Review of Economic Dynamics 26 (2017) 301–326

Contents lists available at ScienceDirect

Review of Economic Dynamics

www.elsevier.com/locate/red

Persistent heterogeneous returns and top end wealth inequality ☆,☆☆

Dan Cao [a,*], Wenlan Luo [b]

[a] Georgetown University, United States
[b] Tsinghua University, China

ARTICLE INFO

Article history:
Received 27 July 2016
Received in revised form 3 October 2017
Available online 13 October 2017

Keywords:
Pareto wealth distribution
Sufficient statistics
r–g
Heterogeneous returns
Financial deregulation
Corporate tax cuts

ABSTRACT

We document in US data that returns to wealth across households are significantly heterogeneous, and persistently so. Motivated by this observation, we build a tractable general equilibrium model where households face persistent idiosyncratic returns to study the US wealth distribution. We show theoretically that the wealth distribution in the model admits a Pareto tail and characterize how the tail index depends on salient equilibrium variables including capital-output ratio, labor share, interest rate, and growth rate. Quantitatively, to match the observed US wealth distribution it requires significant heterogeneity in returns, consistent with our empirical findings. Finally, we show in the model that financial deregulation and a reduction in US corporate tax rates can generate the joint evolution of rising wealth inequality, rising capital-output ratio and declining labor share since the 1980s.

1. Introduction

The recent empirical work by Piketty (2014), Saez and Zucman (2016), and others has found that the wealth distribution in major developed countries is highly unequal, especially in the U.S. The wealth distribution has a thick right tail, which is well approximated by Pareto power law, with very high top wealth shares, i.e., very high *top end wealth inequality*.[1] In addition, top end wealth inequality has risen rapidly since the early 1980s. For example, according to Saez and Zucman (2016), the top 0.1% owned around 8% of total wealth in 1980 but this share increased to more than 20% in 2010. Understanding the mechanisms that lead to the high level of wealth inequality and its rapid rise is therefore of utmost importance. In this paper, we focus on one of these mechanisms: persistent heterogeneous returns to wealth.

☆ First version: July 2016. It was circulated as "Sufficient Statistics for Top End Wealth Inequality."

☆☆ For useful comments and discussions we thank Daron Acemoglu, George Akerlof, Jinhui Bai, Alberto Bisin, Mary-Ann Bronson, Paco Buera, Martin Evans, Garance Genicot, Pedro Gete, Mark Huggett, Roger Lagunoff, Atif Mian, Ben Moll, Jonathan Parker, Thomas Piketty, Martin Ravallion, Richard Rogerson, John Rust, and participants at Georgetown macroeconomic seminar. We also thank Son Le for excellent research assistance and his initial involvement in the project. We are grateful to the editor, Vincenzo Quadrini, and two anonymous referees whose comments and suggestions help improve the paper greatly.

* Corresponding author.
E-mail addresses: dc448@georgetown.edu (D. Cao), luowenlan@gmail.com (W. Luo).

[1] A Pareto power law distribution with tail index θ has the following complementary cdf above a certain wealth level $w_{\min}$: $\Pr(W \geq w | W \geq w_{\min}) = \left(\frac{w_{\min}}{w}\right)^{-\theta}$. For the U.S., using the wealth data from the Survey of Consumer Finances in 2010, we find that the distribution of wealth beyond the top 10% is well approximated by a Pareto distribution (see Online Appendix H). In this case, the tail index can be computed simply using top wealth shares: $\theta \approx \frac{1}{1+\log_{10}\frac{S(1)}{S(10)}}$ where $S(p)$ stands for the share of wealth held by the top p%, or more generally $\frac{1}{1+\log_{(p_2/p_1)}\frac{S(p_1)}{S(p_2)}}$ for any $p_1 < p_2 \leq 10$.

https://doi.org/10.1016/j.red.2017.10.001
1094-2025/

First, we use the PSID to calculate the returns to wealth of households. We find that returns to wealth are heterogeneous across households: the cross-sectional standard deviation of annualized returns is 11.3% without capital gains and 27.1% with capital gains. The heterogeneous returns are also persistent over time with five-year correlation of more than 0.22 and ten year correlation of more than 0.18.

Motivated by these facts, we construct a tractable general equilibrium model which features heterogeneity in returns. The model is cast in continuous time and agents face idiosyncratic investment risk as in Quadrini (2000) and Angeletos (2007). In the model, the agents share the same utility function but may differ in their investment returns, which are determined by their investment productivity.[2] For tractability, we assume that at any instant, there are at most two levels of investment returns, high or low. In the special case of homogeneous returns that we also consider, the two returns are the same. The agents are subject to idiosyncratic shocks, as in Huggett (1993) and Aiyagari (1994), with Poisson arrival rates. If an agent is hit by a shock, her return switches from high to low, or low to high. In addition, to capture earnings inequality, we allow for permanent heterogeneity in labor productivity across agents. There is a bond market that allows the agents to save or borrow, subject to a borrowing constraint. Given the idiosyncratic investment returns and the financial market structure, the agents make the consumption, saving, and investment decisions to maximize their inter-temporal expected utility, taking interest rate and wage rate as given. We study the properties of the (recursive) competitive equilibrium of the model, in which prices are determined such that bond and labor markets clear at all time.

We show that in the balanced growth path of the model, the stationary wealth distribution indeed has a Pareto right tail. In addition, the tractability of the model allows us to express the Pareto tail index as a function of the model parameters and the salient aggregate statistics including interest rate, growth rate, capital-output ratio, and labor share. Piketty (2014) and Piketty and Zucman (2014, 2015) argue, rather informally, for the importance of the aggregate statistics in shaping wealth inequality.[3] We validate their argument by showing formally that in our model there exist structural relations between these statistics and top end wealth inequality. Through these structural relations, observable aggregate statistics provide valuable information on return heterogeneity and saving rates, which are hard to directly and accurately estimate in the data.

To demonstrate the importance of return heterogeneity, we first show that the special case of the model with homogeneous returns, when calibrated to match the salient aggregate statistics, produces a tail index that is of an order of magnitude too high compared to the one in the data (*higher* tail index corresponds to *lower* top end wealth inequality). We then calibrate the general model with return heterogeneity. The model requires substantial return heterogeneity to match the low Pareto index of around 2 observed in the data. The more productive agents enjoy more than 4% in unlevered returns and 6% in levered returns compared to the less productive agents. In addition, in this calibration, varying earnings inequality or initial wealth distribution does not significantly change wealth inequality at the very top, beyond the top 10% or the Pareto tail index, even though it strongly affects wealth inequality at lower wealth percentiles.

Given the calibration, we also look at the joint evolution of top end wealth inequality, capital-output ratio, and labor share after structural changes in the model. First, we find that a uniform corporate tax cut of the magnitude observed in the data – from 40% to 25% – generates a decreasing Pareto tail index of the wealth distribution, increasing capital-output ratio, and decreasing labor share as observed in the data from the 1980s to recent years. However, the uniform corporate tax cut fails to generate increasing top wealth shares. A non-uniform corporate tax cut which gives the more productive agents a larger cut brings the top wealth share dynamics closer to the empirical ones.

Second, a relaxation of the borrowing constraints, also of the magnitude in the data generated by waves of financial deregulation in the late 1980s, leads to a rapid rise in top wealth shares and a rapid decline in Pareto tail index, quantitatively close to the dynamics in the data. However, the model produces counter factual dynamics for capital-output ratio and labor share. When combined with a uniform corporate tax cut, the relaxation of borrowing constraint produces a realistic joint evolution of top end wealth inequality, capital-output ratio, and labor share. One lesson we draw from these exercises is that a single policy change, either a uniform corporate tax cut or financial deregulation, is unlikely to produce the joint evolution in the U.S. since the 1980s. In addition, when we seek causes for the changes in wealth inequality, the dynamics of the aggregate variables provide disciplines that can not be ignored and our general equilibrium framework highlights these linkages.

The rest of this paper is organized as follows. The next section presents the empirical evidence for persistent heterogeneous returns. Section 3 presents a general model with capital accumulation, production, and labor. Section 4 presents several sufficient statistics results that link top end wealth inequality to aggregate statistics. In Section 5, we calibrate the main model to the U.S. economy and investigate the quantitative implications of the model in stationary balanced growth paths. Section 6 investigates transitional paths. Section 7 discusses the related literature and concludes.

[2] Heterogeneous investment returns in our model can have several interpretations. They might arise from different degrees of financial sophistication while investing in publicly traded financial instruments, as documented in Calvet et al. (2007), differential access to high return investments in private businesses, or from entrepreneurship as documented and modeled in Quadrini (2000) and Cagett and De Nardi (2006), or both as reported in Fagereng et al. (2016) and Bach et al. (2017). While we do not take a stand on the precise channels leading to heterogeneous returns, our model has a direct entrepreneurship interpretation.

[3] A priori, one should not expect any relationship between the aggregate statistics and wealth inequality. For example, in the standard neoclassical growth model presented in Barro and Sala-i Martin (2004) and Acemoglu (2009), changing capital-output ratio or labor share has no effect on wealth inequality since all agents have exactly the same wealth (see Online Appendix I).

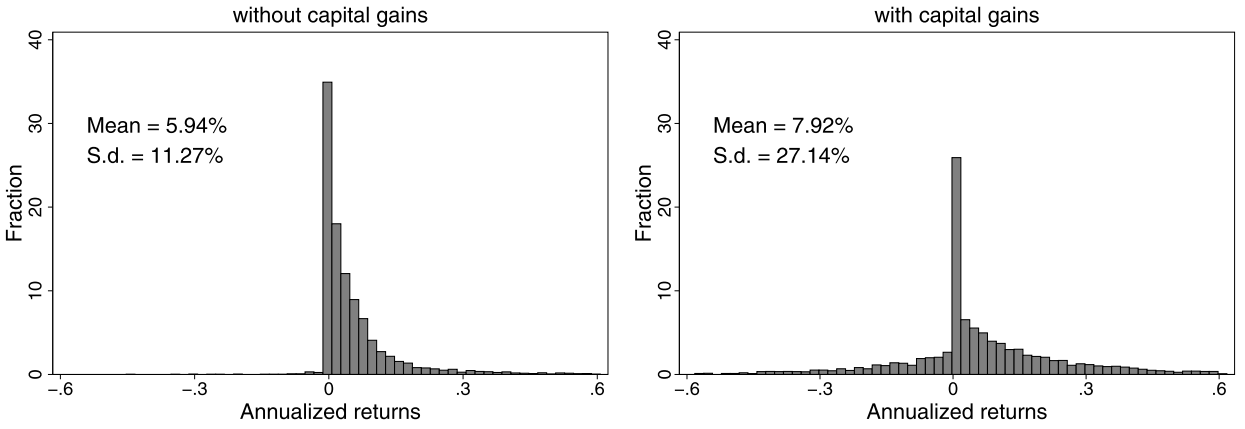

Fig. 1. Histograms of returns to wealth.

2. Empirical analysis

Using PSID (The Panel Study of Income Dynamics), this section documents two facts that motivate our theoretical analysis: (1) there is substantial heterogeneity in returns to wealth across households; (2) the returns to wealth between years are positively correlated across households, which we interpret as evidence that returns to wealth are persistent.

PSID keeps track more than 5,000 families and their descendants. Besides its excellent coverage on demographics and income information, the three waves in 1984, 1989, and 1994 include wealth supplements that provide information on household wealth and changes in wealth composition. Starting from 1999, PSID becomes biennial and surveys wealth information every wave. Using wealth information, we define the following key variables:

- **Wealth**: The sum of net worths[4] of (1) home equity; (2) other real estate; (3) business and farms; (4) vehicles; (5) stocks; (6) savings, checking accounts, and certificate of deposits; (7) bonds, insurance and other assets; (8) IRA accounts; (9) minus other debts.
- **Core asset:** Wealth defined above, excluding home equity and vehicles.
- **Capital income:** The following income components from head and wife: asset income from farm and business, rent, interest, dividends, income from royalty and trust funds.
- **Capital gains:** Computed using differences between changes in asset values and reported flows, i.e., capital gains from t to $t+s=$ (asset value$_{t+s}$ − asset value$_t$) − (asset bought from t to $t+s$ − asset sold from t to $t+s$). Capital gains are computed for business, stocks, and other real estate as households reported flows for these asset categories.

Though in early waves of the survey, capital income is reported every year, since capital gains are computed based on wealth information, in our benchmark results, we compute total returns from one survey year that contains wealth information to the next survey year that contains wealth information. In particular, we define the following:

- **Wealth return without capital gains from t to $t+s$:** $\sum_{\tau=t}^{t+s}$ capital income$_\tau$/core asset$_t$.
- **Wealth return with capital gains from t to $t+s$:** ($\sum_{\tau=t}^{t+s}$ capital income$_\tau$ + capital gains from t to $t+s$)/core asset$_t$.

We choose core asset instead of total wealth as the denominators when computing wealth returns since home equity and vehicles are durable consumption and mostly do not generate monetary returns.[5] Notice that the above calculations require us to observe capital income every year from t to $t+s$. Therefore we choose $t \in \{1984, 1989, 1994\}$ and $s=5$ as our benchmark when the capital income is surveyed annually and wealth information is surveyed every five years. We also impute capital income for years not covered when the survey becomes biennial after 1999 and report the results in Online Appendix H. For the following analysis, we remove outliers that reported annualized returns higher than 300% or lower than −100%, which consist around 2.5% of total samples.

Fig. 1 plots the distributions of 5-year returns (annualized) pooling observations from the years 1984, 1989, and 1994.[6] The left panel plots the distribution of returns without capital gains. As shown, the standard deviation of returns is 11.27%. The majority of households report little or no capital income but the distribution exhibits a long right tail. Around 10% of households earn an annualized return higher than 15% and 5% of households earn an annualized return higher than

[4] PSID does not report worths and debts separately until 2013. Therefore, while in our theoretical analysis return is defined on asset, we measure return on net wealth in the data.

[5] We remove samples with negative or very low wealth (core asset <$1000).

[6] The horizontal axes are truncated at −0.6 and 0.6 for better presentations. The means and standard deviations are based on total selected samples.

Table 1
Correlation of returns to wealth.

Corr.	5-Year correlation		10-Year correlation
	$(R_{84\text{–}89}, R_{89\text{–}94})$	$(R_{89\text{–}94}, R_{94\text{–}99})$	$(R_{84\text{–}89}, R_{94\text{–}99})$
Returns without capital gains	0.224^{***}	0.299^{***}	0.189^{***}
	(0.04)	(0.062)	(0.05)
Returns with capital gains	−0.022	−0.024	0.185^{***}
	(0.036)	(0.065)	(0.062)

Notes: Standard errors in parentheses are computed based on bootstrap procedures with 200 repetitions. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

30%. The distribution of returns with capital gains, depicted in the right panel, exhibits even more dispersion. The standard deviation of annualized returns is as high as 27.14%. While there are masses of households who earn low returns close to zero, around 10% of households enjoy an annualized return higher than 30%. In addition, there is a significant fraction of households who report capital losses. Both panels speak consistently that there is substantial heterogeneity in returns to wealth across households.

Table 1 tabulates the correlation of returns to wealth between different periods across households. The first row reports the correlation of returns without capital gains. As shown in the first two columns, the correlation of returns between 1984–1989 and 1989–1994, between 1989–1994 and 1994–1999 are 0.224 and 0.299 respectively. Both correlations are accurately estimated as shown by standard errors in the parentheses. As shown in the third column, returns to wealth remain highly persistent even after 10 years: the correlation of returns between 1984–1989 and 1994–1999 is at a significant level of 0.189.

The second row reports the correlation of returns with capital gains. Since we define capital gains as the differences between changes in asset values and asset flows, for two adjacent periods, the measurement error in asset values in the last year of the first period enters in the numerator of the first-period return and it enters in the denominator of the second-period return. Therefore the measurement error would artificially bias downward the correlation of returns. This source of bias might explain the slightly negative (and insignificant) estimates for the 5-year correlations. But as shown in the third column, for the 10-year correlation which is less affected by the measurement error, we estimate a positive correlation of 0.185. This is also close to the 10-year correlation estimated without capital gains as reported in the first row.

In Online Appendix H we report additional estimates of correlation of returns using data after 1999 where both capital income and wealth information are surveyed once every two years and imputations are needed. We also provide evidence on wealth mobility conditional on wealth return groups (in the spirit of Quadrini, 2000) which is suggestive of persistent returns. The analyses in this section are kept simple and transparent on purpose but suffice to highlight the heterogeneity and persistence in returns to wealth which serve as the core building blocks in our theoretical analysis.

3. A neoclassical economy in continuous time

Motivated by the empirical evidence in the last section, we develop a neoclassical growth model in continuous time as presented in Barro and Sala-i Martin (2004) and Acemoglu (2009) but with heterogeneous agents facing persistent heterogeneous investment shocks, which determine the productivity of their capital investment. The agents can finance their projects only by issuing risk-free debt, collateralized by their capital.

3.1. The environment

Time – t – is continuous and runs from 0 to ∞.[7] The economy is populated by a continuum of agents that are indexed by $h \in \mathcal{N}_t = [0, N_t]$ where N_t is the population size at time t.

Investment productivity Let i_t^h denote the individual state of agent h at time t. We assume that i_t^h follows a two-state Markov chain, $i_t^h \in \mathcal{I} = \{L, H\}$, which capture low ($i_t^h = L$) and high investment productivity ($i_t^h = H$), with Poisson switching rates λ_{LH} and λ_{HL} from one state to the other.

By the law of large numbers, the fraction of agents with high investment returns (H-agents) and low investment returns (L-agents) are respectively:

$$M_H = \frac{\lambda_{LH}}{\lambda_{HL} + \lambda_{LH}} \text{ and } M_L = \frac{\lambda_{HL}}{\lambda_{HL} + \lambda_{LH}} = 1 - M_H.$$

In each instant, the agents can produce output using a constant return to scale technology, which depends on the idiosyncratic state, using capital and labor[8]

[7] We also studied and solved the model in discrete time. However, the solution for the equilibrium wealth distribution is much simpler in continuous time.

[8] The production function includes depreciation, i.e., Y_t is net output. The CES production function in Section 5 is an example.

$$Y_t = F_{i_t^h}(k_t^h, l_t^h).$$

We assume that the H-agents are weakly more productive than the L-agents.

Assumption 1. For any $k, l > 0$,

$$F_H(k,l) \geq F_L(k,l).$$

While in this paper we focus on the case in which there are only two-level of investment productivities, we can easily extend the model to allow for more heterogeneity in this dimension. In addition, when agents with lower productivity actively produce in equilibrium, this model is equivalent to one in which a competitive representative firm uses the less productive production function and entrepreneurs have access to the more productive production. The macroeconomic literature on entrepreneurship (such as Quadrini, 2000 and Cagett and De Nardi, 2006) often embraces this interpretation.

Labor productivity Each agent is also endowed with x_t^h efficiency units of labor. We assume that x_t^h grows at the rate g_x common across households: $x_{t+t'}^h = x_t^h \exp(g_x t')$.[9] We assume that the initial labor productivity of an agent is determined when she is born. Normalized by a constant trend $G_{x,t} = \exp(g_x t)$, $\frac{x_t^h}{G_{x,t}}$ is initially drawn from a distribution defined over $\mathbb{R}_+$ with c.d.f $\Phi(x)$ and mean 1.[10]

Death shocks, population growth, and redistribution Population N_t grows at a constant rate $n \geq 0$ and the initial population is normalized to 1. We also assume that agents are hit by "death shocks," arriving at Poisson rate $\bar{\lambda} > 0$. Therefore, there are $(n+\bar{\lambda})N_t \Delta t$ newborns in each infinitesimal $[t, t+\Delta t]$ time interval. The investment productivity of the newborns can be H or L with the fraction M_H and M_L respectively.

The total wealth of the dying agents is redistributed to the newborns according to a redistribution function (a complementary cdf):

$$\Gamma : \mathbb{R} \longmapsto [0,1],$$

that is, each newborn receives a draw of fraction ψ from the distribution Γ: $\Pr\left(\psi \geq \tilde{\psi}\right) = \Gamma(\tilde{\psi})$. The newborn then obtains a fraction ψ of the aggregate wealth of the dying agents and corporate tax revenue *per newborn*.

Since in each instant, there is a mass $\bar{\lambda}\Delta t N_t$ of dying agents and a mass $\left(\bar{\lambda}+n\right)\Delta N_t$ of newborns, and the aggregate wealth of the dying agents and corporate tax revenue is redistributed fully, we must have

$$\int \psi d\Gamma(\psi) = 1.$$

Death shocks and redistribution are crucial in preventing the wealth distribution from ever expanding and give rise to a stationary distribution.[11]

Utility, constraints, and optimization Let $\{r_t, e_t\}_{t=0}^{\infty}$ denote the sequence of interest rates and wages. Agent h chooses a sequence of capital holding k_t^h, bond holding b_t^h, and consumption c_t^h to maximize the inter-temporal expected utility:[12]

$$\max_{c_t^h, k_t^h, b_t^h} \mathbb{E}\left[\int_0^{\infty} \exp\left(-\left(\rho+\bar{\lambda}\right)t\right) \log\left(c_t^h\right) dt\right] \tag{1}$$

subject to

$$\frac{dw_t^h}{dt} = \left(1-\tau_{i_t^h}\right) \max_{l_t^h} \left\{F_{i_t^h}\left(k_t^h, l_t^h\right) - e_t l_t^h\right\} + r_t b_t^h + e_t x_t^h - c_t^h \tag{2}$$

$$w_t^h = k_t^h + b_t^h, \tag{3}$$

where w_t^h denote the financial wealth of agent h at time t. The budget constraint (2), and portfolio constraint (3), are standard. At time t, change in financial wealth $\frac{dw_t^h}{dt}$, consists of return from capital investment subject to corporate tax τ_i,

[9] It is straightforward to extend the model to allow for heterogeneous but deterministic growth rate of labor productivity across agents. As shown in Gabaix et al. (2016), heterogeneous growth rates of labor productivity can potentially help explain high labor income inequality and a rapid rise in labor income inequality. However, we lose tractability if we allow for idiosyncratic productivity shocks over the agents' lifetime since the optimization problem of the agents would no longer be homogeneous in total wealth.

[10] The model does not exhibit the scale effect, so normalizing the mean of Φ to 1 is without loss of generality.

[11] See Gabaix (2009) for other types of assumptions that guarantee the existence of a stationary wealth distribution with Pareto tail such as a reflecting barrier.

[12] To simplify the notations, we assume log utility but many of our analyses and results can be extended to allow for general CRRA utilities. We present them Online Appendix B.

interest from bond holding $r_t b_t$, labor earning $e_t x_t^h$, minus consumption c_t^h. Financial wealth w_t^h, is allocated to capital holding k_t^h and bond holding b_t^h. The agent is also subject to the borrowing constraint (5) defined below.

Let Q_t^h denote the present discounted value of future labor income for agent h, i.e., her human wealth,

$$Q_t^h = \int_0^\infty \exp\left(-\int_0^{t'} r_{t+t_1'} dt_1'\right) e_{t+t'} x_{t+t'}^h dt'.$$

Since x_t^h grows at the rate g_x, we also have $Q_t^h = q_t x_t^h$, where

$$q_t = \int_0^\infty \exp\left(-\int_0^{t'} \left(r_{t+t_1'} - g_x\right) dt_1'\right) e_{t+t'} dt'.$$

The dynamics of Q_t^h and q_t are then given by

$$\frac{dQ_t^h}{dt} = r_t Q_t^h - e_t x_t^h$$

and

$$\frac{dq_t}{dt} = (r_t - g_x) q_t - e_t. \tag{4}$$

Given Q_t^h, we assume that the borrowing constraint takes the form

$$0 \le k_t^h \qquad \text{and} \qquad 0 \le m k_t^h + \left(b_t^h + Q_t^h\right), \tag{5}$$

where $m \in (0, 1)$. It is important that the borrowing constraint is in terms of borrowing plus the present discounted value of labor income so that the optimization problem of the agent is homogeneous, as we will see below.[13]

The equilibrium definition is standard.

Definition 1. For any initial distribution of wealth $\left\{w_0^h\right\}_{h\in\mathcal{N}_0}$, a competitive equilibrium is described by stochastic processes: the processes for interest rates and wage rates $\{r_t, e_t\}_{t=0}^\infty$, wealth distribution $\left\{w_t^h\right\}_{t=0,h\in\mathcal{N}_t}^\infty$, capital and bond holdings $\left\{k_t^h, b_t^h\right\}_{t=0,h\in\mathcal{N}_t}^\infty$ and consumption $\left\{c_t^h\right\}_{t=0,h\in\mathcal{N}_t}^\infty$; such that

(i) The allocation $\left\{w_t^h, k_t^h, b_t^h, c_t^h\right\}$ solves agent h's maximization problem (1) given the processes for interest rates and wage rates.

(ii) Bond and labor markets clear in each instant

$$\int_{h\in\mathcal{N}_t} b_t^h dh = 0,$$

and

$$\int_{h\in\mathcal{N}_t} l_t^h dh = \int_{h\in\mathcal{N}_t} x_t^h dh.$$

Notice that by Walras' Law, the market clearing condition in the bond market implies that the market for consumption goods also clears, i.e., total output equals to total consumption plus total investment:

$$\int_{h\in\mathcal{N}_t} F_{i_t^h}\left(k_t^h, l_t^h\right) dh = \int_{h\in\mathcal{N}_t} c_t^h dh + \int_{h\in\mathcal{N}_t} \frac{dw_t^h}{dt} dh.$$

Together with the portfolio choice constraint (3), the market clearing condition in the bond market implies market clearing in the capital market:

$$\int_{h\in\mathcal{N}_t} k_t^h dh = \int_{h\in\mathcal{N}_t} w_t^h dh.$$

[13] Angeletos (2007) makes a similar assumption on borrowing constraint, except for $m = 0$.

Let W_t denote the common value which corresponds to aggregate wealth:

$$W_t = \int_{h\in\mathcal{N}_t} w_t^h dh,$$

which we will use extensively later on.

Given the homogeneous structure of the model, in the following subsection, we show that a competitive equilibrium has a simpler representation. Indeed, the definition of competitive equilibrium suggests that we need to keep track of the whole wealth distribution to solve for the equilibrium. However, taking advantage of the homogeneity of the agents' optimization problem, we show that only a subset of aggregates variables are sufficient to determine the equilibrium in the economy. Having solved for the equilibrium prices and policy functions, we can then go back to solve for the equilibrium wealth distribution. In this sense, solving for an equilibrium is "decoupled" into solving for the equilibrium aggregates and solving for the equilibrium wealth distribution.

3.2. Markov equilibrium

To obtain sharper characterizations of a competitive equilibrium, we first characterize the agents' optimal dynamic strategies using the Hamilton–Jacobi–Bellman (HJB) equation for their value functions in a competitive equilibrium.

To simplify the notations, we use the following auxiliary functions, $R_i(e)$ and $S_i(e)$:

$$R_i(e)k = \max_l F_i(k,l) - el$$

and $S_i(e)k$ denote the maximizer. Because F_i has constant returns to scale, we have

$$F_i(k, S_i(e)k) = R_i(e)k + eS_i(e)k.$$

$R_i(e)$ and $S_i(e)$ are basically the payments to capital and labor per unit of capital used in the production function F_i.

Let $V(t, i_t^h, x_t^h, w_t^h)$ denote the value function for the maximization problem (1) of agent h. The following lemma provides the PDE that characterizes V.

Lemma 1. *The HJB equation for V is*

$$\begin{aligned}\left(\rho+\bar{\lambda}\right)V - \frac{\partial V}{\partial t} = &\max_{c,k,b} \log(c) + \frac{\partial V}{\partial x} g_x x_t + \frac{\partial V}{\partial w}\left(\left(1-\tau_{i_t^h}\right)R_{i_t^h}(e_t)k + r_t b - c + e_t x_t\right)\\ &+ \lambda_{i_t^h,-i_t^h}(V(t,-i_t^h,x_t^h,w_t^h) - V(t,i_t^h,x_t^h,w_t^h))\end{aligned} \tag{6}$$

where the maximization problem is subject to the constraints $w_t^h = k+b$ and $0\le k$ and $0\le mk + \left(b+Q_t^h\right)$ (we adopt the standard notation that $-H = L$ and $-L = H$ for i_t^h and $-i_t^h$).

Proof. Online Appendix A. □

We conjecture and verify that the value function V has the form:

$$V^h(t, i_t^h, x_t^h, w_t^h) = v(t, i_t^h) + \frac{1}{\rho+\bar{\lambda}}\log\left(w_t^h + x_t^h q_t\right). \tag{7}$$

In addition, we work with $\tilde{b}_t = b_t + Q_t^h$. Given the functional form for V, the HJB equation for the agents, (6), becomes

$$(\rho+\bar{\lambda})v(t,i_t^h) - \frac{\partial v(t,i_t^h)}{\partial t} = \max_{c,k,\tilde{b}} \mathcal{H}_{i_t^h}(c,k,\tilde{b};t,v(t,i_t^h)) + \lambda_{i_t^h,-i_t^h}(v(t,-i_t^h) - v(t,i_t^h)) \tag{8}$$

where

$$\mathcal{H}_i(c,k,\tilde{b};t,v) = \log(c) + \frac{1}{\rho+\bar{\lambda}}\left((1-\tau_i)R_i(e_t)k + r_t\tilde{b} - c\right) \tag{9}$$

and the maximization problem is subject to

$$1 = k+\tilde{b} \quad \text{and} \quad 0\le k \quad \text{and} \quad 0\le mk+\tilde{b}. \tag{10}$$

This implies that the policy functions are linear in total wealth, i.e., financial wealth plus human wealth:

$$k_t^h = k^*_{i_t^h}(t)\left(w_t^h + Q_t^h\right)$$
$$\tilde{b}_t^h = b^*_{i_t^h}(t)\left(w_t^h + Q_t^h\right)$$
$$c_t^h = c^*_{i_t^h}(t)\left(w_t^h + Q_t^h\right).$$

The maximization problem (8) together with log utility in the Hamiltonian (9) implies that

$$c_i^* \equiv \rho + \lambda. \tag{11}$$

Let $W_{i,t} = \int_{h:i_t^h=i} w_t^h dh$ denote the aggregate wealth conditional on investment productivity type i. By definition, $W_{H,t} + W_{L,t} = W_t$, where W_t is the aggregate wealth defined above. Let Q_t denote the aggregate human wealth:

$$Q_t = \int_{h\in\mathcal{N}_t} Q_t^h dh = G_t N_t q_t.$$

The market clearing condition in the bond market becomes:

$$b_H^*(t)\left(W_{H,t} + M_H Q_t\right) + b_L^*(t)\left(W_{L,t} + M_L Q_t\right) - Q_t = 0, \tag{12}$$

and in the labor market becomes:

$$S_H(e_t)k^*_{H,t}(W_{H,t} + M_H Q_t) + S_L(e_t)k^*_{L,t}(W_{L,t} + M_L Q_t) = G_t N_t, \tag{13}$$

which depend only on $W_{H,t}, W_{L,t}, q_t$. The following lemma provides the equations that determine the dynamics of these three variables.

Lemma 2. *In a competitive equilibrium, the dynamics of aggregate conditional wealth satisfy, for $i \in \{H, L\}$,*

$$\begin{aligned}\frac{dW_{i,t}}{dt} &= ((1-\tau_i)R_i(e_t)k^*_{i,t} + r_t b^*_{i,t} - c^*_{i,t})\left(W_{i,t} + M_i Q_t\right)\\ &\quad - \frac{dq_t}{dt}M_i G_t N_t - g_x q_t M_i G_t N_t - \lambda_{i,-i}W_{i,t} + \lambda_{-i,i}W_{-i,t} + M_i Z_t - \bar{\lambda}W_{i,t},\end{aligned} \tag{14}$$

where

$$Z_t = \bar{\lambda}W_t + \sum_j \tau_j R_j(e_t)k^*_{j,t}\left(W_{j,t} + M_j Q_t\right)$$

is the total wealth of the dying agents and corporate tax revenue, and q_t satisfies (4).

Proof. Online Appendix A. □

Given $(W_{H,t}, W_{L,t}, q_t)$, the market clearing conditions, (12) and (13), and the agents' optimality conditions (from the HBJ equations) determine interest rate r_t, wage rate e_t, and agents' decisions, $c^*_{i,t}, k^*_{i,t}, b^*_{i,t}$. In other words, $\left(W_{H,t}, W_{L,t}, q_t\right)$ are sufficient state variables. We focus on equilibria in which this is indeed the case.

Definition 2. A Markov equilibrium is a competitive equilibrium in which the equilibrium prices r_t, e_t and policy functions, $c^*_{i,t}, k^*_{i,t}, b^*_{i,t}$ depend only on the aggregate state variables $(W_{H,t}, W_{L,t}, q_t)$.

In a Markov equilibrium, since r_t, e_t and $c^*_{i,t}, k^*_{i,t}, b^*_{i,t}$ are functions of the state variables, $\left(W_{H,t}, W_{L,t}, q_t\right)$, the three equations (4) and (14) form a system of three ODEs that fully characterizes the aggregate dynamics of the economy without the knowledge of the entire wealth distribution among the agents in the economy.

Given the aggregate dynamics, the following lemma provides the PDEs that characterize the evolution of the wealth distribution over time. Let ω_t^h denote the total wealth (financial plus human wealth) of agent h relative to the average total wealth in the population at time t:

$$\omega_t^h = \frac{w_t^h + q_t x_t^h}{\frac{W_t + Q_t}{N_t}},$$

and let $p_i(t,\omega)$ denote:

$$p_i(t,\omega) = \Pr\left(\omega_t^h \geq \omega \middle| i_t^h = i\right).$$

By the law of large numbers, $p_H(t,\omega)$ ($p_L(t,\omega)$), is also the fraction of agents with high investment return (low investment return) and with relative total wealth exceeding ω. The dynamics for p_i depend on g_t – the growth of the aggregate total wealth, which is given by

$$g_t = \frac{1}{W_{H,t} + W_{L,t} + Q_t} \frac{d\left(W_{H,t} + W_{L,t} + Q_t\right)}{dt},$$

and $g^*_{i,t}$ – the relative wealth growth rate of type i agents:

$$g^*_{i,t} \equiv (1-\tau_i)\, R_i(e_t) k^*_{i,t} + r_t b^*_{i,t} - c^*_{i,t} - g_t.$$

Given $g^*_{i,t}$, $p_H(t,\omega)$ and $p_L(t,\omega)$ satisfy the following Kolmogorov forward equations.

Lemma 3. *p_H and p_L satisfy the following PDEs:*

$$\begin{aligned}\frac{\partial p_i(t,\omega)}{\partial t} = &-\frac{\partial p_i(t,\omega)}{\partial \omega}\omega\left(g^*_{i,t}+n\right) - \left(\lambda_{i,-i}+\bar{\lambda}+n\right) p_i(t,\omega) + \lambda_{i,-i} p_{-i}(t,\omega) \\ &+ \left(\bar{\lambda}+n\right) J\left(\frac{\omega(W_t+Q_t)}{Z_t/(\bar{\lambda}+n)}, \frac{q_t G_t N_t}{Z_t/(\bar{\lambda}+n)}\right),\end{aligned} \tag{15}$$

where

$$J(y_1,y_2) \equiv \int \Gamma(y_1 - y_2 x)\, d\Phi(x)$$

and p_i also satisfies the boundary conditions:

$$\lim_{\omega\to 0} p_i(t,\omega) = 1 \qquad \textit{and} \qquad \lim_{\omega\to\infty} p_i(t,\omega) = 0.$$

Proof. Online Appendix A. □

3.3. Stationary balanced growth path

In this subsection, we look for a stationary equilibrium as in Huggett (1993) and Aiyagari (1994), in which interest rate, wage rate, rates of return on investment, wealth redistribution functions, and (relative) wealth distributions remain unchanged over time. However, the economy as a whole grows at a constant rate.

Definition 3. A stationary balanced growth path, or balanced growth path (BGP) for short, is a Markov equilibrium in which interest rate, wage rate, rates of return on investment, growth rate, and relative wealth distribution are constant over time.

The concept of stationary balanced growth path is standard (see for example Huggett et al., 2011 and Acemoglu and Cao, 2015). With strictly concave production functions, in a BGP the economy grows at the rate $\bar{g} = n + g_x$ and

$$W_{i,t} = \bar{W}_i \exp(\bar{g}t),$$

which implies $W_t = \bar{W}\exp(\bar{g}t)$ and $Z_t = \bar{Z}\exp(\bar{g}t)$ for some $\bar{W}_i, \bar{W}, \bar{Z}$ that are endogenously determined in equilibrium.

Since interest rate and wage rate are constant over time, $q_t \equiv \bar{q}$, and by (4), $\bar{q}$ is determined by:

$$0 = (\bar{r} - g_x)\,\bar{q} - \bar{e}. \tag{16}$$

The value functions in (8) become $v(t,i) \equiv \bar{v}_i$, where $\bar{v}_i$ satisfies

$$(\rho + \bar{\lambda})\bar{v}_i = \max_{c,k,\tilde{b}} \mathcal{H}_i(c,k,\tilde{b};\bar{v}_i) + \lambda_{i,-i}(\bar{v}_{-i} - \bar{v}_i) \tag{17}$$

where

$$\mathcal{H}_i(c,k,\tilde{b};v) = \log(c) + \frac{1}{\rho+\bar{\lambda}}(R_i(\bar{e})k + \bar{r}\tilde{b} - c) \tag{18}$$

and the maximization is subject to the constraints (10).

The dynamics of aggregate wealth, (14), simplify to

$$(g_x+n)\bar{W}_i = ((1-\tau_i)R_i(\bar{e})\bar{k}^*_i + \bar{r}\bar{b}^*_i - \bar{c}^*_i)\left(\bar{W}_i + \bar{q}M_i\right) - \left(\lambda_{i,-i}+\bar{\lambda}\right)\bar{W}_i + \lambda_{-i,i}\bar{W}_{-i} + M_i\bar{Z} - g_x\bar{q}M_i. \tag{19}$$

The market clearing conditions (12) and (13) become

$$\bar{b}^*_H\left(\bar{W}_H + \bar{q}M_H\right) + \bar{b}^*_L\left(\bar{W}_L + \bar{q}M_L\right) = \bar{q}, \tag{20}$$

and

$$S_H(\bar{e})\bar{k}_H^*\left(\bar{W}_H+\bar{q}M_H\right)+S_L(\bar{e})\bar{k}_L^*\left(\bar{W}_L+\bar{q}M_L\right)=1. \tag{21}$$

The following theorem characterizes stationary balanced growth paths.

Theorem 1 *(Stationary Balanced Growth Path). Assume Assumption 1 holds and $\tau_H \le \tau_L$. Depending on the parameters of the model, a stationary BGP is characterized by four equations: the two equations determining $\bar{W}_H$ and $\bar{W}_L$, (19), the two market clearing conditions, (20) and (21).*

However, the unknowns differ in three different cases:

***Case 1**: $\bar{r}=(1-\tau_L)R_L(\bar{e})<(1-\tau_H)R_H(\bar{e})$. Then $\left(\bar{k}_H^*,\bar{b}_H^*\right)=\left(\frac{1}{1-m},-\frac{m}{1-m}\right)$ and $\bar{b}_L^*=1-\bar{k}_L^*$. The four unknowns are $\bar{k}_L^*,\bar{e},\bar{W}_H,\bar{W}_L$.*

***Case 2:** $(1-\tau_L)R_L(\bar{e})<\bar{r}<(1-\tau_H)R_H(\bar{e})$. Then $\left(\bar{k}_H^*,\bar{b}_H^*\right)=\left(\frac{1}{1-m},-\frac{m}{1-m}\right)$ and $\left(\bar{k}_L^*,\bar{b}_L^*\right)=(0,1)$. The four unknowns are $\bar{r},\bar{e},\bar{W}_H,\bar{W}_L$.*

***Case 3**: $\bar{r}=(1-\tau_H)R_H(\bar{e})$. The four unknowns are $\bar{k}_H^*,\bar{e},\bar{W}_H,\bar{W}_L$.*

Lastly, in Case 1 and Case 2:

$$(1-\tau_H)R_H(\bar{e})\bar{k}_H^*+\bar{r}\bar{b}_H^*-\bar{c}_H^*>(1-\tau_L)R_L(\bar{e})\bar{k}_L^*+\bar{r}\bar{b}_L^*-\bar{c}_L^*.$$

Proof. Appendix. □

In Case 1, given that $\bar{r}=(1-\tau_L)R_L(\bar{e})$, L-agents are indifferent between producing using their production function and lending to the H-agents at interest rate $\bar{r}$. In equilibrium, they do both, i.e., $0\le\bar{k}_L^*\le 1$ and $\bar{k}_L^*,\bar{b}_L^*$ are determined by the equilibrium conditions. The same logic applies to Case 3. Case 3 also applies under homogeneous rate of returns, i.e., $F_H\equiv F_L$ and $\tau_H=\tau_L$.

One important implication of this theorem is that in Case 1 and Case 2, the agents with the higher rate of returns on investment (H-agents) save more. The result is due to the fact that the agents have the same saving rate given by (11) and the more productive agents earn strictly higher returns on their wealth.[14] This result is consistent with Saez and Zucman (2016) who find that saving rates tend to rise with wealth.

Theorem 1 does not tell us the conditions under which the equilibrium belongs to Case 1, 2, or 3. For a special case of this model, the AK model presented in Online Appendix C, we show that Case 1 happens when m is low, Case 2 when m is intermediate, and Case 3 when m is high. We have verified numerically that the same pattern holds for the current model with production.

In a BGP, the aggregate labor share and capital-output ratio are given by:

$$EY\equiv\frac{\bar{e}}{\sum_{i\in\{H,L\}}\left(F_i(1,S_i(\bar{e}))+\delta\right)\bar{k}_i^*\left(\bar{W}_i+M_i\bar{Q}\right)}, \tag{22}$$

and

$$KY\equiv\frac{\bar{W}_H+\bar{W}_L}{\sum_{i\in\{H,L\}}\left(F_i(1,S_i(\bar{e}))+\delta\right)\bar{k}_i^*\left(\bar{W}_i+M_i\bar{Q}\right)}, \tag{23}$$

where δ stands for the depreciation rate of capital.

How are KY and EY related to each other? By looking at the distribution of output between wages and returns to capital, we obtain

$$\underbrace{\left(\frac{\bar{k}_H^*\bar{X}}{\bar{k}_H^*\bar{X}+\bar{k}_L^*(1-\bar{X})}(R_H(\bar{e})+\delta)+\frac{\bar{k}_L^*(1-\bar{X})}{\bar{k}_H^*\bar{X}+\bar{k}_L^*(1-\bar{X})}(R_L(\bar{e})+\delta)\right)}_{\text{average (before-tax) rate of return to capital}}KY=\underbrace{1-EY}_{\text{capital share}},$$

where $\bar{X}=\frac{\bar{W}_H+\bar{q}M_H}{\bar{W}+\bar{q}}$ denotes the total wealth shares of H-agents. Notice that this identity is a generalization of the Piketty (2014)'s First Fundamental Law of Capitalism: "$r\beta=\alpha$," in his model with homogeneous returns. In our model with het-

[14] In Online Appendix B we show that this result holds even for general CRRA utility functions, $\frac{c^{1-\sigma}}{1-\sigma}$. In this case, consumption can be higher or lower due to the opposite income and substitution effects, but the effect of higher rates of returns on saving is unambiguous. The proof of this property is not straightforward due to the switching probability between the two rates of returns. When $\sigma<1$, we can show that the substitution effect on consumption dominates the income effects, therefore $\bar{c}_H^*<\bar{c}_L^*$ so the result on saving rate is immediate. It is more difficult to prove the result when $\sigma>1$, so that $\bar{c}_H^*>\bar{c}_L^*$.

erogeneous returns, capital share is equal to the product of a weighted average of the rates of return to capital and the capital-output ratio.[15]

4. Sufficient statistics for top end wealth inequality

In this section, we use the characterization of stationary BGPs above to show that top end wealth inequality, as characterized by Pareto tail index of the wealth distribution, is structurally linked to salient aggregate statistics through sufficient statistics formulae. First, the following lemma shows that the Pareto tail index for the distribution of total wealth is also the (approximate) index for the distribution of financial wealth. To simplify the analysis, we make the following assumption.

Assumption 2. Φ and Γ have bounded supports.

The lemma is stated as follows.

Lemma 4. *Assume that Assumption 2 holds and the distribution of total wealth (financial plus human wealth) has a right Pareto tail with the tail index θ, that is,*

$$\Pr\left(\omega_t^h \geq \omega\right) = \Psi\omega^{-\theta} \tag{24}$$

for all ω beyond a threshold for some $\Psi > 0$. Then the distribution of financial wealth follows an approximate power law with the same tail index as the one for the distribution of total wealth, i.e., there exist $\bar{d} > \underline{d} > 0$ such that

$$\bar{d}w^{-\theta} \geq \Pr\left(\frac{w_t^h}{W_t/N_t} \geq w\right) \geq \underline{d}w^{-\theta},$$

when $w \to \infty$.

Proof. Appendix. □

The result still applies if the supports of Φ and Γ are unbounded but the distributions have thin tails, or Pareto tail with the tail index strictly higher than θ. However, we need to relax the definition of Pareto tail to "asymptotic Pareto tail" as in Acemoglu and Cao (2015).[16]

Following Piketty (2014), we define *top end wealth inequality* as the upper tail index, θ in Theorem 3. A lower θ corresponds to a fatter tail of the wealth distribution, i.e., a higher degree of top end wealth inequality. In the remaining of this section, we show that θ can be expressed as a function of aggregate statistics.

4.1. Homogeneous returns

In this subsection, we assume $F_H \equiv F_L$ and $\tau_H = \tau_L$ so that all agents in the economy face the same rate of returns to their investment. Under this assumption, we obtain the first sufficient statistics result that links top end wealth inequality to aggregate variables.

Theorem 2 *(Sufficient statistics I). Assume that Assumption 2 holds and $F_H \equiv F_L$ and $\tau_H = \tau_L = \tau \in [0, 1)$, in a stationary BGP the distribution of total wealth has a right Pareto tail, as defined in (24), with the tail index θ given by*

$$\left(1 + \frac{\bar{\lambda}}{n}\right)\left(1 + \frac{EY}{KY}\frac{1}{\bar{r} - g_x}\right)\frac{1}{1 - \frac{\tau}{1-\tau}\frac{\bar{r}}{n}\left(1 + \frac{EY}{KY}\frac{1}{\bar{r}-g_x}\right)}, \tag{25}$$

which depends on exogenous model parameters $\tau, g_x, n, \bar{\lambda}$ and endogenous variables $EY, KY, \bar{r}$.

Proof. Appendix. □

Formula (25) shows that top end inequality is decreasing in corporate tax, (25) is increasing in τ, with the lower bound at zero corporate tax given by:

[15] The production functions F_i implicitly incorporate depreciation. So to measure total output – gross output – we add back depreciation to the net output given by F_i. This practice is more consistent with the modern formulation of the neoclassical growth models such as Cass (1965). Krusell and Smith (2015) show that the distinctions between gross output versus net output and gross saving rate versus net saving rate are important for the predictions of the neoclassical growth models.

[16] A distribution density φ has asymptotic Pareto tail χ if for any $\xi > 0$ there exists $\bar{B}$, $\underline{B}$ and $\underline{x}$ such that $\underline{B}x^{-(\chi-1+\xi)x} < \varphi(x) < \bar{B}x^{-(\chi-1-\xi)x}$ for all $x \geq \underline{x}$.

$$\left(1+\frac{\bar{\lambda}}{n}\right)\left(1+\frac{EY}{KY}\frac{1}{\bar{r}-g_x}\right). \tag{26}$$

In Online Appendix I, we also provide a similar formula and more direct derivation for the standard neoclassical growth model with labor augmenting technological progress as in Barro and Sala-i Martin (2004) and Acemoglu (2009).

Formulae (25) and (26) show that there is a structural relation between top end wealth inequality and salient aggregate statistics such as capital-output ratio and labor share besides "r-g" emphasized in Piketty (2014) and Piketty and Zucman (2014, 2015). The authors argue informally for the importance of these other aggregates on wealth inequality but stop short of making explicit and formal statements as we do. Notice also that (26) is increasing in EY, decreasing in KY and in $\bar{r}-g_x$, consistent with the informal discussions in Piketty (2014) that the wealth distribution is more unequal in economies with lower EY, higher KY, or higher "r-g" (for example, Piketty suggests that higher $r-g$ implies a higher growth rate of wealth relative to labor income, leading to higher wealth inequality, at least between capitalists (wealth holders) and workers; and a lower EY corresponds to a higher capital income share, leading to a higher r).

Using the standard values for parameters $\bar{\lambda}, n, g_x$ and the historical values of EY and KY in the U.S. data (see Table 2), (25) implies a lower bound on the tail index of

$$\left(1+\frac{1/75}{0.01}\right)\left(1+\frac{0.65}{3}\frac{1}{0.03-0.02}\right)\approx 53.89$$

which is too high compared to those observed in the U.S. data (which are between 1.4 and 2 over the years). We will show below that allowing for heterogeneous returns in the model can help match the empirical tail index.

4.2. Heterogeneous returns

Going back to the general model with heterogeneous returns, in the BGP characterized by Theorem 1, the PDEs for stationary wealth distribution, (15), become

$$0=-\frac{\partial \bar{p}_i(\omega)}{\partial \omega}\omega\left(\bar{g}_i^*+n\right)-\left(\lambda_{i,-i}+\bar{\lambda}+n\right)\bar{p}_i(\omega)+\lambda_{i,-i}\bar{p}_{-i}(\omega)+\left(\bar{\lambda}+n\right)J\left(\frac{\omega\left(\bar{W}+\bar{q}\right)}{\bar{Z}/(\bar{\lambda}+n)},\frac{\bar{q}}{\bar{Z}/(\bar{\lambda}+n)}\right), \tag{27}$$

where J is defined in Lemma 3.

The following theorem characterizes the stationary wealth distribution.

Theorem 3 *(Sufficient statistics II). Assume that Assumption 2 holds. In a stationary BGP in which the H-agents earn strictly higher returns to wealth (Case 1 and Case 2 in Theorem 1), we have the following results.*

1) The wealth distribution has a right Pareto tail with the tail index θ where $-\theta$ is the negative root (η) of the following quadratic equation:

$$\left(\frac{\lambda_{HL}+\bar{\lambda}+n}{\bar{g}_H^*+n}+\eta\right)\left(\frac{\lambda_{LH}+\bar{\lambda}+n}{\bar{g}_L^*+n}+\eta\right)=\frac{\lambda_{HL}\lambda_{LH}}{\left(\bar{g}_H^*+n\right)\left(\bar{g}_L^*+n\right)}. \tag{28}$$

2) The tail index, which we call top end wealth inequality, is a function of aggregate statistics – labor share, capital-output ratio, interest rate- and a subset of model parameters:

$$\theta=\hat{\theta}(EY, KY, \bar{r}; g_x, n, \delta, \lambda_{HL}, \lambda_{LH}, \bar{\lambda}, \tau_L, \tau_H, m). \tag{29}$$

Proof. Appendix. □

The equation for the tail index (28) in Part 1 contains several endogenous, and potentially non-observable, variables such as the growth rates $\bar{g}_H^*, \bar{g}_L^*$. Part 2 uses equilibrium conditions to show that the tail index can be expressed as functions of observable aggregate statistics as in Theorem 2. The result allows for any production functions. For example, changing production functions certainly changes the equilibrium stationary BGP. However, top end wealth inequality only changes to the extent that EY, KY, and $\bar{r}$ change due to the change in the production functions.

The second part of Theorem 3 generalizes formula (25) to cases with strictly heterogeneous returns and again provides a structural relation between top end wealth inequality and salient aggregate statistics. Under further assumptions, the following theorem shows how top end wealth inequality covaries with these aggregates.

Theorem 4 *(Comparative statics for top end wealth inequality). Assume that Assumption 2 holds and $n=0$ and $\tau_H=\tau_L=0$. In a stationary BGP with $\bar{k}_L^*>0$, i.e., Case 1 in Theorem 1, we have the following results.*

1) Top end wealth inequality θ is a function of the relative growth rate of the high type and the persistent parameters:

$$\theta=\hat{\theta}_a(\bar{g}_H^*; \lambda_{HL}, \lambda_{LH}, \bar{\lambda}).$$

In addition, top end wealth inequality increases in $\bar{g}_H^*$, i.e., $\frac{\partial \hat{\theta}_b}{\partial \bar{g}_H^*} < 0$.

2) Top end wealth inequality is a function of the aggregate statistics together with the primitive parameters $g_x, \delta, \lambda_{HL}, \lambda_{LH}, \bar{\lambda}$:

$$\theta = \hat{\theta}_b \left(EY, KY, \bar{r}; g_x, \delta, \lambda_{HL}, \lambda_{LH}, \bar{\lambda} \right).$$

In addition, $\frac{\partial \hat{\theta}_b}{\partial KY} > 0$, $\frac{\partial \hat{\theta}_b}{\partial EY} > 0$ and $\frac{\partial \hat{\theta}_b}{\partial g_x} > 0$.

Proof. Appendix. □

The results in Theorem 4 deserve some discussion. With heterogeneous investment returns, top end wealth inequality depends on the relative growth rate of wealth of the high type agents, which in turn depends on their investment returns, equilibrium interest rate, and endogenous consumption and portfolio choice decisions:

$$\bar{g}_H^* = (1 - \tau_H) R_H(\bar{e}) \bar{k}_H^* + \bar{r} \bar{b}_H^* - \bar{c}_H^* - \bar{g}.$$

The first two terms in the right hand side capture the rate of returns on the optimal portfolio of the H-agents. The third term captures the fact that a fraction of the returns is consumed. This term is implicitly ignored by Piketty (2014) when he focuses on the gap "r-g" as the most important determinant of wealth inequality. This emphasis is criticized by Mankiw (2015) and Ray (2015).

The second part of the theorem links the top end wealth inequality to the two aggregate statistics.[17] The last two comparative statics in this part of the theorem are consistent with Piketty (2014)'s discussion: higher labor income share, or higher growth rate of the economy (since $n = 0$, $\bar{g} = g_x$), corresponds to higher Pareto tail index and thus lower top end wealth inequality. However, the first comparative statics differ from Piketty (2014)'s prediction that higher capital-output ratio is associated with higher wealth inequality.[18]

In the following sections, we use the general formula provided in Theorem 3 to calibrate the model with heterogeneous returns and examine its quantitative properties.

5. Quantitative implications

In this section, we calibrate our model using the sufficient statistic results in the previous section and draw quantitative implications. The calibration targets include top end wealth inequality as captured by the Pareto tail index and the salient aggregate statistics including capital-output ratio and labor share. First of all, the simple calibration presented in Subsection 4.1 shows that the special case of the model with homogeneous returns cannot match the tail index observed in the U.S. data. Therefore, we need to use the model with heterogeneous returns studied in Subsection 4.2. Formula (29) for the Pareto tail index of the stationary distribution presented in Theorem 3 allows us to target the tail index in the data. Given the calibrated model, we ask how labor earnings inequality and initial wealth distribution matter quantitatively for top end wealth inequality. In the next section, we also investigate the joint dynamics of top end wealth inequality and the aggregate capital to output ratio and labor share over transitional paths of the calibrated economy after corporate tax cuts and financial deregulation.

5.1. Baseline calibration

For the main calibration, we assume that the production function has constant elasticity of substitution between labor and capital described as below,

$$F_i(k, l) = A_i \left(\alpha k^{\frac{\gamma-1}{\gamma}} + (1 - \alpha) l^{\frac{\gamma-1}{\gamma}} \right)^{\frac{\gamma}{\gamma-1}} - \delta k,$$

where A_i is the productivity of agents of type $i \in \{L, H\}$ with $A_H > A_L$. α and γ are the technology parameters governing factor shares and the elasticity of substitution between capital and labor. Capital depreciates at a constant rate δ.

Table 2 summarizes the parameters and the associated moments targeted, which are from the U.S. economy in the 1980s. The calibrated economy is at the equilibrium where the low type is actively producing. Therefore, the interest rate is equal to the after-tax return to capital for the L-agents.

[17] Notice also that the formula does not depend on the degree of financial friction m, compared to the formula in Theorem 3 (Part 2). As for production functions, changes in m should lead to changes in equilibrium BGP. But top end wealth inequality changes only to the extent that the changes in m lead to changes in KY, EY, or $\bar{r}$.

[18] A casual observation of cross-country data on capital to income ratio and top end wealth inequality from Piketty (2014), Piketty and Zucman (2014), Alvaredo and Saez (2009), and Saez and Zucman (2016) suggests that countries with higher top end wealth inequality might have lower capital income ratio. For example the U.S. has the highest level of top end wealth inequality, but has lower capital income ratio, compared to France, Spain, and the U.K. Our preliminary cross-country regression of top 1% wealth share on KY, controlling for other factors, shows that the coefficient on KY is negative and significant.

Table 2
Calibration targets and results for the neoclassical economy.

Parameters	Value	Target/Source	Target	Model
m	1/3	Evans and Jovanovic (1989)		
γ	1.5	Piketty and Zucman (2015)		
n	0.01	Population growth rate		
g_x	0.02	Labor productivity growth rate		
δ	0.06	Depreciation rate, King and Rebelo (1999)		
$\bar{\lambda}$	1/75	Life span		
$\tau_H = \tau_L$	0.4	Corporate tax rate		
A_L	1	Normalization		
A_H	1.1587	Interest rate	0.03	0.03
α	0.2421	Labor share, Piketty and Zucman (2015)	0.65	0.65
$\rho + \bar{\lambda}$	0.0107	Capital-output ratio, Piketty and Zucman (2015)	3.0	3.0
λ_{HL}	0.1000	Average duration of being of high type	10 yrs	10 yrs
λ_{LH}	1.4240e−04	Pareto tail index of wealth distribution	2.00	2.00

We set the maximum fraction of capital that can be collateralized $m = 1/3$, taken from the estimates by Evans and Jovanovic (1989). We set the elasticity of substitution between labor and capital $\gamma = 1.5$, which is in middle of the range documented by Piketty and Zucman (2015). As emphasized by Piketty and Zucman (2015), this elasticity allows for the positive co-movement between capital-output ratio and capital share over transitional paths.[19] We set the annual depreciation rate $\delta = 0.06$, which is the common value used in the growth and business cycle literature (see e.g. King and Rebelo, 1999). We normalize the productivity of the low type $A_L = 1$, and calibrate the productivity of the high type A_H to match the annual interest rate of 0.03. Corporate tax is set at 40% in line with the numbers documented in McGrattan and Prescott (2005).

We calibrate the capital share α in the production function to match the aggregate labor income share of 0.65. We set the discount rate ρ to match the aggregate capital-output ratio 3.0. As we have shown in Theorem 4, the labor share and capital-output ratio help determine top end wealth inequality, i.e., the Pareto tail index. We target the Pareto tail index in the 1980s for the U.S. economy inferred from Saez and Zucman (2016)'s top wealth shares. Note that, as highlighted in formula (29), with the other parameters set in Table 2, the only remaining parameters that pin down the Pareto tail index are the degrees of persistence of investment returns given by λ_{HL} and λ_{LH}. In order to generate a strong concentration of wealth (low Pareto tail index) we need some persistence in the high return (low λ_{HL}) and few agents with high returns (very low λ_{LH}). In other words, very few agents have access to high returns investments but once they are there, they tend to stay there for a while. This gives them time to accumulate a lot of wealth given the higher incentives to save. In fact, if returns were not persistent, the incentive to save would be much smaller. We choose λ_{HL} so that the average duration of being more productive (type H) is 10 years and λ_{LH} so that the Pareto tail index of the wealth distribution is 2.[20]

In Table 2, we have solved for the aggregate variables and the tail index without having to know the whole stationary distribution of wealth. To solve for the stationary distribution of wealth as a solution to the PDEs (27), we need to choose the distribution of initial labor productivity Φ as well as the redistribution function Γ. We assume that x_t follows a log-normal distribution with standard deviation σ_x:

$$\Phi \propto \exp\left(\mathcal{N}\left(-\frac{\sigma_x^2}{2}, \sigma_x^2\right)\right).$$

In the baseline calibration, we choose $\sigma_x = 0.98$ so that the top 1% income share is 8.2% as documented in Piketty and Saez (2003, Table II) for 1980.

We also assume that the redistribution function follows a log normal distribution:

$$\Gamma \propto \exp\left(\mathcal{N}\left(-\frac{\sigma_\Gamma^2}{2}, \sigma_\Gamma^2\right)\right).$$

The redistribution function corresponds to the wealth inequality at the beginning of the agents' life. We set $\sigma_\Gamma^2 = 0.86$ to match the top 1% wealth share of the 1980 US economy, 24.3%, as reported by Saez and Zucman (2016).[21]

Column 3 (Baseline) in Table 3 shows other statistics for the wealth distribution in the baseline calibration. Notice that our model not only matches the degree of wealth inequality observed in the U.S. data at the very top (Pareto tail index, top 0.1% and top 1% wealth shares) but also matches the statistics at lower wealth levels (top 5% and top 10% wealth shares).

[19] In Online Appendix F, we explore alternative calibrations with different elasticities of substitution.

[20] The switching rate from high to low returns λ_{HL} of 0.1 is consistent with the spells of high returns and high wealth growth rates observed for the richest Americans from Forbes' billionaire lists, for example, Mark Zuckerberg or Bill Gates. In Online Appendix D, we also explore calibrations with different values of λ_{HL}. We find that higher λ_{HL} (shorter duration) leads to faster speed of transition for wealth inequality.

[21] Interestingly, the value of σ_Γ^2 is close to the estimate (0.849) in Huggett et al. (2011) for the variance of log wealth at age 20–25 using PSID data.

Table 3
Wealth inequality statistics as functions of initial wealth distribution.

Wealth statistics	Standard deviation of $\log \psi$			Data (1980)
	$\sigma_\Gamma^2 = 0.1$	0.86 (Baseline)	2.0	
Top 10% wealth share	38.2%	64.7%	92.4%	67.1%
Top 5% wealth share	28.0%	47.5%	72.3%	50.7%
Top 1% wealth share	18.1%	24.3%	37.5%	24.3%
Top 0.1% wealth share	11.5%	11.6%	13.8%	8.0%

These moments of the wealth distribution are "out-of-sample" in our calibration in the sense that we do not target them while calibrating the parameters of the model. This result is a by-product of our calibration that matches exactly the tail index of the wealth distribution. To the extent that the right tail of the wealth distribution in the data can be approximated by a Pareto distribution beyond a certain wealth level, if we match the top wealth share at that wealth level, we closely match top wealth shares at all higher wealth levels.[22]

The high level of top end wealth inequality comes from the persistent difference in the rates of returns. In the calibration, the after-tax rate of returns of the high type is 7.14% (around 9.21% levered returns) compared to 3% interest rate earned by the low type (or by producing themselves). The difference in levered returns is less than the standard deviation of returns from PSID data reported in Section 2.[23] On average this return difference lasts for 10 years, i.e., $\frac{1}{\lambda_{HL}}$. Starting from high returns, the model-implied 5-year and 10-year correlations of returns to wealth are 0.602 and 0.390 respectively.[24] While these correlations are higher than the estimates from PSID, the ratio of the two correlations (≈ 0.65) is close to the ratio from PSID estimates.[25] In summary, the return difference and its persistence are reasonable, yet they are able to deliver high top end wealth inequality.

5.2. The decomposition of wealth inequality

Given the calibrated model in the previous subsection, we can look at different factors that determine wealth inequality. In some sense, we carry out a "decomposition" exercise. We ask, among the factors that potentially affect wealth inequality – return heterogeneity, earnings inequality, and initial wealth distribution – which factors contribute the most to wealth inequality. Because the aggregate dynamics are not affected by changing earnings inequality (the distribution Φ of initial productivity) and initial wealth distribution (the redistribution functions Γ), we are able to provide a very clean answer to this question in our model. First and foremost, top end wealth inequality as measured by the Pareto tail index is not affected by these changes, a direct application of Theorem 3. But we are also interested in other moments of the wealth distribution including top wealth shares. In this subsection, we carry out this investigation.

In our model, earnings shocks are permanent, i.e. deterministic over the life-cycle. Therefore, they do not affect saving rates. The high growth rates of wealth come instead from high investment returns. Consequently, changes in earnings inequality matter little for top end wealth inequality in our model, as characterized by the Pareto tail index or by the top wealth shares. We demonstrate this point in Online Appendix G.1 (Table G.1). However, when earnings shocks are not permanent, changes in the earnings inequality have stronger effects on wealth inequality. In Online Appendix G.2, we present a stochastic life-cycle model, in which earnings shocks are not permanent, and show that this is indeed the case (Table G.2).[26]

Wealth redistribution has stronger effects on top end wealth inequality. Recall that in our model, the wealth redistribution function, Γ, also corresponds to wealth inequality at the beginning of the agents' lifetime. Table 3 shows the statistics of the stationary wealth distribution when we vary σ_Γ^2 around the baseline value of 0.86, which was estimated to match the top 1% wealth share of U.S. economy in 1980. We can see that varying the variance of the redistribution function does not affect top 0.1% wealth share significantly. However, at lower percentiles, such as the top 1% to the top 10%, changing wealth redistribution significantly affects these statistics.

[22] Quantitative papers which do not match the tail index such as Castaneda et al. (2003) and Kaymak and Poschke (2016) but target only certain top wealth share, say top 1% wealth share, tend to under-estimate wealth shares at higher wealth percentiles.

[23] The unlevered return difference of around 4% is close to the difference between the 99th percentile and median returns estimated by Fagereng et al. (2016) using Norwegian household data.

[24] To compute the conditional correlations of returns to wealth in the model, we use Monte Carlo simulation and draw 10^6 samples of agents in the stationary distribution and paths of shocks and simulate their returns and wealth over time. We then construct the 5-year returns using the exact same procedure and sample selection criterion as in Section 2 for PSID data and compute the 5-year and 10-year correlations of the 5-year returns. We condition on current high return, i.e., type H, because the low return state is extremely persistent due to very low λ_{LH}, which yields 5-year and 10-year correlation close to 1.

[25] To the extent that persistently high returns are important in leading to very high wealth and PSID data does not contain many households with high wealth (The household at the 99th percentile in the PSID 2011 wealth distribution is worth only $2.7 millions, compared to $5.5 millions in SCF 2010, and $5.8 millions in Saez and Zucman's 2010 wealth distribution), we expect that the correlation estimates from PSID should be significantly smaller than population-wide correlations.

[26] These analyses also explain the seemingly contradictory results in Benhabib et al. (2015) and Kaymak and Poschke (2016). The former finds negligible effects of earnings inequality on wealth inequality (with deterministic earnings) while the later finds significant effects (with stochastic earnings).

6. Transitional paths and dynamics of wealth inequality

Piketty (2014) and Saez and Zucman (2016) show that top end wealth inequality in the major advanced economies including the U.S. since 1900 until present follows U-shape patterns, i.e., wealth inequality was very high in the beginning of the century then declined through the Great Depression, WWI and WWII, until the late 1970s, and has been rising since then. At the aggregate level, Piketty (2014) and Piketty and Zucman (2014) find that the aggregate wealth to income ratio and capital share have increased in these countries since the late 1970s. In Section 4, we argue that, in a stationary balanced growth path, top end inequality and these aggregate variables are closely related. In this section, we investigate the joint dynamics of top end wealth inequality and the aggregate variables over transitional paths, starting from the calibration in the last section and compare those to their counterparts in the U.S. data from the 1980s until recent years.

Our solution of the transitional paths is facilitated by the decoupling of the dynamics of the aggregate variables and the distribution of wealth. As shown in Lemma 2, we can first solve for the dynamics of $(W_{H,t}, W_{L,t}, q_t)$ over the transitional paths. The solution gives us the dynamics of $g_{i,t}$'s and we use them to solve for the dynamics of wealth distribution using Lemma 3. Wealth distributions over the transitional paths are computed using the finite difference method presented in Achdou et al. (2015).

6.1. Corporate tax cuts

Our first experiment is an unexpected cut of corporate tax rate from 40% to 25%. The cut is in a similar magnitude as the combined decreases in corporate income tax and corporate distribution tax in the 1980s documented in McGrattan and Prescott (2005). The solid lines in Fig. 2 depict the dynamics of top end wealth inequality and the aggregate variables over the transitional path, compared to the empirical counterparts – dotted lines. The empirical tail index at 1% is given by

$$\frac{1}{1+\log_{10}\frac{S(1)}{S(10)}}$$

where $S(1)$ and $S(10)$ are top 1% and top 10% wealth shares,[27] and the wealth to income ratio is taken from Piketty and Zucman (2014), computed using net output:

$$\tilde{KY} = \frac{K}{Y-\delta K} = \frac{KY}{1-\delta KY},$$

where KY is defined as capital to (gross) output ratio as in (23). Because the data on labor to income ratio computed using net output in Piketty and Zucman (2014) is too noisy, we use the standard definition of labor share given in (22) using gross output.

We see that as in the data, the empirical tail index decreases over time, indicating rising top end wealth inequality, and the wealth to income ratio increases and labor share decreases over time. However, the model produces decreasing 1% top wealth share, unlike what is observed in the data. The reason for this outcome is that a corporate tax cut increases the relative growth rate of both H- and L-agents' wealth. To get around this issue, we consider an alternative scenario in which the corporate tax rate for H-agents, τ_H, is cut from 40% to 25% but the rate for L-agents is cut only from 40% to 35%. Following the study by McGrattan and Prescott (2005), we can think of the differential reductions in corporate tax as all agents enjoy a 5%-cut in corporate income tax while only the H-agents enjoy the additional 10%-cut in corporate distribution tax.[28] The dashed lines in Fig. 2 depict the dynamics of top end wealth inequality and the aggregate variables in this scenario. We see that all model variables move in the same direction as those in the data. Quantitatively, the model generates fast speed of transition for wealth inequality – close to half of the speed observed in the data. This is because the model features type-dependence growth rates, i.e., the H-agents experience a higher wealth growth rate than the L-agents do. This type-dependence is crucial for fast transition as emphasized in Gabaix et al. (2016). In Online Appendix D, we show that increasing λ_{HL}, i.e., reducing the average duration of high returns which Gabaix et al. call "live-fast-die-young" dynamics, fastens the transition even more.

Welfare analysis Given the significant increase in top end wealth inequality, one might ask whether the tax cuts might be politically feasible. To answer this question, we look at the welfare changes of agents after the tax cuts using the value function (7). The model admits very tractable welfare analysis and allows us to decompose the welfare changes after reforms into changes due to investment returns and human wealth. The details of the analyses are presented in Online Appendix E. The non-uniform and uniform corporate tax cuts generate similar welfare consequences qualitatively so we just discuss the former in this section.

[27] When the wealth distribution follows exactly a Pareto law then the formula yields the tail index. See Jones (2015) and Gabaix et al. (2016) for papers using this formula to study income distribution.

[28] For example, corporate profits can be paid out as dividends to H-agents and benefit from lower top marginal income tax which was significantly cut by the Tax Reform Act in 1986 (McGrattan and Prescott, 2005, p. 774).

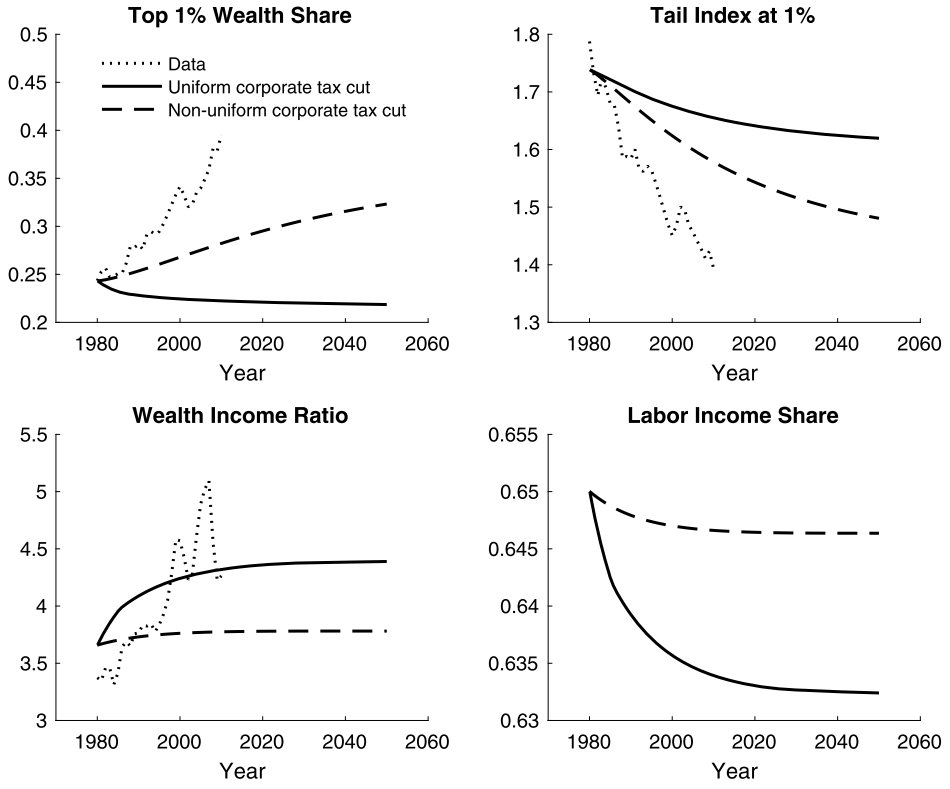

Fig. 2. Transition paths after unexpected corporate tax cuts.

Table 4
Consumption equivalent variation after non-uniform corporate tax cuts.

Income ↓, wealth →	Bottom 50%	50% to 10%	10% to 1%	1% to 0.1%
Bottom 1/3	2.0%	2.9%	3.3%	4.0%
Middle 1/3	1.9%	2.6%	2.8%	3.9%
Top 1/3	1.8%	2.4%	2.5%	3.8%

First, agents benefit from higher investment returns for both investment productivity types. Second, even though the cuts in corporate tax rates result in a higher steady state aggregate capital and a higher wage level eventually, in the short run interest rate increases and that lowers q_t – the present value of human wealth per labor efficiency unit. Since the value function (7) depends on q_t, the agents suffer from welfare loss from the decline in q_t. The average consumption equivalent welfare gain is 2.1%, of which 3.8% is due to increased returns and −1.7% is due to decline in human wealth. Table 4 reports the consumption equivalent variation after the non-uniform corporate tax cut reform for different wealth and income classes. Welfare gain is unevenly distributed across population. Wealth-rich agents benefit more from the increased returns. High-income agents lose more from the decline in human wealth and thus face lower total welfare gains. But *all* households support the tax reform – despite the resulting increase in wealth inequality. The welfare analysis tells us that policies which lead to increases in wealth inequality might not all be detrimental to the welfares of agents in an economy and sometimes might even be beneficial to the majority of agents.

6.2. Financial deregulation

In this subsection, we examine the transitional path after m increases to $1/2$, from its initial value of $1/3$ in the baseline calibration. This relaxation of the H-agents' borrowing constraint captures the deregulation of the financial system in the U.S. in the late 1980s, such as the repeal of the Glass–Steagall act, leading to a rapid rise of credit extension to non-financial corporations (and other sectors in the economy).[29] The solid lines in Fig. 3 again show the dynamics of top end wealth

[29] See the time series for total credit for non-financial corporation in Fred (QUSNAM770A). Credit extension to non-financial corporations increases from roughly 40% of GDP before 1980s to 60% of GDP afterwards. Relatedly, Zetlin-Jones and Shouride (2016, Fig. 16) shows that the aggregate debt-to-total assets of non-farm non-financial corporations in the U.S. increased from around 30% before the 1980s to more than 50% afterwards.

Fig. 3. Transition path after financial deregulation with or without a corporate tax cut.

inequality and the aggregate variables over the transitional path, compared to the empirical counterparts – dotted lines. We observe that wealth inequality, captured by the top 1% wealth shares and the empirical tail index at 1%, increases rapidly closes to what they do in the data.[30] However, the wealth to income ratio decreases and labor share increases – opposite to what we see in the data. The reason is that, since H-agents are more productive, their optimal capital-output ratio is lower than the L-agents' ratio. As the H-agents can leverage more, they hold a larger share of capital for production. Consequently, the aggregate capital-output ratio tilts toward the H-agents ratio and thus decreases.

Now, as we see in the last subsection, a decrease in corporate tax leads to a higher capital-output ratio. So a combination of two changes – financial deregulation and a corporate tax cut – can potentially lead to both rising wealth inequality and increasing capital-output ratio, and decreasing labor share. The dashed lines in Fig. 3 depict the dynamics of top end wealth inequality and the aggregate variables over the transitional path after m increases from $1/3$ to $1/2$ and corporate tax rate decreases from 40% to 25% and show that it is indeed the case. The combination of the two policy changes also lead to faster transition for wealth inequality.

7. Related literature and conclusion

We document empirical evidence supporting persistent differences in returns to wealth from PSID surveys.[31] Motivated by this finding, we put forth a tractable neoclassical growth, incomplete markets model with persistent heterogeneous investment risk. The model delivers a fat-tailed distribution of wealth. We show that there is a structural relation between top end wealth inequality, characterized by the right Pareto tail index of the wealth distribution, and the observables aggregate statistics including capital-output ratio, labor share, interest rate, and growth rate of the economy. The model can be calibrated to match the Pareto tail index estimated from the U.S. data and, consistent with the empirical evidence, requires significant heterogeneous returns. Over the transitional paths, the model can produce the joint dynamics of wealth inequality, capital-output ratio and labor share observed in the data, after corporate tax cuts and financial deregulation.

[30] We show in Online Appendix C that varying m might lead to non-monotone changes in top end wealth inequality (Financial Kuznets Curve) due to opposing "leverage" and "returns equalization" effects. In our calibration, the former effect dominates and hence an increase in m leads to an increase in top end wealth inequality.

[31] Fagereng et al. (2016) find similar evidence in Norwegian household-level data. Saez and Zucman (2016) also find that, for foundations, total rates of returns – including unrealized capital gains – rise sharply with foundation wealth. Using Swedish household-level data, Calvet et al. (2007) document that more financially sophisticated households earn higher mean returns from investments in publicly traded financial instruments, and Bach et al. (2017) find that wealthier households earn higher returns, on all types of financial assets, to wealth.

The current paper belongs to the theoretical literature on Pareto distribution for wealth. Early work started with Stiglitz (1969); more recently, Benhabib et al. (2011, 2015, 2016), Moll (2012), Toda (2014), Achdou et al. (2015), Nirei and Aoki (2016), and others have provided interesting and important mechanisms to generate stationary wealth distribution with Pareto tails.[32] The Pareto tail index in these papers is given by formulae involving unobservable parameters and variables. In this paper, we not only show that wealth distribution has a right Pareto tail as found in the earlier literature, but we also attempt to go a step further by characterizing how the tail parameters vary with the underlying parameters and observable aggregate variables in the economy. In general equilibrium, observable aggregates provide information about unobservable parameters that help determine and forecast wealth inequality. This mechanism is absent in partial equilibrium papers, such as Benhabib et al. (2011, 2015, 2016).

Our paper is also related to the growing literature studying idiosyncratic investment risks. In this literature, most papers, including Angeletos and Calvet (2005, 2006), Angeletos (2007), Toda (2014), Benhabib et al. (2015, 2016), Piketty and Zucman (2015), and Nirei and Aoki (2016), assume I.I.D. investment risks, or, more precisely, investment returns.[33,34] We extend their analysis to allow for *persistent* idiosyncratic investment risks. From this perspective, our paper is closely related to Benhabib et al. (2011), Buera and Shin (2011), Moll (2012, 2014) who consider persistent idiosyncratic investment risks. Our paper can be seen as a simplification of these papers. We use a simple two-level idiosyncratic investment risk Poisson process, which allows us to obtain sharp analytical characterizations of top end wealth inequality and transitional paths in *general equilibrium*.[35]

Heterogeneous returns in our paper can be interpreted as returns to entrepreneurs versus workers, as in the literature on entrepreneurship, including Quadrini (2000), Cagett and De Nardi (2006), and Buera (2009). In these papers, entrepreneurs have access to more productive production technologies than non-entrepreneurs. This assumption can be mapped into heterogeneous returns to investment in our model. However, unlike these papers, we do not assume fixed costs of starting up a business or decreasing returns to scale. This assumption allows us to preserve the homogeneity of the optimization problem of the agents, which simplifies the characterization of wealth distribution and transitional paths.

Another strand of literature starting with Huggett (1996) attempts to explain the high degree of wealth inequality using models with idiosyncratic earnings shocks over the lifecycle of households. However, this class of models cannot match wealth inequality at the very top unless one assumes very large temporary earnings shocks as in Castaneda et al. (2003). Using a similar model, Kaymak and Poschke (2016) show that increases in (temporary) earnings inequality have accounted for most of the increase in wealth inequality in the U.S. since the 1970s. In our model, in which wealth inequality is driven by persistent heterogeneous investment returns and earnings inequality is permanent, changing earnings inequality has negligible effects on wealth inequality. Benhabib et al. (2011) and Benhabib et al. (2015) also emphasize this point in partial equilibrium models.

Over transitional paths, Gabaix et al. (2016) find that models with type-dependent growth rates can generate fast transition for income or wealth distribution. Type dependence arises in our model since agents with different returns have different wealth growth rates. Consequently, our model also generates fast transition for wealth inequality – close to the transition speed observed in the U.S. data since the 1980s – after a reduction in corporate tax rate and financial deregulation which relaxes agents' borrowing constraint.

The literature on wealth inequality is large and fast growing. But there are many open questions begging for further research. For example, as we have argued for the importance of heterogeneous returns, it is natural to ask where the heterogeneity comes from and endogenize it in a model. Relatedly, in our model, we assume that capital is accumulable but in reality, large components of aggregate wealth are in relatively fixed supply such as land and natural resources. The returns on these assets can be very different from the returns on accumulable capital. We believe that some extensions of our model can help study these questions.

Appendix A

Proof of Theorem 1. From the HBJ equations (17), we can solve separately for the portfolio choice of the agents with investment productivity i as:

$$\max_{k,b}(1-\tau_i)R_i(\bar{e})k+b \tag{30}$$

subject to constraints (10).

[32] Benhabib and Bisin (2016) provide an excellent survey of the literature.

[33] Piketty and Zucman (2015) assume I.I.D. saving rates, but this assumption is isomorphic to I.I.D. investment risks assumption. Krusell and Smith (1998) and more recently Hubmer et al. (2016) also generate large wealth inequality from heterogeneous saving rates, which in turn arise from heterogeneous discount factors.

[34] The analytical results in Nirei and Aoki (2016) rely on I.I.D. investment shocks, but the authors relax this assumption in their quantitative investigation.

[35] Benhabib et al. (2011) and Gabaix et al. (2016, Online Appendix) provide analytical characterizations of the Pareto tail index for wealth distribution under persistent investment shocks as we do, but only in partial equilibrium settings in which the rates of returns are exogenous. General equilibrium might lead to interesting comparative statics, such as a non-monotone effect of financial development on the tail index, as shown numerically in Moll (2012) and analytically in our Online Appendix C (Proposition C.2).

In a BGP, if $\bar{r} > (1-\tau_H)R_H(\bar{e}) > (1-\tau_L)R_L(\bar{e})$, then (30) implies that

$$\left(\bar{k}_H^*, \bar{b}_H^*\right) = \left(\bar{k}_L^*, \bar{b}_L^*\right) = (0,1).$$

This, however, contradicts the labor market clearing condition, (21).

Similarly, if $\bar{r} < (1-\tau_L)R_L(\bar{e})$, (30) implies that

$$\left(\bar{k}_H^*, \bar{b}_H^*\right) = \left(\bar{k}_L^*, \bar{b}_L^*\right) = \left(\frac{1}{1-m}, -\frac{m}{1-m}\right),$$

which also contradicts the bond market clearing condition. Therefore in a BGP, $\bar{r} \in [(1-\tau_L)R_L(\bar{e}), (1-\tau_H)R_H(\bar{e})]$.

Case 1: If $\bar{r} = (1-\tau_L)R_L(\bar{e}) < (1-\tau_H)R_H(\bar{e})$, then (30) implies that $\left(\bar{k}_H^*, \bar{b}_H^*\right) = \left(\frac{1}{1-m}, -\frac{m}{1-m}\right)$.

Case 2: If $\bar{r} \in ((1-\tau_L)R_L(\bar{e}), (1-\tau_H)R_H(\bar{e}))$, then (30) implies that $\left(\bar{k}_H^*, \bar{b}_H^*\right) = \left(\frac{1}{1-m}, -\frac{m}{1-m}\right)$ and $\left(\bar{k}_L^*, \bar{b}_L^*\right) = (0,1)$.

Case 3: if $\bar{r} = (1-\tau_H)R_H(\bar{e}) > (1-\tau_H)R_L(\bar{e})$, then $\left(\bar{k}_L^*, \bar{b}_L^*\right) = (0,1)$.

In Case 1 and Case 2, since $\bar{c}_H^* = \bar{c}_L^* = \rho + \bar{\lambda}$ and $(1-\tau_H)R_H(\bar{e}) > \bar{r}$, it is immediate that $\frac{(1-\tau_H)R_H(\bar{e}) - m\bar{r}}{1-m} - \bar{c}_H^* > \bar{r} - \bar{c}_L^*$. □

Proof of Lemma 4. In a stationary BGP, $W_t = N_t G_t \bar{W}$ and $Q_t = N_t G_t \bar{q}$. From the definition of (relative) total wealth,

$$\omega_t^h = \frac{w_t^h + q_t x_t^h}{(W_t + Q_t)/N_t} = \frac{w_t^h + \bar{q}x_t^h}{G_t\bar{W} + G_t\bar{q}}.$$

Therefore

$$\Pr\left(\frac{w_t^h}{W_t/N_t} \geq w\right) = \int_x \Pr\left(\frac{w_t^h}{G_t\bar{W}} \geq w, \frac{x_t^h}{G_t} = x\right) d\Phi(x) = \int_x \Pr\left(\omega_t^h \geq \frac{w\bar{W} + \bar{q}x}{\bar{W} + \bar{q}}, \frac{x_t^h}{G_t} = x\right) d\Phi(x)$$

Now

$$\Pr\left(\omega_t^h \geq \frac{w\bar{W} + \bar{q}\underline{x}}{\bar{W} + \bar{q}}\right) \geq \int_x \Pr\left(\omega^h \geq \frac{w\bar{W} + \bar{q}x}{\bar{W} + \bar{q}}, \frac{x_t^h}{G_t} = x\right) d\Phi(x) \geq \Pr\left(\omega_t^h \geq \frac{w\bar{W} + \bar{q}\bar{x}}{\bar{W} + \bar{q}}\right)$$

By the assumption, when w is sufficiently high,

$$\Pr\left(\omega_t^h \geq \frac{w\bar{W} + \bar{q}\underline{x}}{\bar{W} + \bar{q}}\right) = \Psi\left(\frac{w\bar{W} + \bar{q}\underline{x}}{\bar{W} + \bar{q}}\right)^{-\theta}$$

and

$$\Pr\left(\omega_t^h \geq \frac{w\bar{W} + \bar{q}\bar{x}}{\bar{W} + \bar{q}}\right) = \Psi\left(\frac{w\bar{W} + \bar{q}\bar{x}}{\bar{W} + \bar{q}}\right)^{-\theta}.$$

Taking $w \to \infty$, we obtain the desired bounds. □

Proof of Theorem 2. Since the agents have the same returns to wealth, the dynamics for aggregate wealth, (19), simplifies to:

$$(g_x + n)\bar{W} = \left(R(\bar{e}) - \bar{c}^*\right)\left(\bar{W} + \bar{q}\right) - g_x\bar{q} \tag{31}$$

and the labor market clearing condition becomes $\bar{W}S(\bar{e}) = 1$. Lastly, $\bar{q} = \frac{\bar{e}}{(1-\tau)\bar{R}(\bar{e}) - g_x}$.

The PDE for the stationary distribution of total wealth, (3), also simplifies to:

$$0 = -\bar{p}'(\omega)\omega\left(\bar{g}^* + n\right) - \left(\bar{\lambda} + n\right)\bar{p}(\omega) + (\bar{\lambda} + n)J\left(\frac{\omega\left(\bar{W} + \bar{q}\right)}{\bar{Z}/(\bar{\lambda} + n)}, \frac{\bar{q}}{\bar{Z}/(\bar{\lambda} + n)}\right).$$

Because Γ and Φ have bounded support, $\bar{p}(\omega) \propto \omega^{-\theta}$ where the tail index θ is given by

$$\frac{\bar{\lambda} + n}{\bar{g}^* + n} = \frac{\bar{\lambda} + n}{\bar{r} - \bar{c}^* - g_x}, \tag{32}$$

where $\bar{r} = (1-\tau)R(\bar{e})$ and $\bar{c}^* = \rho + \lambda$.

From (31), we have

$$\bar{r}-\bar{c}+\frac{\tau}{1-\tau}\bar{r}=(g_x+n)\frac{\bar{W}}{\bar{W}+\bar{q}}+g_x\frac{\bar{q}}{\bar{W}+\bar{q}}=n\frac{\bar{W}}{\bar{W}+\bar{q}}+g_x=n\frac{1}{1+\frac{\bar{e}}{W(\bar{r}-g_x)}}+g_x.$$

Therefore

$$\bar{r}-\bar{c}^*-g_x=n\frac{\bar{r}-g_x}{\bar{r}-g_x+\frac{EY}{KY}}-\frac{\tau}{1-\tau}\bar{r}.$$

Plugging this expression for $\bar{r}-\bar{c}^*-g_x$ into (32) and simplifies, we obtain (25). □

Proof of Theorem 3 (Part 1). Since the supports of Φ and Γ are bounded, there exists $\bar{x}$ such that $\Phi'(x)=0$ for all $x\geq\bar{x}$ and $\bar{\psi}$ such that $\Gamma(\psi)=0$ for all $\psi\geq\bar{\psi}$ and $t\geq 0$. Let $\bar{\omega}$ be defined by

$$\frac{\bar{\omega}\left(\bar{W}_H+\bar{W}_L+\bar{q}\right)}{\left(\bar{W}_H+\bar{W}_L\right)\frac{\bar{\lambda}}{\bar{\lambda}+n}}-\frac{\bar{q}}{\left(\bar{W}_H+\bar{W}_L\right)\frac{\bar{\lambda}}{\bar{\lambda}+n}}\bar{x}=\bar{\psi}$$

Then from the definition of J

$$J\left(\frac{\omega\left(\bar{W}_H+\bar{W}_L+\bar{q}\right)}{\left(\bar{W}_H+\bar{W}_L\right)\frac{\bar{\lambda}}{\bar{\lambda}+n}},\frac{\bar{q}}{\left(\bar{W}_H+\bar{W}_L\right)\frac{\bar{\lambda}}{\bar{\lambda}+n}}\right)=0$$

for all $\omega\geq\bar{\omega}$.

The PDEs (27) then become, for $\omega\geq\bar{\omega}$ and $i\in\{H,L\}$

$$0=-\omega(\bar{g}_i^*+n)\frac{dp_i^*(\omega)}{d\omega}-p_i^*(\omega)\left(\lambda_{i,-i}+\bar{\lambda}\right)+p_{-i}^*(\omega)\lambda_{i,-i}. \tag{33}$$

We use the following change of variable

$$z=\log(\omega)$$
$$\tilde{p}_i(z)=p_i^*(e^z).$$

We rewrite (33) as

$$\tilde{p}_H'(z)=\frac{\lambda_{HL}}{\bar{g}_H^*+n}\tilde{p}_L(z)-\frac{\lambda_{HL}+\bar{\lambda}}{\bar{g}_H^*}\tilde{p}_H(z)$$
$$\tilde{p}_L'(z)=\frac{\lambda_{HL}}{\bar{g}_L^*}\tilde{p}_H(z)-\frac{\lambda_{LH}+\bar{\lambda}}{\bar{g}_L^*}\tilde{p}_L(z),$$

or equivalently,

$$\begin{bmatrix}\tilde{p}_H'(z)\\ \tilde{p}_L'(z)\end{bmatrix}=\begin{bmatrix}-\frac{\lambda_{HL}+\bar{\lambda}+n}{\bar{g}_H^*+n} & \frac{\lambda_{HL}}{\bar{g}_H^*+n}\\ \frac{\lambda_{LH}}{\bar{g}_L^*+n} & -\frac{\lambda_{LH}+\bar{\lambda}+n}{\bar{g}_L^*+n}\end{bmatrix}\begin{bmatrix}\tilde{p}_H(x)\\ \tilde{p}_L(x)\end{bmatrix}. \tag{34}$$

The eigenvalues of the matrix are the root of the quadratic equation given in the statement of the theorem.

When n is close to 0, the quadratic equation has two distinct real roots, one positive and one negative. Denote them by $\eta_1<0,\eta_2>0$. Given η_1 and η_2, the general solution of (34) is

$$\begin{cases}\tilde{p}_H(x)=\Psi_H\exp(\eta_1 x)+\tilde{\Psi}_H\exp(\eta_2 x)\\ \tilde{p}_L(x)=\Psi_L\exp(\eta_1 x)+\tilde{\Psi}_L\exp(\eta_2 x)\end{cases}$$

Since $\begin{cases}\lim_{x\to+\infty}\tilde{p}_H(x)=0\\ \lim_{x\to+\infty}\tilde{p}_L(x)=0\end{cases}$, we must have $\begin{cases}\tilde{p}_H(x)=\Psi_H\exp(\eta_1 x)\\ \tilde{p}_L(x)=\Psi_L\exp(\eta_1 x)\end{cases}$ with Ψ_H,Ψ_L positive. This is equivalent to $\begin{cases}p_H^*(\omega)=\Psi_H\omega^{\eta_1}\\ p_L^*(\omega)=\Psi_L\omega^{\eta_1}\end{cases}$. □

Proof of Theorem 3 (Part 2). Part 1 of the theorem shows that θ is a function of $\bar{g}_H^*,\bar{g}_L^*$ beside n and λ's. To derive formula (29), we needs to show that $\bar{g}_H^*,\bar{g}_L^*$ are themselves functions of the aggregate statistics and model parameters. We derive these functions below.

A BGP is characterized by the following system of equations:

$$KY = \frac{\bar{k}_L^*(1-\bar{X}) + \bar{k}_H^*\bar{X}}{[\bar{k}_L^*(\bar{R}_L+\delta)(1-\bar{X}) + \bar{k}_H^*(\bar{R}_H+\delta)\bar{X}]/(1-EY)} \tag{35a}$$

$$(1-\bar{k}_H^*)\bar{X} + (1-\bar{k}_L^*)(1-\bar{X}) - \frac{EY}{KY\cdot(\bar{r}-g_x)+EY} = 0 \tag{35b}$$

$$\bar{r} = \bar{R}_L(1-\tau_L) \tag{35c}$$

$$\bar{g}_L^* = (1-\tau_L)\bar{R}_L\bar{k}_L^* + \bar{r}(1-\bar{k}_L^*) - (\rho+\bar{\lambda}) - (n+g_x) \tag{35d}$$

$$\bar{g}_H^* = (1-\tau_H)\bar{R}_H\bar{k}_H^* + \bar{r}(1-\bar{k}_H^*) - (\rho+\bar{\lambda}) - (n+g_x) \tag{35e}$$

$$\bar{g}_H^*\bar{X} + \bar{g}_L^*(1-\bar{X}) + n\frac{EY}{KY\cdot(\bar{r}-g_x)+EY} + \tau_L\bar{R}_L\bar{k}_L^*(1-\bar{X}) + \tau_H\bar{R}_H\bar{k}_H^*\bar{X} = 0 \tag{35f}$$

$$\begin{aligned}(\bar{g}_H^* - \bar{g}_L^*)\bar{X}(1-\bar{X}) - \lambda_{HL}\bar{X} + \lambda_{LH}(1-\bar{X}) \\ + [\tau_L\bar{R}_L\bar{k}_L^*(1-\bar{X}) + \tau_H\bar{R}_H\bar{k}_H^*\bar{X} + n\frac{EY}{KY\cdot(\bar{r}-g_x)+EY} + \bar{\lambda}](M_H-\bar{X}) = 0,\end{aligned} \tag{35g}$$

where the first equation comes from the definition of KY, the second equation comes from the bond market clearing conditions, and the last two equations come from the dynamics of g_t and X_t given in Lemma A.1 (Online Appendix A). Given the observable variables $(KY, EY, \bar{r}, g_x)$ and given $\bar{k}_H^* = 1/(1-m)$, this is a system of seven equations for seven unknowns $(\bar{k}_L^*, \bar{X}, \bar{R}_L, \bar{R}_H, \bar{g}_L^*, \bar{g}_H^*, \rho)$.

From (35b) and (35a), we have

$$\bar{k}_H^*\bar{X} + \bar{k}_L^*(1-\bar{X}) = 1 - \frac{EY}{KY\cdot(\bar{r}-g_x)+EY} = \frac{KY\cdot(\bar{r}-g_x)}{KY\cdot(\bar{r}-g_x)+EY} \tag{36}$$

and

$$KY = \frac{[\bar{k}_L^*(1-\bar{X}) + \bar{k}_H^*\bar{X}](1-EY)}{\bar{k}_L^*\bar{R}_L(1-\bar{X}) + \bar{k}_H^*\bar{R}_H\bar{X} + \delta[\bar{k}_H^*\bar{X} + \bar{k}_L^*(1-\bar{X})]}.$$

Therefore

$$\bar{k}_L^*\bar{R}_L(1-\bar{X}) + \bar{k}_H^*\bar{R}_H\bar{X} = \frac{[\bar{k}_L^*(1-\bar{X}) + \bar{k}_H^*\bar{X}](1-EY)}{KY} - \delta[\bar{k}_L^*(1-\bar{X}) + \bar{k}_H^*\bar{X}] = \frac{(\bar{r}-g_x)(1-EY-\delta KY)}{KY\cdot(\bar{r}-g_x)+EY}. \tag{37}$$

(36) and (37) basically say that aggregate statistics KY, EY, $\bar{r}$ put discipline on the fraction of wealth held by high type and investment returns of the two types.

Equations (35d), (35e), and (35f) yield:

$$\bar{k}_L^*\bar{R}_L(1-\bar{X}) + \bar{k}_H^*\bar{R}_H\bar{X} + \bar{r}[1-(\bar{k}_H^*\bar{X} + \bar{k}_L^*(1-\bar{X}))] - (\rho+\bar{\lambda}) - (n+g_x) + n\frac{EY}{KY\cdot(\bar{r}-g_x)+EY} = 0.$$

Plugging (36) and (37) into the last equation, we arrive at

$$\rho + \bar{\lambda} = \frac{(\bar{r}-g_x)(1-EY-\delta KY) + \bar{r}EY + nEY}{KY\cdot(\bar{r}-g_x)+EY} + (n+g_x). \tag{38}$$

Finally, combine (35c) and (35d), we have

$$\bar{g}_L^* = \bar{r} - (\rho+\bar{\lambda}) - (n+g_x).$$

Substituting $\rho+\bar{\lambda}$ in using (38), we obtain:

$$\begin{aligned}\bar{g}_L^* &= \bar{r} - \frac{(\bar{r}-g_x)(1-EY-\delta KY) + \bar{r}EY + nEY}{KY\cdot(\bar{r}-g_x)+EY} \\ &= \bar{r} - g_x - \frac{(1-KY(g_x+\delta))(\bar{r}-g_x) + nEY}{KY(\bar{r}-g_x)+EY}.\end{aligned} \tag{39}$$

Next, (35a) and (35b) can be viewed as a system of linear equations for $\bar{X}$ and $\bar{k}_L^*(1-\bar{X})$, which gives the unique solution for $\bar{X}$:

$$\bar{X} = \frac{(1-KY(\bar{R}_L+\delta)-EY)(\bar{r}-g_x)(1-m)}{(\bar{R}_H-\bar{R}_L)(KY(\bar{r}-g_x)+EY)}. \tag{40}$$

At this point, we have already solved out $(\rho, \bar{R}_L)$ as functions of observables. From (35e), $\bar{R}_H$ is an affine transformation of $\bar{g}_H^*$ with intercept and slope that are functions of observables only, hence combining (40), $\bar{g}_H^*\bar{X}$ is an affine transformation of $\bar{X}$ with intercept and slope that are functions of observables only.

Now from (35f)

$$n\frac{EY}{KY\cdot(\bar{r}-g_X)+EY}+\tau_L\bar{R}_L\bar{k}_L^*(1-\bar{X})+\tau_H\bar{R}_H\bar{k}_H^*\bar{X}=-\bar{g}_H^*\bar{X}-\bar{g}_L^*(1-\bar{X}).$$

Plugging this into (35g), we have

$$(\bar{g}_H^*-\bar{g}_L^*)\bar{X}(1-\bar{X})-\lambda_{HL}\bar{X}+\lambda_{LH}(1-\bar{X})+[-\bar{g}_H^*\bar{X}-\bar{g}_L^*(1-\bar{X})+\bar{\lambda}](M_H-\bar{X})=0,$$

which is further simplified to

$$\bar{g}_H^*\bar{X}-\lambda_{HL}\bar{X}+\lambda_{LH}(1-\bar{X})+[-\bar{g}_H^*\bar{X}-\bar{g}_L^*(1-\bar{X})]M_H+\bar{\lambda}(M_H-\bar{X})=0 \tag{41}$$

Since $\bar{g}_H^*\bar{X}$ is an affine transformation of $\bar{X}$ with intercept and slope that are functions of observables only, (41) is a linear equation for $\bar{X}$ with coefficients that are functions of observables. Therefore $\bar{X}$ can be uniquely solved as a function of observables. And we can then solve out $\bar{g}_H^*$ from the expression of $\bar{g}_H^*\bar{X}$.

Combining the expression for g_H^* and the earlier expression for g_L^*, (39), with the explicit solution for the tail index given in Part 1 of the theorem, we arrive at formula (29). □

Proof of Theorem 4 (Part 1). Since $n=0$, from the equations (A.3) and (A.4) in Online Appendix A, evaluated at the steady-state, we have

$$\bar{g}_H^*\bar{X}+\bar{g}_L^*\left(1-\bar{X}\right)=0$$

and

$$\left(\bar{g}_H^*-\bar{g}_L^*\right)\left(1-\bar{X}\right)\bar{X}-\bar{X}\left(\lambda_{HL}+M_L\bar{\lambda}\right)+\left(1-\bar{X}\right)\left(\lambda_{LH}+M_H\bar{\lambda}\right)=0.$$

We also have $\bar{g}_H^*\geq 0\geq\bar{g}_L^*$. If $\bar{g}_L^*=0$ then $\bar{g}_H^*=\bar{g}_L^*=0$. Consequently there is no wealth inequality in a stationary equilibrium, i.e. $\eta_1=-\infty$. We now show the result in this theorem assuming $\bar{g}_H^*>0>\bar{g}_L^*$.

Denote $\tilde{\lambda}_{HL}=\lambda_{HL}+M_L\bar{\lambda}$ and $\tilde{\lambda}_{LH}=\lambda_{LH}+M_H\bar{\lambda}$. After simplification, we can write $\bar{g}_H^*$ as a function of $\bar{g}_L^*$:

$$\bar{g}_H^*=\hat{g}_H(\bar{g}_L^*):=\frac{\tilde{\lambda}_{HL}}{1-\frac{\tilde{\lambda}_{LH}}{\bar{g}_L^*}}.$$

Plugging this expression for $\bar{g}_H^*$ into the equations that determines the tail index, (28) with $n=0$, we obtain $\hat{f}(\eta_1^*,\bar{g}_L^*)=0$, where

$$\hat{f}(\eta,g_L)=\eta^2+\eta\left(1+\frac{\bar{\lambda}(g_L+\hat{g}_H(g_L))}{g_L\hat{g}_H(g_L)}\right)+\frac{\bar{\lambda}(\tilde{\lambda}_{LH}+\tilde{\lambda}_{HL})}{g_L\hat{g}_H(g_L)}. \tag{42}$$

By the Implicit Function Theorem: $\frac{\partial\hat{\eta}_1}{\partial\bar{g}_L^*}=-\frac{\partial\hat{f}}{\partial g_L}/\frac{\partial\hat{f}}{\partial\eta}$. We have $\frac{\partial\hat{f}}{\partial\eta}|_{\eta=\eta_1^*}<0$. Therefore, to prove that $\frac{\partial\hat{\eta}_1}{\partial\bar{g}_L^*}<0$, it suffices to show $\frac{\partial\hat{f}}{\partial g_L}<0$.

Differentiating (42), we obtain

$$\frac{\partial\hat{f}}{\partial g_L}=-\eta_1\bar{\lambda}\left(\frac{\hat{g}_H^2+g_L^2\frac{\partial\hat{g}_H}{\partial g_L}}{(g_L\hat{g}_H)^2}\right)-\frac{\bar{\lambda}(\tilde{\lambda}_{LH}+\tilde{\lambda}_{HL})}{(g_L\hat{g}_H)^2}(\hat{g}_H+g_L\frac{\partial\hat{g}_H}{\partial g_L}).$$

From the expression for $\hat{g}_H$: $\frac{\partial\hat{g}_H}{\partial g_L}=-\frac{\tilde{\lambda}_{LH}\tilde{\lambda}_{HL}}{(g_L-\lambda_{LH})^2}$. Therefore,

$$\begin{aligned}\hat{g}_H+g_L\frac{\partial\hat{g}_H}{\partial g_L}&=\frac{\tilde{\lambda}_{LH}\tilde{\lambda}_{HL}(g_L-\tilde{\lambda}_{LH})+\tilde{\lambda}_{HL}(g_L-\tilde{\lambda}_{LH})^2-\tilde{\lambda}_{LH}\tilde{\lambda}_{HL}g_L}{(g_L-\lambda_{LH})^2}\\&=\frac{\tilde{\lambda}_{HL}(g_L-\tilde{\lambda}_{LH})^2-\tilde{\lambda}_{LH}^2\tilde{\lambda}_{HL}}{(g_L-\lambda_{LH})^2}>0\end{aligned}$$

since $g_L<0$.

Now

$$\frac{\partial\hat{f}}{\partial g_L}<0\Leftrightarrow\eta_1\bar{\lambda}\left(\frac{\hat{g}_H^2+g_L^2\frac{\partial\hat{g}_H}{\partial g_L}}{(g_L\hat{g}_H)^2}\right)>-\frac{\bar{\lambda}(\tilde{\lambda}_{LH}+\tilde{\lambda}_{HL})}{(g_L\hat{g}_H)^2}(\hat{g}_H+g_L\frac{\partial\hat{g}_H}{\partial g_L}).$$

If $\frac{\hat{g}_H^2+g_L^2\frac{\partial \hat{g}_H}{\partial g_L}}{(g_L\hat{g}_H)^2}\leq 0$, this trivially holds since $\eta_1<-0$.

If $\frac{\hat{g}_H^2+g_L^2\frac{\partial \hat{g}_H}{\partial g_L}}{(g_L\hat{g}_H)^2}>0$, this is equivalent to

$$\eta_1>\eta^*=-\frac{(\tilde{\lambda}_{LH}+\tilde{\lambda}_{HL})\left(\hat{g}_H+g_L\frac{\partial \hat{g}_H}{\partial g_L}\right)}{\hat{g}_H^2+g_L^2\frac{\partial \hat{g}_H}{\partial g_L}}.$$

Simplify we get

$$\eta^*=-(\tilde{\lambda}_{LH}+\tilde{\lambda}_{HL})\frac{\frac{\tilde{\lambda}_{HL}g_L(g_L-2\tilde{\lambda}_{LH})}{(g_L-\tilde{\lambda}_{LH})^2}}{\frac{\tilde{\lambda}_{HL}g_L^2(\tilde{\lambda}_{HL}-\tilde{\lambda}_{LH})}{(g_L-\tilde{\lambda}_{LH})^2}}=-(\tilde{\lambda}_{LH}+\tilde{\lambda}_{HL})\frac{g_L-2\tilde{\lambda}_{LH}}{g_L(\tilde{\lambda}_{HL}-\tilde{\lambda}_{LH})}.$$

Now $\eta_1>\eta^*$ is equivalent to $\hat{f}(\eta^*,\bar{g}_L^*)>0$. Plug $\eta=\eta^*$ into $\hat{f}(\eta,\bar{g}_L^*)$, and simplify we get, at $g_L=\bar{g}_L^*$,

$$\begin{aligned}\hat{f}(\eta^*,g_L)=&\frac{g_L^2[(\tilde{\lambda}_{LH}+\tilde{\lambda}_{HL})((1+M_H)\tilde{\lambda}_{LH}\tilde{\lambda}_{HL}+M_H\tilde{\lambda}_{HL}^2+\bar{\lambda}M_H^2\tilde{\lambda}_{LH}-M_LM_H\tilde{\lambda}_{HL}\bar{\lambda})]}{g_L^2(M_L\tilde{\lambda}_{HL}-M_H\tilde{\lambda}_{LH})^2\tilde{\lambda}_{HL}}\\&-\frac{g_L[-2\tilde{\lambda}_{LH}\tilde{\lambda}_{HL}M_H\bar{\lambda}+6\tilde{\lambda}_{LH}^2\tilde{\lambda}_{HL}+2\tilde{\lambda}_{LH}\tilde{\lambda}_{HL}^2+2M_H\tilde{\lambda}_{LH}^2\bar{\lambda}](\tilde{\lambda}_{LH}+\tilde{\lambda}_{HL})}{g_L^2(M_L\tilde{\lambda}_{HL}-M_H\tilde{\lambda}_{LH})^2\tilde{\lambda}_{HL}}\\&+\frac{\tilde{\lambda}_{HL}(\tilde{\lambda}_{LH}+\tilde{\lambda}_{HL})^2 4\tilde{\lambda}_{LH}^2+\bar{\lambda}\tilde{\lambda}_{LH}(\tilde{\lambda}_{LH}+\tilde{\lambda}_{HL})(M_L\tilde{\lambda}_{HL}-M_H\tilde{\lambda}_{LH})^2}{g_L^2(M_L\tilde{\lambda}_{HL}-M_H\tilde{\lambda}_{LH})^2\tilde{\lambda}_{HL}}.\end{aligned}$$

Notice that

$$M_H\tilde{\lambda}_{HL}^2-M_LM_H\tilde{\lambda}_{HL}\bar{\lambda}=M_H\tilde{\lambda}_{HL}(\tilde{\lambda}_{HL}-M_L\bar{\lambda})>0$$

and

$$-2\tilde{\lambda}_{LH}\tilde{\lambda}_{HL}M_H\bar{\lambda}+2\tilde{\lambda}_{LH}^2\tilde{\lambda}_{HL}=2\tilde{\lambda}_{LH}\tilde{\lambda}_{HL}(\tilde{\lambda}_{LH}-M_H\bar{\lambda})>0$$

Since every term is strictly positive (recall $\bar{g}_L^*<0$), we have $\hat{f}(\eta^*,\bar{g}_L^*)>0$. Therefore, $\eta_1>\eta^*$, which implies $\frac{\partial \hat{f}}{\partial g_L}<0$ and thus, $\frac{\partial \hat{\eta}_1}{\partial \bar{g}_L^*}<0$.

Now, we have $\theta=-\eta_1$ and $\bar{g}_H^*=\frac{\tilde{\lambda}_{HL}}{1-\frac{\tilde{\lambda}_{LH}}{\bar{g}_L^*}}$. Therefore

$$\theta=\hat{\theta}_a(\bar{g}_H^*;\lambda_{HL},\lambda_{LH},\bar{\lambda})=-\hat{\eta}_1\left(\frac{\tilde{\lambda}_{LH}}{1-\frac{\tilde{\lambda}_{HL}}{\bar{g}_H^*}};\lambda_{HL},\lambda_{LH},\bar{\lambda}\right).$$

So

$$\frac{\partial \hat{\theta}_a}{\partial \bar{g}_H^*}=-\frac{\partial \hat{\eta}_1}{\partial \bar{g}_L^*}\cdot\left(-\frac{\tilde{\lambda}_{LH}}{\left(1-\frac{\tilde{\lambda}_{HL}}{\bar{g}_H^*}\right)^2}\frac{\tilde{\lambda}_{HL}}{\left(\bar{g}_H^*\right)^2}\right)<0.\qquad\square$$

Proof of Theorem 4 (Part 2). Setting $\tau_H=\tau_L=n=0$ in (39), we arrive at:

$$\bar{g}_L^*=\bar{r}-g_x-\frac{(1-KY(g_x+\delta))(\bar{r}-g_x)}{(\bar{r}-g_x)KY+EY}.\tag{43}$$

As shown in the proof of Part 1, because $n=0$, $\bar{g}_H^*=\frac{\tilde{\lambda}_{HL}}{1-\frac{\tilde{\lambda}_{LH}}{\bar{g}_L^*}}$. Plugging this expression for g_H^* into the equations that determines the tail index, (28) with $n=0$, we obtain $\hat{f}(\eta_1^*,\bar{g}_L^*)=0$, where $\hat{f}$ is given by (42). Since $\bar{g}_L^*$ is a function of $KY,EY,\bar{r},g_x,\delta$, the tail index is a function of these variables in addition to $\lambda_{HL},\lambda_{LH},\bar{\lambda}$ that determines $\hat{f}$.

From expressions for $\bar{g}_L^*$ in (43), we have

$$\frac{\partial \bar{g}_L^*}{\partial KY}=(\bar{r}-g_x)\frac{(g_x+\delta)EY+\bar{r}-g_x}{((\bar{r}-g_x)KY+EY)^2},$$

since $\bar{r} - g_x = \frac{\bar{e}}{\bar{q}} > 0$, it is immediate that $\frac{\partial \bar{g}_L^*}{\partial KY} > 0$.

Similarly,

$$\frac{\partial \bar{g}_L^*}{\partial EY} = (\bar{r} - g_x)\frac{(1 - KY(g_x + \delta))}{(\bar{r}KY + EY)^2}.$$

From the equation for X^*, (40), since $X^* > 0$ and $\bar{r} - g_x > 0$, $1 - KY(g_x + \delta) > 1 - KY(\bar{r} + \delta) - EY > 0$. Therefore, $\frac{\partial \bar{g}_L^*}{\partial EY} > 0$.

Lastly, we rewrite $\bar{g}_L^*$ in (43) as

$$\bar{g}_L^* = \bar{r} + \delta - \frac{1 - EY}{KY} + \frac{\frac{EY}{KY}(1 - (\bar{r} + \delta)KY - EY)}{(\bar{r} - g_x)KY + EY}.$$

So

$$\frac{\partial \bar{g}_L^*}{\partial g_x} = \frac{EY(1 - (\bar{r} + \delta)KY - EY)}{((\bar{r} - g_x)KY + EY)^2} > 0$$

since $(1 - (\bar{r} + \delta)KY - EY) > 0$, from $X^* > 0$ as argued above.

The last three results, combined with Part 1, give us the desired comparative statics stated in this Part 2 of the theorem. □

Appendix B. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.red.2017.10.001.

References

Acemoglu, D., 2009. Introduction to Modern Economic Growth. Princeton University Press.

Acemoglu, D., Cao, D., 2015. Innovation by entrants and incumbents. Journal of Economic Theory 157, 255–294.

Achdou, Y., Han, J., Lasry, J.-M., Lions, P.-L., Moll, B., 2015. Heterogeneous agent models in continuous time. Mimeo.

Aiyagari, S.R., 1994. Uninsured idiosyncratic risk and aggregate saving. The Quarterly Journal of Economics 109 (3), 659–684.

Alvaredo, F., Saez, E., 2009. Income and wealth concentration in Spain from a historical and fiscal perspective. Journal of the European Economic Association 7 (5), 1140–1167.

Angeletos, G.-M., 2007. Uninsured idiosyncratic investment risk and aggregate saving. Review of Economic Dynamics 10 (1), 1–30.

Angeletos, G.-M., Calvet, L.-E., 2005. Incomplete-market dynamics in a neoclassical production economy. Journal of Mathematical Economics 41 (4), 407–438. Special Issue on the Conferences at Davis, Tokyo and Venice.

Angeletos, G.-M., Calvet, L.-E., 2006. Idiosyncratic production risk, growth and the business cycle. Journal of Monetary Economics 53 (6), 1095–1115.

Bach, L., Calvet, L., Sodini, P., 2017. Rich pickings? risk, return, and skill in the portfolios of the wealthy. Mimeo.

Barro, R., Sala-i Martin, X., 2004. Economic Growth, 2nd ed. MIT Press.

Benhabib, J., Bisin, A., 2016. Skewed Wealth Distributions: Theory and Empirics. Working Paper 21924. National Bureau of Economic Research.

Benhabib, J., Bisin, A., Luo, M., 2015. Wealth Distribution and Social Mobility in the US: A Quantitative Approach. Working Paper 21721. National Bureau of Economic Research.

Benhabib, J., Bisin, A., Zhu, S., 2011. The distribution of wealth and fiscal policy in economies with finitely lived agents. Econometrica 79 (1), 123–157.

Benhabib, J., Bisin, A., Zhu, S., 2015. The wealth distribution in Bewley economies with capital income risk. J. Econ. Theory, Part A 159, 489–515.

Benhabib, J., Bisin, A., Zhu, S., 2016. The distribution of wealth in the Blanchard–Yaari model. Macroeconomic Dynamics 20, 466–481.

Buera, F.J., 2009. A dynamic model of entrepreneurship with borrowing constraints: theory and evidence. Annals of Finance 5 (3), 443–464.

Buera, F.J., Shin, Y., 2011. Self-insurance vs. self-financing: a welfare analysis of the persistence of shocks. Journal of Economic Theory 146 (3), 845–862. Incompleteness and Uncertainty in Economics.

Cagett, M., De Nardi, M., 2006. Entrepreneurship, frictions, and wealth. Journal of Political Economy 114 (5), 835–870.

Calvet, L.E., Campbell, J.Y., Sodini, P., 2007. Down or out: assessing the welfare costs of household investment mistakes. Journal of Political Economy 115 (5), 707–747.

Cass, D., 1965. Optimum growth in an aggregative model of capital accumulation. The Review of Economic Studies 32 (3), 233–240.

Castaneda, A., Diaz-Gimenez, J., Rios-Rull, J.-V., 2003. Accounting for the U.S. earnings and wealth inequality. Journal of Political Economy 111 (4), 818–857.

Evans, D.S., Jovanovic, B., 1989. An estimated model of entrepreneurial choice under liquidity constraints. Journal of Political Economy 97 (4), 808–827.

Fagereng, A., Guiso, L., Malacrino, D., Pistaferri, L., 2016. Heterogeneity and Persistence in Returns to Wealth. Working Paper 22822. National Bureau of Economic Research.

Gabaix, X., 2009. Power laws in economics and finance. Annual Review of Economics 1 (1), 255–294.

Gabaix, X., Lasry, J.-M., Lions, P.-L., Moll, B., 2016. The dynamics of inequality. Econometrica 84 (6), 2071–2111.

Hubmer, J., Krusell, P., Smith, A., 2016. The Historical Evolution of the Wealth Distribution: A Quantitative-Theoretic Investigation. Working Paper 23011. National Bureau of Economic Research.

Huggett, M., 1993. The risk-free rate in heterogeneous-agent incomplete-insurance economies. Journal of Economic Dynamics and Control 17 (5–6), 953–969.

Huggett, M., 1996. Wealth distribution in life-cycle economies. Journal of Monetary Economics 38 (3), 469–494.

Huggett, M., Ventura, G., Yaron, A., 2011. Sources of lifetime inequality. The American Economic Review 101 (7), 2923–2954.

Jones, C.I., 2015. Pareto and Piketty: the macroeconomics of top income and wealth inequality. The Journal of Economic Perspectives 29 (1), 29–46.

Kaymak, B., Poschke, M., 2016. The evolution of wealth inequality over half a century: the role of taxes, transfers and technology. Journal of Monetary Economics 77, 1–25.

King, R.G., Rebelo, S.T., 1999. Resuscitating real business cycles, Chapter 14. In: Woodford, J.B.T.a.M. (Ed.), Handbook of Macroeconomics, Volume 1, Part B. Elsevier, pp. 927–1007.

Krusell, P., Smith, A., 1998. Income and wealth heterogeneity in the macroeconomy. Chapter 14. Journal of Political Economy 106 (5), 867–896.

Krusell, P., Smith, A.A., 2015. Is Piketty's second law of capitalism fundamental? Journal of Political Economy 123 (4), 725–748.

Mankiw, N.G., 2015. Yes, r > g. so what? The American Economic Review 105 (5), 43–47.

McGrattan, E.R., Prescott, E.C., 2005. Taxes, regulations, and the value of U.S. and U.K. corporations. The Review of Economic Studies 72 (3), 767.
Moll, B., 2012. Inequality and Financial Development: A Power-Law Kuznets Curve. Working Paper. Princeton University.
Moll, B., 2014. Productivity losses from financial frictions: can self-financing undo capital misallocation? The American Economic Review 104 (10), 3186–3221.
Nirei, M., Aoki, S., 2016. Pareto distribution of income in neoclassical growth models. Review of Economic Dynamics 20, 25–42.
Piketty, T., 2014. Capital in the 21st Century. Harvard University Press.
Piketty, T., Saez, E., 2003. Income inequality in the United States, 1913–1998. The Quarterly Journal of Economics 118 (1), 1–41.
Piketty, T., Zucman, G., 2014. Capital is back: wealth-income ratios in rich countries, 1700–2010. The Quarterly Journal of Economics.
Piketty, T., Zucman, G., 2015. Wealth and inheritance in the long run, Chapter 15. In: Atkinson, A.B., Bourguignon, F. (Eds.), Handbook of Income Distribution, vol. 2. Elsevier, pp. 1303–1368.
Quadrini, V., 2000. Entrepreneurship, saving, and social mobility. Review of Economic Dynamics 3 (1), 1–40.
Ray, D., 2015. Nit-Piketty: a comment on Thomas Piketty s capital in the twenty first century. CESifo Forum 16 (1), 19–25.
Saez, E., Zucman, G., 2016. Wealth inequality in the united states since 1913: evidence from capitalized income tax data. The Quarterly Journal of Economics 131 (2), 517–578.
Stiglitz, J.E., 1969. Distribution of income and wealth among individuals. Econometrica 37 (3), 382–397.
Toda, A.A., 2014. Incomplete market dynamics and cross-sectional distributions. Journal of Economic Theory 154, 310–348.
Zetlin-Jones, A., Shouride, A., 2016. External Financing and the Role of Financial Frictions over the Business Cycle: Measurement and Theory. CMU Working Paper.