}

\textsuperscript{41}There is some difference in insurance needs arising from differential labor responses to the markup shock.
Figure 7: Optimal monetary response to a markup shock. The bold blue and red lines are the calibrated HANK and RANK responses respectively. The dashed black lines with circles are responses under HANK when the shares of labor and dividend income are aligned.
Figure 8: Optimal monetary response to a markup shock. The bold blue lines are the responses for the baseline calibration. The dashed black lines with squares and circles are responses under a calibration with higher and lower labor taxes respectively.

C.3 Sensitivity with respect to choice of Pareto weights

Here we present sensitivity to the choice of Pareto weights. As mention in the main text, we set Pareto weights using a thee parameter exponential specification, which loads on the three dimensions of initial heterogeneity and maps to optimal levels of tax rates $\bar{\Upsilon}$, on labor income, dividend income, and bond income. For the purpose of sensitivity, we vary these implied tax rates in a large range: the labor tax from half to double of its baseline value; the dividend tax rate also from half to double its baseline value; and the tax rate on bond from zero to a rate of $50$. In addition, we also study a Utilitarian planner that weights all agents equally.

We start with the experiments that vary the labor income tax rate and the responses are depicted in 8 and 9 to markup and productivity shocks, respectively. We see that the responses to both the shocks are larger when labor tax rates are higher and lower when labor taxes are lower. Raising the labor tax compresses labor shares pushing the economy further away from full insurance, while decreasing the labor tax pushes the economy closer to full insurance. In line with this, we see that the increasing the labor tax leads to a stronger policy response while decreasing the labor tax diminishes the response.

Next we vary the tax on dividend income and report the results in figures 10 and 11. Our baseline calibration exhibits are far more unequal distribution of dividend share than
Figure 9: Optimal monetary response to a TFP shock. The bold blue lines are the responses for the baseline calibration. The dashed black lines with squares and circles are responses under a calibration with higher and lower labor taxes respectively.

labor shares. Increasing the dividend tax brings the economy closer to full insurance while decreasing the dividend tax pushes the economy away from full insurance. As such, we see that the response of inflation and other variables is stronger when the dividend tax is low and weaker when it is high when we look at markup shocks. Since productivity shocks affect wages and dividends symmetrically, we should expect that the responses are not very different across cases that vary in the level of tax on dividend income. This prior is confirmed in figure 11.

On the contrary, a bond tax directly controls the dispersion in after-tax bond income which is key statistic for insurance against a productivity shock. In figures 12 and 13, we see that a higher bond tax lowers the response to the productivity shock and leaves the response to markup shock barely unchanged. To make our plots comparable with the baseline case we report impulse responses to the after-tax nominal and real interest rates.

Finally, we study the utilitarian Planner who sets Pareto weights equal. In our setup a utilitarian planner would set labor tax rate of $68\%$, a dividend tax rate of $116\%$ and a bond tax rate of $118\%$. These effects go in offsetting directions but overall we find little deviation from the baseline responses. The results are summarized in figures 14 and 15.
Figure 10: Optimal monetary response to a markup shock. The bold blue lines are the responses for the baseline calibration. The dashed black lines with squares and circles are responses under a calibration with higher and lower dividend taxes respectively.

Figure 11: Optimal monetary response to a TFP shock. The bold blue lines are the responses for the baseline calibration. The dashed black lines with squares and circles are responses under a calibration with higher and lower dividend taxes respectively.
Figure 12: Optimal monetary response to a markup shock. The bold blue lines are the responses for the baseline calibration. The dashed black lines with squares are responses under a calibration with higher bond taxes.

Figure 13: Optimal monetary response to a TFP shock. The bold blue lines are the responses for the baseline calibration. The dashed black lines with squares are responses under a calibration with higher bond taxes.
Figure 14: Optimal monetary response to a markup shock. The bold blue lines are the responses for the baseline calibration. The dashed black lines with squares are responses under utilitarian Pareto weights.

Figure 15: Optimal monetary response to a TFP shock. The bold blue lines are the responses for the baseline calibration. The dashed black lines with squares are responses under utilitarian Pareto weights.
C.4 Sensitivity with respect to the period of the shock

In this section, we compare optimal responses to a shock that occurs at $t = 25$ as well as $t = 50$ with our baseline in which the shocks occur at $t = 1$. For brevity we only report the optimal monetary response. Figure 16 plots the response to a markup shock, and figure 17 plots a response to a TFP shock. We find the responses to be very similar. The response to the TFP shock are slightly larger with time because the distribution of risk-free debt spreads out with idiosyncratic shocks and therefore there is a larger role for providing insurance.

C.5 Sensitivity with respect to choice of initial conditions

In the main text, we set the initial distribution of productivities, risk-free nominal bonds claims, and equity claims using the observed SCF distribution. Here we redo the optimal policy starting at a joint distribution of wealth and productivities that arises after simulating 100 years in the calibrated competitive equilibrium with fixed policies. The results are summarized in figures 18 and 19. The response to a markup shock is a balance of two forces. On the one hand the passage of time diminishes the correlation between stock holdings and labor earnings which renders inequality more misaligned according to our distance measure. That increases the planner’s gains from providing insurance. On the other
hand the correlation of shares of equities and bond holdings diminishes which diminishes the insurance gains from the unanticipated inflation. The increase in the spread of nominal debt leads the planner to be more responsive to a TFP shock.

D Additional details for section 6

D.1 Example with poor hand-to-mouth agents

We can also consider an alternative calibration of the hand-to-mouth agents were we restrict bottom 15% of the cash-in-hand distribution to be hand-to-mouth. This environment is similar in spirit to what would arise in an standard Aiyagari model as the new hand to mouth agents more homogeneous and are almost entirely reliant on labor income. We plot the optimal policy response with only poor hand-to-mouth agents using the dashed red line in figure 20. As opposed to the hand-to-mouth setting in the main text that is calibrated to the evidence in Jappelli and Pistaferri (2014), the optimal policy with only poor hand to mouth agents is almost identical to that of the baseline economy as the government can construct a transfer scheme to smooth the consumption of the hand to mouth agents by mirroring the path of wages.
Figure 18: Optimal monetary response to a markup shock. The bold blue lines are the responses for the baseline calibration. The dashed lines are from the policies after initializing the Ramsey allocation with $t = 100$ years of the competitive equilibrium.

Figure 19: Optimal monetary response to a TFP shock. The bold blue lines are the responses for the baseline calibration. The dashed lines are from the policies after initializing the Ramsey allocation with $t = 100$ years of the competitive equilibrium.
Figure 20: Optimal monetary responses with hand-to-mouth agents. The top panel plots responses to a markup shock and the bottom panel plots responses to a productivity shocks.
D.2 Optimal monetary-fiscal response with mutual fund

In this section, we present optimal monetary response to productivity shock, as well as the optimal monetary-fiscal response to both under the mutual fund calibration. The optimal monetary response to the productivity shock is in figure 21.

One aspect of the mutual fund calibration is that it enforces a perfect correlation between bond and dividend wealth following any history of shocks. As a result, the optimal policy the bond and dividend tax rates are indeterminate as the planner can achieve the same effective returns with either instrument. To make the results comparable with our benchmark calibration, we assume that the planner adjusts the dividend tax in response to a markup shock and the bond rate in response to a markup shock. The results are plotted in figures 22 and 23. In both cases the optimal policy under the mutual fund is almost identical to the benchmark calibration.

D.3 Optimal monetary and monetary-fiscal response to a TFP shock with heterogeneous labor income exposures

As noted in section 6.3 we calibrate the coefficients of $f(\theta) = f_0 + f_1 \theta + f_2 \theta^2$ by simulating the competitive equilibrium for 30 periods and extracting “recessions” as consecutive periods
Figure 22: Optimal monetary-fiscal response to a markup shock. The bold red are the benchmark response while the bold blue lines are the responses for the mutual fund calibration.

Figure 23: Optimal monetary-fiscal response to a TFP shock. The bold blue lines are the response under the baseline while the dashed black lines are the responses for the mutual fund calibration.
where the growth rate of output one standard deviation below zero. Following the empirical procedure in Guvenen et al. (2014), we rank workers by percentiles of their average log labor earnings 5 years prior to the shock and compute the percent earnings loss for each percentile relative to the median. The parameters $f_1, f_2$ are set to match earnings losses of the 5th, 95th percentiles. The parameter $f_0$ is set so that agent with the median productivity faces a drop similar to the aggregate TFP. Figure 24 plots the earnings losses by percentile of the income distribution relative to those found by Guvenen et al. (2014).

Figure 25 plots the responses of the monetary-fiscal policy. When the government has access to fiscal policy, it no longer needs to rely solely on monetary policy. In figure 25 we see that in response to an inequality shock the planner raises the labor tax rate by nearly 1% and then allows it to mean revert back as the TFP shock dissipates. This mean reversion arises because the level of inequality loads on TFP and partly captures the forces laid out in Werning (2007) where the planner responds to changes in relative labor productivity though changes in the labor tax rate. Unlike in baseline, case when nearly all insurance can be provided through a surprise tax on bond income the planner must also rely on a surprise increase in the dividend tax rate to partially provide insurance. This highlights a feature of heterogeneous agent models. Unlike representative agent models where a single tax on returns can complete markets, with heterogeneous agents one tax may not provide insurance.
for all agents and the planner may exploit multiple different asset taxes.

In figure 26, we apply our decomposition to the monetary response with setting with heterogeneous exposures. The small difference in the rate of inflation in the HANK complete market relative to RANK case captures the redistribution. In figure 25, we saw that labor income taxes are used to respond to inequality even with complete markets. When the planner cannot adjust labor income tax, wages and thereby inflation is used to attain similar objectives.
Figure 26: Decomposition of the optimal monetary response to a TFP shock with heterogeneous exposures. The bold blue and red lines are the calibrated HANK and RANK responses respectively. The dashed black lines with squares and circles are responses under HANK with idiosyncratic shocks shutdown and with complete markets, respectively.