# Redistribution and Reallocation:
## Monetary Policy with Heterogeneous Returns*

Nick Hoffman†

*June 1, 2026*

**Abstract**

I study how heterogeneity in rates of return across households determines the economy's response to monetary policy. In an economy with return heterogeneity, increases in prices and demand following a monetary expansion redistribute assets towards high-return households. This redistribution rationalizes two empirical facts: monetary expansions increase both aggregate productivity and wealth inequality. Because productivity is endogenous to monetary policy, the overall effect of a rate change is determined by the amount of redistribution that it induces. I show that, as a result of two countervailing forces, the potency of monetary policy is hump-shaped in the degree of wealth inequality: when inequality is very low or very high, monetary policy affects output and investment by less than when inequality takes a moderate value. Calibrating my model to the data, I find that the increase in wealth inequality over the past fifty years can help explain the empirically observed decrease in the effect of monetary policy on output.

Keywords: Monetary policy, Wealth distribution, Return heterogeneity, Entrepreneurs, Financial frictions
JEL Codes: D31, D53, E32, E44, E52

*This paper is a revised version of the first chapter of my PhD Thesis at Carnegie Mellon University. I am deeply indebted to Ali Shourideh, Laurence Ales, and Ariel Zetlin-Jones for invaluable guidance. I also extend my sincere gratitude to Andre Sztutman, Liyan Shi, Daniel Carroll, Eric Young, Kevin Mott, Brandon Pecoraro, Martin Lopez-Daneri, and participants at the 2024 Midwest Macroeconomics Meetings, the 2024 Washington University in St. Louis Economics Graduate Student Conference, the University of Pittsburgh macroeconomics brownbag, the CMU Marvin Goodfriend Economics Lunch Seminar, and in the Carnegie Mellon Macroeconomics Reading Group for helpful discussions and feedback.

†Joint Committee on Taxation, United States Congress. Email: nicholas.hoffman@jct.gov. The views expressed in this paper are solely those of the author and thus should not be attributed to the Joint Committee on Taxation nor any member of the United States Congress.

# 1 Introduction

Wealth concentration in the U.S. results largely from households earning different rates of return on their assets.[1] Wealthy households generate high returns by owning productive assets such as stocks and businesses, and amplify these returns through leverage, capitalizing on a spread between their private returns and borrowing costs. Expansionary monetary policy lowers borrowing costs and boosts asset values, directly benefiting these wealthy households. This channel is supported in the data: wealth inequality rises following a monetary expansion.[2] These distributional effects have critical but under-examined implications for aggregate outcomes.

Because wealthy households own productive assets, when monetary policy redistributes to these households it reallocates capital to more productive uses. In the data, monetary expansions increase aggregate productivity (Baqaee et al., 2021)—an empirical fact largely unexplained by existing Heterogeneous Agent New Keynesian (HANK) models. This literature has primarily focused on heterogeneity in incomes and marginal propensities to *consume*; less is understood about the importance of differences in returns and marginal propensities to *invest.* Understanding how productivity responds is essential for the conduct of monetary policy: productivity determines the output-inflation tradeoff, which is *the* central tradeoff that policymakers face.

I address this gap and show that redistribution of assets toward high-return households is not simply a side effect of monetary policy; rather, it is a key channel by which policy affects the economy. In my model, households operate private businesses and differ in the productivity with which they do so. Households choosing to invest in their business may issue debt–up to a collateral constraint–to finance their operations, and nominal rigidities give the monetary authority the ability to set the interest rate on this debt. The core determinant of wealth inequality in my model is the *persistence* in households' private returns: the longer a household remains productive, the more capital it is able to accumulate to scale up its operations. In this environment, a fall in the nominal interest rate redistributes capital from low-return to high-return households through two channels. Increases in aggregate demand increase both prices and profit margins; the former devalues high-return households' debt while the latter increases their real returns. As a result, my model accounts for the joint increase in wealth inequality and aggregate productivity following monetary expansions in the data—an empirical pattern that existing frameworks have struggled to explain.

An immediate implication of this result is that the overall effect of monetary policy on the economy depends on how much the policy change alters productivity by redistributing assets. I show further that the amount of redistribution that occurs depends on the *initial* level of wealth inequality present when the policy change occurs. Changes in wealth inequality influence the efficacy of monetary policy through two channels, which oppose one another: the first implies that greater inequality *amplifies* policy, and the second that greater inequality *dampens* it. As a result, the extent of redistribution, and thus of amplification, is hump-shaped in the degree of wealth inequality at the time of the policy change. In very equal and very unequal economies this amplification is minimal; it is larger when wealth inequality takes an intermediate value.

The first channel implies that the effect of a monetary policy shock rises with the wealth inequality present when the shock hits. Intuitively, greater inequality arises primarily from more persistent returns, so the

[1] See, for instance, Benhabib et al., 2019.

[2] See a review in Colciago et al. (2019), or examples in Feilich (2021) and Medlin (2023).

initial redistribution shifts assets towards households that will remain highly productive, raising aggregate productivity. A second channel, however, countervails as wealth inequality increases. As wealth becomes more concentrated in the hands of high-return households, the return on capital–a wealth-weighted average of individual investors' returns–aligns with the private returns of wealthy households. Following a monetary easing, this alignment of returns results in *less* redistribution to high-return households: their wealth grows at a similar rate to aggregate wealth, so their share of total wealth changes by less. The reduction in redistribution implies that monetary policy exerts less of an effect on productivity, and thus less of an effect on aggregates.

Understanding how productivity responds in a monetary expansion is critical because productivity alters the output-inflation tradeoff. Higher productivity implies lower marginal costs, reducing the inflation that accompanies a given increase in output. Redistribution in my model amplifies the aggregate effect of a monetary shock in a novel way that is more in line with the data. In existing HANK models, household heterogeneity amplifies policy on the *demand* side: expansions increase the incomes of high-MPC households, increasing demand and inflation more than in a representative-agent model. By contrast, amplification in my model occurs on the *supply* side, as assets shift to more productive investors. This reallocation mutes the response of inflation, flattening the Phillips curve. In the data, the Phillips curve has indeed flattened over time (Del Negro et al., 2020; Hazell et al., 2022). Redistribution in my baseline calibration flattens the Phillips curve by about 45% relative to the standard representative-producer framework, more realistically capturing the policy-relevant tradeoff.

Quantitatively, my model can help understand a puzzle in the data. Since the 1970s, wealth inequality has increased, while the effect of monetary policy on output has decreased (Boivin et al., 2010). The existing HANK literature has not yet addressed these coinciding trends; instead, it has focused primarily on the *amplification* of policy shocks through heterogeneity—the first of my two channels. By formulating a model that captures the joint distribution of wealth and returns while remaining analytically tractable, I am able to isolate the second channel and show that it reconciles the two trends in the data. Calibrating my model to capture the evolution in wealth and returns over time, I find that the increase in wealth inequality since the 1970s helps account for the decline in the potency of monetary policy observed over the same period.

*Related Literature.* I contribute to the literature on monetary policy in economies with household heterogeneity.[3] This literature has emphasized the role of *labor* income risk (McKay et al., 2016; Kaplan et al., 2018; Auclert et al., 2020), focusing on the presence of borrowing-constrained agents with higher marginal propensities to consume. This emphasis on the role of poor households has enhanced the role of heterogeneity in MPCs on monetary transmission, but has left relatively less explored the role of *wealthy* households. My contribution is to fill in this gap, and to characterize the role played by wealthy, high-return households in translating a policy shock to aggregate output. I focus on heterogeneity in marginal propensities to invest, rather than to consume, in transmitting shocks.

My paper builds upon the "Financial Accelerator" literature (Carlstrom and Fuerst, 1997; Bernanke et al., 1999), which links aggregate activity to financial market fluctuations via entrepreneurs. Kiyotaki (1998) mentions but does not study a simple version of the mechanism which lies at the heart of my paper: with

[3]See Auclert et al., 2025 for a review.

entrepreneurs who are *ex-ante* heterogeneous, a monetary shock can potentially redistribute assets between entrepreneurs. While both channels in my model were implicitly present in these papers, until now their relevance for monetary policy has not been uncovered.

A nascent strand of the HANK literature has studied monetary transmission with heterogeneous firms (Ottonello and Winberry, 2020; Jeenas, 2023), investors (Melcangi and Sterk, 2024), and entrepreneurs (Matusche and Wacks, 2021; González et al., 2024). Across these models, the conclusion is consistent: higher wealth concentration *amplifies* monetary policy's effect in output and investment. The intuition is that households with high returns have a high propensity to invest out of transitory income shocks, and thus shifting resources to these households increases the economy's propensity to invest in the aggregate.

This intuition contradicts the data: as wealth inequality has increased since the early 1970's, the measured effects of monetary policy on output have decreased, not increased (Boivin and Giannoni, 2002; Boivin et al., 2010). I show that this intuition in the literature is incomplete: it captures one channel, but misses a second, countervailing channel. My model transparently uncovers both and shows that as wealth inequality increases past threshold level, monetary transmission is *dampened* as the redistribution channel shuts down. Empirically, I find that the second channel has dominated since the 1970s, muting the efficacy of monetary policy and reconciling the observed empirical patterns.

# 2 Model

This section describes the model economy in my paper. I construct a model in which households earn different returns on their investment in equilibrium — returns which are endogenous to both monetary policy and the underlying distribution of wealth itself. To do so, I use a familiar setting of *entrepreneurship* and *financial frictions*: some households have the ability to operate private firms with idiosyncratic productivity, and they may issue debt to fund their operations up to a collateral constraint.

My model is sufficiently parsimonious to admit analytical results while still being able to capture the joint distribution of wealth and returns in the data. The two primary determinants of wealth inequality in the data are the *persistence* and *volatility* in investors' returns. By varying these underlying parameters, I can transparently expose the channels by which monetary policy redistributes assets across households, and how the strengths of these channels change with underlying inequality.

## 2.1 Investors' Problem and Financial Frictions

Investor households, indexed my $i \in [0, 1]$, accumulate two assets: capital and risk-free nominal bonds.[4] Entrepreneurs choose consumption, capital investment, and single-period bonds to maximize their lifetime expected utility:

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \ln(c_{it}) \tag{1}$$

[4]As will become clear shortly, these bonds are risk-free in the sense that their *nominal* rate of return between periods $t$ and $t+1$ is predetermined in period $t$. However, their *real* return is subject to risk in the event of unanticipated inflation between these two periods.

Investors split their total income between purchases of consumption goods $c_{it}$ and investment (capital) goods $q_{it}$; all goods are identical and hence share a nominal price $P_t$. Investor households earn income from two sources. These households trade in nominal bonds $D_{i,t}$, which they can issue ($D_{i,t} > 0$) to borrow or purchase ($D_{i,t} < 0$) to lend. Investor households can also operate a private business, which they run by combining capital with labor hired on the spot market to maximize profits in a manner described below. I refer to households operating a business in the current period as *entrepreneurs*, and those saving exclusively in bonds as *lenders*. Denoting entrepreneurs' nominal profit income as $P_t \Upsilon_{i,t}$, investors' budget constraint is

$$P_t \left(c_{i,t} + q_{i,t}\right) = P_t \Upsilon_{i,t} + \left(1 + i_t\right) D_{i,t} - D_{i,t+1} \tag{2}$$

where $P_t$ is the aggregate price level and $i_t$ the nominal interest rate between $t-1$ and $t$, set by the monetary authority at time $t-1$. Investor $i$'s capital stock evolves according to $k_{it+1} = \left(1-\delta\right) k_{it} + q_{it}$, where $q_{it}$ is the quantity of investment goods purchased.

Households who operate their private businesses may issue debt to do so, up to a limit. Following Buera and Moll (2015), I assume that households are subject to a collateral constraint of the form

$$D_{it+1} \leq \theta P_t k_{it+1}, \qquad \theta \in [0,1] \tag{3}$$

This collateral constraint implies that only a proportion $\theta$ of the nominal value of the next-period capital stock may be externally financed.

Each entrepreneur household operates a firm that produces homogeneous goods, which are sold to retailers (described below) at nominal price $P_{t,x}$. Entrepreneurs maximize their nominal profit income from running its firm:

$$P_t \Upsilon_{it} = \max_{n_{it}^d} \; P_{tx} \left(z_{it} k_{it}\right)^{\alpha} \left(n_{it}^d\right)^{1-\alpha} - W_t n_{it}^d \tag{4}$$

The following lemma is a well-known result in problems such as (4):

**Lemma 1.** *Entrepreneurial labor demand is linear in capital, and given by:*

$$n_{it}^d = \left(\frac{1-\alpha}{W_t/P_t} \frac{P_{tx}}{P_t}\right)^{\frac{1}{\alpha}} z_{it} k_{it} \tag{5}$$

Lemma 1 results from the fact that the problem in (4) is static: entrepreneurs hire labor on the spot market to maximize their profit given their state $(z_{it}, k_{it})$ and the wage and price $W_t$ and $P_{tx}$, which they take as given. Defining

$$\omega_t \equiv \alpha \left(\frac{1-\alpha}{w_t}\right)^{\frac{1-\alpha}{\alpha}} p_{tx}^{\frac{1}{\alpha}} \tag{6}$$

where $w_t$ is the real wage and $p_{tx}$ the real entrepreneurs' price, the budget constraint can be written as

$$P_t c_{it} + P_t k_{it+1} = P_t \left[\omega_t z_{it} + \left(1-\delta\right)\right] k_{it} - \left(1 + i_t\right) D_{it} + D_{it+1} \tag{7}$$

It is useful to write the entrepreneurs' budget constraint in real terms. To do so, define real bond issuance as $d_{i,t+1} \equiv D_{i,t+1}/P_t$. With this definition, the budget constraint for entrepreneurial household $i$ in real terms is

$$c_{it} + k_{it+1} = [\omega_t z_{it} + 1 - \delta] \, k_{it} - (1 + r_t) \, d_{it} + d_{it+1} \tag{8}$$

Here, $r_t$ is the time-$t$ ex-post real interest rate, defined by the Fisher equation:

$$1 + r_t = (1 + i_t) \frac{P_{t-1}}{P_t} = \frac{1 + i_t}{1 + \pi_t} \tag{9}$$

Note that this interest rate depends on the realized inflation rate $\pi_t$. Defining the real net worth as $a_{i,t} \equiv k_{i,t} - d_{i,t}$ I can write the borrowing constraint in real terms:

$$d_{t+1} \leq \theta k_{t+1} \tag{10}$$

or substituting the definition of net worth $a_t$,

$$k_{t+1} \leq \lambda a_{t+1}, \qquad \lambda \equiv \frac{1}{1-\theta} \tag{11}$$

The timing in the economy works as follows. Following production in period $t$, investor households realize their next-period productivity $z_{i,t+1}$, and choose their next-period capital stock $k_{i,t+1}$ and bond holdings $d_{i,t+1}$ — equivalently, whether to be entrepreneurs or lenders. Because returns are linear in wealth, the choice to be an entrepreneur only depends on idiosyncratic productivity $z_{i,t}$. Furthermore, households who do choose to be entrepreneurs will choose to invest as much as possible:

**Lemma 2.** *Entrepreneurs' capital and bond choices are corner solutions:*

$$k_{t+1} = \begin{cases} \lambda a_{t+1} & z_{t+1} \geq \underline{z}_{t+1} \\ 0 & z_{t+1} < \underline{z}_{t+1} \end{cases} \qquad d_{t+1} = \begin{cases} (\lambda - 1) \, a_{t+1} & z_{t+1} \geq \underline{z}_{t+1} \\ -a_{t+1} & z_{t+1} < \underline{z}_{t+1} \end{cases} \tag{12}$$

*where the cutoff $\underline{z}_{t+1}$, which solves $\omega_{t+1}\underline{z}_{t+1} = r_{t+1} + \delta$, is the productivity at which an investor is indifferent between private production and public bonds.*

Lemma 2 shows that investors are divided into two groups: those above the productivity threshold $\underline{z}_t$, who are entrepreneurs, and those below it, who are lenders. Active entrepreneurs, who earn excess returns on their investment above the risk-free rate, borrow up to their limit and are thus bound by the collateral constraint. Lenders save at the risk-free rate by lending to active entrepreneurs. Due to the linearity of the production technology, the cutoff productivity $\underline{z}_t$ is independent of wealth; instead, $\underline{z}_t$ is a linear function of the risk-free rate $r_t$ and $\omega_t$, which can be thought of as the private return per effective unit of capital $zk$.

Finally, the assumption of log utility among investor households implies that these households will choose to save a constant fraction $\beta$ of their financial wealth, post-returns:

$$a_{i,t+1} = \beta R_t(z) a_{i,t} \tag{13}$$

I refer to the object $\beta R_t(z)$ as the marginal propensity to save for type $z$, $MPS_t(z)$. This object captures the propensity of a household of type $z$ to accumulate out of an increase in *pre-return* wealth. As such, the distribution of $MPS(z)$ will be crucial in determining how redistribution across households translates into aggregate outcomes.x

### 2.2 Nominal Rigidities

To introduce nominal rigidities while maintaining tractability, I follow Bernanke et al. (1999) in assuming a three-tiered production structure. Entrepreneurs produce a homogenous good $x_t$, which is then sold to retailers. Retailers, a continuum of whom are indexed by $j \in [0, 1]$, in turn costlessly differentiate these goods. Retailers sell their output $y_{tj}$ to a final good producer, who aggregates them using a CES technology:

$$Y_t = \left[\int_0^1 y_{tj}^{\frac{\varepsilon-1}{\varepsilon}} \, \mathrm{d}j\right]^{\frac{\varepsilon}{\varepsilon-1}} \tag{14}$$

This assumption on the structure of production allows me to introduce price stickiness in a way that preserves the tractability of the entrepreneurs' problem. It is analytically convenient to assume that entrepreneurs are price takers; otherwise, their investment and savings choices would be intermingled with a forward-looking pricing problem, which would complicate the model without providing any additional insights.

Optimal behavior by the final good aggregator in (14) implies that the demand for variety $j$ is

$$y_{t,j} = \left(\frac{P_{t,j}}{P_t}\right)^{-\varepsilon} Y_t \tag{15}$$

where

$$P_t = \left(\int P_{t,j}^{1-\varepsilon} \, \mathrm{d}j\right)^{\frac{1}{1-\varepsilon}} \tag{16}$$

is the overall price level. Retailer $j$'s output is then simply $y_{tj} = x_{tj}$, and their marginal cost is $m_t = p_{tx}$. In addition, retailers incur Rotemberg (1982)-style quadratic adjustment costs to change their price:

$$\Theta\left(P_{tj}, P_{t-1,j}\right) = \frac{\theta}{2}\left(\frac{P_{tj}}{P_{t-1,j}} - 1\right)^2 Y_t \tag{17}$$

The retailer incurs an adjustment cost when it wants to update its price relative to its own lagged price. Defining inflation as

$$\pi_t \equiv \frac{P_t}{P_{t-1}} - 1 \tag{18}$$

the following lemma describes the behavior of inflation over time:

**Lemma 3.** *Inflation evolves according to the New Keynesian Phillips Curve, which arises under optimal behavior by retailers:*

$$\pi_t = \frac{\varepsilon}{\theta}\left(p_{tx} - m^*\right) + \beta \mathbb{E}_t \pi_{t+1} \tag{19}$$

*Here, $m^* = \varepsilon / \left(\varepsilon - 1\right)$ is the inverse of the optimal markup in the absence of price rigidities.*

The intuition in Lemma 3 is standard. In the presence of price stickiness, retailers raise their prices when they believe that future marginal costs will exceed their long-run optimal level—equivalently, retailers raise current prices when they believe that, in the future, markups will fall *below* their long-run optimal level.

### 2.3 Equilibrium

There are two actors needed to close the model: workers, and a monetary authority. I assume that workers—who supply their labor to entrepreneurs at real wage $w_t$—cannot borrow or save, and are thus constrained to be hand-to-mouth. Workers are, however, free to adjust their labor supply in response to movements in the real wage. Worker households are all identical, and have preferences as in Greenwood et al. (1988):

$$U^w\left(C_t^w, N_t^w\right) = \frac{1}{1-\gamma}\left(C_t^w - \frac{\left(N_t^w\right)^{1+\eta}}{1+\eta}\right)^{1-\gamma} \tag{20}$$

In addition, worker households own the retailers, and receive the profits of these firms as real dividends $T_t$. Workers' budget constraint in real terms is simply $C_t = w_t N_t + T_t$. The labor supplied by the households is given by

The monetary authority sets the nominal interest rate $i_t$ according to a Taylor rule:

$$i_{t+1} = \overline{r} + \phi_\pi \pi_t + \nu_t \tag{21}$$

Recall that $i_t$ dictates the nominal cost that an entrepreneur pays for outside financing. In order to ensure notational consistency, I date all interest rates according to when they are earned, rather than when they are set. As such, the nominal rate $i_{t+1}$ in (21) is set at time $t$, and dictates the interest rate on nominal debt issued in period $t$ and maturing in period $t+1$. The term $\nu_t$ is an exogenous, stochastic innovation; I will use this shock to measure the impact of monetary policy in my model.

With these remaining actors thus established, I can define an equilibrium in my model economy:

**Definition 1.** An equilibrium is a sequence of prices $\{P_t, P_{tx}, W_t\}$, aggregates $\{C_t^e, C_t^w, N_t, Y_t, K_t, Z_t\}$, interest rates $\{i_t, r_t\}$, a path for inflation $\pi_t$, a sequence of aggregate shocks $\nu_t$, and a sequence of distributions $\{g_t\left(a, z\right)\}$ over the idiosyncratic states for entrepreneurs such that:

1. Entrepreneurs, workers, retailers, and the final good producer all maximize their respective objectives,

2. The monetary authority sets the nominal interest rate in accordance with the Taylor rule in (21), given an exogenous sequence for the shock $\nu_t$,

3. Prices clear markets:

$$K_t = \int_0^{\overline{z}} \int_0^\infty a g_t\left(a, z\right)\, \mathrm{d}a\, \mathrm{d}z \tag{22}$$

$$N_t^w = N_t^d \tag{23}$$

$$C_t^e + C_t^w + K_{t+1} + \Theta\left(\pi_t\right) = Y_t + \left(1-\delta\right) K_t \tag{24}$$

To track the equilibrium distribution of wealth, I use *wealth shares*:

$$s_t(z) \equiv \frac{1}{K_t} \int_0^\infty a g_t(a, z) \, \mathrm{d}a \tag{25}$$

As in Moll (2014), among others, the wealth share $s_t(z)$ denotes the share of aggregate wealth held by agents of type $z$. There are a number of reasons why these objects are a convenient tool for studying the behavior of the model. First, the shares $s_t(z)$ can be thought of as a density: they are nonnegative for all $z$, and integrate to one: $\int_0^{\overline{z}} s_t(z) \, dz = 1$ for all $t$. As such, I can define the analogous cumulative share:

$$S_t(z) \equiv \int_0^z s_t(\hat{z}) \, \mathrm{d}\hat{z} \tag{26}$$

Second, note that because returns are linear in wealth, individual wealth follows a random growth process.[5] As a result, the joint distribution $g_t(a, z)$ does *not* admit a stationary measure: the log of individual wealth $a_t$ follows a random walk, and thus the cross-sectional variance of $a_t$ grows without bound in $t$. However, it can be demonstrated that the wealth shares $s_t(z)$ *do* admit a stationary measure. This result allows me to study the evolution of the wealth distribution without needing to augment the model with an assumption to deliver a stationary measure over wealth, such as random death and annuity markets (as in Gouin-Bonenfant and Toda, 2019) or a hard borrowing limit.

Using the definition of wealth shares in (25), aggregate quantities are as follows:

**Proposition 1.** *Aggregate quantities satisfy*

$$Y_t = (Z_t K_t)^\alpha N_t^{1-\alpha} \tag{27}$$

$$K_{t+1} = \beta \left\{ \alpha p_{tx} Y_t + (1 - \delta) K_t \right\} \tag{28}$$

*Aggregate productivity is a function of the wealth distribution $s_t(z)$:*

$$Z_t = \frac{\int_{\underline{z}_t}^{\overline{z}} z s_t(z) \, \mathrm{d}z}{\int_{\underline{z}_t}^{\overline{z}} s_t(z) \, \mathrm{d}z} = \mathbb{E}_{s_t}\left[ z | z > \underline{z}_t \right] \tag{29}$$

*Given a path for wealth shares $s_t(z)$, the cutoff productivity $\underline{z}_t$ is pinned down by capital market clearing:*

$$1 = \lambda \left( 1 - S_t(\underline{z}_t) \right) \tag{30}$$

[5] See Gabaix (2009) for a study of random growth processes in economics, and Benhabib et al. (2015) for an example of how this process gives rise to wealth distributions in models that share the "fat-tailed" (Pareto) nature of their empirical counterparts.

*Factor prices are*

$$w_t = (1-\alpha)\, p_{tx} \left(\frac{Z_t K_t}{N_t}\right)^{\alpha} \tag{31}$$

$$\mathbb{E}_{t-1} r_t = \alpha p_{tx} Z_t^{\alpha} \left(\frac{N_t}{K_t}\right)^{1-\alpha} \frac{\underline{z}_t}{Z_t} - \delta \tag{32}$$

*Returns are given by*

$$\omega_t = \alpha p_{tx} \left(\frac{N_t}{Z_t K_t}\right)^{1-\alpha} \tag{33}$$

$$R_{tK} = 1 - \delta + \alpha p_{tx} Z_t^{\alpha} \left(\frac{N_t}{K_t}\right)^{1-\alpha} \tag{34}$$

*The return to an entrepreneur of type $z$ is given by*

$$R_t(z) = 1 + i_{t-1} - \pi_t + \lambda \begin{cases} 0 & z < \underline{z}_t \\ \omega_t z - (i_{t-1} - \pi_t + \delta) & z > \underline{z}_t \end{cases} \tag{35}$$

Proposition 1 is an analogue of similar results in Moll (2014) and Buera and Moll (2015), adapted to an economy with nominal rigidities. Equations (27) and (28) show that this economy behaves as one with a representative firm, with a key difference: aggregate TFP is an endogenous result of the wealth distribution among investor households, per (29). Per equation (31), the real wage is given by the real value of the aggregate marginal product of labor hired by entrepreneurs. The same is generally *not* true for the ex-ante expected risk-free real rate $\mathbb{E}r_t$: in equation (32), this object is equal to the aggregate marginal return to capital, weighted by $\underline{z}_t/Z_t$.[6] In this economy, capital market frictions prevent the return from investment to be equated with outside savings. As a corollary, the two are equated in the case that $\underline{z}_t = Z_t$, which is the case when capital markets are frictionless, $\lambda = \infty$. Note as well that both factor prices, as well as the returns $R_{tK}$ and $\omega_t$, move with the price paid to entrepreneurs by retailers for their goods.

In equation (34), the aggregate return to capital is derived from

$$\begin{aligned} R_{tK} &= \int_0^{\overline{z}} R_t(z)\, s_t(z)\ \mathrm{d}z \\ &= \mathbb{E}_{s_t}\left[R_t(z)\right] \end{aligned}$$

Thus, the aggregate return on capital is the average return across all entrepreneurs, weighted by their respective wealth shares. Finally, entrepreneurs' returns, per equation (35), exhibit a few key properties that will later drive my results. First, entrepreneurs with $z > \underline{z}_t$ are able to earn excess returns above the ex-ante risk-free rate $i_{t-1} - \pi_t$, due to their ability to make leveraged investments into their inside firm. In partial

[6] The expectations operator indicates that the ex-post real rate is subject to inflation risk. In the absence of nominal rigidities, Equation (32) would always hold. In order to study the redistributive effects of unanticipated inflation, I leave open the possibility that the ex-post real rates may differ from their ex-ante expectations. In the event that inflation is not equal to its ex-ante expectation, this equation will hold for the expected risk-free rate, upon which time-$t$ contracts are based, but *not* for the ex-post rate $r_{t+1}$.

equilibrium, equation (35) previews the differential impact of a change in real rates on entrepreneurs of different productivity. Returns can also be written as

$$R_t(z) = \begin{cases} 1 + i_{t-1} - \pi_t & z \leq \underline{z}_t \\ 1 - (\lambda - 1)(i_{t-1} - \pi_t) + \lambda(\omega_t z - \delta) & z > \underline{z}_t \end{cases} \tag{36}$$

From (36) it is immediately obvious that a fall in the real rate $i_{t-1} - \pi_t$ lowers the returns of savers, who *earn* this rate, and raises that of active entrepreneurs, who *pay* this rate to borrow capital. Additionally, the expression of returns in (36) makes clear the role of inflation in driving redistribution in this model: an unexpected increase in $\pi_t$ reduces nominal obligations, redistributing from low types (lenders) to high types (borrowers). Crucially, $\pi_t$ is realized one period *after* $\underline{z}_t$ has been determined: entrepreneurs operate their firms at time $t$ with capital installed at time $t-1$, and thus agents cannot switch from being inactive to active following unexpected inflation.

Before studying the properties of this equilibrium, I wish to emphasize its generality in understanding the role of return heterogeneity in transmitting monetary policy. From the perspective of policy, the key features of this model are that (i) agents earn heterogeneous returns and (ii) take on leverage, (iii) the aggregate return to capital is a wealth-weighted average of returns above a cutoff for participation in the risky asset, and (iii) this cutoff is endogenously determined so as to clear the capital market. The most restrictive assumption that I make is that these firms are unable to issue equity, and thus must fund all investments above their existing wealth with debt. Nevertheless, these forces are likely to be present even when households have risky returns resulting from ownership of *public* firms as well.

### 2.4 Distribution of Returns

To study analytically the effects of a monetary shock in my model economy, I make the following assumption on individual entrepreneurs' productivity:

**Assumption 1.** *With probability $p$, an entrepreneur will maintain his productivity from one period to the next, $z_{t+1} = z_t$. With probability $1-p$, meanwhile, he draws his next-period productivity at random from the time-invariant distribution $F(z)$. Furthermore, I assume $F(z)$ is lognormal, and parameterized by $\sigma_z$:*

$$\ln z \sim \mathcal{N}(-\sigma_z^2/2, \sigma_z^2)$$

Assumption 1 allows for gains in tractability while maintaining rich heterogeneity in wealth and returns. With this assumption, the underlying mass of households in the economy will be given by $F(z)$, and $p$ is the autocorrelation of $z_t$ and $z_{t+1}$. Persistence in returns is the empirically relevant case: autocorrelated returns have been documented in the micro data (see, e.g., DeBacker et al., 2023). Furthermore, persistence in returns allows the model to capture the empirical correlation between returns and wealth (Xavier, 2021): if investors keep their returns over time, then high-productivity entrepreneurs can accumulate savings out of perennial high returns to ease the financial constraint in (11). In the long run, then, persistent returns give rise to an equilibrium in which wealthier households earn higher returns — as in the data.

Under assumption 1, wealth shares $s_t(z)$ defined in (25) evolve as follows:

**Proposition 2.** *The wealth share of type $z$, $s_t(z)$, evolves according to*

$$s_{t+1}(z) = p\frac{R_t(z)}{R_{tK}}s_t(z) + (1-p)f(z) \tag{37}$$

*where $R_{tK}$ is the aggregate (wealth-share weighted average) return to capital, as defined in Proposition 1.*

Proposition (2) has an intuitive interpretation. There are two sources of change in the wealth share $s_{t+1}(z)$: entrepreneurs who retain their type $(z_{t+1} = z_t)$, and entrepreneurs who transition to type $z$ at $t+1$ from some other type $(z_t \neq z_{t+1})$. For each source of change, the sign of its contribution (whether it increases or decreases $s_{t+1}(z)$ relative to $s_t(z)$) depends on the returns of the agents in question, relative to the aggregate return on capital. For agents who retain their type: if the time-$t$ return $R_t(z)$ is greater than the aggregate return to capital, then the wealth of agents of type $z$ grows faster than the overall capital stock, and their share increases. Agents transitioning to type $z$ from some other $z'$ – the $(1-p)f(z)$ term in (37) – on average earn the aggregate return on capital $R_{tK}$ by definition, hence the coefficient of one on this term.

## 2.5 Long-Run Wealth Inequality

One of the primary questions of my paper is how the distribution of wealth at the time of a policy shock influences the economy's aggregate response to that shock. Here, I analyze properties of the long-run steady state equilibrium that will be critical to answering this question. I assume that prior to the unexpected change in monetary policy $\nu_0$, the economy is in its *long-run, zero-inflation steady state*. Equation (25) implies that in the steady state, wealth shares are given by

$$s(z) = \frac{1-p}{1-\beta pR(z)}f(z) \tag{38}$$

Equation (38) uses the fact that in the steady state, the return to capital is pinned down by the entrepreneurs' discount factor:

$$R_K = 1-\delta+\alpha p_x Z^\alpha (N/K)^{1-\alpha} = \frac{1}{\beta} \tag{39}$$

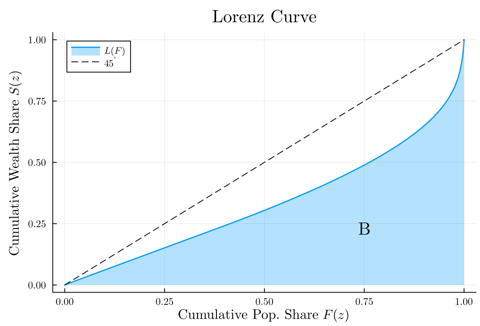
The wealth shares $s(z)$ are a mixture of the marginal distribution $f(z)$, which determines the mass of agents of type $z$, and the steady-state returns earned by type $z$, given by

$$R(z) = 1+\alpha p_x Z^\alpha \left(\frac{N}{K}\right)^{1-\alpha}\frac{z_\lambda(z)}{Z}-\delta \tag{40}$$

where

$$z_\lambda(z) \equiv \begin{cases} z+(\lambda-1)(z-\underline{z}) & z>\underline{z} \\ \underline{z} & z\leq\underline{z} \end{cases} \tag{41}$$

Figure 1: Lorenz Curve




is the effective return to an entrepreneur of type $z$, taking into account leverage. The real interest rate is the effective productivity of the *marginal* entrepreneur:

$$r = \alpha p_x Z^{\alpha} \left(\frac{N}{K}\right)^{1-\alpha} \frac{\underline{z}}{Z} - \delta \tag{42}$$

The price for entrepreneurial goods is equal to its optimal level in the absence of nominal rigidities:

$$p_x = 1/\mathcal{M}^* = \frac{\varepsilon - 1}{\varepsilon} \tag{43}$$

Aggregate productivity is determined by the allocation of wealth among entrepreneurs, as described by the shares $s\left(z\right)$:

$$Z = \lambda \int_{\underline{z}}^{\overline{z}} z s\left(z\right) \, \mathrm{d}z \tag{44}$$

Equation (44) uses the fact that capital market clearing again implies $1 = \lambda\left(1 - S\left(\underline{z}\right)\right)$.

To measure inequality, I use the Gini coefficient, defined at time $t$ by

$$G_t = 1 - 2 \int_0^1 S_t \left[F^{-1}\left(x\right)\right] \, \mathrm{d}x \tag{45}$$

In the context of this model, the Gini coefficient measures the discrepancy between $F\left(z\right)$, which determines the incidence of each type $z$ in the population, and the wealth shares $S\left(z\right)$, which measure the distribution of wealth among these types. Visually, the Gini measures the area between the Lorenz curve, which is constructed by plotting the cumulative wealth shares $S$ against the population shares $F$. Figure 1 illustrates the Lorenz curve; the Gini coefficient is $1 - 2B$, where $B$ is the shaded blue area. This measure is convenient, as it can be calculated in the steady state and along transition paths using the wealth shares directly, without needing to infer the underlying wealth distribution.

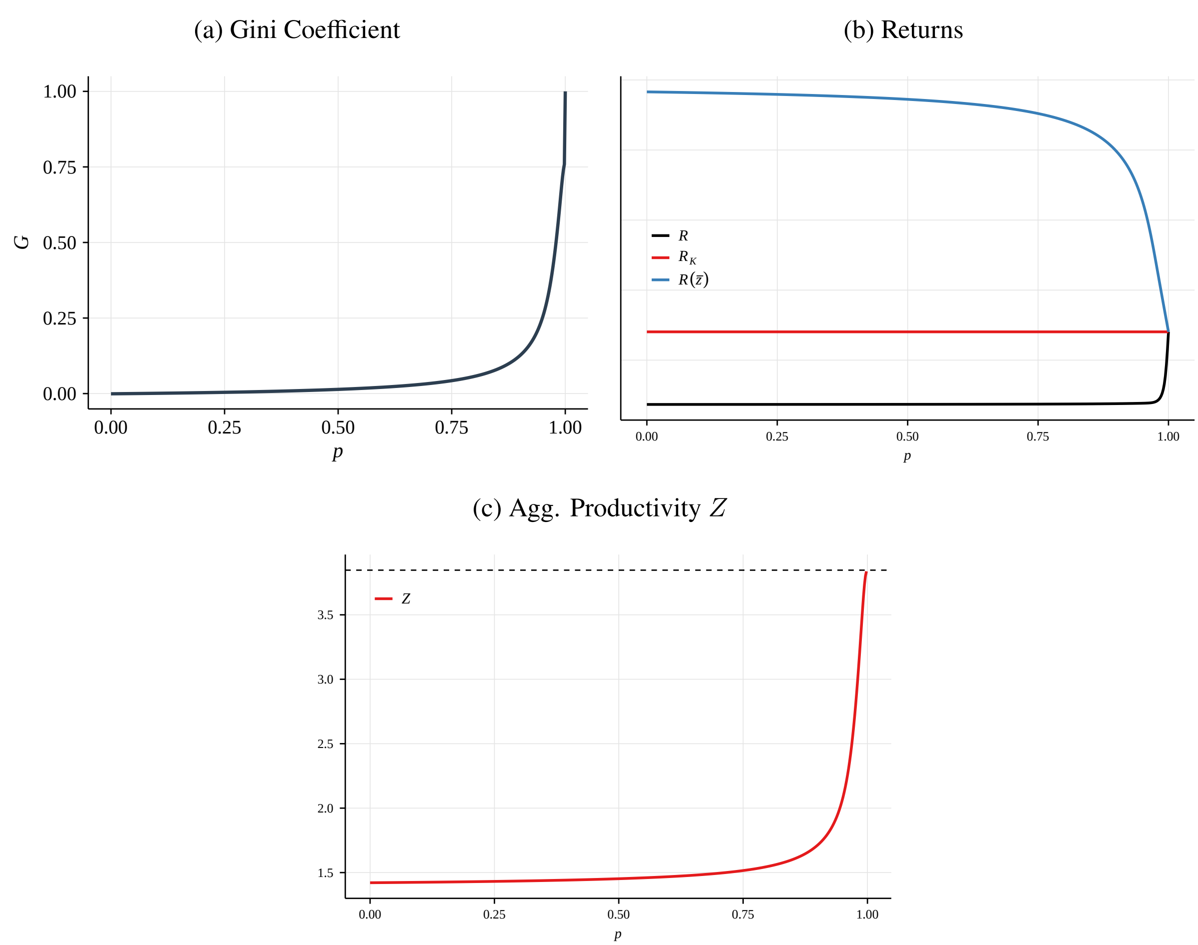
#### 2.5.1 Determinants of Wealth Inequality

Equations (38)-(40) show the forces that give rise to the steady state wealth distribution. First and foremost among these forces if $p$, the persistence in returns. At $p = 0$, productivity shocks are IID: each period, entrepreneurs draw a new $z_{it+1}$ that is independent of their current productivity $z_{it}$. When this is the case, $s(z) = f(z)$: the wealth shares are equal to the marginal density, the Lorenz curve aligns with the 45°line, and the Gini coefficient is zero. In this case, returns and wealth are uncorrelated: entrepreneurs who receive a high productivity shock today are unlikely to receive another high shock tomorrow, and are thus unable to accumulate wealth from a series of shocks.

At the other extreme, if $p = 1$ then return heterogeneity is permanent. When this is the case, the wealth distribution will be $s(z) = \delta_z(\overline{z})$, where $\delta(\cdot)$ is the Dirac measure. In the long run, entrepreneurs of type $z$ end up holding all of the wealth in the economy. As a result, the long-run risk-free rate and return to capital will equate with the returns of entrepreneurs of type $\overline{z}$, $R_K = R = R(\overline{z})$

For intermediate values of persistence, inequality is strictly increasing in $p$. Figure 2 shows how steady state inequality, returns, and productivity vary in the autocorrelation $p$. Panel 2a demonstrates that the relationship between inequality and persistence is convex in $p$. Panels 2b and 2c, meanwhile, shows how aggregate productivity and returns change in persistence. As $p \to 1$, a greater and greater share of aggregate wealth is held by the top type $\overline{z}$, and the risk-free rate $R$ and the return $R(\overline{z})$ earned by these types converge to $R_K = \beta^{-1}$, while aggregate productivity converges to $\overline{z}$.

Figure 2: Steady State Comparative Statics in $p$



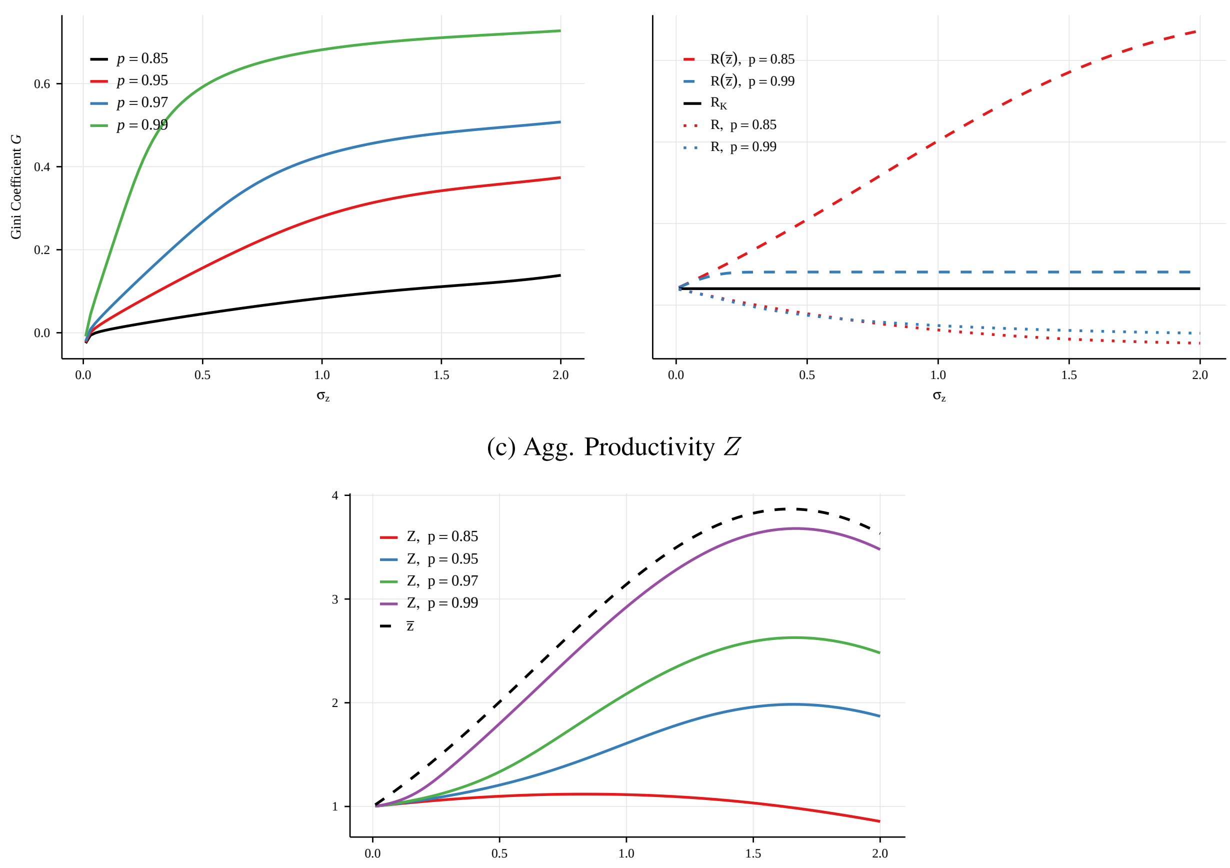
Wealth inequality in the steady state is also determined by the *size* of idiosyncratic return shocks, as measured by the standard deviation $sigma_z$ of the distribution $F(z)$. Figure 3 shows how wealth inequality, returns, and productivity vary with $\sigma_z$ across levels of persistence $p$. At the extreme of no variance, the model once again collapses to one with a representative producer with productivity $z = 1$. Increases in the variance of the idiosyncratic shocks $z$ from this extreme increase the level of wealth inequality, as heterogeneity among households becomes more salient and their wealth separates over time. However, as is clear from Figure 3, the marginal effect of increases in $\sigma_z$ on inequality diminishes rapidly, and the overall effect of increases to the variance in shocks depends heavily on how persistent the shocks are. Intuitively, beyond a certain point increasing the size of the largest shocks can only increase wealth inequality by a small amount if these shocks do not persist over time.

Figure 3: Steady State Comparative Statics in $\sigma_z$

(a) Gini Coefficient

(b) Returns

(c) Agg. Productivity $Z$

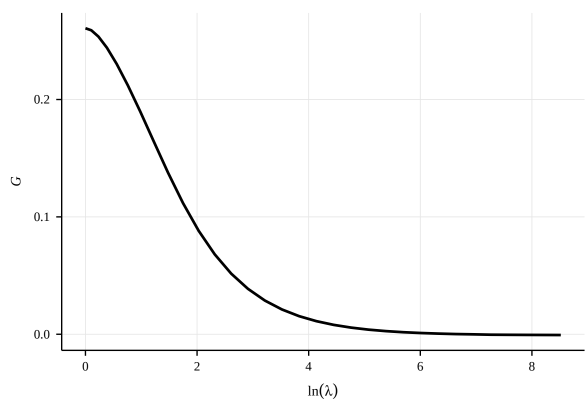



An increase in the collateral constraint $\lambda$, meanwhile, *decreases* wealth inequality while *increasing* productivity. Whereas an increase in persistence undoes financial frictions via wealth accumulation, an increase $\lambda$ reduces misallocation by increasing the scope for *within-period* wealth transfers via the capital market. At the limit of $\lambda = \infty$, entrepreneurs are unconstrained in their ability to take on leverage, and can borrow as much as the market will bear. In this case, the real interest rate is again bid up to the marginal product of the highest-type entrepreneurs, $\overline{z}$, and the economy once again attains its first-best productivity.

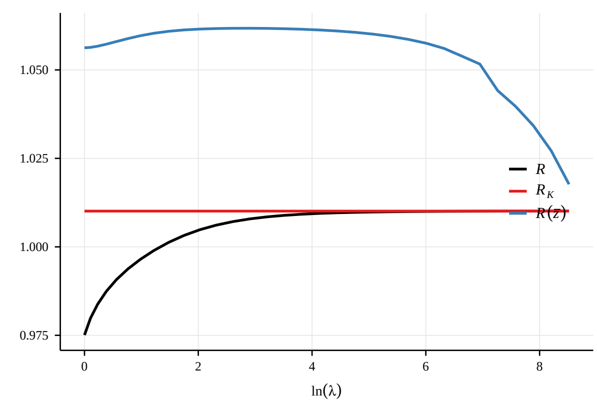
Figure 4 shows how productivity, returns, and wealth inequality are affected by an increase in the quality of financial markets, modeled as an increase in $\lambda$. Here again, $Z$ is increasing in $\lambda$, and returns display a similar pattern, converging to the return of the high types $R(\overline{z})$. The wealth distributions underlying these changes, however, are markedly different. An increase in $\lambda$ benefits low and high-productivity agents, shifting wealth away from the middle. Intuitively, a loosening of credit constraints allows for high-type entrepreneurs to borrow more capital from inactive entrepreneurs ($z < \underline{z}$). This has the effect of increasing the risk-free real rate: in the absence of capital market frictions ($\lambda = \infty$), the return to outside savings is equated with the marginal product of the highest types $\overline{z}$, who undertake all of the investment. The increase in the risk-free rate benefits types who do not invest, who now earn a higher return on their savings. The increase in $\lambda$ also benefits high-$z$ types, who still earn excess returns, but are now able to take on greater leverage to increase their capital income. Looser financial frictions shift wealth away from those with $z$ in the middle of the support, near the cutoff. These types earn small excess returns, and to them, the benefit of looser capital markets is undone by the concomitant increase in the cost of external financing.

Figure 4: Steady State Comparative Statics in $\sigma_z$

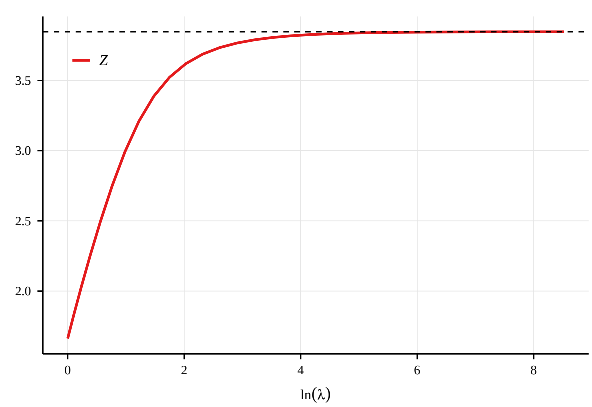
(a) Gini Coefficient




(b) Returns




(c) Agg. Productivity $Z$




The discussion here presages the importance of the wealth distribution in determining the effects of monetary policy. As will become clear in Section 3, different causes for an increase in wealth inequality imply different responses to a policy change: it matters whether high productivity entrepreneurs accumulate capital over time, or borrow it on spot markets. The wealth distribution offers a way to disentangle whether aggregate returns and productivity are being driven by credit markets or by wealth accumulation, and thus what sort of response to policy we may expect.

## 3 Effect of a Monetary Shock

Here, I consider the effect of an unanticipated change in the stance of monetary policy. I assume that, prior to the shock, the economy is in its long-run, zero-inflation steady state as outlined in Section (2.5). Then, at time $t = 0$, there is an unanticipated innovation $\nu_0 < 0$ to the Taylor rule:

$$i_{t+1} = \overline{r} + \phi_\pi \pi_t + \nu_t$$

The shock $\nu_0$ decays at rate $\rho_\nu$, so $\nu_t = \rho_\nu \nu_{t-1}$. Although agents do not anticipate the initial shock $\nu_0$, they understand that it will decay according to the process above, and thus for $t > 0$ we return to a *perfect foresight* equilibrium, where there are no further aggregate shocks and agents perfectly anticipate the evolution of all aggregate variables. For a given variable $X_t$, I denote by $\hat{X}_t \equiv \ln X_t - \ln X$ its log-deviation from steady state.

I begin by focusing on changes in the distribution of wealth, as measured by changes in the wealth shares $\hat{s}_t(z)$ across the space of types $z$. In period 0, when the shock hits, prices rise, so $\pi_0 > 0$. The rise in prices increases aggregate returns to effective capital: $\hat{\omega}_0 > 0$. Intuitively, increased demand for consumption and capital goods pushes up prices. Due to nominal rigidities, not all retailer firms can adjust their prices, and must instead adjust the quantity of goods $x_t$ that they purchase from entrepreneurs. This increase in demand for entrepreneurial goods pushes up marginal costs ($\hat{p}_{tx} > 0$), which increases entrepreneurs' return to effective capital.

The following Lemma describes how these changes manifest in changes to wealth shares $\hat{s}_1(z)$ in the period immediately following the monetary policy shock:

**Lemma 4.** *The changes in wealth shares at* $t = 1$ *immediately following the shock are*

$$\hat{s}_1(z) = \begin{cases} p\beta R(z)\left\{\frac{\lambda-1}{R(z)}\pi_0 + \hat{\omega}_0\left(\frac{\lambda\omega z}{R(z)} - \frac{\omega Z}{R_K}\right)\right\} & z > \underline{z} \\ p\beta R\left\{-\frac{1}{R}\pi_0 - \hat{\omega}_0\frac{\omega Z}{R_K}\right\} & z \leq \underline{z} \end{cases} \tag{46}$$

*where* $\hat{\omega}_0 = \hat{\omega}(\pi_0)$*. Furthermore,* $\hat{s}_1(z) > 0$ *if* $z > \underline{z}$*,* $< 0$ *otherwise.*

Equation (46) reveals two channels by which active entrepreneurs benefit from these higher prices. The first channel relates to the change in inflation $\pi_0$, and is known in the literature as the "Fisher" channel (see, e.g., Auclert, 2019). This channel is familiar: the rise in inflation devalues nominal debts, redistributing real assets from lenders to borrowers. The difference in my model is that this channel now benefits high-productivity entrepreneurs ($z > \underline{z}$) – who are borrowers – at the expense of lower-productivity lenders.

The second channel relates to rising profit margins, $\hat{\omega}(\pi_0)$. The increase in demand leads to a rise in aggregate profit margins $\hat{\omega}_0$, which benefits active entrepreneurs, who produce in the period of the shock ($t = 0$). Intuitively, the Fisher channel redistributes *existing* wealth (carried forward from the period before the shock), while the profit channel dictates that *newly created* wealth in the expansion is distributed in a pattern that favors higher-productivity entrepreneurs.

To shed further light on how this profit channel operates, and how its benefits are tilted towards higher-productivity entrepreneurs, note that the immediate change in wealth share for active entrepreneurs ($z > \underline{z}$) can be written as

$$\hat{s}_1(z) = p\beta R(z)\left\{\varepsilon_{R(z),r}\pi_0 + \hat{\omega}_0\left(\varepsilon_{R(z),\omega} - \varepsilon_{R_K,\omega}\right)\right\} \tag{47}$$

The extent to which active entrepreneurs benefit from the two redistributive channels depends on the elasticity of their returns to aggregate prices. Active entrepreneurs' wealth is increased by inflation – which devalues their debts – in proportion to the elasticity of their returns to the risk-free rate $r$.

Similarly, active entrepreneurs benefit from increases in profits $\hat{\omega}_0$ in proportion to the elasticity of their returns to aggregate profits in the steady state. Here, the relevant measure is the *excess* elasticity $\varepsilon_{R(z),\omega} - \varepsilon_{R_K,\omega}$ , which measures how elastic their returns are to $\omega$ in excess of the elasticity of the aggregate return on capital to $\omega$. idiosyncratic returns – augmented by leverage – imply that active entrepreneurs' returns are *more* elastic to aggregate profit margins than the aggregate return to capital. As a result, when a monetary expansion pushes profit margins up, these households' returns increase by more than the aggregate return. Their wealth in turn grows faster than aggregate wealth, and so their share of aggregate wealth increases further. This channel is stronger, the higher is a household's type $z$.

Figure (5) shows the pattern in $\hat{s}_1(z)$ across $z$: inactive types experience a uniform percentage decline in their wealth shares as a result of inflation, which devalues the nominal contracts written in the period before the shock. Active entrepreneurs benefit both from the inflationary shock, which devalues their *debt*, and from the increase in aggregate demand, visible in $\hat{\omega}_0$. As seen in Figure (5), this change benefits higher-$z$ entrepreneurs more than lower.

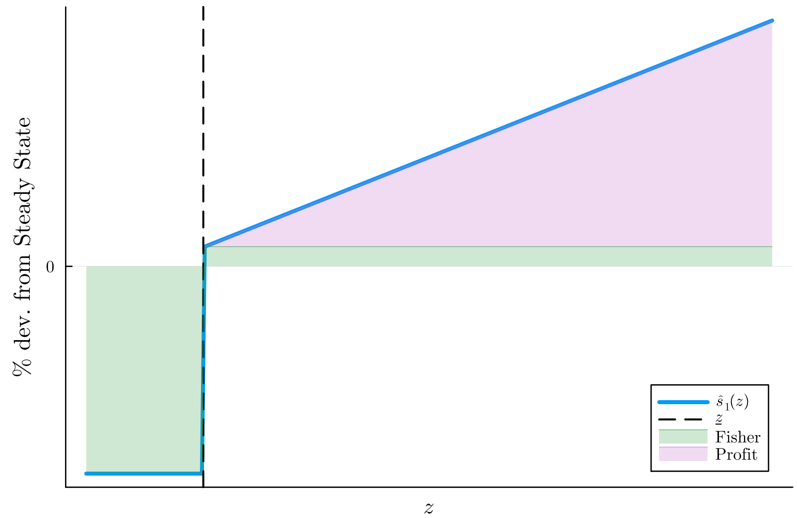
This redistributive channel has the effect, of course, of raising aggregate productivity, further amplifying the effect on output. The redistribution of wealth toward high-types, who benefit from the shock, raises overall productivity. Linearizing the expression for TFP $Z_t$ in Equation (29) and combining with capital market clearing (30), I have the following lemma:

**Lemma 5.** *The change in TFP following the shock $\nu_0$ can be written as*

$$\begin{aligned}\hat{Z}_1 \propto\ & \pi_0\beta p\underbrace{\frac{\lambda-1}{\lambda}}_{D/K}(Z-\underline{z}) && \textit{Inflation}\\ & +\hat{\omega}_0 p\mathbb{E}_s\left[(z-\underline{z})\,MPS(z)\left(\varepsilon_{R(z),\hat{\omega}}-\varepsilon_{R_K,\hat{\omega}}\right)\right] && \textit{Profit}\end{aligned}$$

Lemma 5 shows that the two channels in Lemma 4 both act to increase productivity by redistributing wealth to active entrepreneurs. The inflation channel, which shifts assets from lenders to borrowers (entrepreneurs), increases productivity in proportion to aggregate debt-to-assets in the economy — the more levered production

Figure 5: Wealth Share Change $\hat{s}_1(z)$, Following Shock




is, the larger of an effect this channel has. The increase in profit margins depends on the distribution of excess elasticities $\varepsilon_{R(z),\hat{\omega}} - \varepsilon_{R_K,\hat{\omega}}$ among active entrepreneurs, weighted by their $MPS$ and steady-state wealth shares $s(z)$. Both channels *increase* aggregate TFP, by shifting capital to more productive uses.

By capital market clearing, the cutoff $\hat{\underline{z}}_t$ must rise as well, in order to keep the mass of wealth above the cutoff constant. Mechanically, the entrepreneurs who benefit ($\hat{s}_1(z) > 0$) from the change in policy wish to invest a portion of their new wealth, in doing so they bid up the cost of capital and thus crowd out marginal entrepreneurs, who instead switch to lending. It is here that we see the importance of the wealth distribution in determining the overall effect of the change in policy: not only does investment increase (an effect which obtains even with identical entrepreneurs, or IID shocks), productivity also changes as the composition of investment is altered, with the additional investment being carried out by entrepreneurs who are more productive than average.

Turning to output, the linearized production function has the usual form:

$$\hat{Y}_t = \alpha \left( \hat{Z}_t + \hat{K}_t \right) + (1 - \alpha) \hat{N}_t \tag{48}$$

This expression for output makes clear the amplification in this model, and the surrounding discussion. If shocks are IID, and wealth and productivity are uncorrelated in the steady state, then $\hat{\underline{z}}_t = \hat{s}_t(z) = 0$ for all $t$, and thus $\hat{Z}_t = 0$ as well. In this case, Equation (48) becomes $\hat{Y}_t = \alpha \hat{K}_t + (1 - \alpha) \hat{N}_t$, the standard form with fixed productivity. In this world, there is still amplification through investment: $\hat{K}_t$ will increase, as in e.g. Bernanke et al. (1999), which creates an elevated and long-lived response of output to the policy shift. In the case that $p > 0$, however, the results of Section (3) show that the pattern of redistribution $\hat{s}_t(z)$ is such that $\hat{Z}_t > 0$ for a period following the shock. This creates further amplification of the shock, through an increase in productivity, which also occurs in the data (Christiano et al. 2005, Baqaee et al. 2021). These results further imply that this additional amplification also leads to longer-lasting effects: as the shock fades, entrepreneurs slowly spend down their wealth to return to steady state, leaving both investment and productivity elevated for some time after the shock has faded.

The effect of the steady-state wealth distribution on the economy's response to monetary policy can also be seen through the lens of inflation. The linearized Phillips curve is

$$\pi_t = \kappa_p \hat{p}_{tx} + \beta \mathbb{E}_t \pi_{t+1} \tag{49}$$

where $\kappa_p > 0$ determines the slope of the Phillips curve as a function of the elasticity of substitution and adjustment costs. Clearing in the markets for labor and entrepreneurial goods together imply that

$$\hat{p}_{tx} = \frac{\alpha + \eta}{1 - \alpha} \hat{Y}_t - \frac{\alpha (\eta + 1)}{1 - \alpha} \left( \hat{Z}_t + \hat{K}_t \right) \tag{50}$$

Combining the two gives

$$\pi_t = \kappa_p \left\{ \frac{\alpha + \eta}{1 - \alpha} \hat{Y}_t - \frac{\alpha (\eta + 1)}{1 - \alpha} \left( \hat{Z}_t + \hat{K}_t \right) \right\} + \beta \mathbb{E}_t \pi_{t+1} \tag{51}$$

Above and beyond the impact of additional investment, the change in productivity $\hat{Z}_t$—itself a function of redistribution, per Lemma 5—has the effect of *lowering* the Phillips curve, implying less inflation for a given pattern of economic activity. The size of this disinflationary force is driven by the response of $\hat{Z}_t$, and so it inherits from Lemma 5 its dependence on the initial distribution. The larger are the gains in productivity from redistribution, the larger is the *downward* shift in the Phillips curve.

# 4 Wealth Inequality and Monetary Policy

Section 3 lays out the fact that in this model, one effect of monetary policy is to redistribute wealth towards entrepreneurs with higher idiosyncratic productivities $z$, thereby increasing *aggregate* productivity $Z$. The increase in aggregate productivity amplifies the effect of the shock, flattening the Phillips curve and increasing the amount of output generated for a given change in capital and labor. The implication of Section 3, then, is this: the overall effect of a change in monetary policy depends on the degree to which this policy change redistributes assets among households. I show now that the extent of redistribution, and thus the overall effect of the monetary shock, depends on the distribution of wealth that is present at the time of the policy change.

## 4.1 Extreme Cases: Establishing the Bounds

In studying the effect of wealth inequality on monetary policy, it is clearest to start with the intermediate cases of zero inequality and perfect inequality, which correspond respectively to $p = 0$ and $p = 1$. In the IID case, with $p = 0$, the law of motion for the wealth shares $s_t(z)$ given in equation (25) implies that $s_t(z) = f(z)$ for all $t$ and $z$; the wealth shares are fixed in time. Although entrepreneurs in this case do earn heterogeneous returns, they are just as likely in the next period to be low productivity as they are to be high, and thus wealth shares are unaffected by returns. As a direct result, aggregate productivity and the cutoff $\underline{z}$ are fixed in time as well, equal to

$$\underline{z} = F^{-1}\left(1 - \frac{1}{\lambda}\right) \tag{52}$$

$$Z = \lambda \int_{\underline{z}}^{\overline{z}} z \, \mathrm{d}F(z) \tag{53}$$

At the other extreme of complete inequality ($G = 1$), which corresponds to perfect persistence $p = 1$, a similar result obtains. Because all of the wealth is held by the highest type, no redistribution is possible, and $\underline{z} = Z = \overline{z}$: only the highest types produce, so aggregate productivity coincides with $\overline{z}$. The upshot is that in both cases, the economy acts as one with a representative producer whose productivity is fixed in time. In both cases, the equations governing the evolution of the economy are:

$$\hat{\omega}_t = \hat{p}_{tx} + (1-\alpha)\left(\hat{N}_t - \hat{K}_t\right) \tag{54}$$

$$\pi_t = \kappa_p \hat{p}_{tx} + \beta_f \mathbb{E}_t \pi_{t+1} \tag{55}$$

$$\hat{i}_{t+1} = \phi_\pi \pi_t + \nu_t \tag{56}$$

$$\hat{r}_t = \hat{i}_t - \pi_t \tag{57}$$

$$\hat{\omega}_{t+1} = \frac{1}{r+\delta}\hat{r}_{t+1} \tag{58}$$

$$\hat{N}_t = \frac{1}{\alpha+\eta}\hat{p}_{tx} + \frac{\alpha}{\alpha+\eta}\hat{K}_t \tag{59}$$

$$\hat{Y}_t = \alpha \hat{K}_t + (1-\alpha)\hat{N}_t \tag{60}$$

$$\hat{K}_{t+1} = [1 - \beta(1-\delta)]\left(\hat{p}_{tx} + \hat{Y}_t\right) + \beta(1-\delta)\hat{K}_t \tag{61}$$

Equation (54) pins down the return to effective capital $\hat{\omega}_t$ as a function of aggregates; equation (58) requires that this return comove with the cost of capital, consistent with a constant cutoff $\underline{z}$. Equations (55) and (56) are the linearized Phillips Curve and Taylor rule, respectively. Equation (59) results from clearing in the labor market. Equations (60) and (61) are the linearized equivalents of the production function and capital accumulation equation.

This system can be solved via the method of undetermined coefficients (see, e.g. Chapter 3 of Galí 2015). However, the key property of its behavior following an accommodative shock ($\nu_0 < 0$) can be inferred without any such solution. Note that when linearized about the steady state, the only aggregate steady-state object which appears is $r$, the steady state real interest rate. As such, all economies featuring constant wealth shares and a common interest rate $r$ will behave *identically* following a monetary shock, regardless of any differences in their steady states. In particular: assuming that both economies are calibrated to match the US economy, such that they replicate the real interest rate observed in the data[7], the two economies will respond in the same way to a monetary shock, despite displaying polar opposite levels of wealth inequality. Even without re-calibration, the transition paths in these economies will be very similar, as evidenced by the system in (54)-(61). This equivalence between the two extreme economies lays the foundation for the hump-shaped

[7]In the case of IID shocks, this can be done by choice of $F(z)$ or $\lambda$.

relationship between inequality and the effect of monetary policy: because the response is the same at the polar extremes of inequality, to the extent that interest rate changes redistribute in the intermediate cases, the effect will be larger there. I turn to these cases next.

### 4.2 Intermediate Cases: Countervailing Forces

For intermediate values of wealth inequality, the tractability of my model allows for further insights as to the role of wealth inequality in determining the effects of monetary policy. Recall that the extent of redistribution, measured as the change in wealth shares $\hat{s}_1(z)$ immediately following the shock, can be written as

$$\hat{s}_1(z) = p\beta R(z)\left\{\hat{R}_0(z) - \hat{R}_{0K}\right\} \tag{62}$$

The question that I ask now is how this initial redistribution depends on the persistence $p$. To do so, I fix a value of $\pi_0$, to study how redistribution alters the response of the economy to monetary policy for a given path of prices. Later, I solve for the full response of the economy, including that of prices, computationally.

The channels of redistribution and reallocation are laid out in the following Proposition:

**Proposition 3.** *Given a price change $\pi_0$, for $z > \underline{z}$*

$$\frac{d}{dp}\hat{s}_1(z) = \underbrace{\beta R(z)\left(\hat{R}_0(z) - \hat{R}_{0K}\right)}_{\textit{Term 1}} + p\beta \times \underbrace{\frac{d}{dp}\left\{R(z)\cdot\left(\hat{R}_0(z) - \hat{R}_{0K}\right)\right\}}_{\textit{Term 2}}$$

*Term 1, which is positive for $z > \underline{z}$, captures the redistribution channel. Term 2, which is negative for $z < \underline{z}$, captures the reallocation channel.*

The proof of Proposition 3 can be found in the appendix. This proposition highlights the countervailing forces that affect redistribution as wealth inequality, driven by changes in persistence, increases. The first term, redistribution, is positive. To understand the intuition behind the redistribution channel, fix a type $z$ and a level of excess returns $\widehat{ER}_0(z) = \hat{R}_0(z) - \hat{R}_{0K}$. Term 1 measures the extent to which this change in excess returns filters into wealth shares: the more persistent are returns, the greater is the scope for agents who earn returns in excess of the return on capital following the shock to *accumulate* from those returns. Term 2, however, is negative for $z < \underline{z}$, and captures the reallocation channel. This term captures, for a given level of persistence (and thus scope for accumulation), what level of excess returns the agents can expect to earn following the shock. To understand this term, recall the comparative statics on steady-state returns in Figure 2b. As wealth concentration increases, the idiosyncratic returns $R(z)$ for active entrepreneurs ($z > \underline{z}$) align with the return on capital. Proposition 3 shows that this property of the steady state, also holds along the transition path following a monetary shock: as $p \to 1$, idiosyncratic returns $\hat{R}_0(z)$ align with the return on capital $\hat{R}_{0K}$, and thus $\widehat{ER}_0(z)$ declines in $p$.

Figure 6 illustrates the results of Proposition 3. In each panel, I plot the responses of the variables in question again given a fixed price response $\pi_0$. Panel 6a shows the initial change in the wealth share of the highest type, $\hat{s}_1(\overline{z})$, following the shock across persistence $p$. As illustrated in Lemma 4, the changes for all of the remaining active entrepreneurs will be bounded above by $\hat{s}_1(\overline{z})$. Panel 6a shows the hump-shaped response of wealth shares, as a result of the redistribution and reallocation channels. Panel 6b shows the derivative $d\hat{s}_1(\overline{z})/dp$ in green, and the two terms in blue and orange respectively. As in Proposition 3, the first term is positive, reflecting the benefit of accumulation that comes from higher persistence. The second term is negative, reflecting the logic of the reallocation channel: further wealth concentration in the steady state reduces the scope for active entrepreneurs to earn returns in excess of the return on capital. Panel 6c illustrates this point directly: analogously to Figure 2b, Panel 6c shows the idiosyncratic return of type-$\overline{z}$ agents immediately following the shock, $\hat{R}_0(\overline{z})$, as well as the return on capital $\hat{R}_{0K}$, for fixed $\pi_0$ across $p$. Note that for a given change in prices, the response of the return on capital $\hat{R}_{0K}$ is fixed and independent of steady-state wealth concentration. As in Figure 2b, we see that as $p$ increases, the return earned by these top entrepreneurs converges to the return on capital, and their difference $\widehat{ER}_0(\overline{z})$ in green converges to zero.

Do note that each of the panels in Figure 6 shows a discontinuity that is characteristic of the steady state in this model: for any $p$ arbitrarily close to one, credit markets will be active, and entrepreneurs will borrow up to their debt limit. At $p = 1$, however, the top types hold all of the wealth, and the credit markets vanish—inactive entrepreneurs no longer hold any wealth to lend to the highest types. The discontinuity reflects this fact: for any $p$ near but less than one, active entrepreneurs still benefit from the Fisher channel described in Section 3, and thus still enjoy a small amount of redistribution from inflation. At $p = 1$, meanwhile, credit markets are inactive, and the idiosyncratic return of the top types is by definition equal to the return on capital, thus shrinking redistribution to zero.

With the results of Proposition 3 established, the connection between wealth inequality and monetary policy is immediate. Because the magnitude of redistribution changes with wealth inequality, so too does the magnitude of the response of productivity to the shock. The amplification of the shock, and therefore its overall affect on aggregates, then depends on how unequally wealth is distributed to begin with. To pin down this relationship, the following Corollary calculates the derivative of the initial change in TFP $\hat{Z}_1$ in $p$:

**Corollary 1.** *Given a price change $\pi_0 > 0$,*

$$\frac{d}{dp}\hat{Z}_1 = \underbrace{\mathbb{E}_s\left[(z-\underline{z}) \times \frac{d}{dp}\hat{s}_1(z)\,|z>\underline{z}\right]}_{\lesseqgtr 0} \qquad ①$$

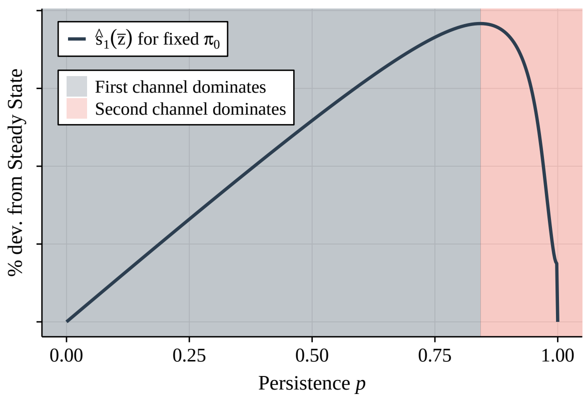
$$+\underbrace{\mathbb{E}_s\left[\frac{1}{s(z)}\left\{\frac{d}{dp}s(z)\right\}(z-\underline{z})\,\hat{s}_1(z)\,|z>\underline{z}\right]}_{>0} \qquad ②$$

$$+\underbrace{\mathbb{E}_s\left[\left(-\frac{d\underline{z}}{dp}\right)\hat{s}_1(z)\,|z>\underline{z}\right]}_{<0} \qquad ③$$

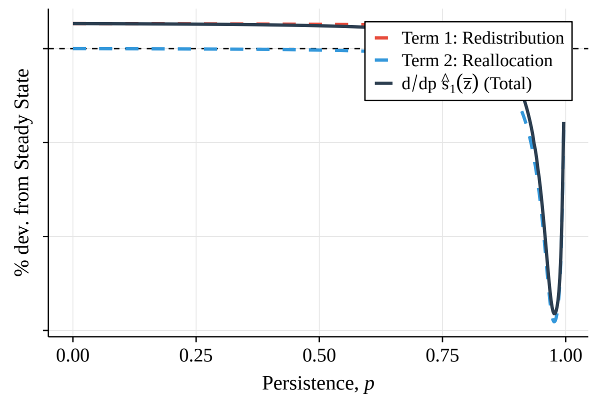
Corollary 1 says that the response of aggregate TFP to the monetary shock is *also* subject to countervailing forces in the persistence of entrepreneurs' returns. First and foremost, whether $\hat{Z}_1$ is increasing or decreasing in wealth inequality depends on the response of the wealth shares, as in Proposition 3. From the lens of

Figure 6: Redistribution across Persistence

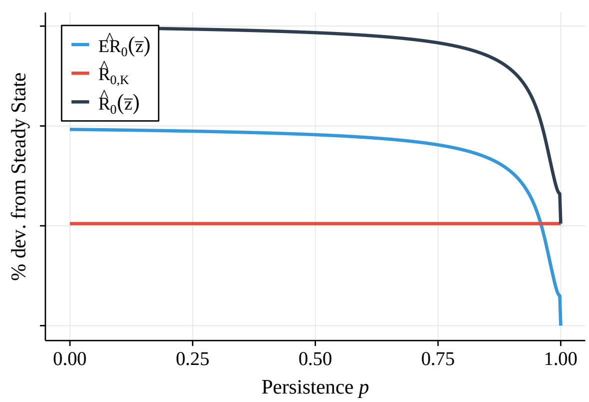
(a) Change in Top Wealth Share $\hat{s}_1(\overline{z})$, fixed $\pi_0$




(b) Derivative $\frac{d}{dp}\hat{s}_1(\overline{z})$




(c) Alignment of Returns in $p$



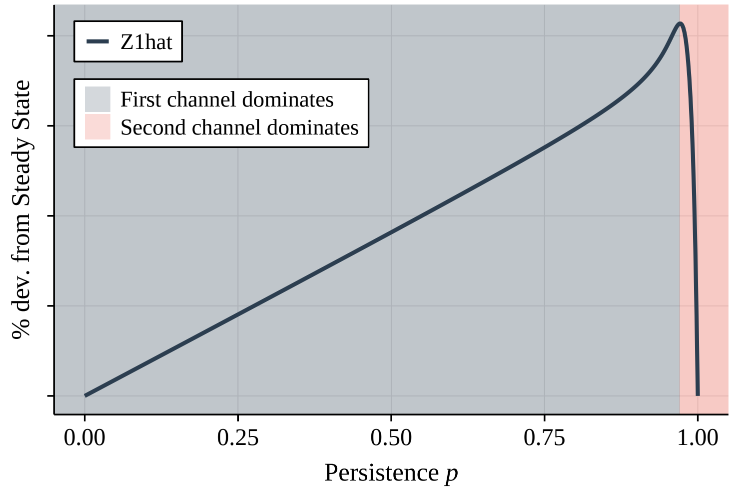
aggregate productivity, higher persistence implies that the initial redistribution—which benefits agents with high $z$ types in the period of the shock—is *targeted*, in the sense that it benefits agents who will have high productivity *tomorrow*. All else being equal, additional persistence implies that redistribution has a larger effect on productivity, as the effect of the shock will be to redistribute wealth to entrepreneurs of higher productivity in the period following the shock, thereby increasing aggregate productivity. Of course, as $p$ increases the size of redistribution $\hat{s}_1(z)$ for $z > \underline{z}$ shrinks with these entrepreneurs' excess returns, as discussed above.

There are two additional forces which determine the magnitude of $\hat{Z}_1$ in Corollary 1. Tern ② captures the fact that increases in $p$ also alter the steady-state distribution $s(z)$, which determines how changes in wealth shares post-shock $\hat{s}_1$ are translated into productivity gains. All else equal, a given pattern of redistribution $\hat{s}_1$ will have a greater effect on productivity if its beneficiaries have greater wealth to begin with, as this gives them more weight in the expectation that determines $\hat{Z}_1$. As discussed in Section 2.5, increases in $p$ shift wealth toward high-$z$ entrepreneurs, which increases the response $\hat{Z}_1$ holding the pattern of redistribution $\hat{s}_1$ fixed. However, Term ③ captures a downward effect similar to that in Proposition 3: as $p$ increases, the cutoff productivity $\underline{z}$ increases as well. This shrinks the response of $\hat{Z}_1$ to redistribution $\hat{s}_1(z)$, for a similar reason as before: as $\underline{z}$ increases, the excess productivity $z - \underline{z}$ shrinks for all active $z$. Effectively, with greater wealth concentration the beneficiaries of redistribution have productivities closer to the cutoff, and so the productivity gains to giving these entrepreneurs a bit more wealth shrink.

Figure 7 illustrates the response of $\hat{Z}_1$ that results from Corollary 1, across $p$ and again for fixed $\pi_0$. Here again, the hump-shaped response in $\hat{Z}_1$ is evident: aggregate TFP does not respond to a monetary shock at $p = 0$ and $p = 1$, and the magnitude of its response peaks between these two values. Notably, the response of $\hat{Z}_1$ peaks at a higher value of $p$ than does the response $\hat{s}_1(\overline{z})$ in Figure 6a. This result is due to Tern ② in Corollary 1: higher persistence shifts steady-state wealth shares towards agents with higher $z$, which increases the response of $\hat{Z}_1$ to any given pattern of redistribution, $\hat{s}_1(z)$. Crucially, it implies that even as the magnitude of $\hat{s}_1(\overline{z})$ begins to decline in $p$, the productivity response continues to rise in $p$: even though redistribution is smaller, it is targeted towards households who exert a greater effect on aggregate productivity due to their higher shares of aggregate wealth. This is the benefit of higher wealth inequality among firms in Baqaee et al. (2021), and households in Colciago et al. (2019) — essentially, this feature results from the redistribution channel, which their results highlight. However, Figure 7 demonstrates that these benefits fade away for high values of wealth inequality: as wealth concentration nears its maximum, the response of TFP to monetary policy begins to fade to zero.

As a result of the hump-shaped pattern of the TFP response $\hat{Z}_1$ in Figure 7, the *overall* effect of monetary policy will be be hump-shaped in wealth inequality. In essence, the degree of wealth inequality present at the time of a change in the policy rate determines how much redistribution the policy rate creates, and thus how much the response is amplified by the resulting change in aggregate productivity. The results here also imply that beyond a certain point, increases in wealth inequality *dampen* the efficacy of monetary policy.

Figure 7: Productivity Response $\hat{Z}_1$ (fixed $\pi_0$)




# 5 Quantitative Results

In Sections 3 and 4, I study theoretically the response of my economy to monetary policy, and measure how that response depends on wealth inequality. Here, I calibrate my model to match moments of the joint distribution of wealth and returns in the US economy as of 2022, to assess the quantitative relevance of the channels studied above. To begin, I discuss my strategy for calibrating my model to the data. Then, with the calibrated model in hand, I ask a few questions about how the model speaks to the data.

First, I show that this model captures the response of the economy to monetary policy fairly well; in particular, it captures movements in TFP and inequality resulting from monetary easings. Then, I show that the degree of amplification resulting from the redistribution in assets is substantial, implying that this is a meaningful channel of monetary policy to capture. My calibrated model also implies a markedly different relationship between output and inflation than in standard RANK and HANK models, again arguing that heterogeneity in household returns is an important source to capture. Finally, I show that my calibration suggests that the increase in wealth inequality since the 1970s can partially account for the decrease in monetary policy that has been documented in the empirical literature over the same time period.

## 5.1 Calibration

The time period in my model is one quarter. My parameters are summarized in Table 1. To begin, in my model the wealth-weighted average return to wealth in the steady state is equal to $1/\beta$, and so I set this to match the average return in Fagereng et al. (2020), approximately 0.9 percent quarterly. Depreciation is then set so that the quarterly capital stock to output ratio is 12. Note that in my model, this ratio is given by

$$\frac{K}{Y} = \frac{\alpha p_x}{\beta^{-1} - (1 - \delta)} \tag{63}$$

Table 1: Quarterly Calibration, 2022

| *Parameter* | *Value* | *Source* |
|---|---|---|
| $\beta$ | 0.99 | Avg. return to wealth, Fagereng et al., 2020 |
| $\delta$ | 0.021 | $K/Y$ of 12 |
| $\eta$ | 2 | Labor elasticity (standard) |
| $\lambda$ | 1.43 | Debt/Assets (FRB) |
| $\varepsilon$ | 10 | Avg. Markup 11% |
| $\theta$ | 90 | Slope of Phillips Curve 0.1 |
| $\alpha$ | 0.4 | Capital share $\alpha p_x = 0.36$ (standard) |
| $\phi_\pi$ | 1.5 | Standard value |
| $\rho_\nu$ | 0.6 | Standard value |
| $\mu_w$ | 0.4 | Lorenz Curve (SCF) |
| $p$ | 0.98 | Wealth Gini of 0.8 (SCF) |
| $\sigma_z$ | 2.75 | Cross-section of returns (see text) |

As a result of my assumption on workers' preferences, workers supply labor with elasticity $1/\eta$. McClelland and Mok (2012) report an estimate of this elasticity of about 0.5, implying $\eta = 2$. These households hold no wealth, and so I assume that they make up 40% of the population ($\mu_w = 0.4$), corresponding to the fact that in the SCF, the bottom 40% of households by wealth cumulatively hold approximately zero net worth.

For $\lambda$: entrepreneurs in my model borrow a proportion $\lambda - 1$ of their assets in debt, meaning for active entrepreneurs, $d/k = (\lambda - 1)/\lambda$. I therefore set $\lambda$ so that firms in equilibrium have a debt to assets ratio of 0.44, which is the long-run average debt-to-assets ratio as measured by the Federal Reserve Board's Financial Accounts of the United States for corporate and non-corporate businesses combined. I choose this particular leverage ratio to capture the debt that backs productive business assets, which are a primary asset class among wealthy households with high returns.

The Phillips curve parameters $\varepsilon$ and $\theta$ follow Kaplan et al. (2018), and imply a Phillips curve slope of 0.1, as in Schorfheide (2008). The capital loading factor $\alpha$ is then set so that the capital share $\alpha p_x$ is 0.36, a standard value. The weight on inflation in the Taylor Rule $\phi_\pi$ and the persistence $\rho_\nu$are assigned standard values from the literature.

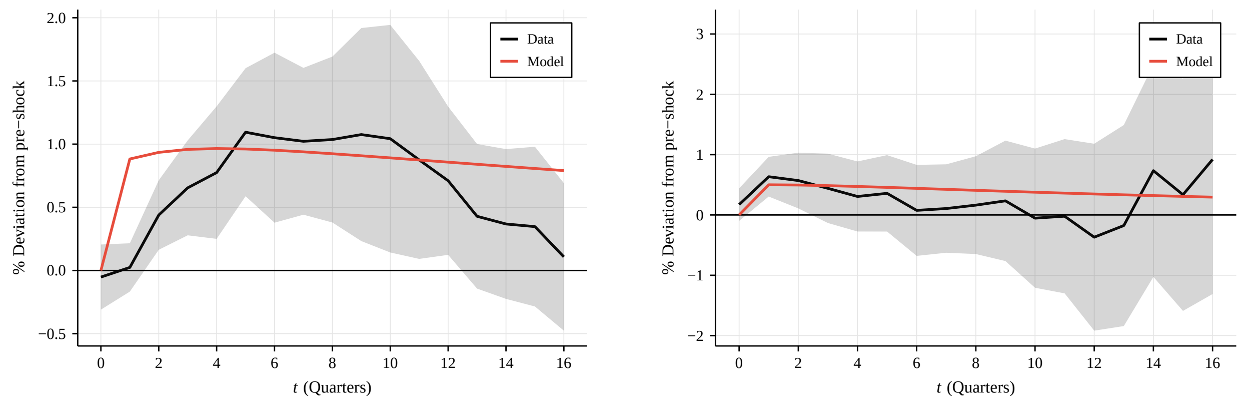
What remains are the persistence of entrepreneurs returns $p$ and the distribution of types $F(z)$. The degree of persistence $p$ matches a wealth Gini of 0.8, corresponding to the average wealth Gini in the SCF from 1970-2022. The volatility in returns $\sigma_z$, meanwhile, targets the wealth-weighted cross sectional standard deviation of returns to wealth of about 1.95% quarterly, as documented in Fagereng et al. (2020).

## 5.2 Impulse Responses in the Calibrated Model

Here, I discuss the response of my economy to a 100bp monetary shock. To calculate the full impulse response, I use a computational strategy, which I outline in Appendix 7.4. To begin, Figure 8 shows the responses of two variables of particular interest in my model: aggregate TFP $\hat{Z}_t$, and the wealth Gini coefficient $\hat{G}_t$. The impulse responses in red are computed from my model; those in black are estimated using Jordà (2005) local projections.

Figure 8: Impulse Responses to 100bp monetary shock

(a) Total Factor Productivity $\hat{Z}_t$ (b) Gini Coefficient $\hat{G}_t$




*Note: grey shaded areas indicate 90% confidence intervals using Newey-West standard errors.*

To estimate the empirical response of total factor productivity, I use data on aggregate TFP from the Federal Reserve Bank of San Francisco and follow Baqaee et al. (2021) in estimating the following specification:

$$z_{t+h} = A + \sum_{k=0}^{4} B_{h,k}\nu_{t-k} + \sum_{k=1}^{4} C_{h,k}z_{t-k} + \epsilon_t \tag{64}$$

Here, the terms $\nu_t$ are monetary shocks – identified via the narrative approach in Romer and Romer (2004) and extended through 2012 by Wieland and Yang (2020) – and $z_t$ is total factor productivity for the U.S. economy. The impulse responses, then, are the $B_{h,0}$ terms, which measure the impact of a shock in the current period on $h$-period ahead forecasts of TFP.

To estimate the response of inequality, I use data on wealth shares from the Federal Reserve Board's Distributional Financial Accounts (DFA) for the U.S. This dataset contains a quarterly time series on wealth shares for the bottom 50% of the wealth distribution, as well as top shares for selected percentile bins. Using these I compute the Gini coefficient by reconstructing the Lorenz curve using the procedure in Medlin (2023). One can verify that the estimates of the Gini coefficient constructed from the DFA data track those implied by the triennial SCF waves in both magnitude and time-series pattern. With the quarterly Gini estimates in hand, I estimate the response of the Gini coefficient to the identified monetary shocks using the same specification as in (64).

At its peak, TFP $\hat{Z}_t$ in Panel 8a increases by about 1% relative to the steady state — a similar magnitude to its response in the data. On inequality $\hat{G}_t$, the model fit is similarly good: the Gini coefficient increases at its peak by about 0.5%, about four-fifths the size of the increase in the Gini Coefficient that I estimate in the data, as well as that in Medlin (2023). In total, Figure 8 suggests that despite being relatively parsimonious, the model can accurately capture the responses of both wealth inequality and productivity to a monetary shock — responses that are not captured in standard HANK models.

Figure 9 shows the responses of output and investment in the calibrated model. Unsurprisingly, both

Figure 9: Impulse Responses to 100bp monetary shock

(a) Output $\hat{Y}_t$ (b) Investment $\hat{I}_t$

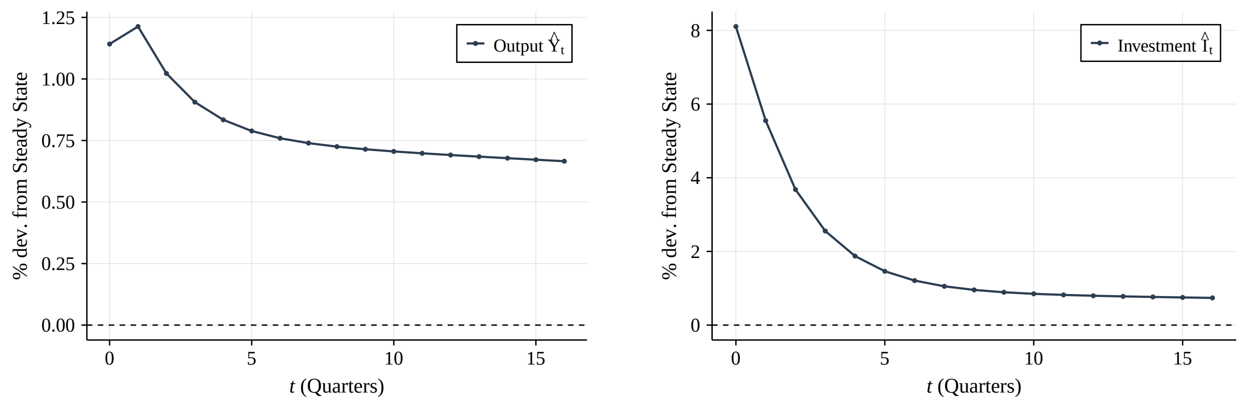
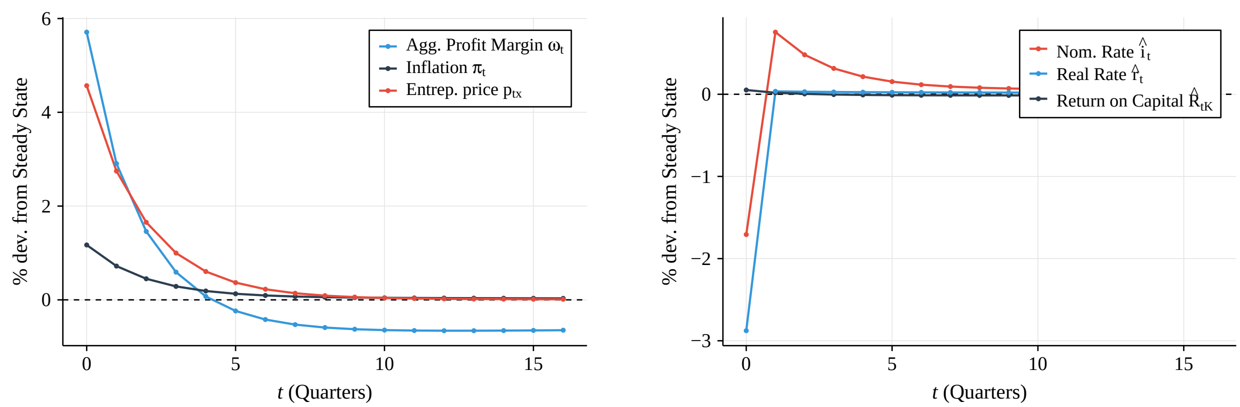



Figure 10: Impulse Responses to 100bp monetary shock

(a) Prices (b) Rates



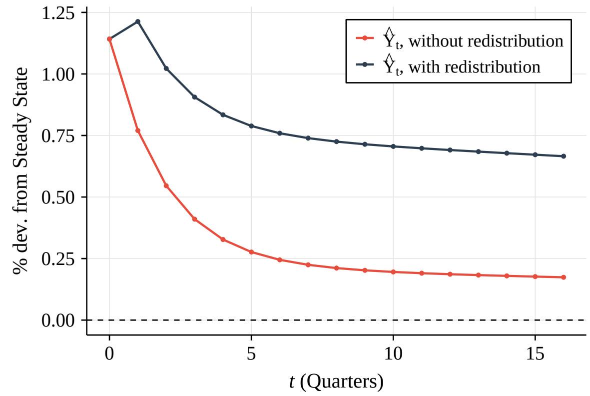

increase, and investment by more than output, as is the case in the data. Figure 10, meanwhile, shows the remaining impulse responses. In particular, Panel 10a shows the response of the overall price level $\pi_t$, as well as the response of entrepreneurs' price $\hat{p}_{tx}$ and profit margin $\hat{\omega}_t$. Here, both the price that entrepreneurs earn on their goods $\hat{p}_{tx}$ and their profit margin per unit of effective capital $\hat{\omega}_t$ both rise in response to the increase in demand. As a result, then, the overall price level rises as well. Panel 10b, meanwhile, shows the impulse responses of the nominal and real interest rates $\hat{i}_t$ and $\hat{r}_t$ respectively, as well as the response of the aggregate return on capital $\hat{R}_{tK}$. The real interest rate falls as the monetary authority loosens, as intended, and then rises following the subsequent tightening that the policymaker enacts to counter inflation. As a result of the rise in prices and returns, as well as the redistribution that follows, the return on capital $\hat{R}_{tK}$ rises as well.

## 5.3 Amplification

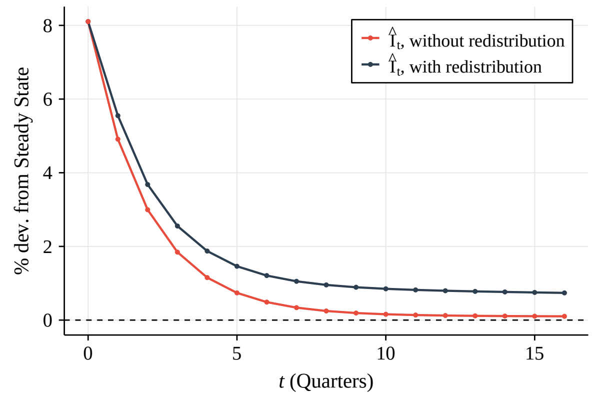
Here, I provide two arguments as to the empirical significance of the redistributive channel of monetary policy. To begin, Figure 11 assesses the effect of redistribution on output $\hat{Y}_t$ and investment $\hat{I}_t$. For each variable,

Figure 11: Amplification through Redistribution

(a) Impulse Response to 100bp shock: Output $\hat{Y}_t$




(b) Impulse Response to 100bp shock: Investment $\hat{I}_t$




I plot two impulse responses to the same 100bp monetary shock as above. In blue, I plot the full impulse response, allowing for redistribution, which is identical to that in Figure 9. In red, meanwhile, I plot the impulse response of an identical economy with redistribution shut down: I begin from the same steady state, but fix wealth shares $s_t(z)$ following the shock. Clearly, the redistributive channel exerts a strong influence on output in particular: the increase in productivity that results from shifting assets towards higher-productivity entrepreneurs implies that over the life of the shock, the cumulative change in output is about fifty percent higher than the world without redistribution.

Figure 12 shows the flattening of the Phillips curve described theoretically in Section 3. For this experiment, xI consider a range of monetary shocks $\nu_0$ ranging from -200bp to 200bp. For each, I calculate the response of output $\hat{Y}_1$ and inflation $\pi_1$ in response to the shock, both in the period *after* the shock in order to account for any potential redistributive effects. I plot the former against the latter in Figure 12. I repeat this experiment for the two economies described immediately above: one with redistribution, and the other without, both starting from the same initial conditions. As demonstrated by Figure 12, allowing for changes in productivity resulting from redistribution meaningfully alters this relationship.

For the theoretical reasons discussed in Section 3, the rise in productivity implies that the marginal cost to producing an additional unit of output falls, which implies less inflation for a given increase in output. Figure 12 shows this directly: the slope of the output-inflation line is meaningfully smaller than in the economy without redistribution. Quantitatively, allowing for redistribution flattens the slope of the Phillips curve by about 45%. This is a relevant comparison for two reasons, one normative and one positive. On the positive side, standard HANK models assume that production is carried out by a representative producer, thereby shutting down the redistributive, supply-side channel in my model—analogous to the orange line in Figure 12. Furthermore, when evaluating *optimal* monetary policy, the tradeoff between output and inflation is the fundamental choice faced by the policymaker. As a result, the optimal response of monetary policy to a given shock will be altered in the presence of redistribution.

Figure 12: Phillips Curve Slope

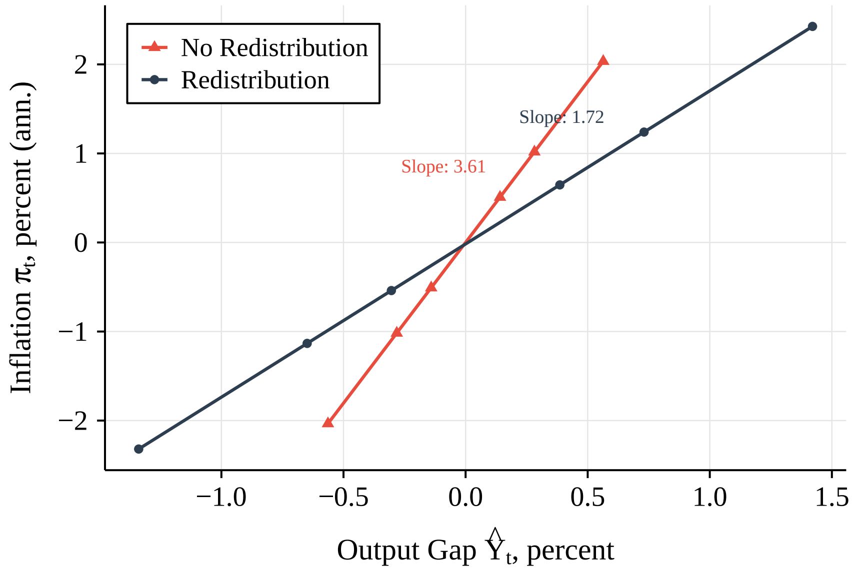



Table 2: Calibration, 1970s and 2010s

| Parameter | Value, 1970s | Value, 2010s |
|---|---|---|
| $p$ | 0.980 | 0.985 |
| $\sigma_z$ | 1.15 | 3.90 |
| Gini | 0.7 | 0.82 |

### 5.4 Increase in Inequality

Since the early 1970s, two trends are present in the data. First, and famously, wealth inequality has increased: I measure using historical SCF data that the wealth Gini coefficient for the US has increased from 0.7 in 1971, to 0.85 in 2022. Concurrently, the efficacy of monetary policy shocks has decreased; for example, Boivin et al. (2010) estimate that the effect of an identified monetary shock on output is about half as large now as in the 1960s and early 1970s. Because my model suggests that increases in inequality can dampen the effects of monetary policy, it offers a natural lens with which to explore whether these two trends are related.

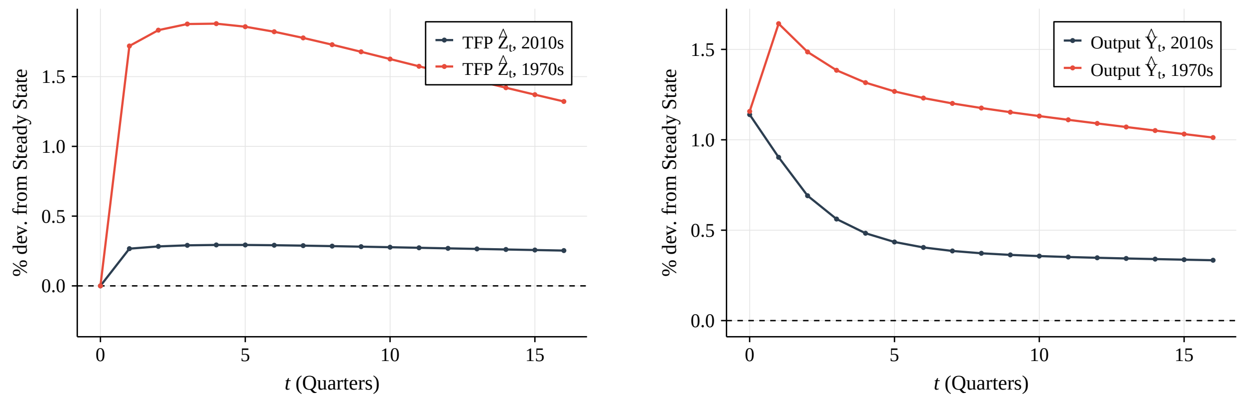
To study this question, I consider two calibrations of my model. Holding all other parameters fixed, I adjust return persistence $p$ and volatility $\sigma_z$ to target two levels of wealth inequality: the Gini Coefficient in 1971, and the Gini coefficient in 2012, both measured in the SCF. Table 2 summarizes the two calibrations. My model attributes the increase in wealth inequality between these two years to an increase in both persistence *and* volatility.

The impulse responses in the two economies, presented in Figure 13, are markedly different. As compared to the 2022 economy, a monetary shock to the 1970s economy generates a much larger increase in productivity, as seen in Panel 13a. Per the results in Section 4, this change in the productivity suggests that as inequality has increased since the 1970s, the *reallocation* channel has dominated, and the efficacy of policy has been muted

Figure 13: Impulse Responses: 1970 and 2022

(a) Productivity (TFP) Impulse Responses, 1970 and 2022

(b) Output Impulse Responses, 1970 and 2022




as the productivity channel has shut down. Because the productivity response is smaller in the 2022 economy, so too is the response of output, as shown in Panel 13b. Thus, my calibration suggests that the increase in wealth inequality since the 1970s is indeed partially responsible for the decrease in the efficacy of policy over the same period. Additionally, one finding in the literature on the changing efficacy of monetary policy over time is that, relative to the 1960s-1970s, the response of output in the post-1970s period to monetary policy peaks at a lower level, but persists for longer. This is exactly the pattern of output in Panel 13b: following the shock, output in the 2022 economy with higher wealth inequality peaks at a lower level, but remains elevated for longer than in the 1970 economy.

# 6 Conclusion

I argue here two key points concerning the effect of monetary policy on economies with unequal wealth distributions generated by households who earn persistently different returns on their investments. First, in this framework, redistribution of wealth among households is a key component of the transmission of monetary policy: in particular, a reduction in interest rates ultimately redistributes from low-return households to those with higher returns. Second, the economy's response to monetary policy is determined by the wealth distribution, and the underlying process for household returns that generates it.

This paper reconciles two findings in the data: that expansionary monetary policy increases productivity and wealth inequality. It also has important implications for the *optimal* conduct of monetary policy (see González et al. 2024 for further discussion in a similar framework). My model can also make some headway in explaining empirical evidence that suggests the effects of monetary shocks have decreased over time (e.g. Canova and Gambetti, 2009; Boivin et al., 2010), as in the US the concentration of wealth in the hands of the most successful entrepreneurs has increased. The most important implication, in my opinion, is this: my paper contributes to a growing notion in the literature on monetary policy that measuring and predicting responses to changes in interest rates cannot be done by observing aggregates alone, and that distributions play an equally important role. Where many early papers in this strain emphasize the importance of heterogeneity in marginal propensities to *consume*, I stress that marginal propensities to *invest*, and their correlation with wealth, are of equal importance.

# References


Auclert, Adrien (2019). "Monetary policy and the redistribution channel". In: *American Economic Review* 109.6, pp. 2333–67.

Auclert, Adrien, Rognlie, Matthew, and Straub, Ludwig (2020). *Micro jumps, macro humps: Monetary policy and business cycles in an estimated HANK model*. Tech. rep. National Bureau of Economic Research.

— (2025). "Fiscal and monetary policy with heterogeneous agents". In: *Annual Review of Economics* 17.

Baqaee, David, Farhi, Emmanuel, and Sangani, Kunal (2021). *The supply-side effects of monetary policy*. Tech. rep. National Bureau of Economic Research.

Benhabib, Jess, Bisin, Alberto, and Luo, Mi (2019). "Wealth distribution and social mobility in the US: A quantitative approach". In: *American Economic Review* 109.5, pp. 1623–47.

Benhabib, Jess, Bisin, Alberto, and Zhu, Shenghao (2015). "The wealth distribution in Bewley economies with capital income risk". In: *Journal of Economic Theory* 159, pp. 489–515.

Bernanke, Ben S, Gertler, Mark, and Gilchrist, Simon (1999). "The financial accelerator in a quantitative business cycle framework". In: *Handbook of macroeconomics* 1, pp. 1341–1393.

Boivin, Jean and Giannoni, Marc P (2002). "Assessing changes in the monetary transmission mechanism: A VAR approach". In: *Economic Policy Review* 8.1.

Boivin, Jean, Kiley, Michael T, and Mishkin, Frederic S (2010). "How has the monetary transmission mechanism evolved over time?" In: *Handbook of monetary economics*. Vol. 3. Elsevier, pp. 369–422.

Buera, Francisco J and Moll, Benjamin (2015). "Aggregate implications of a credit crunch: The importance of heterogeneity". In: *American Economic Journal: Macroeconomics* 7.3, pp. 1–42.

Canova, Fabio and Gambetti, Luca (2009). "Structural changes in the US economy: Is there a role for monetary policy?" In: *Journal of Economic dynamics and control* 33.2, pp. 477–490.

Carlstrom, Charles T and Fuerst, Timothy S (1997). "Agency costs, net worth, and business fluctuations: A computable general equilibrium analysis". In: *The American Economic Review*, pp. 893–910.

Christiano, Lawrence J, Eichenbaum, Martin, and Evans, Charles L (2005). "Nominal rigidities and the dynamic effects of a shock to monetary policy". In: *Journal of political Economy* 113.1, pp. 1–45.

Colciago, Andrea, Samarina, Anna, and Haan, Jakob de (2019). "Central bank policies and income and wealth inequality: A survey". In: *Journal of Economic Surveys* 33.4, pp. 1199–1231.

DeBacker, Jason, Panousi, Vasia, and Ramnath, Shanthi (2023). "A risky venture: Income dynamics among pass-through business owners". In: *American Economic Journal: Macroeconomics* 15.1, pp. 444–474.

Del Negro, Marco, Lenza, Michele, Primiceri, Giorgio E, and Tambalotti, Andrea (2020). *What's up with the Phillips Curve?* Tech. rep. National Bureau of Economic Research.

Fagereng, Andreas, Guiso, Luigi, Malacrino, Davide, and Pistaferri, Luigi (2020). "Heterogeneity and persistence in returns to wealth". In: *Econometrica* 88.1, pp. 115–170.

Feilich, Ethan (2021). "Monetary policy and the dynamics of wealth inequality." In.

Gabaix, Xavier (2009). "Power laws in economics and finance". In: *Annu. Rev. Econ.* 1.1, pp. 255–294.

Galí, Jordi (2015). *Monetary policy, inflation, and the business cycle: an introduction to the new Keynesian framework and its applications*. Princeton University Press.

González, Beatriz, Nuño, Galo, Thaler, Dominik, and Albrizio, Silvia (2024). "Firm heterogeneity, capital misallocation and optimal monetary policy". In.

Gouin-Bonenfant, Emilien and Toda, Alexis Akira (2019). "Pareto extrapolation: An analytical framework for studying tail inequality". In: *Available at SSRN 3260899*.

Greenwood, Jeremy, Hercowitz, Zvi, and Huffman, Gregory W (1988). "Investment, capacity utilization, and the real business cycle". In: *The American Economic Review*, pp. 402–417.

Hazell, Jonathon, Herreno, Juan, Nakamura, Emi, and Steinsson, Jón (2022). "The slope of the Phillips Curve: evidence from US states". In: *The Quarterly Journal of Economics* 137.3, pp. 1299–1344.

Jeenas, Priit (2023). *Firm balance sheet liquidity, monetary policy shocks, and investment dynamics*. Universitat Pompeu Fabra, Department of Economics and Business.

Jordà, Òscar (2005). "Estimation and inference of impulse responses by local projections". In: *American economic review* 95.1, pp. 161–182.

Kaplan, Greg, Moll, Benjamin, and Violante, Giovanni L (2018). "Monetary policy according to HANK". In: *American Economic Review* 108.3, pp. 697–743.

Kiyotaki, Nobuhiro (1998). "Credit and business cycles". In: *The Japanese Economic Review* 49.1, pp. 18–35.

Matusche, Alexander and Wacks, Johannes (2021). "Monetary Policy and Wealth Inequality: The Role of Entrepreneurs". In.

McClelland, Robert and Mok, Shannon (2012). "A review of recent research on labor supply elasticities". In.

McKay, Alisdair, Nakamura, Emi, and Steinsson, Jón (2016). "The power of forward guidance revisited". In: *American Economic Review* 106.10, pp. 3133–3158.

Medlin, Aaron (2023). "Federal Reserve Monetary Policy and Wealth Inequality: An instrumental-variable local projections approach". In: *Available at SSRN 4650805*.

Melcangi, Davide and Sterk, Vincent (2024). "Stock market participation, inequality, and monetary policy". In: *Review of Economic Studies*, pp. 2656–2690.

Moll, Benjamin (2014). "Productivity losses from financial frictions: Can self-financing undo capital misallocation?" In: *American Economic Review* 104.10, pp. 3186–3221.

Ottonello, Pablo and Winberry, Thomas (2020). "Financial heterogeneity and the investment channel of monetary policy". In: *Econometrica* 88.6, pp. 2473–2502.

Romer, Christina D and Romer, David H (2004). "A new measure of monetary shocks: Derivation and implications". In: *American Economic Review* 94.4, pp. 1055–1084.

Rotemberg, Julio J (1982). "Monopolistic price adjustment and aggregate output". In: *The Review of Economic Studies* 49.4, pp. 517–531.

Schorfheide, Frank (2008). "DSGE model-based estimation of the New Keynesian Phillips curve". In: *FRB Richmond Economic Quarterly* 94.4, pp. 397–433.

Wieland, Johannes F and Yang, Mu-Jeung (2020). "Financial dampening". In: *Journal of Money, Credit and Banking* 52.1, pp. 79–113.

Xavier, Inês (2021). "Wealth inequality in the US: the role of heterogeneous returns". In: *Available at SSRN 3915439*.

# 7 Appendix

## 7.1 Proof of Proposition (1)

*Proof.* I begin with aggregate output. Given the optimal choice of labor in Lemma 1, the output of an entrepreneur with productivity $z$and capital $k_t$ is

$$y_t(z, k_t) = \frac{\omega_t}{\alpha} z k_t$$

Integrating over entrepreneurs using the stationary distribution $g_t(a, z)$, and using the fact that $k_t(a, z) = \lambda a$ if $z > \underline{z}_t$ and 0 otherwise, gives

$$\begin{aligned} Y_t &= \frac{\omega_t}{\alpha} \lambda \int_{\underline{z}}^{\infty} \int_{0}^{\infty} a z g_t(a, z) \, \mathrm{d}a \, \mathrm{d}z \\ &= \frac{\omega_t}{\alpha} \lambda K_t \int_{\underline{z}}^{\infty} z s_t(z) \, \mathrm{d}z \\ &= \frac{\omega_t}{\alpha} \lambda K_t X_t \end{aligned}$$

where

$$X_t = \int_{\underline{z}}^{\infty} z s_t(z) \, dz$$

From the labor market, we have

$$\begin{aligned} N_t &= \left(\frac{\omega_t}{\alpha}\right)^{\frac{1}{1-\alpha}} \lambda \int_{\underline{z}}^{\infty} \int_{0}^{\infty} a z g_t(a, z) \, \mathrm{d}a \, \mathrm{d}z \\ &= \left(\frac{\omega_t}{\alpha}\right)^{\frac{1}{1-\alpha}} \lambda K_t X_t \end{aligned}$$

which implies

$$\omega_t = \alpha \left(\frac{N_t}{\lambda K_t X_t}\right)^{1-\alpha}$$

so production is

$$\begin{aligned} Y_t &= \frac{\omega_t}{\alpha} \lambda K_t X_t \\ &= \left(\frac{N_t}{\lambda K_t X_t}\right)^{1-\alpha} \lambda K_t X_t \\ &= (\lambda X_t K_t)^{\alpha} N_t^{1-\alpha} \end{aligned}$$

In order to eliminate $\lambda$, note that capital market clearing requires

$$\begin{aligned}
K_t &= \int_0^\infty \int_0^\infty k_t(a,z)\, g_t(a,z)\, \mathrm{d}a\, \mathrm{d}z \\
&= \int_{\underline{z}}^\infty \int_0^\infty \lambda a g_t(a,z)\, \mathrm{d}a\, \mathrm{d}z \\
&\downarrow \\
1 &= \lambda \int_{\underline{z}}^\infty s_t(z)\, \mathrm{d}z
\end{aligned}$$

and thus

$$\lambda = \frac{1}{\int_{\underline{z}}^\infty s_t(z)\, dz}$$

Replacing this into production gives

$$Y_t = (Z_t K_t)^\alpha N_t^{1-\alpha}$$

where

$$\begin{aligned}
Z_t &= \lambda X_t \\
&= \frac{\int_{\underline{z}}^\infty z s_t(z)\, \mathrm{d}z}{\int_{\underline{z}}^\infty s_t(z)\, \mathrm{d}z} \\
&= \mathbb{E}_\omega\left[z | z > \underline{z}\right]
\end{aligned}$$

Now, the law of motion for the aggregate capital stock is

$$\begin{aligned}
K_{t+1} &= \int\int a_{t+1}(a,z)\, g_t(a,z)\, \mathrm{d}a\, \mathrm{d}z \\
&= \int\int \beta R_t(z)\, a g_t(a,z)\, \mathrm{d}a\, \mathrm{d}z \\
&= \beta K_t \int R_t(z)\, s_t(z)\, \mathrm{d}z
\end{aligned}$$

Recall that

$$R(z_t) = 1 + r_t + \lambda \max\{\omega_t z_t - r_t - \delta, 0\}$$

and so I can write this as

$$\begin{aligned}
K_{t+1} &= \beta K_t \int \left[1 + r_t + \lambda \max\{\omega_t z_t - r_t - \delta, 0\}\right] s_t(z)\, \mathrm{d}z \\
&= \beta K_t \left(1 + r_t + \lambda \int_{\underline{z}}^\infty (\omega_t z - r_t - \delta)\, s_t\, \mathrm{d}z\right) \\
&= \beta K_t \left(1 + r_t + \lambda \int_{\underline{z}}^\infty z \omega_t s_t(z)\, dz - \lambda (r_t + \delta) \int_{\underline{z}}^\infty s_t\, \mathrm{d}z\right)
\end{aligned}$$

The two integrals are

$$\omega_t \lambda \int_{\underline{z}}^{\infty} z s_t(z)\,\mathrm{d}z = \omega_t X_t \lambda$$
$$= \omega_t Z_t$$

and

$$\lambda \int_{\underline{z}}^{\infty} s_t\,\mathrm{d}z = \lambda\left(1 - S\left(\underline{z}\right)\right)$$
$$= 1$$

and so the LoM is

$$K_{t+1} = \beta K_t \left(1 + \omega_t Z_t - \delta\right)$$

From labor market clearing,

$$\omega_t Z_t = \alpha p_{tx} Z_t^{\alpha} \left(\frac{N_t}{K_t}\right)^{1-\alpha}$$

and so the LoM becomes

$$\begin{aligned} K_{t+1} &= \alpha\beta p_{tx} \left(Z_t K_t\right)^{\alpha} N_t^{1-\alpha} + \beta\left(1-\delta\right) K_t \\ &= \alpha\beta p_{tx} Y_t + \beta\left(1-\delta\right) K_t \end{aligned}$$

The return on capital equals the average return across all entrepreneurs, calculated above:

$$\begin{aligned} R_{tK} &= \int_0^{\overline{z}} R_t(z)\, s_t(z)\,\mathrm{d}z \\ &= 1 - \delta + \omega_t Z_t \\ &= 1 - \delta + \alpha p_{tx} Z_t^{\alpha} \left(\frac{N_t}{K_t}\right)^{1-\alpha} \mathrm{d}z \end{aligned}$$

Finally, the factor prices. The wage can be calculated from labor market clearing: recall that

$$N_t = \left(\frac{1-\alpha}{w_t} p_{tx}\right)^{\frac{1}{\alpha}} \lambda K_t X_t$$

Rearranging gives

$$w_t = \left(1-\alpha\right) p_{tx} \left(\frac{Z_t K_t}{N_t}\right)^{\alpha}$$

as in the text. The net real interest rate comes from the definition of the cutoff $\underline{z}_t$:

$$r_t = \omega_t \underline{z}_t - \delta$$

Substituting the definition of $\omega_t$ from labor market clearing gives the form in Equation (32). Note that this only holds in expectation, as nominal debt contracts are negotiated at the end of period $t$, and the ex-post real return $r_t$ depends on the realization of inflation. □

## 7.2 Proof of Proposition (2)

*Proof.* By definition, the law of motion for the *cumulative* wealth share $S_t(z)$ is

$$S_{t+1}(z) = \frac{1}{K_{t+1}} p \int_0^z \int_0^\infty a'(a,\hat{z})\, g_t(a,\hat{z})\ \mathrm{d}a\,\mathrm{d}\hat{z} + \frac{1}{K_{t+1}}(1-p)\,F(z) \int_0^\infty \int_0^\infty a'(a,\hat{z})\, g_t(a,\hat{z})\ \mathrm{d}a\,\mathrm{d}\hat{z}$$

Differentiating with respect to $z$ gives

$$S'_{t+1}(z) = \frac{1}{K_{t+1}} p \int_0^\infty a'(a,z)\, g_t(a,z)\ \mathrm{d}a + \frac{1}{K_{t+1}}(1-p)\,f(z) \int_0^\infty \int_0^\infty a'(a,\hat{z})\, g_t(a,\hat{z})\ \mathrm{d}a\,\mathrm{d}\hat{z}$$

With the policy functions:

$$\begin{aligned} s_{t+1}(z) &= \frac{1}{K_{t+1}} p \int_0^\infty \beta R(z)\, a g_t(a,z)\ \mathrm{d}a + \frac{1}{K_{t+1}}(1-p)\,f(z) \int_0^\infty \int_0^\infty \beta R(\hat{z})\, a g_t(a,\hat{z})\ \mathrm{d}a\,\mathrm{d}\hat{z} \\ &= \frac{K_t}{K_{t+1}} p\beta R(z)\, s_t(z) + \frac{K_t}{K_{t+1}}(1-p)\,f(z) \int_0^\infty \beta R(\hat{z})\, s_t(\hat{z})\ \mathrm{d}\hat{z} \end{aligned}$$

Leibniz rule implies:

$$\frac{d}{dz}\int_0^z \int_0^\infty \beta R(\hat{z})\, a g_t(a,\hat{z})\ \mathrm{d}a\,\mathrm{d}\hat{z} = \int_0^\infty \beta R(z)\, a g_t(a,z)\ \mathrm{d}a$$

Furthermore,

$$\int R(z)\, s_t(z)\ \mathrm{d}z = 1-\delta+\alpha p_{tx} Z_t^\alpha \left(\frac{N_t}{K_t}\right)^{1-\alpha}$$

as derived above, i.e. the RoK. So the LoM is

$$s_{t+1}(z) = \frac{K_t}{K_{t+1}}\beta\left[pR(z)\, s_t(z) + (1-p)\,f(z)\left(1-\delta+\alpha p_{tx} Z_t^\alpha \left(\frac{N_t}{K_t}\right)^{1-\alpha}\right)\right] \tag{65}$$

Recall that from the law of motion for capital,

$$\frac{K_{t+1}}{\beta K_t} = 1 - \delta + \alpha p_{tx} Z_t^{\alpha} \left(\frac{N_t}{K_t}\right)^{1-\alpha} = R_{tK}$$

Substituting this into (65) gives Equation (37) in the text. □

## 7.3 Full Log-Linearized System

The full system is

$$\hat{s}_{t+1}(z) = p\frac{R(z)}{R_K}\left\{\hat{R}_t(z) - \hat{R}_{tK} + \hat{s}_t(z)\right\} \tag{66}$$

$$\hat{R}_t(z) = \begin{cases} \frac{1}{R(z)}\left\{(1-\lambda)\left(\hat{i}_t - \hat{\pi}_t\right) + \lambda\omega z\hat{\omega}_t\right\} & z > \underline{z}_t \\ \frac{1}{R}\left\{\hat{i}_t - \hat{\pi}_t\right\} & z < \underline{z}_t \end{cases} \tag{67}$$

$$\hat{R}_{tK} = r_K\left\{\hat{\omega}_t - \hat{Z}_t\right\} \tag{68}$$

$$\hat{\omega}_t = \hat{p}_{tx} + (1-\alpha)\left(\hat{N}_t - \hat{Z}_t - \hat{K}_t\right) \tag{69}$$

$$\hat{w}_t = \eta\hat{N}_t \tag{70}$$

$$\pi_t = \kappa_p\hat{p}_{tx} + \beta_f\mathbb{E}_t\pi_{t+1} \tag{71}$$

$$\hat{i}_t = \phi_\pi\pi_t + \nu_t \tag{72}$$

$$\hat{r}_t = \hat{i}_{t-1} - \pi_t \tag{73}$$

$$\hat{\underline{z}}_{t+1} = \frac{1}{r+\delta}\hat{r}_{t+1} - \hat{\omega}_{t+1} \tag{74}$$

$$\hat{N}_t = \frac{1}{\alpha+\eta}\hat{p}_{tx} + \frac{\alpha}{\alpha+\eta}\left(\hat{Z}_t + \hat{K}_t\right) \tag{75}$$

$$\hat{Y}_t = \alpha\left(\hat{Z}_t + \hat{K}_t\right) + (1-\alpha)\hat{N}_t \tag{76}$$

$$\hat{K}_{t+1} = \left[1 - \beta(1-\delta)\right]\left(\hat{p}_{tx} + \hat{Y}_t\right) + \beta(1-\delta)\hat{K}_t \tag{77}$$

$$\hat{Z}_{t+1} = \lambda\left\{\underline{z}s(\underline{z})\left(1 - \frac{\underline{z}}{Z}\right)\hat{\underline{z}}_{t+1} + \int_{\underline{z}}^{\overline{z}}\left(\frac{z}{Z} - 1\right)s(z)\,\hat{s}_{t+1}(z)\,\mathrm{d}z\right\} \tag{78}$$

$$0 = \underline{z}s(\underline{z})\,\hat{\underline{z}}_t + \int_0^{\overline{z}} s(z)\,\hat{s}_t(z)\,\mathrm{d}z \tag{79}$$

### 7.4 Computational Strategy

#### 7.4.1 Steady State

The equations pinning down the steady state are as follows:

$$s(z) = \frac{1-p}{1-p\beta R(z)} f(z) \tag{80}$$

$$R(z) = \begin{cases} 1 + r + \lambda(\omega z - r - \delta) & z > \underline{z} \\ 1 + r & z \leq \underline{z} \end{cases} \tag{81}$$

$$\omega = \alpha p_x \left(\frac{N}{ZK}\right)^{1-\alpha} \tag{82}$$

$$p_x = \frac{\varepsilon - 1}{\varepsilon} \tag{83}$$

$$\frac{1}{\beta} = 1 + \alpha p_x Z^{\alpha} \left(\frac{N}{K}\right)^{1-\alpha} \tag{84}$$

$$N = [(1-\alpha) p_x]^{\frac{1}{\alpha+\eta}} (ZK)^{\frac{\alpha}{\alpha+\eta}} \tag{85}$$

$$\underline{z} = \frac{r+\delta}{\omega} \tag{86}$$

$$Z = \lambda \int_{\underline{z}}^{\overline{z}} z s(z) \, \mathrm{d}z \tag{87}$$

$$1 = \lambda \int_{\underline{z}}^{\overline{z}} s(z) \, \mathrm{d}z \tag{88}$$

My assumption on the process for $z_t$ allows the computation of the steady state to be reduced to a system of two equations in two unknowns. I guess a pair $(\omega^0, r^0)$. From there, I calculate

$$\underline{z}^0 = \frac{r^0 + \delta}{\omega^0}$$
$$Z^0 = \frac{1/\beta - (1-\delta)}{\omega^0}$$

where the second follows from stationarity of the capital stock. Aggregate productivity then pins down the capital-labor ratio:

$$\frac{K^0}{N^0} = \left(\frac{\alpha p_x (Z^0)^{\alpha}}{\frac{1}{\beta} - (1-\delta)}\right)^{\frac{1}{1-\alpha}}$$
$$\equiv \xi^0$$

and then labor market clearing pins down equilibrium labor $N^0$:

$$N^0 = \left[(1-\alpha) p_x \left(Z^0 \xi^0\right)^{\alpha}\right]^{1/\eta}$$

and then $K^0 = \xi^0 N^0$. Now, I check errors on the implied definitions of $Z^0$ and $\underline{z}^0$. I calculate $\underline{z}^1$ using capital market clearing: given $r^0$ and $\omega^0$, I can calculate the wealth shares $s^0(z)$, and then $\underline{z}^1$ solves

$$1 - \frac{1}{\lambda} = \int_{\underline{z}^1}^{\overline{z}} s^0(z)\, \mathrm{d}z$$

and $Z^1$ is given by

$$Z^1 = \int_{\underline{z}^0}^{\overline{z}} z s^0(z)\, \mathrm{d}z$$

I define a discrete grid for $z$ on $[0, \overline{z}]$, and compute the above integrals using the Trapezoidal rule. As a starting guess, I set

$$r^0 = r^{FB} = \frac{1}{\beta}$$

and

$$\omega^0 = \omega^{FB} = \frac{1/\beta - (1-\delta)}{\overline{z}}$$

equal to their values under the first-best equilibrium.

### 7.4.2 Transitions

I use the following algorithm to compute impulse responses to a monetary shock in my linearized model:

1. Start with $\hat{K}_0, \hat{s}_0(z), \hat{\underline{z}}_0, \hat{Z}_0, \hat{i}_0$ all equal to zero.
2. Guess a path for $\left\{\hat{i}_{t+1}\right\}_{t=0}^{T-1}$. For boundary conditions, we have $\hat{i}_0 = \hat{i}_{T+1} = 0$.
3. Calculate the path for $\pi_t$ from the Taylor Rule:

$$\pi_t = \frac{\hat{i}_{t+1} - \nu_t}{\phi_t}$$

4. Using $\pi_t$ and $\pi_{t+1}$, calculate $\hat{p}_{tx}$ from the Phillips curve:

$$\pi_t = \kappa_p \hat{p}_{tx} + \beta_f \mathbb{E}_t \pi_{t+1}$$

$$\hat{p}_{tx} = \frac{1}{\kappa_p}\left\{\pi_t - \beta \pi_{t+1}\right\}$$

5. Calculate $\hat{N}_t$:

$$\hat{N}_t = \frac{1}{\alpha+\eta}\hat{p}_{tx} + \frac{\alpha}{\alpha+\eta}\left(\hat{Z}_t + \hat{K}_t\right)$$

6. Calculate $\hat{\omega}_t, \hat{r}_t, \hat{R}_{tK}, \hat{R}_t(z)$:

$$\hat{R}_t(z) = \begin{cases} \frac{1}{R(z)} \left\{ (1-\lambda) \left( \hat{i}_t - \hat{\pi}_t \right) + \lambda \omega z \hat{\omega}_t \right\} & z > \underline{z}_t \\ \frac{1}{R} \left\{ \hat{i}_t - \hat{\pi}_t \right\} & z < \underline{z}_t \end{cases}$$
$$\hat{R}_{tK} = r_K \left\{ \hat{\omega}_t - \hat{Z}_t \right\}$$
$$\hat{\omega}_t = \hat{p}_{tx} + (1-\alpha) \left( \hat{N}_t - \hat{Z}_t - \hat{K}_t \right)$$
$$\hat{r}_t = \hat{i}_t - \pi_t$$

- Note the notation in $R_t(z)$: it's important that we not let them switch ex-post to investing after the initial unanticipated $\pi_0$.

7. Using the returns, update the wealth shares:

$$\hat{s}_{t+1}(z) = p \frac{R(z)}{R_K} \left\{ \hat{R}_t(z) - \hat{R}_{tK} + \hat{s}_t(z) \right\}$$

8. Using the continuation shares $\hat{s}_{t+1}(z)$, find $\hat{\underline{z}}_{t+1}$ from capital market clearing:

$$\hat{\underline{z}}_{t+1} = \frac{1}{s(\underline{z})} \int_{\underline{z}}^{\overline{z}} \hat{s}_{t+1}(z)\, dz$$

9. Repeat steps 3-7 for all $t$. Then, evaluate the $\underline{z}_{t+1}$ values implied by capital market clearing against their definition: for $t = 0, ..., T-1$:

$$\hat{\underline{z}}_{t+1} \stackrel{?}{=} \frac{1}{r+\delta} \hat{r}_{t+1} - \hat{\omega}_{t+1}$$

If these are all satisfied, stop. Otherwise, update the path $\left\{ \hat{i}_t \right\}$ and return to step 2.