Pareto Extrapolation: Bridging Theoretical and Quantitative Models of Wealth Inequality[*]

Émilien Gouin-Bonenfant[†] and Alexis Akira Toda[‡]

Department of Economics, University of California San Diego

March 20, 2019

Abstract

We propose a new, systematic approach for analyzing and solving heterogeneous-agent models with fat-tailed wealth distributions. Our approach exploits the asymptotic linearity of policy functions and the analytical characterization of the Pareto exponent to make the solution algorithm more transparent, efficient, and accurate with zero additional computational cost. As an application, we solve a heterogeneous-agent model that features persistent earnings and investment risk, borrowing constraint, portfolio decision, and endogenous Pareto-tailed wealth distribution. We find that a wealth tax is a "lose-lose" policy: the introduction of a 2% wealth tax (with extra tax revenue used as consumption rebate) decreases wage by 12%, welfare (in consumption equivalent) by 14.4%, and total tax revenue by 1.1%.

Keywords: asymptotic linearity, Bewley-Huggett-Aiyagari model, Pareto exponent, power law, solution accuracy, wealth tax.

JEL codes: C63, D31, D58, E21.

1 Introduction

Macroeconomic models increasingly incorporate heterogeneity. Doing so allows researchers to identify who gains and who loses from new policies, but also to assess how the effectiveness of policies depend on the nature of heterogeneity. Since the first generation of heterogeneous-agent models such as Huggett (1993) and Aiyagari (1994), one challenge has been to generate a realistic wealth distribution. Empirically, it is well known since Pareto (1895, 1896, 1897)'s seminal work that the wealth distributions obey the power law: the fraction of agents with wealth

[*] We thank Mark Bils, Fedor Iskhahov, Narayana Kocherlakota, Alisdair McKay, Ben Moll, John Stachurski, and seminar participants at Australian National University, Rochester, and University of New South Wales for comments.

[†] Email: egouinbo@ucsd.edu.

[‡] Email: atoda@ucsd.edu.

w or larger decays like a power function $w^{-\zeta}$, where ζ is called the Pareto exponent. Thus successful economic models of wealth inequality should endogenously generate fat-tailed wealth distributions.

We are now much closer to understanding the economic forces that determine wealth inequality, and many models have been proposed that can account for the extreme concentration of wealth we observe in the data. Two parallel literatures have emerged. The first studies relatively simple models and provides theoretical characterizations of the power law behavior of the wealth distribution, often relying on analytical solutions. The second studies rich general equilibrium models and conducts quantitative analysis and experiments that rely heavily on numerical methods. However, there has always been the trade-off between analytical tractability and the richness of models. The former requires strong (and oftentimes unrealistic) assumptions. The latter relies on numerical methods, which are in general not well-suited for studying the tail behavior of the wealth distribution because models are commonly solved on a finite grid and hence misses the top tail *by definition*. What is lacking in the current literature is a systematic approach for analyzing and solving heterogeneous-agent models that (potentially) generate fat-tailed wealth distributions but do not admit closed-form solutions.

In this paper, we propose a simple, systematic approach for tackling heterogeneous-agent models with fat-tailed wealth distributions numerically. Our approach builds on the conventional solution algorithm but extends it with two additional steps: (i) the "asymptotic analysis" of the individual optimization problem to compute the Pareto exponent of the wealth distribution and (ii) the "Pareto extrapolation" of the wealth distribution off the grid to compute the equilibrium and wealth distribution accurately. Our approach enables researchers to easily and accurately analyze rich heterogeneous-agent models that feature persistent earnings and investment risk, borrowing constraint, portfolio choices, recursive utility, and endogenous Pareto wealth distributions, etc.

In the "asymptotic analysis" step we solve, given the candidate equilibrium object (risk-free rate, wage, etc.), a simplified, or "asymptotic" individual optimization problem semi-analytically. Roughly speaking, this problem ignores all additive elements and focuses on proportional elements. For example, consider the income fluctuation problem, which is a building block of Bewley-Huggett-Aiyagari models. The asymptotic problem in this case is one with no income (i.e., consumption is financed only through savings), which can be solved analytically as in Merton (1969) and Samuelson (1969). The benefit of studying the asymptotic problem is threefold. First, its solution determines the behavior of rich agents, which governs the tail property of the wealth distribution. This enables researchers to determine whether the model generates a fat-tailed wealth distribution, and if so, to compute the theoretical Pareto exponent explicitly. Second, the analysis of the asymptotic problem places parametric restrictions on the equilibrium object through equilibrium considerations such as the existence of a solution to the individual optimization problem and a wealth distribution with a finite mean. This enables researchers to narrow down the set of equilibrium object and search for the equilibrium more efficiently. Third, the solution to the asymptotic problem can be used as an initial guess for solving the actual individual optimization problem, making the algorithm more efficient and stable.

In the "Pareto extrapolation" step, we extrapolate the wealth distribution off the grid using the theoretical Pareto exponent computed from the asymptotic analysis to correct for the truncation error when constructing the transition probability matrix governing the state variables and computing aggregate quantities from the actual optimization problem. The benefit of Pareto extrapolation is twofold. First, it makes the solution more accurate at no additional computational cost. This is because the correction terms take care of the truncation error, and these terms are introduced only at the largest grid point, which is negligible compared with the total number of grid points. Second, and more importantly, Pareto extrapolation enables researchers to avoid making mistakes. While it is true that we can solve models to any accuracy if we use sufficiently large and fine grids and sufficiently strong computing power, with existing methods one can never be sure whether the truncation error is small enough. As an illustration, suppose some researcher says "I truncate the grid so that there is less than (say) 10^{-4} of the probability mass at the top grid point". This is not a good idea because (i) the mass at the largest grid point is severely biased downwards with existing methods, and (ii) even with mass 10^{-4} at the largest grid point, there can be substantial amount of wealth held by those agents.

In summary, our approach makes the solution and analysis of heterogeneous-agent models with fat-tailed wealth distributions (i) more transparent (because it exploits economic theory as much as possible), (ii) more efficient (because it narrows down the equilibrium object and uses good initial guesses), and (iii) more accurate (because it corrects for truncation errors). Furthermore, we achieve all of that with zero additional computational cost because the asymptotic analysis is semi-analytical and the correction terms in the Pareto extrapolation are introduced only at one grid point.

To illustrate the usefulness of our approach, we present two exercises. First, using a simple heterogeneous-agent model that admits a closed-form solution (and a Pareto wealth distribution) as a laboratory, we show that the error with existing methods can be substantial, while it is minimal with our approach. Specifically, we find that the error with our method (with an exponentially-spaced grid with 100 points) ranges from 0.002 to 0.6% depending on the choice of the largest grid point, whereas it is 3.3–21% with the usual ("Truncation") method that does not introduce correction terms. Simulation-based methods with 10,000 agents also have about 11% of error. The errors in the existing methods are especially severe when the Pareto exponent is smaller than 2, which is typical for wealth (1.5) and firm size (close to 1, Zipf's law).

Second, we develop a Merton-Bewley-Aiyagari (MBA) model that features persistent idiosyncratic endowment and investment risk, borrowing constraint, portfolio decision, recursive utility, and endogenous Pareto-tailed wealth distribution. To the best of our knowledge, our paper is the first to solve such a complicated model without relying on (restrictive) closed-form solutions, and we obtain tractability through the asymptotic analysis of the individual optimization problem and the Pareto exponent formula. We provide a step-by-step approach to solving the model and use it as a laboratory to conduct a counterfactual experiment. We find that a wealth tax is a "lose-lose" policy: the introduction of a 2% wealth tax (with extra tax revenue used as consumption rebate) decreases wage and output by 12%, welfare (in consumption equivalent) by 14.4%, and total tax revenue by 1.1%.

1.1 Related literature

Our paper is related to a large literature that spans across many disciplines, including quantitative macroeconomics, economic theory on consumption and portfolio choices, mathematical and statistical results on Pareto tails, and numerical analysis.

It is well-known in the quantitative macroeconomics literature that idiosyncratic unemployment risk and incomplete financial markets alone are insufficient to generate a sufficiently dispersed wealth distribution (Krueger, Mitman, and Perri, 2016). Recently, Stachurski and Toda (2018) have theoretically proved that in canonical Bewley-Huggett-Aiyagari models in which agents are infinitely-lived, have constant discount factors, and can invest only in a risk-free asset, the wealth distribution necessarily inherits the tail property of the income distribution. Therefore canonical heterogeneous-agent models cannot explain the wealth distribution. They also argue that introducing other ingredients such as random discount factors (Krusell and Smith, 1998), idiosyncratic investment risk (Quadrini, 2000; Cagetti and De Nardi, 2006), and random birth/death (Carroll, Slacalek, Tokuoka, and White, 2017; McKay, 2017) can generate fat tails. However, because these papers are all numerical, it is not clear how to build and solve general heterogeneous-agent models that feature fat-tailed wealth distributions. Our paper contributes to the quantitative macroeconomics literature by showing the usefulness of the theoretical analysis of the asymptotic problem and providing a general solution algorithm for such models.

As mentioned in the introduction, since numerical methods are in general not well-suited for studying the tail behavior of the wealth distribution, most papers that study the power law behavior in the wealth distribution use analytical solutions. Nirei and Souma (2007) and Benhabib, Bisin, and Zhu (2011) solve growth models with idiosyncratic investment risk and use the properties of Kesten (1973) processes to obtain a Pareto wealth distribution. Moll (2014), Toda (2014), Arkolakis (2016), Benhabib, Bisin, and Zhu (2016), and Nirei and Aoki (2016) consider stochastic birth/death and obtain the double Pareto wealth distribution based on the mechanism of Reed (2001).[1] For reviews of generative mechanisms of Pareto tails used in these papers, see Gabaix (2009). Our paper bridges this literature on power law in economics and quantitative macroeconomics by showing that the theoretical insight carries over to rich quantitative models.

Toda (2018b) pointed out the usefulness of the asymptotic problem for computing the Pareto exponent in general models that admit no closed-form solutions.[2] However, he does not consider the solution algorithm for general equilibrium models with fat-tailed wealth distributions. The asymptotic linearity of consumption policies has been known for a long time since at least Huggett (1993) and Krusell and Smith (1998), among others. Benhabib, Bisin,

[1] Other recent applications include firm dynamics (Acemoglu and Cao, 2015), asset pricing (Toda and Walsh, 2015, 2017), dynamics of inequality (Gabaix, Lasry, Lions, and Moll, 2016; Aoki and Nirei, 2017; Cao and Luo, 2017; Kasa and Lei, 2018), bequests (Zhu, 2018), and entrepreneurship (Jones and Kim, 2018).

[2] The asymptotic problem is related to the "method of moderation" in Carroll, Tokuoka, and Wu (2012), who bound the consumption policy function from above and below by closed-form solutions to improve accuracy and stability.

and Zhu (2015) show the asymptotic linearity when earnings and investment risk are mutually independent and jointly i.i.d. over time and obtain a Pareto lower bound for the wealth distribution. In Appendix A, we argue that similar results should hold in richer models. To analytically characterize the Pareto exponent of the wealth distribution in a general Markovian environment, we apply the recent results from Beare and Toda (2017).

Our paper is also related to the literature on solution methods for heterogeneous-agent models such as Krusell and Smith (2006), Algan, Allais, and Den Haan (2008), Reiter (2009, 2010), Den Haan (2010a,b), and Algan, Allais, Den Haan, and Rendahl (2014), among others. In particular, we use the insight from Algan, Allais, and Den Haan (2008) and Winberry (2018), who approximate cross-sectional distributions using finite-dimensional parametric families. In our case, because economic theory suggests that the upper tail of the wealth distribution is Pareto and it is possible to compute the Pareto exponent from the solution to the asymptotic problem, we use this Pareto distribution to approximate the upper tail. Although we do not take a stance on how to deal with the rest of the distribution, we use Young (2010)'s non-stochastic simulation to compute the wealth distribution from the transition probability matrix implied by the law of motion.

The closest paper to ours in spirit is Achdou, Han, Lasry, Lions, and Moll (2017). They recast the Bewley-Huggett-Aiyagari model in continuous-time, which allows them to obtain a number of novel characterizations and results, including closed-form expressions for the stationary wealth distribution (in a special case) and the marginal propensity to consume of agents close to the borrowing constraint. They also prove that a stationary equilibrium exists and is unique when the intertemporal elasticity of substitution is weakly above one. Finally, they leverage finite-difference methods and propose a fast solution algorithm that can be applied to much more general heterogeneous-agent models in continuous time. While our paper is different—we focus on the complications arising with fat-tailed wealth distributions—we share the same goal of bridging the gap between theoretical and quantitative work in macroeconomics.

2 Solving heterogeneous-agent models with Pareto tails

In this section we propose a new solution algorithm for heterogeneous-agent models with fat-tailed wealth distributions based on Pareto extrapolation. We first point out the issues with existing solution algorithms, and then outline our new method.

2.1 Issues with existing algorithms

Suppose that we want to solve a Bewley (1977, 1983)-Huggett (1993)-Aiyagari (1994) model numerically when the wealth distribution could be fat-tailed. The conventional solution algorithm (which we refer to as the "Truncation" method throughout the paper) would be roughly as follows.

1. The researcher sets up a finite grid for wealth denoted by $\mathcal{W}_N = \{w_n\}_{n=1}^N$, where N is the number of grid points and $w_1 < \cdots < w_N$. Suppose there are also other exogenous

state variables (e.g., income, return on wealth, etc.), which can take S possible values indexed by $s = 1, \dots, S$. Given the guess of the equilibrium object (e.g., interest rate, wage, etc.), we can solve the individual optimization problem on the $S \times N$ grid using dynamic programming.

2. Having solved the individual optimization problem and obtained the law of motion for individual wealth, the researcher constructs the $SN \times SN$ transition probability matrix P of all state variables. The stationary distribution $\pi \in \mathbb{R}_+^{SN}$ is obtained by solving $P'\pi = \pi$ (so π is an eigenvector of P' corresponding to the eigenvalue 1).

3. Finally, the researcher imposes the market clearing condition by integrating the individual decision rules (capital, labor, etc.) over the grid using the stationary distribution π to find the equilibrium objects (interest rate, wage, etc.).

There are two potential issues with this truncation method when the stationary wealth distribution is fat-tailed, both of which are related. First, consider the largest grid point w_N. This grid point in principle does not represent just the point $w = w_N$, but the entire interval $w \in [w_N, \infty)$. Therefore when we construct the transition probability from w_N to other grid points, instead of assuming that the current wealth state w is concentrated at w_N, we need to take into account that w is really distributed over the interval $[w_N, \infty)$ according to the (true) stationary distribution. Since the interval $[w_N, \infty)$ contains substantial probability mass when the wealth distribution is fat-tailed, failing to account for this will overestimate the transition probability to lower wealth states, and hence underestimate the top tail probability.

Second, suppose that we use the stationary distribution $\pi = (\pi_{sn})$ to compute aggregate quantities used in market clearing conditions. For concreteness, consider the aggregate wealth

$$W = \sum_{s=1}^{S} \sum_{n=1}^{N} \pi_{sn} w_n. \tag{2.1}$$

The right-hand side of (2.1) essentially supposes that the top tail is concentrated on the grid point w_N, whereas in fact it is distributed over the interval $[w_N, \infty)$. Thus failing to account for this will underestimate the aggregate wealth, which affects the computation of equilibrium through market clearing conditions.

Of course, one may choose a very large truncation point w_N (say, one million times the aggregate wealth) to reduce the truncation error, but that is computationally inefficient because it will either increase the number of grid points (making the solution algorithm slower) or decrease the grid density (making the solution less accurate). One may also argue that the above two issues are specific to the particular algorithm that involves truncation, and other methods such as simulation (Aiyagari, 1994; Krusell and Smith, 1998) may not be subject to those issues. As we see below, however, the situation is equally problematic. Simulation-based methods essentially use the law of large numbers to approximate the market clearing condition. Suppose we simulate I agents and compute the sample mean of wealth $\frac{1}{I}\sum_{i=1}^{I} w_i$. The question is how fast the sample mean converges to the population mean. If the Pareto exponent ζ exceeds 2, then wealth has finite variance and we can apply the Central Limit Theorem. In this case the sample mean converges at rate $I^{1/2}$. If $\zeta < 2$ on the other hand, it is well-known that the rate of convergence to the stable law is only $I^{1-1/\zeta}$.[3] Therefore solving a model accurately may require an impractically large number of agents.

As an illustration, Table 1 shows the order of error $I^{\max\{-1/2,1/\zeta-1\}}$ in the sample mean for various sample size I and Pareto exponent ζ.[4] If $\zeta \geq 2$ and we use 10,000 agents (the number used in Aiyagari (1994)), then the order of the error in the sample mean is $10000^{-1/2} = 1/100 = 1\%$. However, the error order is much larger if the Pareto exponent is smaller. With $\zeta = 1.5$ (a typical number for the wealth distribution according to Pareto (1897), Klass, Biham, Levy, Malcai, and Solomon (2006), and Vermeulen (2018)), the error order with 10,000 agents is 4.6%, which is substantial. If the Pareto exponent is 1.1 (a typical number for the firm size distribution, which obeys Zipf's law (Axtell, 2001)), then even with ten billion agents ($I = 10^{10}$), which is about the same order of magnitude as the world population, the error order is still 12.3%. To drive the error down to 1%, quite a modest number, the required sample size for $\zeta = 1.1$ is $I = 100^{\frac{\zeta}{\zeta-1}} = 10^{22}$ (ten sextillion), which is about the same order of magnitude as the number of stars in the universe or sand grains on earth.[5] Therefore we cannot expect to solve such models accurately using simulation.

Table 1: Order of error $I^{\max\{-1/2,1/\zeta-1\}}$ in sample mean.

Sample size I	Pareto exponent ζ			
	≥ 2	1.5	1.3	1.1
$10^0 = 1$	1.00000	1.00000	1.00000	1.00000
10^2	0.10000	0.21544	0.34551	0.65793
10^4	0.01000	0.04642	0.11938	0.43288
10^6	0.00100	0.01000	0.04125	0.28480
10^8	0.00010	0.00215	0.01425	0.18738
10^{10}	0.00001	0.00046	0.00492	0.12328

2.2 The Pareto extrapolation algorithm

Our new solution algorithm for heterogeneous-agent models, which we call the "Pareto extrapolation" method, differs from the usual "truncation" method only when computing the stationary distribution and aggregating individual behavior. Therefore we focus on the description of the algorithm only at this aggregation step.

The Pareto extrapolation method consists of three main sub-steps that correct the truncation errors in the standard algorithm:

The Pareto extrapolation algorithm.

1. Solve an "asymptotic" individual optimization problem semi-analytically and compute the theoretical Pareto exponent ζ.

[3] See, for example, Durrett (2010, Theorem 3.7.2) for an accessible proof. Based on this insight, Gabaix (2011) argues that a substantial fraction of aggregate fluctuations is due to idiosyncratic shocks to large firms.

[4] In Table 16 in Appendix C.3, we assess the accuracy of the Aiyagari model in Section 3 using simulation and obtain similar results to Table 1.

[5] http://www.abc.net.au/science/articles/2015/08/19/4293562.htm

2. Construct the $SN \times SN$ transition probability matrix by approximating the wealth distribution for $w \geq w_N$ by a Pareto distribution with exponent ζ.

3. Compute the aggregate wealth by approximating the wealth distribution for $w \geq w_N$ by a Pareto distribution with exponent ζ.

Below, we explain each step in more detail.

2.2.1 Computing the theoretical Pareto exponent

Our method uses the theoretical Pareto exponent ζ to correct for the truncation errors. Thus the first step is to compute the theoretical Pareto exponent of the wealth distribution implied by individual behavior.

For this purpose we can use the insight from Toda (2018b). Since the tail property of the wealth distribution depends on the behavior of wealthy agents, and for those agents labor income is negligible compared to capital income (because labor income enters additively to the budget constraint, whereas capital income is proportional to wealth), we can consider a simplified problem where the labor income is zero. Assuming that agents have homothetic preferences (e.g., additive CRRA, Epstein-Zin, etc.), which is almost always the case in applications, this simplified problem becomes a homogeneous problem in the sense that all control variables scale with wealth. We refer to this problem as the "asymptotic" problem. Such problems can be solved semi-analytically even in a Markovian (non-i.i.d.) environment as shown by Toda (2014, Theorem 5), and the decision rules become linear in wealth. (Appendix A formally defines the asymptotic problem and discusses the asymptotic linearity of policy functions in an abstract dynamic programming setting.)

For concreteness, suppose that agents solve an optimal consumption-savings problem of the form

$$\text{maximize} \quad \mathrm{E}_0 \sum_{t=0}^{\infty} [\beta(1-p)]^t \frac{c_t^{1-\gamma}}{1-\gamma} \tag{2.2a}$$

$$\text{subject to} \quad w_{t+1} = R_{s_t}(w_t - c_t + y_{s_t}), \tag{2.2b}$$

$$w_t \geq \underline{w}. \tag{2.2c}$$

Here $\beta > 0$ is the discount factor, $p \in [0,1)$ is the birth/death probability (infinitely-lived case corresponds to $p = 0$), $\gamma > 0$ is the relative risk aversion, c_t is consumption, w_t is wealth at the beginning of period t excluding current labor income, s_t is some Markov state, y_s is income in state s, $R_s > 0$ is the gross return on wealth in state s, and $\underline{w}$ is minimum wealth. By definition, the asymptotic problem studies the limiting case when $w \to \infty$. In this case, income y and minimum wealth are negligible, so we replace the budget constraint (2.2b) and borrowing constraint (2.2c) by

$$w_{t+1} = R_{s_t}(w_t - c_t), \tag{2.3a}$$

$$w_t \geq 0, \tag{2.3b}$$

respectively. Note that the problem is now homogeneous because the utility function is homothetic: an agent twice as rich will consume twice as much, state-by-state. We can maximize a homothetic function subject to homogeneous constraints of the form (2.3) semi-analytically quite efficiently, as explained in Toda (2014) in detail. After solving this problem, the law of motion for wealth becomes linear, $w' = G_s w$ for some gross growth rate $G_s > 0$.

Now let us go back to the general case and write the law of motion of the asymptotic problem as

$$w_{i,t+1} = G_{i,t+1} w_{it},$$

where $G_{i,t+1} > 0$ is the gross growth rate of wealth between time t and $t + 1$ for individual i and w_{it} is wealth. Thus, in the asymptotic problem, the law of motion for wealth necessarily satisfies Gibrat (1931)'s law of proportional growth. Assuming that agents enter/exit the economy at constant probability $p > 0$, Beare and Toda (2017) show that under mild conditions the stationary wealth distribution has a Pareto upper tail and characterize the Pareto exponent ζ, as follows. Suppose that there are finitely many idiosyncratic states (other than wealth) denoted by $s = 1, \dots, S$ and let $P = (p_{ss'})$ be the transition probability matrix, which we assume to be irreducible. For $z \in \mathbb{R}$, let

$$D(z) = \operatorname{diag}\left(\mathrm{E}\left[\mathrm{e}^{z \log G_{i,t+1}} \,\middle|\, s_{it} = 1\right], \dots, \mathrm{E}\left[\mathrm{e}^{z \log G_{i,t+1}} \,\middle|\, s_{it} = S\right]\right) \tag{2.4}$$

be the diagonal matrix consisting of the conditional moment generating functions of the log growth rate $\log G_{i,t+1}$. For a square matrix A, let $\rho(A)$ denote its spectral radius (the maximum modulus of all eigenvalues of A). Then under mild conditions Beare and Toda (2017) show that the equation

$$\rho(PD(z)) = \frac{1}{1-p} \tag{2.5}$$

has a unique positive solution $z = \zeta > 0$, and that the stationary wealth distribution has a Pareto upper tail with exponent ζ. Toda (2018b) argues that if agents are infinitely lived but there exists a stationary distribution due to other mechanisms than random entry/exit (e.g., borrowing constraint), then we can just set $p = 0$ in (2.5) to compute the theoretical Pareto exponent.

We can summarize this step as follows.

Computing the theoretical Pareto exponent.

1. Verify that the utility function is asymptotically homothetic (e.g., CRRA, HARA, Epstein-Zin) in consumption.

2. Define the asymptotic problem (e.g., no labor income, no borrowing constraint).

3. Solve the asymptotic problem semi-analytically and derive the law of motion $w' = G_s w$.

4. Define the diagonal matrix (2.4) of conditional moment generating functions of log growth rates.

5. The theoretical Pareto exponent $\zeta > 0$ is the solution to (2.5).

See Toda (2018b) for more details about the actual implementation of this step.

2.2.2 Constructing the transition probability matrix

Having characterized the theoretical Pareto exponent $\zeta > 0$, the next step is to construct the transition probability matrix for all state variables.

Let $w' = g(w, s)$ be the law of motion for wealth for the original (non-asymptotic) problem, which can be obtained by numerically solving the individual optimization problem using dynamic programing on the grid $S \times \mathcal{W}_N$, where $\mathcal{W}_N = \{w_n\}_{n=1}^N$ is the grid for wealth. Let

$$I_n = [w_n, w_{n+1}), \quad n = 1, \dots, N-1$$

be the half-open interval with endpoints w_n and w_{n+1}. Let $I_N = [w_N, \infty)$. For $n = 1, \dots, N$, let us construct the transition probability as follows.

Case 1: $n < N$. Take the lower grid point of I_n, which is w_n. If $g(w_n, s) \in I_k$ for some $k < N$, then we can take $\theta \in [0, 1)$ such that

$$g(w_n, s) = (1-\theta)w_k + \theta w_{k+1} \iff \theta = \theta_{nk} = \frac{g(w_n, s) - w_k}{w_{k+1} - w_k}. \tag{2.6}$$

We can then assign probabilities $1 - \theta, \theta$ to the grid points w_k, w_{k+1} (i.e., states k and $k+1$), respectively (Figure 1). If $g(w_n, s) < w_1$ or $g(w_n, s) \geq w_N$, then just assign probability 1 to state 1 or N. (Assigning probabilities to neighboring grid points to match the law of motion this way is essentially the same as what Young (2010) calls "non-stochastic simulation".)

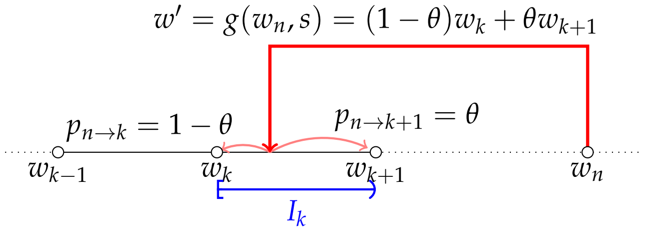

$w' = g(w_n, s) = (1-\theta)w_k + \theta w_{k+1}$
$p_{n \to k} = 1 - \theta$
$p_{n \to k+1} = \theta$
w_{k-1}
w_k
w_{k+1}
w_n
I_k

Figure 1: Construction of transition probabilities from a grid point.

Case 2: $n = N$. Suppose for the moment that there is an untruncated grid $\mathcal{W}_\infty = \{w_n\}_{n=1}^\infty$, and for $n \geq N$ we know the probability of $w = w_n$ conditional on $w \in I_N \cap \mathcal{W}_\infty$. Let this probability be denoted by q_n. By definition, we have $\sum_{n=N}^{\infty} q_n = 1$. Now for each $n \geq N$, we can do precisely as in the previous case, and add probabilities $(1-\theta_{nk})q_n$ and $\theta_{nk}q_n$ (where θ_{nk} is defined by (2.6)) to the grid points w_k, w_{k+1} whenever $w' = g(w_n, s) \in I_k$ for $k < N$ (Figure 2). If $g(w_n, s) < w_1$ or $g(w_n, s) \geq w_N$, then just add probability q_n to the transition to state 1 or N. The nice thing is that for large enough n, the next period's state $w' = g(w_n, s)$ is likely large (contained in I_N), so we only need to compute θ_{nk} for finitely many n (say $n = N, \dots, N'$).

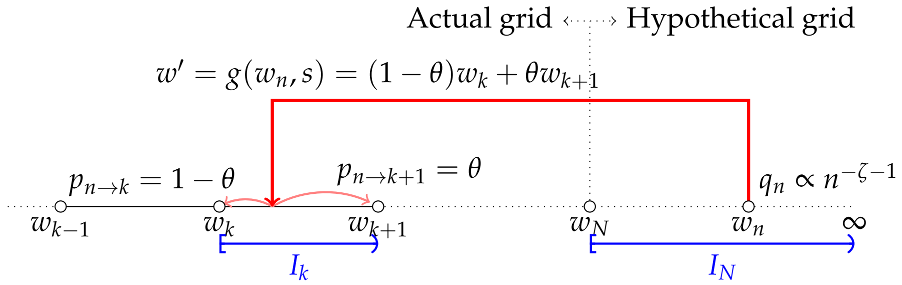


Figure 2: Construction of transition probabilities from a hypothetical grid point.

Now it remains to compute the conditional probability q_n. Assume that for $n \geq N$, the grid spacing $w_{n+1} - w_n$ is some constant $h > 0$. Assuming that the stationary distribution has a Pareto upper tail with exponent $\zeta > 1$, we can set $q_n \propto n^{-\zeta-1}$ for $n \geq N$. (If $\zeta \leq 1$, then the mean is infinite, which is impossible in equilibrium. In this case we exit from the loop and use a different guess for the equilibrium object.) As mentioned before, for $n \geq N'$ the next state will always be N ($w' = g(w_n, s) \in I_N$), so there is no need to compute q_n individually. Approximating the infinite sum by an integral, we can compute

$$\sum_{n=N'}^{\infty} q_n \propto \sum_{n=N'}^{\infty} (nh)^{-\zeta-1} \approx h^{-\zeta-1} \int_{N'}^{\infty} x^{-\zeta-1}\,\mathrm{d}x = h^{-\zeta-1}\frac{1}{\zeta}(N')^{-\zeta}.$$

Imposing the condition $\sum_{n=N}^{\infty} q_n = 1$, we can obtain

$$\begin{cases} q_n = Cn^{-\zeta-1}, & (N \leq n < N') \\ \sum_{n=N'}^{\infty} q_n = \frac{C}{\zeta}(N')^{-\zeta}, \end{cases}$$

where the constant of proportionality C is given by

$$\frac{1}{C} = \frac{1}{\zeta}(N')^{-\zeta} + \sum_{n=N}^{N'-1} n^{-\zeta-1}.$$

We can summarize this step as follows.

Constructing the transition probability matrix.
Let $\zeta > 1$ be the theoretical Pareto exponent, $\mathcal{W}_N = \{w_n\}_{n=1}^{N}$ be the grid, and $\{g(w,s)\}_{s=1}^{S}$ be the law of motion.

1. Choose a spacing parameter $h > 0$.

2. Imagine a hypothetical infinite grid $\mathcal{W}_\infty = \{w_n\}_{n=1}^\infty$, where $w_n = w_N + (n - N)h$ for $n > N$, and $\mathcal{W}_\infty$ agrees with $\mathcal{W}_N$ for $n \le N$.

3. Linearly extrapolate the law of motion $g(w, s)$ for $w > w_N$ (using the slope between the last two grid points, or better yet, the theoretical slope from the solution to the asymptotic problem). Take $N' \ge N$ such that $g(w_{N'}, s) > w_N$ for all s:

   $$N' = \min\left\{n \ge N \,|\, \forall s, g(w_N + (n - N)h, s) > w_N\right\}.$$

   Thus, if $w' = G_s w$ is the asymptotic law of motion, after some algebra we obtain

   $$N' = N + \max_s \left\lceil \frac{w_N - g(w_N, s)}{G_s h} \right\rceil, \tag{2.7}$$

   where $\lceil x \rceil$ denotes the smallest integer exceeding x.

4. For each $(s, n), (s', n') \in \{1, \dots, S\} \times \{1, \dots, N\}$, compute the transition probability from state (s, n) to (s', n') using the non-stochastic simulation described above: use Case 1 for $n < N$ and Case 2 for $n = N$. Collect the probabilities into an $SN \times SN$ transition probability matrix P.

A few remarks are in order. First, the above algorithm has essentially zero additional computational cost, despite its complicated appearance. The reason is that extrapolation from the Pareto distribution is used *only* at the largest grid point w_N. Thus, although we are computing transition probabilities from SN points, which the usual truncation algorithm needs to compute anyway, the Pareto extrapolation algorithm requires only $S \times 1 = S$ additional operations, which is negligible. In our numerical implementation in Section 3, we find that the computing time of this step is trivial, and therefore we do not report it.

Second, the $SN \times SN$ transition probability matrix P is sparse. To see this, let us evaluate the number of nonzero elements of P. For each s, s' and $n < N$, there are at most two states the next wealth can take. For $n = N$, in principle the next wealth state can be anything. Therefore the number of nonzero elements of P is at most

$$2S^2(N - 1) + S^2 N = S^2(3N - 2).$$

Thus the fraction of nonzero elements of P is bounded above by

$$\frac{S^2(3N - 2)}{(SN)^2} = \frac{3N - 2}{N^2} \to 0$$

as $N \to \infty$, so P is sparse. Achdou, Han, Lasry, Lions, and Moll (2017) mention that "[c]ontinuous time imparts a number of computational advantages relative to discrete time [. . . , which] relate to [. . .] the fact that continuous-time problems with discretized state space are, by construction, very sparse." While it is true that continuous-time problems have some advantages over discrete-time problems (e.g., partial differential equations versus nonlinear difference equations), discrete-time problems also do possess sparsity if appropriately solved.

Third, although we have implicitly assumed that N' in (2.7) is larger than N, for particular models it may be $N' \leq N$, which is true if and only if $g(w_N, s) \leq w_N$ for all s. In that case we do not need to consider any extrapolation since the true distribution is not fat-tailed, and the algorithm becomes identical to the truncation method.

Finally, the Pareto extrapolation method requires the spacing parameter $h > 0$. Since the Pareto extrapolation algorithm uses a hypothetical evenly-spaced grid (with grid spacing h) beyond the largest grid point w_N, the most natural choice for h is $w_N - w_{N-1}$, the distance between the two largest actual grid points. Conducting numerical experiments similar to those in Section 3 and Appendix C, we have found that this choice is numerically optimal.

2.2.3 Computing the aggregate wealth

When computing the equilibrium, we need to impose the market clearing condition in some way or another. In Aiyagari models the relevant market clearing condition is for capital, and the demand side is often trivial. To compute the supply of capital (wealth), we can do as follows. First let P be the $SN \times SN$ transition probability matrix computed above. Let $\pi = (\pi_{sn})$ be its stationary distribution, where π_{sn} is the probability of being in state (s, n). Note that since π is the (unique) eigenvector of P' corresponding to the eigenvalue 1 and the matrix P is sparse by construction, computing π is not an issue. The aggregate wealth is (in principle) then $\sum_{s,n} \pi_{sn} w_n$, as in (2.1). The only caveat is that for state N, the probability is not concentrated on the grid point w_N but it is a Pareto distribution with exponent ζ and minimum size w_N. Since its density conditional on $x \geq w_N$ is

$$f(x) = \zeta w_N^{\zeta} x^{-\zeta-1},$$

the conditional mean of wealth is

$$\int_{w_N}^{\infty} x \zeta w_N^{\zeta} x^{-\zeta-1} \, \mathrm{d}x = \frac{\zeta}{\zeta - 1} w_N.$$

Thus we can compute the aggregate wealth as

$$\mathrm{E}[w] \approx \sum_{s=1}^{S} \left(\sum_{n=1}^{N-1} \pi_{sn} w_n + \pi_{sN} \frac{\zeta}{\zeta - 1} w_N \right). \tag{2.8}$$

Comparing (2.8) to (2.1), we can see that the usual truncation method introduces an error because the last term is w_N instead of $\frac{\zeta}{\zeta-1} w_N$. Since w_N is typically large, if ζ is close to 1 (Zipf's law), then failing to account for the term $\frac{\zeta}{\zeta-1}$ will introduce significant error.

The same correction applies to computing the integral of more general functions. Suppose we would like to compute the expectation of the power function w^{ν} for some power ν. For example, $\nu = 1$ corresponds to aggregate wealth, $\nu = 2$ the variance of wealth, and $\nu = 1 - \gamma$ with $\gamma > 0$ appears in calculating the welfare for CRRA preferences with relative risk aversion $\gamma > 0$. Assuming that wealth has a Pareto upper tail with exponent $\zeta > \nu$, then the conditional expectation of the upper tail is

$$\mathrm{E}\left[w^\nu \,|\, w \ge w_N\right] = \int_{w_N}^\infty \zeta w_N^\zeta x^{\nu-\zeta-1}\,\mathrm{d}x = \frac{\zeta}{\zeta-\nu} w_N^\nu.$$

Therefore the analog of (2.8) is

$$\mathrm{E}[w^\nu] \approx \sum_{s=1}^S \left(\sum_{n=1}^{N-1} \pi_{sn} w_n^\nu + \pi_{sN} \frac{\zeta}{\zeta-\nu} w_N^\nu \right), \tag{2.9}$$

which implies that we need to multiply the value for the largest grid point by the factor $\frac{\zeta}{\zeta-\nu}$. Similarly, noting that

$$\mathrm{E}\left[w^\nu \log w \,|\, w \ge w_N\right] = \mathrm{E}\left[\frac{\mathrm{d}}{\mathrm{d}\nu} w^\nu \,\middle|\, w \ge w_N\right] = \frac{\zeta}{(\zeta-\nu)^2} w_N^\nu + \frac{\zeta}{\zeta-\nu} w_N^\nu \log w,$$

setting $\nu = 0$ we obtain

$$\mathrm{E}[\log w] \approx \sum_{s=1}^S \left(\sum_{n=1}^{N-1} \pi_{sn} \log w_n + \pi_{sN} \left(\log w_N + \frac{1}{\zeta} \right) \right). \tag{2.10}$$

Therefore we need to add $1/\zeta$ to the value for the largest grid point when computing the expectation of log wealth.

3 Evaluating solution accuracy

As in any new numerical method, the first order of business is to evaluate the solution accuracy. In this regard, Den Haan, Judd, and Juillard (2010) "find it troublesome that [. . .] the accuracy of numerical solutions obtains so little attention by so many authors these days." One reason why accuracy gets little attention may be due to the lack of benchmark closed-form solutions for heterogeneous-agent models.[6] For this purpose, we present a simple (minimal) heterogeneous-agent model with idiosyncratic investment risk that admits a semi-analytical solution, which we use as a benchmark for evaluating numerical solutions.

3.1 Model

We consider a standard Aiyagari (1994) model, except that the model features no idiosyncratic labor income risk (to make the model analytically tractable) but only investment risk (to generate a fat-tailed wealth distribution). The production side is completely standard: there is a representative firm with Cobb-Douglas production function $F(K, L) = AK^\alpha L^{1-\alpha}$, where $A > 0$ is productivity and $\alpha \in (0, 1)$ is the capital share. Capital depreciates at rate δ each period. There are two types of agents, capitalists and workers, of whom there is a mass 1 continuum each. Workers are identical, supply one unit of labor inelastically, and consume the entire wage

[6] In the context of representative-agent asset pricing models, several authors such as Collard and Juillard (2001), Schmitt-Grohé and Uribe (2004), and Farmer and Toda (2017) use the closed-form solution of Burnside (1998) to evaluate the solution accuracy.

(hand-to-mouth).[7]

For reasons that will become clear shortly, capitalists are born and go bankrupt with probability p each period (Yaari (1965)–Blanchard (1985) perpetual youth model). Newborn agents are exogenously endowed with initial wealth $w_0 > 0$, and capital is destroyed after bankruptcy. Capitalists have constant relative risk aversion (CRRA) utility as in (2.2a) and supply capital to the firm. Importantly, the gross return on capital is *not* risk-free as

$$R_f = F_K(K, 1) + 1 - \delta = A\alpha K^{\alpha-1} + 1 - \delta, \tag{3.1}$$

but rather $z_s R_f$, where $s = 1, \dots, S$ denotes the exogenous Markov state and $z_s > 0$ is the gross return on capital *relative* to the risk-free rate (essentially the excess return). Let $P = (p_{ss'})$ be the transition probability matrix, which we assume to be irreducible. We assume that $\mathrm{E}[z_s] = 1$, so capital income is just a zero-sum redistribution of aggregate capital income across capitalists. An interpretation is that capitalists earn persistent heterogeneous returns (Fagereng, Guiso, Malacrino, and Pistaferri, 2016a; Cao and Luo, 2017) because some are more skillful in using capital (or just lucky) than others. The initial state of a newborn capitalist is drawn from the stationary distribution $\pi = (\pi_1, \dots, \pi_S)'$ of the transition probability matrix P.

The timing is as follows. A capitalist enters period t with some resource (units of consumption good) w_t. He decides how much to consume c_t, and the remaining amount $k_{t+1} := w_t - c_t$ is installed as capital. At the beginning of period $t+1$, production takes place by pooling all capital, and the capitalist receives the proceed $w_{t+1} = z_{s_t} R_f k_{t+1}$, where R_f is the gross risk-free rate in (3.1) and z_{s_t} is the predetermined gross excess return.[8] Thus the budget constraint of a capitalist is

$$w' = z_s R_f (w - c). \tag{3.2}$$

A stationary equilibrium consists of aggregate capital K, gross risk-free rate R_f, optimal consumption rule $\{c_s(w)\}_{s=1}^S$, and a stationary distribution $\Gamma(w, s)$ such that (i) given R_f, the optimal consumption rule maximizes the utility (2.2a) subject to the budget constraint (3.2), (ii) firms maximize profits, so (3.1) holds, (iii) the capital market clears, so

$$K = \int (w - c_s(w)) \, \mathrm{d}\Gamma(w, s), \tag{3.3}$$

and (iv) $\Gamma(w, s)$ is the stationary distribution of the law of motion

$$(w, s) \mapsto \begin{cases} (z_s R_f (w - c_s(w)), s'), & \text{with probability } (1-p)p_{ss'}, \\ (w_0, s'), & \text{with probability } p\pi_{s'}. \end{cases}$$

By exploiting homotheticity, we can solve the model semi-analytically as discussed in Appendix B. We also prove that a stationary equilibrium exists and the wealth distribution has a Pareto upper tail.

We use a numerical example to evaluate the solution accuracy. We consider the parameter

[7] The hand-to-mouth assumption is only for simplicity. Although we can also assume that workers behave optimally, it is inessential for our purpose of discussing numerical algorithms and evaluating the solution accuracy.

[8] We can also allow for the possibility that the gross excess returns are risky by using $z_{s_t s_{t+1}}$ instead of z_{s_t}.

values in Table 2. By solving the equilibrium conditions discussed in Appendix B, we obtain the gross risk-free rate $R_f = 1.0972$, aggregate capital $K = 3.4231$, and Pareto exponent $\zeta = 1.2826$.

Table 2: Parameter values of the Aiyagari model.

Parameter	Symbol	Value
Discount factor	β	0.96
Relative risk aversion	γ	2
Bankruptcy probability	p	0.025
Gross excess return	z	(0.95,1.05)
Transition probability matrix	P	$\begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}$
Productivity	A	1
Capital share	α	0.38
Capital depreciation rate	δ	0.08
Initial wealth	w_0	1

For the numerical solution, we consider both the conventional truncation method as well as the proposed Pareto extrapolation method with various wealth grid, truncation point, and number of grid points. For the Pareto extrapolation spacing parameter h, we always take $h = w_N - w_{N-1}$, the distance between the two largest grid points.

3.2 Solution accuracy in partial equilibrium

We first evaluate the solution accuracy in partial equilibrium. To ensure that all the differences of the numerical solutions from the analytical one are entirely due to the construction of the transition probability matrix, instead of solving for the equilibrium numerically for each method, we use the equilibrium risk-free rate and consumption policies from the semi-analytical solution to compute the stationary distribution on the wealth grid, and then compute the implied aggregate capital using (2.1) and (2.8) for the Truncation and Pareto extrapolation methods, respectively. For this exercise, our primary interest is the relative error $\widehat{K}/K - 1$, where K and $\widehat{K}$ are the aggregate capital from the semi-analytical and numerical solutions, respectively.

To implement our algorithm, we first need to specify the wealth grid $\{w_n\}_{n=1}^N$. Two natural candidates are the evenly- and exponentially-spaced grids, which we discuss in detail in Appendices C.1 and C.2, respectively. These two grids have both advantages and disadvantages. An ideal grid is one such that

1. the grid spacing $w_n - w_{n-1}$ is sufficiently small in the bulk of the wealth distribution so that we can approximate the law of motion accurately, and

2. the largest point w_N is sufficiently large so that the grid covers the (potentially) nonlinear part of the policy functions.

The evenly-spaced grid achieves the first objective but fails the second, while the opposite is true for the exponentially-spaced grid.

As a compromise, we suggest using the hybrid (affine-exponential) grid: construct the exponentially-spaced grid as discussed in Appendix C.2, but replace the bottom by an evenly-spaced grid. This way, we can choose a relatively large truncation point w_N, while keeping the grid spacing $w_n - w_{n-1}$ small for at least the bottom points, which contain the bulk of the wealth distribution. In particular, we construct the grid as follows. First, we compute the aggregate capital in a corresponding representative-agent model, which is

$$K_{\mathrm{RA}} = ((1/(\beta(1-p)) - 1 + \delta)/(A\alpha))^{\frac{1}{\alpha-1}} = 4.5577. \tag{3.4}$$

K_{RA} serves as the transition point between the evenly- and exponentially-spaced grids. Second, we construct an N-point exponential grid on $(0, \bar{w}]$ such that the median grid point corresponds to K_{RA}, and we replace the points on $(0, K_{\mathrm{RA}}]$ by an evenly-spaced grid. Table 3 shows the relative error $\widehat{K}/K - 1$ in the aggregate capital using this affine-exponential grid for various truncation point $\bar{w}$ and number of points N, both for the truncation and Pareto extrapolation methods.

Table 3: Relative error (%) in aggregate capital for the truncation and Pareto extrapolation methods with the affine-exponential grid.

Method:	Truncation			Pareto extrapolation		
$\bar{w}/K_{\mathrm{RA}}$	$N = 25$	50	100	25	50	100
10^1	-31.240	-27.620	-26.250	-1.110	0.292	0.422
10^2	-21.500	-16.860	-14.480	-2.172	-0.642	0.128
10^3	-16.510	-11.210	-8.590	-2.303	-0.827	-0.141
10^4	-13.610	-7.950	-5.360	-2.234	-0.804	-0.205
10^5	-11.750	-5.930	-3.490	-2.125	-0.727	-0.200
10^6	-10.530	-4.610	-2.360	-2.029	-0.643	-0.174

Note: N: number of grid points; $\bar{w}$: wealth truncation point.

We can make a few observations from Table 3. First, the conventional truncation method is extremely poor at calculating the aggregate capital with a moderate truncation point $\bar{w}$: the relative error is about 26% with $\bar{w} = 10K_{\mathrm{RA}} = 45.6$ and $N = 100$, which is similar to the case with an evenly-spaced grid with $\bar{w} = 40$ and $N = 100$ in Table 14 (27%). On the other hand, the Pareto extrapolation method is astonishingly more accurate, with relative errors ranging from 0.13% to 2.3% depending on the specification. Second, for the truncation method, choosing a larger truncation point $\bar{w}$ improves the accuracy because it misses less of the upper tail. However, even with a huge truncation point such as $\bar{w}/K_{\mathrm{RA}} = 10^6$, the errors exceed the largest errors with Pareto extrapolation. Finally, the accuracy of the Pareto extrapolation method is almost independent of $\bar{w}$. This is probably because the upper tail is well-approximated by a Pareto distribution and our method corrects for the truncation error using the theoretical Pareto exponent, so the choice of $\bar{w}$ is not so important for accuracy.

3.3 Solution accuracy in general equilibrium

So far we have evaluated the accuracy of each solution method by comparing the implied aggregate capital to that from the semi-analytical solution. However, aggregate capital is usually not a quantity of interest. Therefore we now evaluate the solution accuracy by solving for the entire equilibrium.

We consider both the truncation and Pareto extrapolation methods with the affine-exponential grid, where the number of grid points is $N = 100$, the median grid point is K_{RA}, and the largest grid point $\bar{w} = w_N$ is such that $\bar{w}/K_{\text{RA}} = 10^1, 10^2, 10^3, 10^4, 10^5, 10^6$ as before. To ensure that all the differences of the numerical solutions from the analytical one are entirely due to the construction of the transition probability matrix, for each guess of the equilibrium risk-free rate, we use the semi-analytical solution to the optimal consumption-savings problem to compute the law of motion for wealth. Table 4 shows the relative errors in equilibrium quantities (gross risk-free rate R_f, aggregate capital K, and Pareto exponent ζ).

Table 4: Relative errors (%) in equilibrium quantities.

Method:	Truncation			Pareto extrapolation		
$\bar{w}/K_{\text{RA}}$	R_f	K	ζ	R_f	K	ζ
10^1	0.761	-7.159	-13.317	-0.005	0.051	0.100
10^2	0.287	-2.801	-5.399	-0.002	0.016	0.031
10^3	0.142	-1.402	-2.732	0.002	-0.017	-0.034
10^4	0.080	-0.793	-1.553	0.003	-0.025	-0.049
10^5	0.049	-0.486	-0.955	0.002	-0.024	-0.048
10^6	0.032	-0.316	-0.622	0.002	-0.021	-0.042

Note: $\bar{w}$: wealth truncation point. The number of grid points is $N = 100$. The actual values are: $R_f = 1.0972$, $K = 3.4231$, $\zeta = 1.2826$.

As opposed to partial equilibrium, in general equilibrium the truncation method with a moderately large truncation point performs reasonably well. For example, the relative errors are 0.1–3% when $\bar{w}/K_{\text{RA}} = 10^3$. However, the relative errors for the Pareto extrapolation method are 0.002–0.03% with the same $\bar{w}$, so it is 100 times more accurate. The reason why the truncation method performs better in general equilibrium than in partial equilibrium is probably due to the general equilibrium effect: because the truncation method underestimates the aggregate capital (given the risk-free rate, as seen in Table 3), the equilibrium risk-free rate must rise to clear the market. In fact, in Table 4, R_f is upwardly biased for the truncation method. Since wealth (and hence aggregate capital) rises with a higher risk-free rate, the downward bias in aggregate capital with the truncation method gets mitigated in general equilibrium.

Given that the truncation method with a moderately large truncation point $\bar{w}$ is reasonably accurate, one may wonder what is the point of improving it further. There are several reasons to prefer the Pareto extrapolation method. First, we can obtain 10–100 times more accurate results at no additional computational cost, so there is just no reason not to use it. Second, and more importantly, the performance of the Pareto extrapolation method is robust across the choice of the truncation point $\bar{w}$, whereas it is sensitive for the truncation method. Thus, with the Pareto extrapolation method, the researcher need not worry about the choice of the truncation point, while extra care is needed with the truncation method. Finally, the interest rate, aggregate capital, and Pareto exponent may not be the only quantities of interest. One may be interested in other quantities, such as the top 1% wealth share.

To address the last point, we compute top wealth shares as follows. For each grid point, we can compute the aggregate wealth held by agents at least as rich as that grid point (using either truncation or Pareto extrapolation, depending on the solution method). Dividing that number by aggregate wealth gives the top wealth share at that grid point. By interpolating between points, we can define the top wealth shares inside the grid. To compute the top wealth shares outside the grid, we do as follows. For the Pareto extrapolation method, we use the theoretical Pareto exponent ζ to extrapolate the wealth share beyond the largest grid point. More precisely, let $\pi_N = \sum_{s=1}^{S} \pi_{sN}$ be the probability mass on the largest grid point w_N in the Pareto extrapolation method. Then the density for $x \geq w_N$ is given by $f(x) = \pi_N \zeta w_N^{\zeta} x^{-\zeta-1}$, where ζ is the Pareto exponent. Using this, the tail probability $\Pr(X \geq x)$ is proportional to $x^{-\zeta}$, whereas the total wealth held by wealthy agents $\mathrm{E}[X; X \geq x]$ is proportional to $x^{-\zeta+1}$. Therefore the wealth share $s(p)$ of the wealthiest fraction $p \in (0,1)$ of agents is given by $s(p) = Cp^{1-1/\zeta}$, where the constant of proportionality C can be easily calculated from π_N, w_N, ζ, and aggregate capital K. Taking the logarithm, we obtain

$$\log s = (1 - 1/\zeta) \log p + \log C, \tag{3.5}$$

so top wealth shares are linear in a log-log scale. For the truncation method, since it is not obvious how to extrapolate the top wealth share beyond the largest grid point, we simply interpolate by a cubic spline using the point $(0,0)$ (by definition, the top 0% wealth share is 0) and all the grid points. Table 5 shows some representative top wealth shares.[9] Because top wealth shares need to be computed only once after solving for the equilibrium, for both truncation and Pareto extrapolation methods, we use a grid that is 10 times finer than the one used for solving the equilibrium. Figure 3 plots the top wealth shares against the truncation point for $\bar{w}/K_{\mathrm{RA}} = 10^1, \dots, 10^6$.

According to Table 5 and Figure 3, the truncation method vastly underestimates top wealth shares when the truncation point $\bar{w}$ is small, as expected. On the other hand, the Pareto extrapolation method gives numbers that are accurate up to two or three significant digits.

3.4 Strategy for solving fat-tailed heterogeneous-agent models

What do we learn from these exercises? The good news is that the conventional truncation method is able to solve models reasonably accurately, provided that we use an exponentially-spaced grid with a large enough number of points and a large enough truncation point—say a million times the typical scale. Thus the conventional solution method will likely give correct answers if the researcher is careful. The bad news are that we do not a priori know how large

[9] Technically, for the semi-analytical solution we cannot compute the exact top wealth shares because the functional form of the wealth distribution is unknown (we only know the tail behavior characterized by the Pareto exponent ζ). For this case, to compute the stationary distribution, we use the Pareto extrapolation method with a highly accurate 2,000-point affine-exponential grid with truncation $\bar{w} = 10^6 \times K_{\mathrm{RA}}$, which we take as the truth.

Table 5: Top wealth shares (%) in equilibrium.

Method:	Truncation				Pareto extrapolation			
$\bar{w}/K_{\text{RA}}$	Top 0.01%	0.1%	1%	10%	Top 0.01%	0.1%	1%	10%
10^1	0.11	1.16	12.76	50.87	13.11	21.81	36.27	60.31
10^2	1.20	10.08	29.07	57.55	13.20	21.92	36.37	60.39
10^3	7.08	17.43	33.76	59.54	13.27	21.99	36.46	60.46
10^4	10.65	20.16	35.51	60.28	13.29	22.01	36.48	60.47
10^5	12.17	21.32	36.23	60.55	13.30	22.03	36.49	60.47
10^6	12.83	21.80	36.50	60.63	13.28	22.00	36.47	60.46
Analytical	13.21	21.92	36.39	60.40	13.21	21.92	36.39	60.40

Note: $\bar{w}$: wealth truncation point; Top x%: wealth share (%) of the wealthiest x%. The number of grid points is $N = 100$ for solving the equilibrium and $10N = 1{,}000$ for computing top wealth shares. "Truncation" and "Pareto extrapolation" refer to the truncation and Pareto extrapolation methods for solving the equilibrium, and "Analytical" shows results from the semi-analytical solution. Top wealth shares for "Analytical" are computed using the grid in Footnote 9.

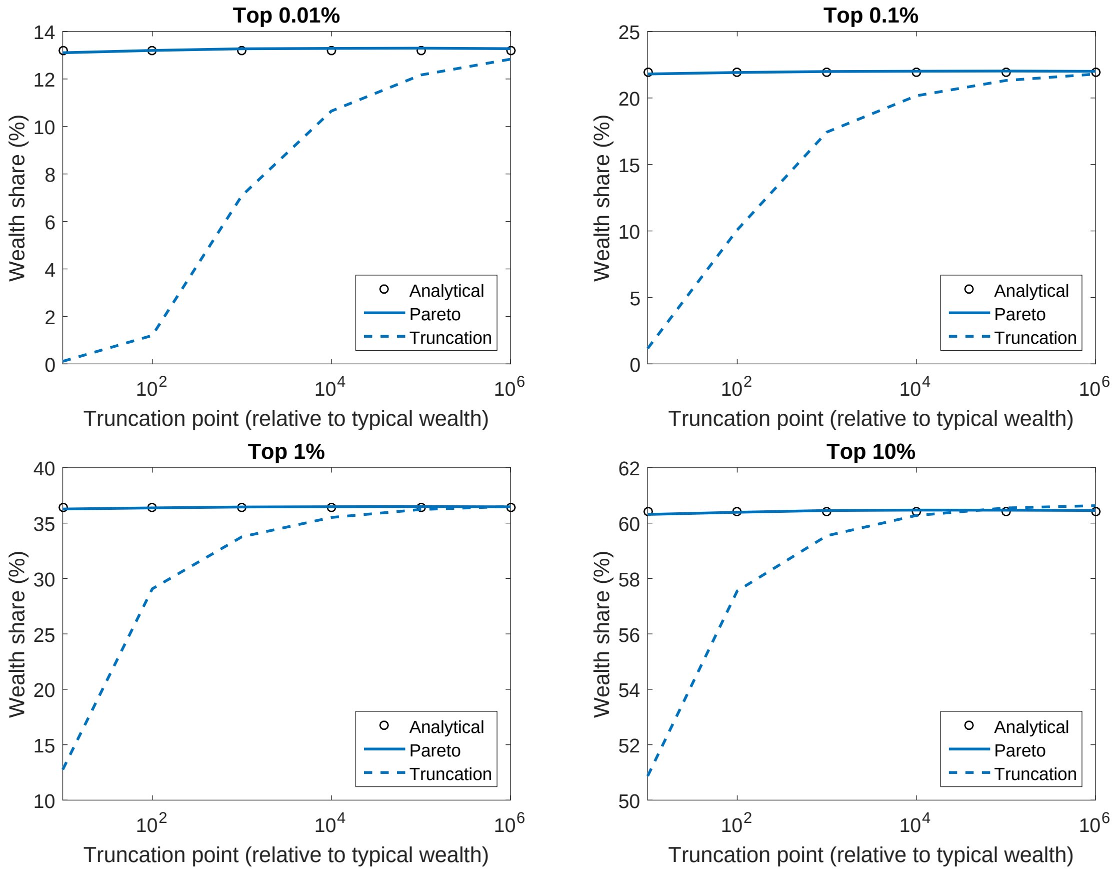


Figure 3: Top wealth shares in the Aiyagari model.

Note: "Analytical", "Pareto", and "Truncation" refer to the semi-analytical solution and the numerical solutions using the Pareto extrapolation and truncation methods, respectively.

is large enough. There is a superior alternative: Pareto extrapolation is far more accurate, it has no additional computational cost, and its performance is robust across the grid specification.

Based on the above observations as well as the results in Appendix C, we recommend the following strategy for solving heterogeneous-agent models with fat-tailed wealth distribution.

Strategy for solving fat-tailed heterogeneous-agent models.

1. Before solving the model, find out a typical scale for the state variable (wealth), perhaps by solving a representative-agent model without any shock.
2. Solve the heterogeneous-agent model using the Pareto extrapolation method with the hybrid affine-exponential grid. More concretely,
   (a) Construct the exponentially-spaced grid with a truncation point about 1,000 times the typical scale for the state variable.
   (b) Replace the bottom half grid points by an evenly-spaced grid.
3. After solving the model, if necessary, recompute the wealth distribution on a finer grid from the already computed equilibrium law of motion.

Note that there are many other possibilities. In the value function iteration step, since we do not need the wealth distribution, we can just use an exponentially-spaced grid with relatively few points to increase the speed. When computing the market clearing condition, we can interpolate the policy functions on a finer grid and then use Pareto extrapolation method for accuracy.

4 Merton-Bewley-Aiyagari model

Having now established that the solution method we propose is accurate, we apply it to study wealth inequality in an incomplete market general equilibrium model in the spirit of Aiyagari (1994). Agents face an income fluctuation problem as in Bewley (1977, 1983) and those who choose to invest face uninsurable investment risk that leads to an investor's problem similar to Merton (1969) and Samuelson (1969), although in a Markovian setting as studied in Krebs (2006) and Toda (2014). The model (which we refer to as the Merton-Bewley-Aiyagari, or MBA model) generates a fat-tailed wealth distribution, where the Pareto exponent is shaped by rich general equilibrium effects. We provide a step-by-step approach to solving the model and analyzing its quantitative implications. We use the model to showcase how the asymptotic analysis of the individual problem can be used to solve for the general equilibrium, and how solution accuracy affects the quantitative implications of the model.

4.1 Model

Time is discrete and denoted by $t = 0, 1, \dots$.

Agents The economy is populated by a unit measure of infinitely-lived agents with Epstein-Zin preferences

$$U_t = \left((1-\beta)c_t^{1-1/\varepsilon} + \beta\,\mathrm{E}_t[U_{t+1}^{1-\gamma}]^{\frac{1-1/\varepsilon}{1-\gamma}}\right)^{\frac{1}{1-1/\varepsilon}}, \tag{4.1}$$

where $c_t > 0$ is consumption, $U_t > 0$ is continuation utility, $\beta \in (0, 1)$ is the discount factor, $\gamma > 0$ is the coefficient of relative risk aversion, and $\varepsilon > 0$ is the elasticity of intertemporal substitution.[10] Agents differ in productivity/ability states denoted by $s \in S = \{1, \dots, S\}$, which evolve over time according to a Markov chain with irreducible transition probability matrix $P = (p_{ss'})$. The idiosyncratic productivity states are independent and identically distributed across agents and we assume the law of large numbers for the continuum as in Sun (2006). Therefore if $\pi = (\pi_1, \dots, \pi_S)'$ denotes the (unique) stationary distribution of the transition probability matrix P, at any point in time exactly fraction $\pi_s > 0$ of agents are in state s.

Labor and financial markets A type s agent has labor productivity $h_s \geq 0$ and earns pre-tax labor income ωh_s, where $\omega > 0$ is the "piece-rate" wage determined in equilibrium. A type s agent has investment ability $z_s > 0$ and earns excess returns in the financial market, as described below. Without loss of generality we assume that $z_1 \leq \cdots \leq z_S$, so the Markov state s indexes the agent's investment ability. There are two types of assets, risk-free and risky. Let R_f be the gross risk-free rate determined in equilibrium. We assume that the ex post pre-tax gross return on risky investment for an investor in state s is

$$R_{sj} = z_{sj} R_f, \tag{4.2}$$

where z_{sj} is defined as the sum of the investment ability z_s and a zero-mean i.i.d. random variable ϵ_j that can take J possible values $\epsilon_1 < \cdots < \epsilon_J$. Let $p_j > 0$ be the probability of state j. Thus high-skilled investors earn higher returns on average (z_s), but there is some element of luck (ϵ_j). We can interpret this as lack of diversification. We assume that $z_1 + \epsilon_1 > 0$, so agents have limited liability even in the worst possible state. To prevent arbitrage, we also assume that $z_S + \epsilon_1 < 1$, so even the most skilled investor underperforms the risk-free asset with positive probability. From the above assumptions, note that (i) z_{sj} is increasing in both s and j, (ii) $z_{sj} > 0$ for all s, j, and (iii) $z_{S1} < 1$.

Technology Technology is represented by a representative firm with a constant-returns-to-scale production function $F(K, L)$. Capital depreciates at rate $\delta \in [0, 1]$. Therefore the firm's problem is

$$\max_{K,L \geq 0} \left[-K + \frac{1}{R_f} (F(K, L) - \omega L + (1 - \delta) K) \right]. \tag{4.3}$$

That is, the firm buys capital K at the end of time t, hires labor to produce, and pays the profit and depreciated capital to the shareholders (who discount using the risk-free rate since there is no aggregate risk).

Budget constraint and taxes Letting w be the financial wealth at the beginning of the period, the budget constraint of an agent is

$$w' = (1 - \tau_w) \Big((1 + (1 - \tau_k)(R_f - 1))(w + (1 - \tau_h)\omega h_s - I - c) + (1 + (1 - \tau_k)(R_{sj} - 1)) I \Big) \geq \underline{w},$$

[10] By considering the limit $\gamma \to 1$, we interpret $\mathrm{E}[U^{1-\gamma}]^{\frac{1}{1-\gamma}}$ as $\exp(\mathrm{E}[\log U])$ if $\gamma = 1$. Similarly, we interpret $((1 - \beta)c^{1-1/\varepsilon} + \beta v^{1-1/\varepsilon})^{\frac{1}{1-1/\varepsilon}}$ as $c^{1-\beta} v^{\beta}$ if $\varepsilon = 1$.

where $\underline{w}$ is an exogenous minimum wealth constraint and $I \geq 0$ is the investment in the risky asset. We assume that there are proportional taxes on labor income (at rate τ_h), capital income (τ_k), and wealth (τ_w). Notice that the capital income tax applies to the net return on the risk-free and risky assets while the wealth tax applies to the beginning of the period wealth. We assume that the tax proceeds are wasted. To simplify the notation, we denote by $\tilde{R}_f, \tilde{R}_{sj}$ the after-tax gross returns on the risk-free and risky assets, respectively:

$$\tilde{R}_f = (1-\tau_w)(1+(1-\tau_k)(R_f-1)), \tag{4.4a}$$

$$\tilde{R}_{sj} = (1-\tau_w)(1+(1-\tau_k)(R_{sj}-1)). \tag{4.4b}$$

The budget constraint thus simplifies to

$$w' = \tilde{R}_f(w+(1-\tau_h)\omega h_s - I - c) + \tilde{R}_{sj} I \geq \underline{w}. \tag{4.5}$$

Equilibrium Our equilibrium concept is the stationary equilibrium defined as follows.

Definition 4.1 (Stationary equilibrium)**.** A stationary equilibrium consists of a gross risk-free rate R_f, a piece-rate wage ω, aggregate capital K, aggregate labor L, optimal decision rules $\{c_s(w), I_s(w)\}_{s=1}^S$, value functions $\{v_s(w)\}_{s=1}^S$, and a stationary distribution $\Gamma(w,s)$ such that

1. given R_f and ω, aggregate capital K and aggregate labor L solves the profit maximization problem (4.3),

2. given R_f and ω, for each s the optimal decision rule $(c_s(w), I_s(w))$ maximizes the recursive utility (4.1) subject to the budget and borrowing constraint (4.5), i.e.,

$$v_s(w) = \max_{c,I\geq 0}\left((1-\beta)c^{1-1/\varepsilon} + \beta\,\mathrm{E}\left[v_{s'}(w')^{1-\gamma}\,\middle|\, s\right]^{\frac{1-1/\varepsilon}{1-\gamma}}\right)^{\frac{1}{1-1/\varepsilon}}, \tag{4.6}$$

3. the capital market clears, so

$$K = \int (w+(1-\tau_h)\omega h_s - I_s(w) - c_s(w))\,\mathrm{d}\Gamma(w,s) + \int z_s I_s(w)\,\mathrm{d}\Gamma(w,s), \tag{4.7}$$

4. the labor market clears, so

$$L = \sum_{s=1}^S \pi_s h_s, \tag{4.8}$$

5. $\Gamma(w,s)$ is the stationary distribution of the law of motion for $(w,s) \in [\underline{w},\infty) \times S$ defined by

$$(w,s) \mapsto \left(\tilde{R}_f(w+(1-\tau_h)\omega h_s - I - c) + \tilde{R}_{sj} I, s'\right) \tag{4.9}$$

with probability $p_{ss'}p_j$, where $\tilde{R}_f$ and $\tilde{R}_{sj}$ are as in (4.4).

Note that an investor who invests I units of wealth supplies $z_s I$ units of capital to the firm (see the right-most term in (4.7)), so investors not only supply funds but also “expertise”. This assumption (together with $\mathrm{E}[\epsilon_j] = 0$) makes the (pre-tax) gross return to investment $R_{sj} = (z_s + \epsilon_j)R_f$ consistent with the aggregate resource constraint.

4.2 Asymptotic analysis

As discussed in Section 2, the solution to the asymptotic problem plays an important role in the Pareto extrapolation algorithm (especially for computing the theoretical Pareto exponent). In this section, we discuss the properties of the asymptotic problem of the MBA model and derive parametric restrictions in general equilibrium.

4.2.1 Asymptotic problem

We first derive the asymptotic problem and convert it in a more convenient form. Since there is no labor income in the asymptotic problem, the budget constraint (4.5) becomes

$$w' = \tilde{R}_f(w - I - c) + \tilde{R}_{sj} I \geq 0, \tag{4.10}$$

where $\tilde{R}_f, \tilde{R}_{sj}$ are as in (4.4). Letting $\theta = \frac{I}{w-c} \geq 0$, (4.10) becomes

$$w' = \left(\tilde{R}_f(1-\theta) + \tilde{R}_{sj}\theta\right)(w - c) \geq 0. \tag{4.11}$$

The following proposition characterizes the solution to the asymptotic problem.

Proposition 4.2 (Asymptotic problem). *Suppose $\gamma \neq 1$. For $s = 1, \dots, S$, define*

$$\rho_s = \max_{0 \leq \theta \leq \bar{\theta}_s} \mathrm{E}\left[\left(\tilde{R}_f(1-\theta) + \tilde{R}_{sj}\theta\right)^{1-\gamma} \,\middle|\, s\right]^{\frac{1}{1-\gamma}}, \tag{4.12}$$

where the upper bound $\bar{\theta}_s := \frac{\tilde{R}_f}{\tilde{R}_f - \tilde{R}_{s1}} > 0$. Then ρ_s is well-defined and there exists a unique maximizer θ_s^*. Letting $D = \operatorname{diag}\left(\rho_1^{1-\gamma}, \dots, \rho_S^{1-\gamma}\right)$, the asymptotic problem has a solution if and only if

$$\beta \rho(DP)^{\frac{1-1/\varepsilon}{1-\gamma}} < 1, \tag{4.13}$$

where $\rho(DP)$ is the spectral radius of DP. Under this condition, the value function of the asymptotic problem is given by $v_s(w) = b_s w$, where $b = (b_1, \dots, b_S) \gg 0$ is the unique positive solution to the system of nonlinear equations

$$b_s = \begin{cases} \left((1-\beta)^\varepsilon + \beta^\varepsilon \left(\rho_s \mathrm{E}\left[b_{s'}^{1-\gamma} \,\middle|\, s\right]^{\frac{1}{1-\gamma}}\right)^{\varepsilon-1}\right)^{\frac{1}{\varepsilon-1}} & (\varepsilon \neq 1) \\ (1-\beta)^{1-\beta}\beta^\beta \left(\rho_s \mathrm{E}\left[b_{s'}^{1-\gamma} \,\middle|\, s\right]^{\frac{1}{1-\gamma}}\right)^\beta & (\varepsilon = 1) \end{cases} \tag{4.14}$$

for $s = 1, \dots, S$. The optimal consumption-investment rules of the asymptotic problem are

$$c_s(w) = \bar{c}_s w := (1-\beta)^\varepsilon b_s^{1-\varepsilon} w, \tag{4.15a}$$

$$I_s(w) = \bar{I}_s w := \theta_s^*(1 - (1-\beta)^\varepsilon b_s^{1-\varepsilon}) w. \tag{4.15b}$$

Anticipating the subsequent experiments on changing tax rates, it is interesting to derive comparative static results for the optimal portfolio.

Proposition 4.3. *Let everything be as in Proposition 4.2 and $\theta_s^* \ge 0$ be the optimal portfolio (fraction of wealth invested in the risky asset) in state s. Then θ_s^* is independent of τ_w and $\theta_s^* > 0$ if and only if $z_s > 1$. Under this condition, we have the following comparative statics:*

$$\frac{\partial \theta_s^*}{\partial R_f} < 0, \quad \frac{\partial \theta_s^*}{\partial z_s} > 0, \quad \frac{\partial \theta_s^*}{\partial \tau_k} > 0.$$

The intuition for this result is as follows. Because the wealth tax applies to the total return on wealth, it does not affect the portfolio choice. On the other hand, because capital income tax is applied to capital gains (and investors can deduct capital losses), capital income tax essentially provides an insurance and makes the risky asset less risky. Consequently, agents invest more in the risky asset under a higher capital income tax.

4.2.2 Parametric restrictions in general equilibrium

Since the diagonal matrix $D = \operatorname{diag}\left(\rho_1^{1-\gamma}, \dots, \rho_S^{1-\gamma}\right)$ depends on the equilibrium interest rate R_f, the spectral condition (4.13) puts a restriction on R_f. The following lemma puts further restrictions based on equilibrium considerations.

Lemma 4.4 (Finite aggregate wealth)**.** *Let everything be as in Proposition 4.2 and suppose that (4.13) holds. Define*

$$G_s := (1 - (1-\beta)^\varepsilon b_s^{1-\varepsilon})(\tilde{R}_f(1-\theta_s^*) + \mathrm{E}\left[\tilde{R}_{sj} \,\middle|\, s\right] \theta_s^*) \tag{4.16}$$

and $G = (G_1, \dots, G_S)'$. Then in equilibrium it must be

$$\rho(P \operatorname{diag} G) < 1. \tag{4.17}$$

The intuition for Lemma 4.4 is as follows. Using the budget constraint and the optimal consumption and portfolio rules for asymptotic agents established in Proposition 4.2, the expected growth rate of wealth in state s becomes G_s in (4.16). The spectral condition (4.17) ensures that the wealth of rich agents does not grow on average and makes the aggregate wealth finite, which must be the case in stationary equilibrium.

Under the assumptions of Lemma 4.4, the following lemma provides an explicit algorithm for computing the Pareto exponent ζ.

Lemma 4.5 (Pareto exponent)**.** *Let everything be as in Lemma 4.4 and*

$$G_{sj} = (1 - (1-\beta)^\varepsilon b_s^{1-\varepsilon})(\tilde{R}_f(1-\theta_s^*) + \tilde{R}_{sj}\theta_s^*) > 0 \tag{4.18}$$

be the ex post gross growth rate of wealth for asymptotic agents in state (s, j). Define the conditional moment generating function of log growth rate by

$$M_s(z) = \mathrm{E}\left[\mathrm{e}^{z \log G_{sj}} \,\middle|\, s\right] = \sum_{j=1}^{J} p_j G_{sj}^z \tag{4.19}$$

and the diagonal matrix $D(z) = \operatorname{diag}(M_1(z), \dots, M_S(z))$. Suppose that $p_{ss} > 0$ for all s and $G_{sJ} > 1$

for some s. Then there exists a unique solution $z = \zeta > 1$ to

$$\rho(PD(z)) = 1.$$

The Pareto exponent of the wealth distribution is $\zeta > 1$. If $G_{sJ} \leq 1$ for all s, then the wealth distribution does not have a Pareto tail.

Finally, it is trivial to show that the equilibrium risk-free rate must satisfy

$$R_f > 1 - \delta, \tag{4.20}$$

for otherwise the demand for capital would be infinite.

4.3 Calibration

A time period represents a year and we calibrate the model to the U.S. economy.

Preferences, technology, and taxes We assume a Cobb-Douglas production function $F(K, L) = K^\alpha L^{1-\alpha}$, where $\alpha \in (0, 1)$ represents the capital share. We set the technological and preference parameters (β, γ, α) to standard values (see Table 6). For the elasticity of intertemporal substitution, most macro papers assume that it is less than 1, while most finance papers assume that it is greater than 1. To be neutral, we set $\varepsilon = 1$, which is also supported by studies using disaggregated data to estimate the elasticity (Mankiw and Zeldes, 1991; Attanasio and Weber, 1993; Beaudry and van Wincoop, 1996; Vissing-Jørgensen, 2002). We set the labor income, capital income, and wealth tax rate respectively to $\tau_h = 0.224$, $\tau_k = 0.25$, $\tau_w = 0$ as in Guvenen, Kambourov, Kuruscu, Ocampo-Diaz, and Chen (2018), who use estimates from McDaniel (2007). Finally, we set the borrowing limit to $1/4$ of average annual labor income as in Kaplan, Moll, and Violante (2018). Our calibration implies that average value of h_s in a stationary equilibrium is 1, so $\underline{w} = -\omega/4$.

Table 6: Parameter values.

Parameter	Symbol	Value
Discount factor	β	0.96
Relative risk aversion	γ	2
Elasticity of intertemporal substitution	ε	1
Capital share	α	0.38
Labor income tax	τ_h	0.224
Capital income tax	τ_k	0.25
Wealth tax	τ_w	0
Borrowing limit	$\underline{w}$	$-\omega/4$

Exogenous individual states We assume that labor productivity $\{h_{s_t}\}_{t=0}^\infty$ and investment ability $\{z_{s_t}\}_{t=0}^\infty$ depend contemporaneously on two Markov state variables s^π and s^τ, with associated transition probability matrices P^π and P^τ. We interpret the first state s^π as a "permanent component", which affects both labor productivity and investment ability and takes three states: low, high, and high-entrepreneur. The second state s^τ, which we call the "transitory component" affects only labor productivity and takes three values: low, average, and high. The index $s = s^\pi \times s^\tau$ can thus take 9 states.

Labor productivity Labor productivity in state s is the product of a permanent component and a transitory component:

$$h_s = h_s^\pi h_s^\tau.$$

The permanent component h_s^π takes two values: 0.3980 (low) and 1.6020 (high and high-entrepreneur). The high state workers (high and high-entrepreneur) thus earn a wage rate 4.03 times higher than low state workers, in line with the ratio of the mean annual income of the top half to the bottom half of full-time workers in the U.S.[11] We interpret agents as dynasties with perfect altruism, and assume that the permanent component of labor productivity is very persistent and changes on average every 40 years. Moreover, we choose transition probabilities for the permanent component as to imply that, in a stationary equilibrium, 50% of the agents are in the low state and 3.7% are in the high-entrepreneur state. We choose the value of 3.7% to match the fraction of households that invest at least half of their net worth in a business (see Footnote 12). The transition probability matrix P^π is thus given by

$$P^\pi = \begin{bmatrix} 0.9875 & 0.0116 & 0.0009 \\ 0.0125 & 0.9866 & 0.0009 \\ 0.0125 & 0.0116 & 0.9759 \end{bmatrix}. \tag{4.21}$$

We model the process for the transitory component of labor productivity h^τ as an AR(1) in logarithm

$$\log h_{s_t}^\tau = \rho \log h_{s_{t-1}}^\tau + \sigma \eta_t, \tag{4.22}$$

where $\eta_t \sim N(0,1)$. Guvenen (2009) obtains values of $\rho = 0.821$ and $\sigma = 0.170$ by estimating a model that allows for heterogeneous lifetime earnings profiles. We discretize the process (4.22) over a three-point grid using the method proposed by Farmer and Toda (2017), while imposing that the unconditional mean of h_s^τ is equal to one. We obtain values for h_s^τ of 0.6584 (low state), 0.9150 (average state), and 1.5115 (high state). The resulting transition probability matrix P^τ is given by

$$P^\tau = \begin{bmatrix} 0.8290 & 0.1630 & 0.0080 \\ 0.0815 & 0.8370 & 0.0815 \\ 0.0080 & 0.1630 & 0.8290 \end{bmatrix}. \tag{4.23}$$

Since h_s^π and h_s^τ are independent and both have an unconditional mean of one, the average labor productivity $h_s = h_s^\pi h_s^\tau$ is one in equilibrium.

Investment returns We assume that investment ability z is fully determined by the permanent component s^π. For the low and high states, we set $z = 1$, which implies that those agents

[11] We use data from the American Community Survey and restrict the sample to employed individuals aged between 20 and 60. For every year, we truncate the sample at the 5th and 95th percentile and then compute the ratio of the mean annual income in the bottom and top half of the sample. We then average the ratio over the 2000-2016 period and obtain an average ratio of 4.03.

do not earn an excess return on their investments (i.e., they earn the risk-free rate R_f). Since investment returns are risky, those agents will never invest. For the high-entrepreneur state, we set $z = 1.0275$, which implies an annual excess return of 2.75%. This is roughly the difference between the average return on financial assets of households at the 90th and 10th percentiles of the financial wealth distribution in Norway as reported in Fagereng, Guiso, Malacrino, and Pistaferri (2016b, Figure 2).

To calibrate the distribution of idiosyncratic investment return shocks ϵ, we use microdata from the Survey of Consumer Finances and construct a measure of the rate of return on business investment (business income over the market value of the business) for each household.[12] Using the nonparametric discretization method proposed by Toda (2018a) on the de-meaned data, we obtain a discrete distribution for ϵ, which takes values $(-0.0836, 0.0761, 0.3795)$ with probability $(0.6345, 0.2822, 0.0833)$.

Putting all the pieces together, we have $S = 9$ exogenous individual states and $J = 3$ idiosyncratic investment return shock states. The transition probability matrix for the exogenous individual states is given by $P = P^\pi \otimes P^\tau$, where $\otimes$ is the Kronecker product and P^π, P^τ are defined in (4.21) and (4.23), respectively. Table 7 summarizes the dependence of labor productivity and investment ability on the exogenous individual state s.

Table 7: Exogenous individual states.

State (s)	Component		Productivity/ability	
	Permanent	Transitory	Labor (h_s)	Investment (z_s)
1	low	low	0.2620	1.00
2	low	average	0.3642	1.00
3	low	high	0.6016	1.00
4	high	low	1.0547	1.00
5	high	average	1.4659	1.00
6	high	high	2.4215	1.00
7	high-entrepreneur	low	1.0547	1.0275
8	high-entrepreneur	average	1.4659	1.0275
9	high-entrepreneur	high	2.4215	1.0275

4.4 Solution algorithm

To solve the calibrated model, we use dynamic programming methods combined with the Pareto extrapolation algorithm. We use a 50-point affine-exponential grid (as described in Section 3.3) to approximate the optimal decision rules $c_s(w)$ and $I_s(w)$ as well as the wealth distribution $\Gamma(w, s)$. Once the model is solved, we recompute the wealth distribution over a finer grid with 1,000 points and report top wealth shares. In both cases, we use a truncation

[12] We use data from the Survey of Consumer Finances for the years 2001, 2004, 2007, 2010, 2013, and 2016 and keep only households who have at least 50% of their net worth invested in businesses. Business income `busseincfarm` is defined as "Income from business, sole proprietorship, and farm" while the market value of the businesses `bus` is defined as "Total value of business(es) in which the household has either an active or nonactive interest". We winsorize the rate of return at the 5% level.

point of $\bar{w} = 10^3 \times K_{\mathrm{RA}}$, where

$$K_{\mathrm{RA}} = ((1/\beta - 1 + \delta)/(A\alpha))^{\frac{1}{\alpha-1}} = 6.2771 \tag{4.24}$$

is the capital stock in the corresponding representative-agent model. The solution algorithm consists of 4 steps.

1. Given a guess of R_f that satisfies the finite capital demand restriction (4.20), solve the problem of the asymptotic agents (4.12), (4.14). If a solution to the asymptotic problem exists (4.13), verify that capital supply is finite (4.17). If there is no solution to the asymptotic problem or if capital supply is infinite, update R_f.

2. Given R_f, compute the capital demand K^d and wage ω implied by profit maximization. Recall that aggregate labor supply is determined only by exogenous parameters and is normalized to $L = 1$. Thus

$$K^d = \left(\frac{\alpha}{R_f - 1 + \delta}\right)^{\frac{1}{1-\alpha}}, \quad \omega = (1-\alpha)\left(\frac{\alpha}{R_f - 1 + \delta}\right)^{\frac{\alpha}{1-\alpha}}.$$

3. Given R_f and ω, solve the individual optimization problem using dynamic programming (see Appendix D for details), compute the stationary distribution from the law of motion (4.9) using the Pareto extrapolation algorithm, and compute the aggregate capital supply K^s (4.7) (with the correction term as in (2.8)).

4. If excess demand $K^d - K^s$ is within error tolerance, stop. Otherwise, update R_f.

4.5 Quantitative results

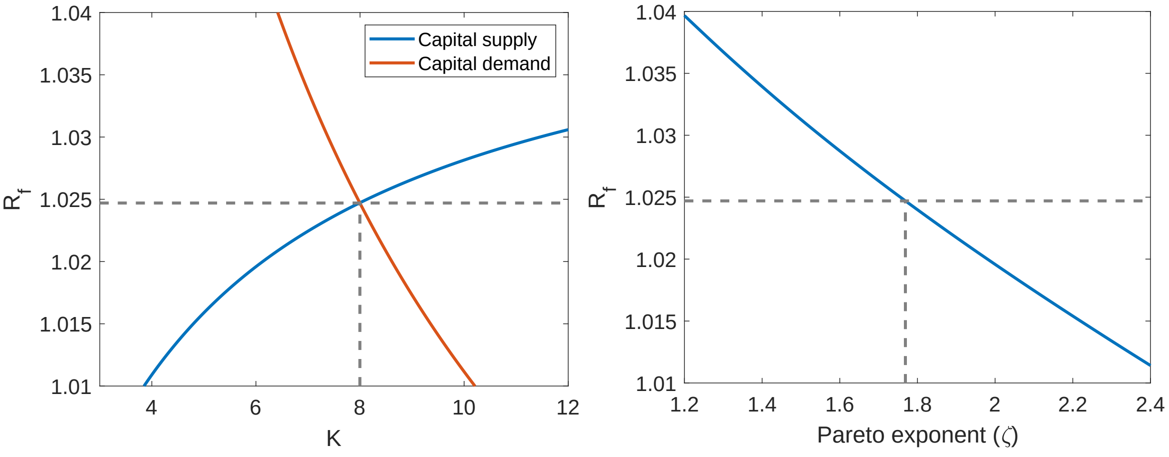
We now discuss the quantitative implications of the calibrated model. Figure 4a shows the aggregate capital supply and demand curves for a range of values of R_f. The demand curve is determined by profit maximization of the representative firm, while the supply curve is obtained by aggregating net capital supply from households, either supplied directly to the firm or intermediated by investors (see (4.7)). The intersection of the two curves pins down the risk-free rate that clears the capital market. We obtain a value of 1.0247 (Table 8), or 2.47%.

The equilibrium interest rate in turn determines the Pareto exponent ζ of the wealth distribution (Figure 4b). Higher interest rates R_f are associated with lower Pareto exponents ζ (higher inequality). The reason is as follows. By homotheticity, the optimal portfolios (θ_s^*) of the asymptotic agents, which solve (4.12), depend only on the ratios $(\tilde{R}_{sj}/\tilde{R}_f)$. By the definition of the after-tax gross returns in (4.4), if there is no capital income tax ($\tau_k = 0$), then

$$\frac{\tilde{R}_{sj}}{\tilde{R}_f} = \frac{1 + (1-\tau_k)(R_{sj} - 1)}{1 + (1-\tau_k)(R_f - 1)} = \frac{R_{sj}}{R_f} = z_{sj},$$

which is exogenous. Thus, by Proposition 4.2, if $\tau_k = 0$ the optimal portfolios are independent of the risk-free rate R_f. Using (4.18), the ex post gross growth rate of wealth when $\varepsilon = 1$ is

$$G_{sj} = \beta(1-\tau_w)R_f(1 - \theta_s^* + z_{sj}\theta_s^*),$$



(a) Capital demand and supply.

(b) Determination of Pareto exponent.

Figure 4: Equilibrium.

which is linear in R_f. Since each moment generating function $M_s(z)$ in (4.19) is increasing in R_f and the Pareto exponent ζ is determined by (2.5), ζ is increasing in R_f in the special case $\tau_k = 0$ and $\varepsilon = 1$.[13] By continuity, the same result holds if τ_k is sufficiently small. Intuitively, the "pace" at which rich agents get richer is increasing in the interest rate, so there is more concentration of wealth at the top of the distribution in economies with higher interest rates. We obtain an equilibrium value of $\zeta = 1.77$ (Table 8), which is close but higher than what is estimated for the U.S. (1.52 according to Table 8 of Vermeulen, 2018).

Table 8: Equilibrium objects.

Object	Symbol	Value
Capital	K	7.99
Labor	L	1.00
Risk-free rate	$R_f - 1$	2.47%
Wage rate	ω	1.37
Pareto exponent	ζ	1.77

While the wealth distribution exhibits a Pareto upper tail, the Pareto exponent ζ does not fully summarize the wealth distribution. For example, the borrowing constraint is an important determinant of the wealth share of the poorest 50% of agents, yet ζ depends only on the behavior of rich agents, who are not affected by the borrowing constraint.[14] Table 9 presents the wealth shares in the model and in the data (Survey of Consumer Finances).[15] The model qualitatively replicates two important features of the data: the poorest 50% of households hold little wealth (2.31% in the model, 1.79% in the data) while the top 1% accounts for a large share of (36.65% in the model, 34.95% in the data).

The last column of Table 9 presents the wealth shares associated with a (pure) Pareto distri-

[13] Actually it is easy to relax the assumption to $\varepsilon \geq 1$. See Proposition 5 of Toda (2018b).

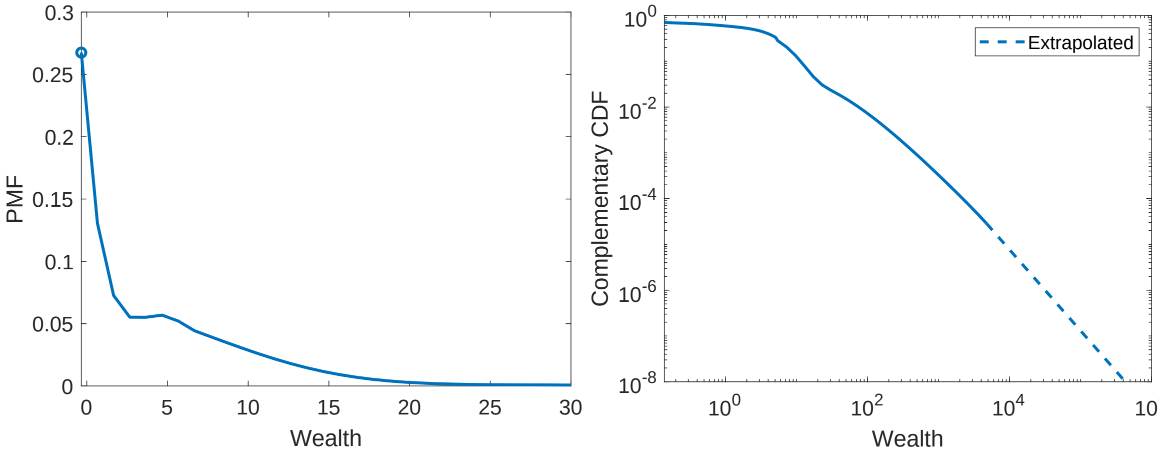
[14] To be precise, the Pareto exponent ζ can be computed using only the solution to the asymptotic problem (see Section 4.2.1, especially Lemma 4.5), which does not depend on the borrowing constraint.

[15] We compute the wealth shares in the data using the Survey of Consumer Finances and average over the survey years 2001, 2004, 2007, 2010, 2013, and 2016.

Table 9: Wealth shares (%).

Groups	Model	Data	Pure Pareto
$[0,50)$	2.31	1.79	26.02
$[50,90)$	33.56	25.09	37.23
$[90,99)$	27.48	38.17	23.25
$[99,100]$	36.65	34.95	13.51

bution with the same exponent ζ as in the model.[16] The Pareto distribution generates a bottom 50% wealth share nearly five times as high as in the model (26.02% versus 4.91%). One drawback of analytical models that imply a "pure-Pareto" wealth distribution such as Aoki and Nirei (2017) is that they do not allow for negative wealth levels, and therefore cannot match the low wealth share of the bottom 50% of agents. In contrast, our model naturally matches both the upper and lower tails of the wealth distribution.



(a) Probability mass function.

(b) Complementary CDF.

Figure 5: Wealth distribution.

Figure 5a presents the probability mass function (PMF) for wealth levels $[\underline{w}, 30]$. Notice that the borrowing constraint binds for 26.8% of households. Visually, it would appear that truncating the distribution at $\bar{w} = 30$ is a reasonable choice.[17] While it is true that 97.1% of households in the model have wealth levels in the range $[\underline{w}, 30]$, the remaining 2.9% of households account for 46% of aggregate wealth. Figure 5b shows the complementary CDF of the wealth distribution in a log-log scale for the range $[10^{-1}, 10^{6}]$. As predicted by theory, the upper tail of the distribution appears to converge to the theoretical Pareto slope.

4.6 Quantitative implication of truncation error

We now solve the model using both the truncation and Pareto extrapolation methods for different truncation points $\bar{w}$ and assess how the choice of the truncation point affects the implied top wealth shares. Figure 6 presents top wealth shares for a range of truncation points

[16] The cumulative distribution function (CDF) of a Pareto distribution with exponent ζ and minimum size $\underline{x} > 0$ is $F(x) = 1 - (x/\underline{x})^{-\zeta}$ over $[\underline{x}, \infty)$. Its top wealth shares are independent of $\underline{x}$.

[17] The "typical scale" computed as in (4.24) is $K_{\text{RA}} = 6.28$, so a truncation point of 30 represents roughly 5 times the typical scale.

$\bar{w}/K_{\text{RA}} = 10^2, 10^3, 10^4, 10^5, 10^6$, where K_{RA} is given by (4.24). As in Section 3.3 (see Figure 3), the truncation method underestimates the top wealth shares and the bias decreases with the truncation point.

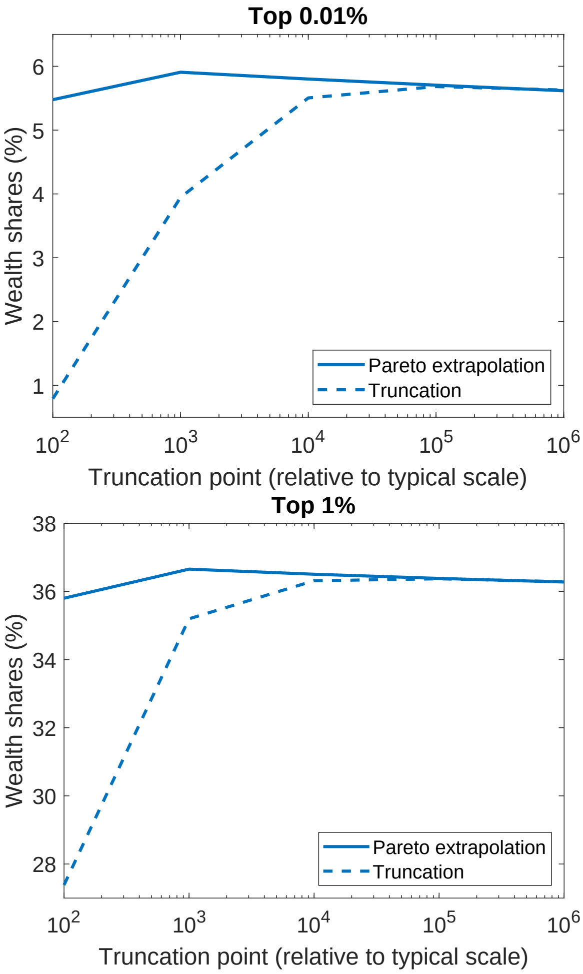


Figure 6: Top wealth shares.

Unlike in the Aiyagari model in Section 3, the graphs of the top wealth shares for the Pareto extrapolation method are not exactly flat. This is likely due to the nonlinearity in the policy functions. In the Aiyagari model, because policy functions are exactly linear, the choice of the truncation point is irrelevant. In the MBA model, however, the policy functions are linear only *asymptotically* (see Figure 7), so the choice of the truncation point matters. Recall that the Pareto extrapolation algorithm exploits the asymptotic linearity of policy functions to approximate the behavior of agents with wealth levels above the truncation point. Therefore, the truncation point must be chosen to be large enough so that policy functions are approximately linear beyond that point. It is not a priori clear which truncation point gives results that are closest to the truth. Choosing a too small truncation point may not fully capture the nonlinearity in the policy functions, while choosing a too large truncation point may compromise accuracy because grid points are less dense.

How should a researcher choose an appropriate truncation point? We suggest choosing a truncation point that implies a small difference between the marginal propensity to consume (MPC) at the largest grid point $(c_{s,N} - c_{s,N-1})/(w_N - w_{N-1})$ and the asymptotic MPC $\bar{c}_s$ (see

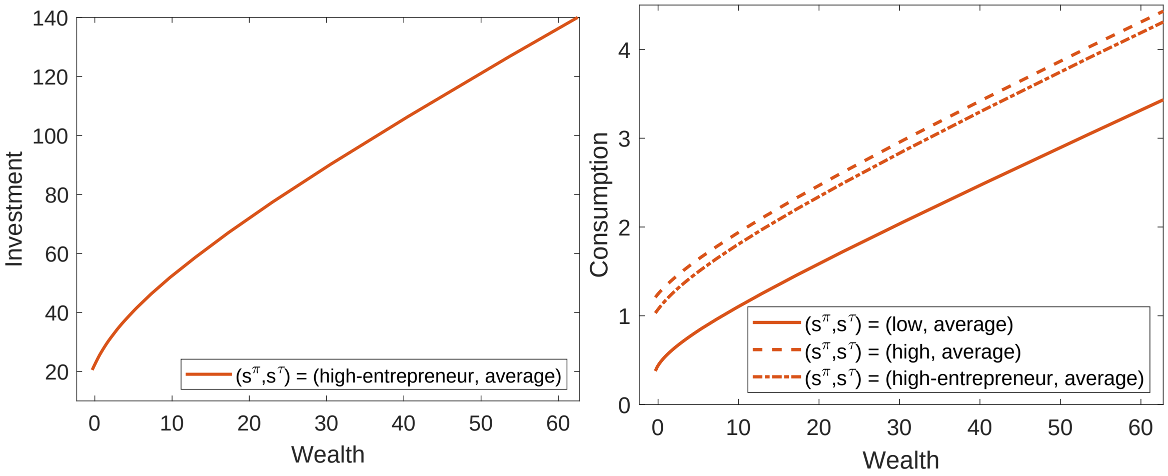


Figure 7: Policy functions.

(4.15)). The idea is that, if the MPC at the largest grid point is "close" to its asymptotic value, then the consumption function should be approximately linear. Table 10 reports the MPC relative error (in percentage points) for a range of truncation points. We define the MPC relative error as

$$100 \times \frac{\frac{c_{s,N}-c_{s,N-1}}{w_N-w_{N-1}} - \bar{c}_s}{\bar{c}_s}, \tag{4.25}$$

where $c_{s,n}$ denotes the policy function in state s at the grid point w_n and $w_N = \bar{w}$ is the largest grid point. The choice of truncation point $\bar{w}/K_{\text{RA}} = 10^3$ that we use to solve the MBA model thus implies that the MPC relative error is at most 0.0153%, suggesting that it is an appropriate choice.

Table 10: Relative error (%) in marginal propensities to consume.

State (s)	Truncation point ($\bar{w}/K_{\text{RA}}$)				
	10^2	10^3	10^4	10^5	10^6
1	0.3652	0.0065	0.0001	0	0
2	0.3677	0.0066	0.0001	0	0
3	0.3706	0.0067	0.0001	0	0
4	0.4560	0.0073	0.0001	0	0
5	0.4584	0.0074	0.0001	0	0
6	0.4596	0.0075	0.0001	0	0
7	0.6848	0.0151	0.0002	0	0
8	0.6894	0.0152	0.0002	0	0
9	0.6937	0.0153	0.0002	0	0

Note: the individual states $s = 1, \ldots, 9$ are defined in Table 7

4.7 Taxing wealth? A bad idea

We now use the model to quantify the welfare effects of introducing a 2% wealth tax. To do so, we first solve the model with $\tau_w = 0.02$. We assume that any increase (decrease) in tax revenue, which we denote ΔT, is rebated to the household in the form of a proportional consumption subsidy (tax) λ. Noting that a typical agent's net capital income is

$$\begin{aligned}(R_f-1)(w+(1-\tau_h)\omega h_s-I_s(w)-c_s(w))+(R_f z_{sj}-1)I_s(w)&\\ =(R_f-1)(w+(1-\tau_h)\omega h_s-c_s(w))+R_f(z_{sj}-1)I_s(w)&\end{aligned}$$

and $\mathrm{E}\left[z_{sj}\,\middle|\,s\right]=\mathrm{E}\left[z_s+\epsilon_j\,\middle|\,s\right]=z_s$, the total tax revenue is

$$\begin{aligned}T=\int\Bigg[&\underbrace{\tau_h\omega h_s}_{\text{Labor income tax}}+\underbrace{\tau_k\big((R_f-1)(w+(1-\tau_h)\omega h_s-c_s(w))+R_f(z_s-1)I_s(w)\big)}_{\text{Capital income tax}}\\ &+\underbrace{\tau_w\big((1+(1-\tau_k)(R_f-1))(w+(1-\tau_h)\omega h_s-c_s(w))+(1-\tau_k)R_f(z_s-1)I_s(w)\big)}_{\text{Wealth tax}}\Bigg]\mathrm{d}\Gamma(w,s).\end{aligned}$$

Due to homothetic preferences, we only need to solve the model without consumption subsidy and then compute aggregate consumption $C := \int c_s(w)\,\mathrm{d}\Gamma(w,s)$ as well as the change in tax revenue ΔT. The consumption subsidy is then computed as

$$\lambda=\frac{\Delta T}{C}, \tag{4.26}$$

and the welfare function as

$$\mathcal{W}=(1+\lambda)\left(\int v_s(w)^{1-\gamma}\,\mathrm{d}\Gamma(w,s)\right)^{\frac{1}{1-\gamma}}, \tag{4.27}$$

which is the certainty equivalent of the value function in the stationary equilibrium. Note that $\mathcal{W}$ is in units of consumption because the Epstein-Zin utility (4.1) is also in units of consumption.[18] Intuitively, the welfare measure (4.27) is the risk-adjusted utility (in units of consumption) of an agent who is randomly thrown into the stationary equilibrium.

Overall, we find that the introduction of a wealth tax is a "lose-lose" policy: wealth concentration remains mostly unchanged, while tax revenue, output, and welfare all decrease. Table 11 shows the percentage change in aggregate variables in response to the introduction of the 2% wealth tax. First, notice that capital decreases by 28.58%. The reason is that the wealth tax reduces the after-tax return on investment, which causes a downward shift of the capital supply curve. Given that capital demand remains unchanged (the wealth tax does not affect the firm's problem), the stock of capital decreases and the risk-free rate increases (+2.43 percentage points). The decrease in the capital stock, combined with the fact that aggregate labor supply is inelastic, leads to a decline in output and wages (−11.97%).

The decline in output alone need not imply a decrease in welfare. For instance, it could be compensated by a reduction in consumption inequality driven by lower wealth concentration. We find that this is not the case. Welfare decreases (in consumption equivalent) by 14.35%. Moreover, wealth inequality remains mostly unchanged. Table 12 shows the change in wealth shares. While it is true that the top 1% wealth share declines modestly (from 36.65% to 35.77%),

[18] To compute (4.27) numerically, we use the correction term from (2.9) with $\nu=1-\gamma$ to extrapolate the term $v_s(w)^{1-\gamma}$ off the grid.

Table 11: Equilibrium objects.

Object	Symbol	No Wealth Tax	2% Wealth Tax	Change (%)
Output	Y	2.20	1.94	-11.97
Capital	K	7.99	5.71	-28.58
Labor	L	1.00	1.00	0.00
Risk-free rate	$R_f - 1$	2.47%	4.90%	2.43
Wage rate	ω	1.37	1.20	-11.97
Welfare	$\mathcal{W}$	0.71	0.61	-14.35

the bottom 50% also declines (from 2.31% to 1.76%).

Table 12: Wealth shares (%).

Groups	No Wealth Tax	2% Wealth Tax	Change
$[0, 50)$	2.31	1.76	-0.55
$[50, 90)$	33.56	34.14	0.58
$[90, 99)$	27.48	28.33	0.85
$[99, 100]$	36.65	35.77	-0.88

A surprising result is that the decline in output and labor income is so large that total tax revenue actually *declines*. Table 13 shows the change in tax revenue by components (labor income tax, capital income tax, and wealth tax). The introduction of the wealth tax provides a new tax revenue of 0.038 (10.5% of initial total tax revenue). Yet, the decline in labor income tax (-12.01%) and capital income tax (-7.35%) dominate, leading to a 1.07% decline in total tax revenue. As a result, the shortfall in tax revenue is compensated by a proportional consumption tax of 0.35% ($\lambda = -0.0035$), which further decreases welfare.

Table 13: Tax revenue.

Tax	No Wealth Tax	2% Wealth Tax	Change (%)
Labor income tax	0.306	0.269	-12.01
Capital income tax	0.053	0.049	-7.35
Wealth tax	0	0.038	-
Total	0.359	0.355	-1.07

5 Concluding remarks

This paper proposes a simple, systematic approach—Pareto extrapolation—to analyze and solve heterogeneous-agent models that endogeneously generate fat-tailed wealth distributions. The core insight that we leverage is due to Pareto, who noticed that wealth data displayed a striking empirical regularity.

> Nous sommes tout de suite frappé du fait que les points ainsi déterminés, ont une tendance très marqué à se disposer en ligne droite.

> (We are instantly struck by the fact that the points determined this way have a very marked tendency to be disposed in straight line.)
>
> —Pareto (1897, pp. 304–305)

We put Pareto's insight to work to tackle *models* of wealth inequality. Our approach makes the solution algorithm more transparent, efficient, and accurate with zero additional computational cost.

References

Daron Acemoglu and Dan Cao. Innovation by entrants and incumbents. *Journal of Economic Theory*, 157:255–294, May 2015. doi:10.1016/j.jet.2015.01.001.

Yves Achdou, Jiequn Han, Jean-Michel Lasry, Pierre-Louis Lions, and Benjamin Moll. Income and wealth distribution in macroeconomics: A continuous-time approach. NBER Working Paper 23732, 2017. URL `http://www.nber.org/papers/w23732`.

S. Rao Aiyagari. Uninsured idiosyncratic risk and aggregate saving. *Quarterly Journal of Economics*, 109(3):659–684, August 1994. doi:10.2307/2118417.

Yann Algan, Olivier Allais, and Wouter J. Den Haan. Solving heterogeneous-agent models with parameterized cross-sectional distributions. *Journal of Economic Dynamics and Control*, 32(3):875–908, March 2008. doi:10.1016/j.jedc.2007.03.007.

Yann Algan, Olivier Allais, Wouter J. Den Haan, and Pontus Rendahl. Solving and simulating models with heterogeneous agents and aggregate uncertainty. In Karl Schmedders and Kenneth L. Judd, editors, *Handbook of Computational Economics*, volume 3, chapter 6, pages 277–324. Elsevier, 2014. doi:10.1016/B978-0-444-52980-0.00006-2.

Shuhei Aoki and Makoto Nirei. Zipf's law, Pareto's law, and the evolution of top incomes in the United States. *American Economic Journal: Macroeconomics*, 9(3):36–71, July 2017. doi:10.1257/mac.20150051.

Costas Arkolakis. A unified theory of firm selection and growth. *Quarterly Journal of Economics*, 131(1):89–155, February 2016. doi:10.1093/qje/qjv039.

Orazio P. Attanasio and Guglielmo Weber. Consumption growth, the interest rate and aggregation. *Review of Economic Studies*, 60(3):631–649, July 1993. doi:10.2307/2298128.

Robert L. Axtell. Zipf distribution of U.S. firm sizes. *Science*, 293(5536):1818–1820, September 2001. doi:10.1126/science.1062081.

Brendan K. Beare and Alexis Akira Toda. Geometrically stopped Markovian random growth processes and Pareto tails. 2017. URL `https://arxiv.org/abs/1712.01431`.

Paul Beaudry and Eric van Wincoop. The intertemporal elasticity of substitution: An exploration using a US panel of state data. *Economica*, 63(251):495–512, August 1996. doi:10.2307/2555019.

Jess Benhabib, Alberto Bisin, and Shenghao Zhu. The distribution of wealth and fiscal policy in economies with finitely lived agents. *Econometrica*, 79(1):123–157, January 2011. doi:10.3982/ECTA8416.

Jess Benhabib, Alberto Bisin, and Shenghao Zhu. The wealth distribution in Bewley economies with capital income risk. *Journal of Economic Theory*, 159(A):489–515, September 2015. doi:10.1016/j.jet.2015.07.013.

Jess Benhabib, Alberto Bisin, and Shenghao Zhu. The distribution of wealth in the Blanchard-Yaari model. *Macroeconomic Dynamics*, 20:466–481, March 2016. doi:10.1017/S1365100514000066.

Truman F. Bewley. The permanent income hypothesis: A theoretical formulation. *Journal of Economic Theory*, 16(2):252–292, December 1977. doi:10.1016/0022-0531(77)90009-6.

Truman F. Bewley. A difficulty with the optimum quantity of money. *Econometrica*, 51(5): 1485–1504, September 1983. doi:10.2307/1912286.

Olivier J. Blanchard. Debt, deficits, and finite horizons. *Journal of Political Economy*, 93(2): 223–247, April 1985. doi:10.1086/261297.

Jaroslav Borovička and John Stachurski. Necessary and sufficient conditions for existence and uniqueness of recursive utilities. 2017. URL `https://arxiv.org/abs/1710.06526`.

Craig Burnside. Solving asset pricing models with Gaussian shocks. *Journal of Economic Dynamics and Control*, 22(3):329–340, March 1998. doi:10.1016/S0165-1889(97)00075-4.

Marco Cagetti and Mariacristina De Nardi. Entrepreneurship, frictions, and wealth. *Journal of Political Economy*, 114(5):835–870, October 2006. doi:10.1086/508032.

Dan Cao and Wenlan Luo. Persistent heterogeneous returns and top end wealth inequality. *Review of Economic Dynamics*, 26:301–326, October 2017. doi:10.1016/j.red.2017.10.001.

Christopher D. Carroll, Kiichi Tokuoka, and Weifeng Wu. The method of moderation. 2012.

Christopher D. Carroll, Jiri Slacalek, Kiichi Tokuoka, and Matthew N. White. The distribution of wealth and the marginal propensity to consume. *Quantitative Economics*, 8(3):977–1020, November 2017. doi:10.3982/QE694.

Fabrice Collard and Michel Juillard. Accuracy of stochastic perturbation methods: The case of asset pricing models. *Journal of Economic Dynamics and Control*, 25(6-7):979–999, June 2001. doi:10.1016/S0165-1889(00)00064-6.

Wouter J. Den Haan. Comparison of solutions to the incomplete markets model with aggregate uncertainty. *Journal of Economic Dynamics and Control*, 34(1):4–27, January 2010a. doi:10.1016/j.jedc.2008.12.010.

Wouter J. Den Haan. Assessing the accuracy of the aggregate law of motion in models with heterogeneous agents. *Journal of Economic Dynamics and Control*, 34(1):79–99, January 2010b. doi:10.1016/j.jedc.2008.12.009.

Wouter J. Den Haan, Kenneth L. Judd, and Michel Juillard. Computational suite of models with heterogeneous agents: Incomplete markets and aggregate uncertainty. *Journal of Economic Dynamics and Control*, 34(1):1–3, January 2010. doi:10.1016/j.jedc.2009.07.001.

Richard Durrett. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, New York, fourth edition, 2010.

Andreas Fagereng, Luigi Guiso, Davide Malacrino, and Luigi Pistaferri. Heterogeneity in returns to wealth and the measurement of wealth inequality. *American Economic Review: Papers and Proceedings*, 106(5):651–655, May 2016a. doi:10.1257/aer.p20161022.

Andreas Fagereng, Luigi Guiso, Davide Malacrino, and Luigi Pistaferri. Heterogeneity and persistence in returns to wealth. NBER Working Paper 22822, 2016b. URL `http://www.nber.org/papers/w22822`.

Leland E. Farmer and Alexis Akira Toda. Discretizing nonlinear, non-Gaussian Markov processes with exact conditional moments. *Quantitative Economics*, 8(2):651–683, July 2017. doi:10.3982/QE737.

Xavier Gabaix. Power laws in economics and finance. *Annual Review of Economics*, 1:255–293, 2009. doi:10.1146/annurev.economics.050708.142940.

Xavier Gabaix. The granular origins of aggregate fluctuations. *Econometrica*, 79(3):733–772, May 2011. doi:10.3982/ECTA8769.

Xavier Gabaix, Jean-Michel Lasry, Pierre-Louis Lions, and Benjamin Moll. The dynamics of inequality. *Econometrica*, 84(6):2071–2111, November 2016. doi:10.3982/ECTA13569.

Robert Gibrat. *Les Inégalités Économiques*. Librairie du Recueil Sirey, Paris, 1931.

Fatih Guvenen. An empirical investigation of labor income processes. *Review of Economic Dynamics*, 12(1):58–79, January 2009. doi:10.1016/j.red.2008.06.004.

Fatih Guvenen, Gueorgui Kambourov, Burhan Kuruscu, Sergio Ocampo-Diaz, and Daphne Chen. Use it or lose it: Efficiency gains from wealth taxation. 2018. URL `https://fguvenendotcom.files.wordpress.com/2018/11/gkkoc-2018_fg_v107.pdf`.

Mark Huggett. The risk-free rate in heterogeneous-agent incomplete-insurance economies. *Journal of Economic Dynamics and Control*, 17(5-6):953–969, September-November 1993. doi:10.1016/0165-1889(93)90024-M.

Charles I. Jones and Jihee Kim. A Schumpeterian model of top income inequality. *Journal of Political Economy*, 126(5):1785–1826, October 2018. doi:10.1086/699190.

Greg Kaplan, Benjamin Moll, and Giovanni L. Violante. Monetary policy according to HANK. *American Economic Review*, 108(3):697–743, March 2018. doi:10.1257/aer.20160042.

Kenneth Kasa and Xiaowen Lei. Risk, uncertainty, and the dynamics of inequality. *Journal of Monetary Economics*, 94:60–78, April 2018. doi:10.1016/j.jmoneco.2017.11.008.

Harry Kesten. Random difference equations and renewal theory for products of random matrices. *Acta Mathematica*, 131(1):207–248, 1973. doi:10.1007/BF02392040.

Oren S. Klass, Ofer Biham, Moshe Levy, Ofer Malcai, and Sorin Solomon. The Forbes 400 and the Pareto wealth distribution. *Economics Letters*, 90(2):290–295, February 2006. doi:10.1016/j.econlet.2005.08.020.

Tom Krebs. Recursive equilibrium in endogenous growth models with incomplete markets. *Economic Theory*, 29(3):505–523, 2006. doi:10.1016/S0165-1889(03)00062-9.

Dirk Krueger, Kurt Mitman, and Fabrizio Perri. Macroeconomics and household heterogeneity. In John B. Taylor and Harald Uhlig, editors, *Handbook of Macroeconomics*, volume 2, chapter 11, pages 843–921. Elsevier, 2016. doi:10.1016/bs.hesmac.2016.04.003.

Per Krusell and Anthony A. Smith, Jr. Income and wealth heterogeneity in the macroeconomy. *Journal of Political Economy*, 106(5):867–896, October 1998. doi:10.1086/250034.

Per Krusell and Anthony A. Smith, Jr. Quantitative macroeconomic models with heterogeneous agents. In Richard Blundell, Whitney K. Newey, and Torsten Persson, editors, *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, volume 1 of *Econometric Society Monograph*, chapter 8, pages 298–340. Cambridge University Press, New York, 2006.

Qingyin Ma and John Stachurski. Dynamic programming deconstructed. 2018. URL `http://johnstachurski.net/_downloads/dpd6.pdf`.

N. Gregory Mankiw and Stephen P. Zeldes. The consumption of stockholders and nonstockholders. *Journal of Financial Economics*, 29(1):97–112, March 1991. doi:10.1016/0304-405X(91)90015-C.

Cara McDaniel. Average tax rates on consumption, investment, labor and capital in the OECD 1950-2003. 2007.

Alisdair McKay. Time-varying idiosyncratic risk and aggregate consumption dynamics. *Journal of Monetary Economics*, 88:1–14, June 2017. doi:10.1016/j.jmoneco.2017.05.002.

Robert C. Merton. Lifetime portfolio selection under uncertainty: The continuous-time case. *Review of Economics and Statistics*, 51(3):247–257, August 1969. doi:10.2307/1926560.

Benjamin Moll. Productivity losses from financial frictions: Can self-financing undo capital misallocation? *American Economic Review*, 104(10):3186–3221, October 2014. doi:10.1257/aer.104.10.3186.

Makoto Nirei and Shuhei Aoki. Pareto distribution of income in neoclassical growth models. *Review of Economic Dynamics*, 20:25–42, April 2016. doi:10.1016/j.red.2015.11.002.

Makoto Nirei and Wataru Souma. A two factor model of income distribution dynamics. *Review of Income and Wealth*, 53(3):440–459, September 2007. doi:10.1111/j.1475-4991.2007.00242.x.

Vilfredo Pareto. La legge della demanda. *Giornale degli Economisti*, 10:59–68, January 1895.

Vilfredo Pareto. *La Courbe de la Répartition de la Richesse*. Imprimerie Ch. Viret-Genton, Lausanne, 1896.

Vilfredo Pareto. *Cours d'Économie Politique*, volume 2. F. Rouge, Lausanne, 1897.

Vincenzo Quadrini. Entrepreneurship, saving, and social mobility. *Review of Economic Dynamics*, 3(1):1–40, January 2000. doi:10.1006/redy.1999.0077.

William J. Reed. The Pareto, Zipf and other power laws. *Economics Letters*, 74(1):15–19, December 2001. doi:10.1016/S0165-1765(01)00524-9.

Michael Reiter. Solving heterogeneous-agent models by projection and perturbation. *Journal of Economic Dynamics and Control*, 33(3):649–665, March 2009. doi:10.1016/j.jedc.2008.08.010.

Michael Reiter. Solving the incomplete markets model with aggregate uncertainty by backward induction. *Journal of Economic Dynamics and Control*, 34(1):28–35, January 2010. doi:10.1016/j.jedc.2008.11.009.

Paul A. Samuelson. Lifetime portfolio selection by dynamic stochastic programming. *Review of Economics and Statistics*, 51(3):239–246, August 1969. doi:10.2307/1926559.

Stephanie Schmitt-Grohé and Martín Uribe. Solving dynamic general equilibrium models using a second-order approximation to the policy function. *Journal of Economic Dynamics and Control*, 28(4):755–775, January 2004. doi:10.1016/S0165-1889(03)00043-5.

John Stachurski and Alexis Akira Toda. An impossibility theorem for wealth in heterogeneous-agent models with limited heterogeneity. 2018. URL `https://arxiv.org/abs/1807.08404`.

Yeneng Sun. The exact law of large numbers via Fubini extension and characterization of insurable risks. *Journal of Economic Theory*, 126(1):31–69, January 2006. doi:10.1016/j.jet.2004.10.005.

Alexis Akira Toda. Incomplete market dynamics and cross-sectional distributions. *Journal of Economic Theory*, 154:310–348, November 2014. doi:10.1016/j.jet.2014.09.015.

Alexis Akira Toda. Data-based automatic discretization of nonparametric distributions. 2018a. URL `http://arxiv.org/abs/1805.00896`.

Alexis Akira Toda. Wealth distribution with random discount factors. *Journal of Monetary Economics*, 2018b. doi:10.1016/j.jmoneco.2018.09.006.

Alexis Akira Toda and Kieran Walsh. The double power law in consumption and implications for testing Euler equations. *Journal of Political Economy*, 123(5):1177–1200, October 2015. doi:10.1086/682729.

Alexis Akira Toda and Kieran James Walsh. Fat tails and spurious estimation of consumption-based asset pricing models. *Journal of Applied Econometrics*, 32(6):1156–1177, September/October 2017. doi:10.1002/jae.2564.

Philip Vermeulen. How fat is the top tail of the wealth distribution? *Review of Income and Wealth*, 64(2):357–387, June 2018. doi:10.1111/roiw.12279.

Annette Vissing-Jørgensen. Limited asset market participation and the elasticity of intertemporal substitution. *Journal of Political Economy*, 110(4):825–853, August 2002. doi:10.1086/340782.

Thomas Winberry. A method for solving and estimating heterogeneous agent macro models. *Quantitative Economics*, 9(3):1123–1151, November 2018. doi:10.3982/QE740.

Menahem E. Yaari. Uncertain lifetime, life insurance, and the theory of the consumer. *Review of Economic Studies*, 32(2):137–150, April 1965. doi:10.2307/2296058.

Eric R. Young. Solving the incomplete markets model with aggregate uncertainty using the Krusell-Smith algorithm and non-stochastic simulations. *Journal of Economic Dynamics and Control*, 34(1):36–41, January 2010. doi:10.1016/j.jedc.2008.11.010.

Shenghao Zhu. A Becker-Tomes model with investment risk. *Economic Theory*, 2018. doi:10.1007/s00199-018-1103-2. Forthcoming.

A Asymptotic homogeneous problem

In this appendix we describe how to derive the asymptotic homogeneous problem in an abstract dynamic programming setting. For the notation, we follow Ma and Stachurski (2018). Let

- X be a set called the *state space*;
- A be a set called the *action space*;
- $\Gamma : X \twoheadrightarrow A$ be a nonempty correspondence called the *feasible correspondence*;
- $g : X \times A \to X$ be a function called the *law of motion*;
- $\mathcal{V}$ be a subset of all functions from X to $\mathbb{R} \cup \{-\infty\}$ called the set of *candidate value functions;*
- $Q : X \times A \times \mathcal{V} \to \mathbb{R} \cup \{-\infty\}$ be a map called the *state-action aggregator*.

Then we say that the value function $v \in \mathcal{V}$ satisfies the Bellman equation if

$$v(x) = \max_{a \in \Gamma(x)} Q(x, a, v(g(x, a))) \tag{A.1}$$

for all $x \in X$.

Definition A.1 (Asymptotic homogeneity)**.** We say that the dynamic programming problem is *asymptotically homogeneous* if has the following properties:

- $X = X_1 \times X_2$, where $\mathbb{R}_+ \subset X_1 \subset \mathbb{R}$;
- $\Gamma(x) = \Gamma_1(x_1, x_2) \times \Gamma_2(x_2)$, where $x = (x_1, x_2) \in X_1 \times X_2$ and $\mathbb{R}^d_+ \subset \Gamma_1(x_1, x_2) \subset \mathbb{R}^d$ for some d;
- $g(x, a) = g_1(x_1, x_2, a_1, a_2) \times g_2(x_2, a_2)$, where $x = (x_1, x_2) \in X_1 \times X_2$ and $(a_1, a_2) \in \Gamma_1(x_1, x_2) \times \Gamma_2(x_2)$;
- $\lim_{\lambda\to\infty} \frac{1}{\lambda}\Gamma_1(\lambda x_1, x_2) = \tilde{\Gamma}_1(x_1, x_2)$ exists for $(x_1, x_2) \in X_1 \times X_2$;
- $\lim_{\lambda\to\infty} \frac{1}{\lambda} g_1(\lambda x_1, x_2, \lambda a_1, a_2) = \tilde{g}_1(x_1, x_2, a_1, a_2)$ exists for $(x_1, x_2) \in X_1 \times X_2$ and $(a_1, a_2) \in \Gamma_1(x_1, x_2) \times \Gamma_2(x_2)$;
- $\lim_{\lambda\to\infty} \frac{1}{\lambda} Q(\lambda x_1, x_2, \lambda a_1, a_2, \lambda v) = \tilde{Q}(x_1, x_2, a_1, a_2, v)$ exists.

Lemma A.2. *Suppose that the dynamic programming problem is asymptotically homogeneous. Then*

(i) $\tilde{\Gamma}_1$ *is homogeneous of degree 1 in* x_1*: for any* $\lambda > 0$ *we have*

$$\tilde{\Gamma}_1(\lambda x_1, x_2) = \lambda \tilde{\Gamma}_1(x_1, x_2).$$

(ii) $\tilde{g}_1$ *is homogeneous of degree 1 in* (x_1, a_1)*: for any* $\lambda > 0$ *we have*

$$\tilde{g}_1(\lambda x_1, x_2, \lambda a_1, a_2) = \lambda \tilde{g}_1(x_1, x_2, a_1, a_2).$$

(iii) $\tilde{Q}$ *is homogeneous of degree 1 in* (x_1, a_1, v)*: for any* $\lambda > 0$ *we have*

$$\tilde{Q}(\lambda x_1, x_2, \lambda a_1, a_2, \lambda v) = \lambda \tilde{Q}(x_1, x_2, a_1, a_2, v).$$

Proof. By the definition of $\tilde{\Gamma}_1$, for any $\lambda > 0$ we have

$$\begin{aligned}\tilde{\Gamma}_1(\lambda x_1, x_2) &= \lim_{\lambda'\to\infty} \frac{1}{\lambda'}\Gamma_1(\lambda'\lambda x_1, x_2) \\ &= \lambda \lim_{\lambda'\to\infty} \frac{1}{\lambda'\lambda}\Gamma_1(\lambda'\lambda x_1, x_2) = \lambda \tilde{\Gamma}_1(x_1, x_2).\end{aligned}$$

The proofs of the other claims are similar. □

When the dynamic programming problem is asymptotically homogeneous, we define the asymptotic problem as follows.

Definition A.3. Suppose that the dynamic programming problem is asymptotically homogeneous. Then the Bellman equation of the asymptotic problem corresponding to (A.1) is defined by

$$v(x_1, x_2) = \max_{(a_1,a_2)\in\tilde{\Gamma}_1(x_1,x_2)\times\Gamma_2(x_2)} \tilde{Q}(x_1, x_2, a_1, a_2, v(\tilde{g}_1(x_1, x_2, a_1, a_2), g_2(x_2, a_2))). \tag{A.2}$$

The following lemma shows that we can reduce the dimension of the asymptotic problem by 1.

Lemma A.4. *Suppose that the dynamic programming problem is asymptotically homogeneous. Consider the following "normalized" Bellman equation:*

$$\tilde{v}(x_2) = \max_{(a_1,a_2)\in\tilde{\Gamma}_1(1,x_2)\times\Gamma_2(x_2)} \tilde{Q}(1,x_2,a_1,a_2,\tilde{g}_1(1,x_2,a_1,a_2)\tilde{v}(g_2(x_2,a_2))). \tag{A.3}$$

If (A.3) *has a solution $\tilde{v}(x_2)$, then $v(x_1,x_2) = x_1\tilde{v}(x_2)$ is a solution to the asymptotic Bellman equation* (A.2)*. Furthermore, letting $\tilde{a} = (\tilde{a}_1,\tilde{a}_2)$ be the policy function of the normalized Bellman equation* (A.3)*, the policy function $a = (a_1,a_2)$ of the asymptotic Bellman equation* (A.2) *is given by $a_1(x_1,x_2) = x_1\tilde{a}_1(x_2)$ and $a_2(x_1,x_2) = \tilde{a}_2(x_2)$.*

Proof. Immediate by multiplying both sides of (A.3) by $x_1 > 0$ and using the homogeneity of $\tilde{\Gamma}_1, \tilde{g}_1, \tilde{Q}$ established in Lemma A.2. □

The following proposition shows that if a dynamic programming problem is asymptotically homogeneous, then the value function and policy functions are asymptotically linear.

Proposition A.5. *Suppose that the dynamic programming problem is asymptotically homogeneous. Suppose that the Bellman equation* (A.1) *has a solution $v(x)$, and it can be computed by value function iteration starting from $v(x) \equiv 0$. Then under some regularity conditions, the value function and policy functions are asymptotically linear: we have*

$$\begin{aligned} v(x_1,x_2) &= x_1\tilde{v}(x_2) + o(x_1), \\ a_1(x_1,x_2) &= x_1\tilde{a}_1(x_2) + o(x_1), \\ a_2(x_1,x_2) &= \tilde{a}_2(x_2) + o(x_1) \end{aligned}$$

as $x_1 \to \infty$, where $\tilde{v}(x_2)$, $\tilde{a}_1(x_2)$, and $\tilde{a}_2(x_2)$ are defined as in the normalized Bellman equation (A.3)*.*

Proof. Define the operator $T : \mathcal{V} \to \mathcal{V}$ by the right-hand side of (A.1). Let $v^{(0)} \equiv 0$ and $v^{(k)} = Tv^{(k-1)} = T^k 0$. Let us show by induction that

$$\lim_{\lambda\to\infty} \frac{1}{\lambda} v^{(k)}(\lambda x_1, x_2) = \tilde{v}^{(k)}(x_1,x_2)$$

exists. If $k = 0$, the claim is trivial since $v^{(0)} \equiv 0$. Suppose the claim holds for some $k-1$. Then by Lemma A.2, we obtain

$$\begin{aligned} \frac{1}{\lambda} v^{(k)}(\lambda x_1,x_2) &= \frac{1}{\lambda}(Tv^{(k-1)})(\lambda x_1,x_2) \\ &= \max_{\substack{(a_1,a_2)\in \\ \frac{1}{\lambda}\Gamma_1(\lambda x_1,x_2)\times\Gamma_2(x_2)}} Q\left(\lambda x_1, x_2, \lambda a_1, a_2, v^{(k-1)}\left(\lambda \frac{1}{\lambda} g_1(\lambda x_1, x_2, \lambda a_1, a_2), g_2(x_2,a_2)\right)\right). \end{aligned}$$

Using the asymptotic homogeneity of Γ_1, g_1, Q established in Lemma A.2, the asymptotic homogeneity of $v^{(k-1)}$, and assuming that we can interchange the limit and maximization (e.g., assuming enough conditions to apply the Maximum Theorem), it follows that $v^{(k)}$ is asymptotically homogeneous. Since by assumption $v^{(k)} \to v$ as $k \to \infty$ point-wise, assuming that the limit of $k \to \infty$ and $\lambda \to \infty$ can be interchanged (which is the case if $v^{(k)}$ converges to v monotonically, which is often the case in particular applications), then v is asymptotically homogeneous in the sense that $\lim_{\lambda\to\infty} \frac{1}{\lambda} v(\lambda x_1, x_2)$ exists.

Now that asymptotic homogeneity of v is established, from (A.1) we obtain

$$v(\lambda x_1, x_2) = \max_{a \in \Gamma(\lambda x_1, x_2)} Q(\lambda x_1, x_2, a, v(g(\lambda x_1, x_2, a))).$$

Dividing both sides by $\lambda > 0$ and letting $\lambda \to \infty$, using the asymptotic homogeneity of Γ_1, g_1, Q, and v, we obtain the asymptotic Bellman equation (A.2). Thus if in particular (A.3) has a unique solution $\tilde{v}(x_2)$, by Lemma A.4 it must be

$$\lim_{\lambda\to\infty} \frac{1}{\lambda} v(\lambda x_1, x_2) = x_1 \tilde{v}(x_2).$$

Consequently, setting $x_1 = 1$ and $\lambda = x_1$, we obtain $v(x_1, x_2) = x_1 \tilde{v}(x_2) + o(x_1)$. The proof for the policy functions is similar. □

B Proofs

B.1 Proof of results in Section 3

We use the following notations. Let $\tilde{\beta} = \beta(1-p)$ be the effective discount factor. For a vector $v = (v_1, \dots, v_S)'$, let $v^{(\alpha)} = (v_1^\alpha, \dots, v_S^\alpha)'$ be the vector of α-th powers and $\operatorname{diag}(v)$ the diagonal matrix whose s-th diagonal element is v_s.

The following proposition characterizes the solution to the capitalist's optimal consumption-savings problem.

Proposition B.1. *Let $z = (z_1, \dots, z_S)'$ be the vector of gross excess returns. A solution to the optimal consumption-savings problem exists if and only if*

$$\tilde{\beta} R_f^{1-\gamma} \rho(\operatorname{diag}(z^{(1-\gamma)}) P) < 1. \tag{B.1}$$

Under this condition, the value function and optimal consumption rule are

$$V_s(w) = b_s \frac{w^{1-\gamma}}{1-\gamma}, \tag{B.2a}$$

$$c_s(w) = b_s^{-1/\gamma} w, \tag{B.2b}$$

where $b = (b_1, \dots, b_S)' \gg 0$ is the smallest solution to the system of nonlinear equations

$$b_s = \left(1 + (\tilde{\beta}(z_s R_f)^{1-\gamma} \operatorname{E}[b_{s'} \mid s])^{1/\gamma}\right)^\gamma, \quad s = 1, \dots, S. \tag{B.3}$$

Proof. Immediate from Toda (2018b, Proposition 1). □

Let us simplify the equilibrium condition (3.3) by exploiting the linearity in Proposition B.1. Using the budget constraint (3.2) and the optimal consumption rule (B.2b), the individual wealth dynamics is

$$w' = z_s R_f(1 - b_s^{-1/\gamma})w =: G_s w.$$

Letting W_s be the aggregate wealth held by agents in state s, by accounting we obtain

$$W_{s'} = p\pi_{s'}w_0 + (1-p)\sum_{s=1}^{S} p_{ss'}G_s W_s.$$

Letting $\pi = (\pi_1, \dots, \pi_S)'$, $G = (G_1, \dots, G_S)'$, and $W = (W_1, \dots, W_S)'$, in matrix form this becomes

$$W = pw_0\pi + (1-p)P'(\operatorname{diag} G)W \iff W = pw_0(I - (1-p)P' \operatorname{diag} G)^{-1}\pi.$$

Let $m_s = b_s^{-1/\gamma} \in (0,1)$ be the marginal propensity to consume out of wealth in state s and $m = (m_1, \dots, m_S)'$. Then the vector of saving rates is given by $1 - m$, where $1 = (1, \dots, 1)'$ is the vector of ones. Using this, the aggregate capital supply is given by

$$K = (1-m)'W = pw_0(1-m)'(I - (1-p)P' \operatorname{diag} G)^{-1}\pi, \tag{B.4}$$

assuming $(1-p)\rho(P' \operatorname{diag} G) < 1$. (If this inequality is violated, we just set $K = \infty$.) On the other hand, by (3.1) the aggregate capital demand is

$$K = \left(\frac{R_f - 1 + \delta}{A\alpha}\right)^{\frac{1}{\alpha-1}}. \tag{B.5}$$

Equating (B.4) and (B.5), the market clearing condition (3.3) becomes

$$0 = f(R_f) := pw_0(1-m)'(I - (1-p)P' \operatorname{diag} G)^{-1}\pi - \left(\frac{R_f - 1 + \delta}{A\alpha}\right)^{\frac{1}{\alpha-1}}. \tag{B.6}$$

The following theorem shows that a stationary equilibrium exists and that the stationary wealth distribution has a Pareto upper tail.

Theorem B.2. *A stationary equilibrium exists if and only if there exists $\underline{R} > 1 - \delta$ such that*

$$\tilde{\beta}\underline{R}^{1-\gamma}\rho(\operatorname{diag}(z^{(1-\gamma)})P) < 1 \tag{B.7}$$

and $f(\underline{R}) < 0$, where f is given by (B.6). The equilibrium is unique if $\gamma < 1$. If in addition $p_{ss} > 0$ for all s and $G_s > 1$ for some s, then the stationary wealth distribution has a Pareto upper tail with exponent $\zeta > 1$ that satisfies

$$\rho(P \operatorname{diag} G^{(\zeta)}) = \frac{1}{1-p}. \tag{B.8}$$

Proof. The existence of equilibrium follows from a continuity argument similar to Toda (2018b, Theorem 3), which we only sketch for space considerations. The condition (B.7) ensures that (B.1) holds for $R_f > \underline{R}$ sufficiently close to $\underline{R}$. Then we can show that the individual optimization problem has a solution and the aggregate wealth is finite for some range $R_f \in [\underline{R}, \bar{R})$, and that the aggregate wealth (as well as supply of capital) diverges to ∞ as $R_f \uparrow \bar{R}$. Since $f(\underline{R}) < 0$ by assumption and $f(\bar{R}) = \infty$, by the intermediate value theorem there exists $R_f \in (\underline{R}, \bar{R})$ that satisfies the market clearing condition (B.6). Uniqueness of equilibrium when $\gamma < 1$ follows by the exact same argument as in Toda (2018b, Theorem 3).

The Pareto tail result follows from the general theorem in Beare and Toda (2017) and a similar argument to Toda (2018b, Theorem 4). □

Numerically solving for the equilibrium is straightforward. Given the guess of the interest rate $R_f > \underline{R}$, solve for the fixed point $b = (b_s)$ using (B.3), and solve (B.6) to obtain the equilibrium risk-free rate.

B.2 Proof of results in Section 4

Proof of Proposition 4.2. The maximization problem (4.12) is equivalent to

$$\max_{0 \le \theta \le \bar{\theta}_s} \frac{1}{1-\gamma} \operatorname{E}\left[(\tilde{R}_f(1-\theta) + \tilde{R}_{sj}\theta)^{1-\gamma} \,\middle|\, s \right].$$

Let $f(\theta)$ be the objective function of this problem. Since by assumption $z_S + \epsilon_1 < 1$ and $z_1 < \cdots < z_S$, we have $z_s + \epsilon_1 < 1$ for all s. Therefore $\bar{\theta}_s = \frac{\tilde{R}_f}{\tilde{R}_f - \tilde{R}_{s1}} > 0$ and

$$f'(\theta) = \operatorname{E}\left[(\tilde{R}_f(1-\theta) + \tilde{R}_{sj}\theta)^{-\gamma} (\tilde{R}_{sj} - \tilde{R}_f) \,\middle|\, s \right] \to -\infty$$

as $\theta \uparrow \bar{\theta}_s$. Furthermore,

$$f''(\theta) = -\gamma \operatorname{E}\left[(\tilde{R}_f(1-\theta) + \tilde{R}_{sj}\theta)^{-\gamma-1} (\tilde{R}_{sj} - \tilde{R}_f)^2 \,\middle|\, s \right] < 0$$

for $\theta \in (0, \bar{\theta}_s)$, so f is strictly concave. Therefore there exists a unique θ_s^* that maximizes (4.12), and hence ρ_s is well-defined.

Assume $\varepsilon \neq 1$. By the discussion in the text, the Bellman equation of the asymptotic problem is

$$v_s(w) = \max_{\substack{0 \le \theta \le \bar{\theta}_s \\ 0 \le c \le w}} \left((1-\beta)c^{1-1/\varepsilon} + \beta \operatorname{E}\left[(v_{s'}(R(\theta)(w-c)))^{1-\gamma} \,\middle|\, s \right]^{\frac{1-1/\varepsilon}{1-\gamma}} \right)^{\frac{1}{1-1/\varepsilon}},$$

where the upper bounds on c, θ ensure that $w' \ge 0$ and

$$R(\theta) = (1-\tau_w)(\tilde{R}_f(1-\theta) + \tilde{R}_{sj}\theta)$$

is the gross portfolio return. By homogeneity, the value function must be of the form $v_s(w) = b_s w$. Substituting into the Bellman equation, we obtain

$$b_s w = \max_{\substack{0 \le \theta \le \bar{\theta}_s \\ 0 \le c \le w}} \left((1-\beta)c^{1-1/\varepsilon} + \beta(w-c)^{1-1/\varepsilon} \operatorname{E}\left[(b_{s'} R(\theta))^{1-\gamma} \,\middle|\, s \right]^{\frac{1-1/\varepsilon}{1-\gamma}} \right)^{\frac{1}{1-1/\varepsilon}}.$$

Noting that $R(\theta)$ does not depend on s' and j is independent of s', using the definition of ρ_s in (4.12), we can rewrite this as

$$b_s w = \max_{0 \le c \le w} \left((1-\beta) c^{1-1/\varepsilon} + \beta \rho_s^{1-1/\varepsilon} (w-c)^{1-1/\varepsilon} \operatorname{E}\left[b_{s'}^{1-\gamma} \,\middle|\, s \right]^{\frac{1-1/\varepsilon}{1-\gamma}} \right)^{\frac{1}{1-1/\varepsilon}}.$$

For notational simplicity let $\kappa_s = \rho_s \operatorname{E}\left[b_{s'}^{1-\gamma} \,\middle|\, s \right]^{\frac{1}{1-\gamma}}$. Then the above problem becomes equivalent to

$$\max_c \frac{1}{1-1/\varepsilon} \left((1-\beta) c^{1-1/\varepsilon} + \beta \kappa_s^{1-1/\varepsilon} (w-c)^{1-1/\varepsilon} \right).$$

Clearly this is a strictly concave function in c. Taking the first-order condition and solving for c, we obtain

$$c = \frac{(1-\beta)^\varepsilon}{(1-\beta)^\varepsilon + \beta^\varepsilon \kappa_s^{\varepsilon-1}} w.$$

Substituting into the Bellman equation, after some algebra we obtain

$$b_s = \left((1-\beta)^\varepsilon + \beta^\varepsilon \rho_s^{\varepsilon-1} \operatorname{E}\left[b_{s'}^{1-\gamma} \,\middle|\, s \right]^{\frac{\varepsilon-1}{1-\gamma}} \right)^{\frac{1}{\varepsilon-1}},$$

which is (4.14). The optimal consumption rule then simplifies to $c = (1-\beta)^\varepsilon b_s^{1-\varepsilon}$ and we obtain the optimal investment rule using $\theta = \frac{I}{w-c}$.

To complete the proof it remains to show that the system of nonlinear equations (4.14) has a solution. For this purpose let us write $\sigma = \frac{1-\gamma}{\varepsilon-1}$ and $x_s = b_s^{1-\gamma}$. Then we can rewrite (4.14) as

$$x_s = \left((1-\beta)^\varepsilon + (\beta^{\varepsilon\sigma} \rho_s^{1-\gamma} \operatorname{E}[x_{s'} \mid s])^{1/\sigma} \right)^\sigma,$$

which is equivalent to

$$x = ((1-\beta)^\varepsilon + (Kx)^{1/\sigma})^\sigma$$

for $x = (x_1, \dots, x_S)'$ and $K = \beta^{\varepsilon\sigma} \operatorname{diag}\left(\rho_1^{1-\gamma}, \dots, \rho_S^{1-\gamma} \right) P$. Since this equation is identical to Equation (12) in Borovička and Stachurski (2017), by their Theorem 2.1, a necessary and sufficient condition for the existence of a unique fixed point is $\rho(K)^{1/\sigma} < 1$, which is equivalent to (4.13).

Finally we briefly comment on the case $\varepsilon = 1$. Although this case requires a separate treatment, it turns out that the equations are valid by taking the limit $\varepsilon \to 1$. To show (4.14), define $g(\varepsilon) = \log((1-\beta)^\varepsilon + \beta^\varepsilon \kappa^{\varepsilon-1})$ for $\kappa > 0$. Then as $\varepsilon \to 1$ we obtain

$$\log((1-\beta)^\varepsilon + \beta^\varepsilon \kappa^{\varepsilon-1})^{\frac{1}{\varepsilon-1}} = \frac{g(\varepsilon)}{\varepsilon-1} = \frac{g(\varepsilon)-g(1)}{\varepsilon-1} \to g'(1).$$

But since

$$g'(\varepsilon) = \frac{(1-\beta)^\varepsilon \log(1-\beta) + \beta^\varepsilon \kappa^{\varepsilon-1} \log(\beta\kappa)}{(1-\beta)^\varepsilon + \beta^\varepsilon \kappa^{\varepsilon-1}} \to (1-\beta)\log(1-\beta) + \beta \log(\beta\kappa)$$

as $\varepsilon \to 1$, it follows that

$$((1-\beta)^{\varepsilon} + \beta^{\varepsilon}\kappa^{\varepsilon-1})^{\frac{1}{\varepsilon-1}} \to (1-\beta)^{1-\beta}(\beta\kappa)^{\beta},$$

which explains (4.14) for $\varepsilon = 1$. The existence and uniqueness of a positive solution can be proved by taking the logarithm of (4.14) and applying a contraction mapping argument to $x = \log b$. □

Proof of Proposition 4.3. By (4.12), θ_s^* maximizes

$$\frac{1}{1-\gamma} \operatorname{E}\left[\left(\tilde{R}_f(1-\theta) + \tilde{R}_{sj}\theta\right)^{1-\gamma} \,\middle|\, s\right].$$

Using (4.4), the gross portfolio return is

$$\begin{aligned}\tilde{R}_f(1-\theta) + \tilde{R}_{sj}\theta &= (1-\tau_w)(1+(1-\tau_k)((R_f-1)(1-\theta)+(R_{sj}-1)\theta)) \\ &= (1-\tau_w)\tau_k(1+\lambda(z_{sj}-1)\theta),\end{aligned}$$

where $\lambda = R_f(1/\tau_k - 1) > 0$. For notational simplicity, let us suppress s, j and write $z_{sj} = z_s + \epsilon_j = z + \epsilon$. Since the objective function is homogeneous of degree $1-\gamma$ in the gross portfolio return, the optimal portfolio maximizes

$$f(\theta) := \frac{1}{1-\gamma}\operatorname{E}[(1+\lambda(z+\epsilon-1)\theta)^{1-\gamma}].$$

Since $f(\theta)$ depends only on λ, z, which are independent of τ_w, the optimal portfolio $\theta = \theta_s^*$ is independent of the wealth tax rate τ_w. Furthermore, clearly $f(\theta)$ is strictly concave and

$$f'(0) = \operatorname{E}[\lambda(z+\epsilon-1)] = \lambda(z-1),$$

so $f'(0) > 0$ if and only if $z > 1$. Therefore f attains the maximum at $\theta > 0$ if $z > 1$ and $\theta = 0$ otherwise.

Suppose $z > 1$, and hence $\theta > 0$. By the first-order condition, we obtain

$$f'(\theta) = \operatorname{E}[(1+\lambda(1+(z+\epsilon-1)\theta))^{-\gamma}\lambda(z+\epsilon-1)] = 0. \tag{B.9}$$

Define

$$F(\theta,\lambda,z) = \operatorname{E}[(1+\lambda(1+(z+\epsilon-1)\theta))^{-\gamma}(z+\epsilon-1)]$$

and $X = 1+\lambda(1+(z+\epsilon-1)\theta) > 0$. Then

$$\frac{\partial F}{\partial \theta} = -\gamma\lambda\operatorname{E}[X^{-\gamma-1}(z+\epsilon-1)^2] < 0.$$

Using the first-order condition (B.9), we obtain

$$\begin{aligned}\frac{\partial F}{\partial \lambda} &= -\gamma \operatorname{E}[X^{-\gamma-1}(1+(z+\epsilon-1)\theta)(z+\epsilon-1)] \\ &= -\frac{\gamma}{\lambda}\operatorname{E}[X^{-\gamma-1}(X-1)(z+\epsilon-1)] && (\because \text{definition of } X) \\ &= \frac{\gamma}{\lambda}\operatorname{E}[X^{-\gamma-1}(z+\epsilon-1)]. && (\because \text{(B.9)})\end{aligned}$$

Since $\theta > 0$, in states such that $z+\epsilon > 1$, we have $X = 1+\lambda(1+(z+\epsilon-1)\theta) > 1+\lambda$. Similarly, we have $X < 1+\lambda$ if $z+\epsilon < 1$. Therefore

$$\left(\frac{X}{1+\lambda}\right)^{-\gamma-1}(z+\epsilon-1) \leq \left(\frac{X}{1+\lambda}\right)^{-\gamma}(z+\epsilon-1)$$

always, with strict inequality if $z+\epsilon \neq 1$. Taking the expectation of both sides, multiplying by $(1+\lambda)^{-\gamma-1}$, and using (B.9), we obtain

$$\operatorname{E}[X^{-\gamma-1}(z+\epsilon-1)] < 0, \tag{B.10}$$

and hence $\partial F/\partial\lambda < 0$. By the Implicit Function Theorem, we obtain

$$\frac{\partial\theta}{\partial\lambda} = -\frac{\partial F/\partial\lambda}{\partial F/\partial\theta} < 0.$$

Since $\lambda = R_f(1/\tau_k - 1)$ and $0 \leq \tau_k < 1$, it follows that $\partial\theta_s^*/\partial R_f < 0$ and $\partial\theta_s^*/\partial\tau_k > 0$. To derive the comparative statics with respect to z, using the chain rule and (B.10), we obtain

$$\frac{\partial F}{\partial z} = -\gamma\lambda\theta\operatorname{E}[X^{-\gamma-1}(z+\epsilon-1)] + \operatorname{E}[X^{-\gamma}] > 0.$$

Again by the Implicit Function Theorem, we obtain

$$\frac{\partial\theta}{\partial z} = -\frac{\partial F/\partial z}{\partial F/\partial\theta} > 0. \quad \square$$

Proof of Lemma 4.4. By Proposition 4.2 and the budget constraint of the asymptotic problem, we obtain the the law of motion

$$w' = (1-(1-\beta)^{\varepsilon}b_s^{1-\varepsilon})(\tilde{R}_f(1-\theta_s^*) + \tilde{R}_{sj}\theta_s^*)w.$$

Taking the expectation conditional on s and using the definition of G_s in (4.16), we obtain $\operatorname{E}[w' \mid s] = G_s w$. By the same derivation as (B.4), a necessary condition for aggregate wealth to be finite is $\rho(P' \operatorname{diag} G) < 1$. Since $\operatorname{diag} G$ is diagonal, we obtain

$$\rho(P \operatorname{diag} G) = \rho((\operatorname{diag} G)'P') = \rho((\operatorname{diag} G)P') = \rho(P' \operatorname{diag} G) < 1. \quad \square$$

Proof of Lemma 4.5. Since by assumption $G_{sJ} > 1$ for some s, we have $M_s(z) \to \infty$ as $z \to \infty$ for this s. Since by assumption $p_{ss} > 0$ for all s, it follows that $\rho(PD(z)) \to \infty$ as $z \to \infty$. Since $D(1) = \operatorname{diag} G$, by (4.17) we obtain $\rho(PD(1)) = \rho(P \operatorname{diag} G) < 1$. By the intermediate value theorem, there exists $\zeta > 1$ such that $\rho(PD(\zeta)) = 1$. Uniqueness follows from the convexity of $\rho(PD(z))$ established in Beare and Toda (2017). The Pareto tail result follows from (2.5) with $p = 0$. □

C Solution accuracy of Aiyagari model

In this appendix we evaluate the solution accuracy of the Aiyagari model in Section 3.1. We consider the evenly- and exponentially-spaced grids as well as simulation.

C.1 Evenly-spaced grid

We first consider an N-point evenly-spaced grid on $(0, \bar{w}]$, where we set the truncation point to $\bar{w} = 10, 20, 40$ and the number of points to $N = 100, 200, 400$. Therefore the wealth grid is $\{nd\}_{n=1}^{N}$, where $d = \bar{w}/N$ is the distance between grid points.[19] Table 14 shows the relative error $\widehat{K}/K - 1$ in the aggregate capital.

Table 14: Relative error (%) in aggregate capital for the truncation and Pareto extrapolation methods with an evenly-spaced grid.

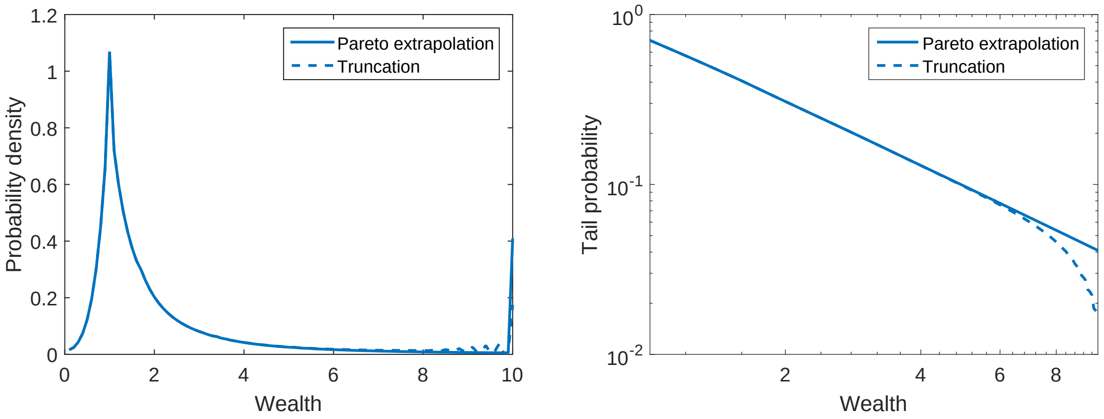
Method:	Truncation			Pareto extrapolation		
$\bar{w}$	$N = 100$	200	400	100	200	400
10	-40.00	-39.69	-39.62	0.214	0.105	0.052
20	-33.58	-32.88	-32.63	0.430	0.097	0.043
40	-26.99	-27.56	-27.03	3.588	0.331	0.046

Note: N: number of grid points; $\bar{w}$: wealth truncation point.

We can make a few observations from Table 14. First, the conventional truncation method is extremely poor at calculating the aggregate capital: the relative error is about 27–40% depending on the specification. On the other hand, the Pareto extrapolation method is astonishingly more accurate. Second, for the truncation method, choosing a larger truncation point $\bar{w}$ somewhat improves the accuracy, probably because it misses less of the upper tail. On the other hand, the accuracy in the Pareto extrapolation method is not necessarily monotonic in $\bar{w}$. There seems to be a trade-off between less truncation (larger $\bar{w}$) and higher density of grid points (smaller $d = \bar{w}/N$).

Figure 8a shows the stationary wealth distribution using $\bar{w} = 10$ and $N = 100$. The two methods are indistinguishable except at the upper tail. To study the tail behavior, Figure 8b plots the tail probability in a log-log scale. As we can see, the graphs show a straight line pattern, which is consistent with the theoretical Pareto distribution. However, the graph for the truncation method becomes concave towards the upper tail, which implies that it underestimates the tail probability. On the other hand, the Pareto extrapolation method shows a straight line pattern including the very top tail.

[19] Note that we exclude 0 from the wealth grid because in equilibrium agents never hit 0 due to the Inada condition.

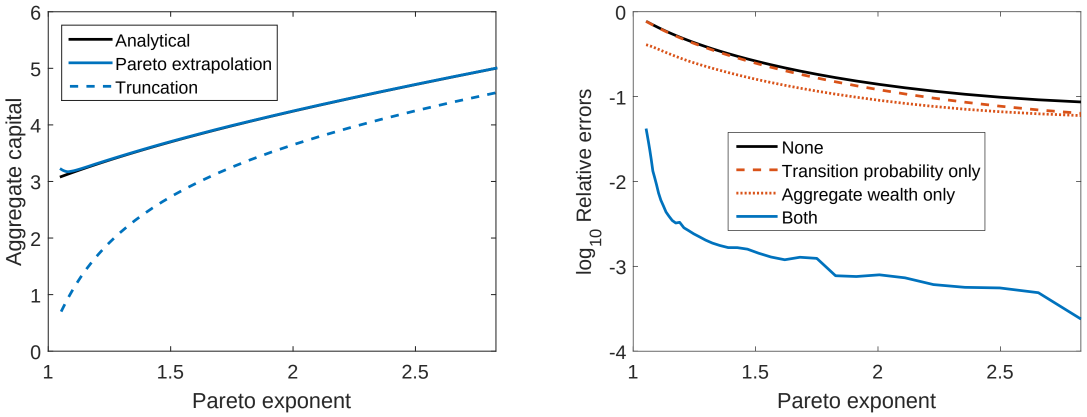


(a) Stationary wealth distribution.

(b) Log-log plot of wealth distribution.

Figure 8: Wealth distribution in the Aiyagari model.

These seemingly small differences have an enormous impact on aggregate quantities, as we have seen in Table 14. To assess the robustness, Figure 9a shows the aggregate capital for the semi-analytical solution as well as the Pareto extrapolation and truncation solutions when we change the initial wealth in the range $w_0 \in [0.2, 5]$. For all cases we set $\bar{w} = 10$ and $N = 100$. The horizontal axis shows the corresponding equilibrium Pareto exponent. The graph for the Pareto extrapolation method is indistinguishable from the semi-analytical solution except when the Pareto exponent is very close to 1 (Zipf's law), in which case the truncation method is especially poor.



(a) Aggregate capital.

(b) Relative errors.

Figure 9: Solution accuracy of the Aiyagari model.

Note: "None", "Transition probability only", "Aggregate wealth only", and "Both" refer to (i) not using Pareto extrapolation and using it to (ii) constructing the transition probability matrix only as in Section 2.2.2, (iii) calculating aggregate wealth only as in Section 2.2.3, and (iv) both, respectively.

Figure 9b shows the relative error $\left|\widehat{K}/K - 1\right|$ in a semi log scale. For this exercise, we consider four solution methods that correspond to using/not using Pareto extrapolation when constructing the transition probability matrix and/or calculating aggregate capital. For example, "None" and "Both" in Figure 9b correspond to the truncation and Pareto extrapolation solutions, respectively. According to the figure, using Pareto extrapolation for only one step (either constructing the transition probability matrix or calculating aggregate capital) improves the accuracy only slightly, and correcting the aggregate capital matters more. However, combining both increases the solution accuracy dramatically.

The intuition for this (surprising) result is as follows. According to (2.8), the two sources of errors introduced by the truncation method (incorrect transition probability matrix *and* incorrect aggregate wealth held by agents at the top grid point) interact with each other. With the truncation method, the last term $\pi_{sN}\frac{\zeta}{\zeta-1}w_N$ in (2.8) becomes $\tilde{\pi}_{sN}w_N$, with typically $\tilde{\pi}_{sN} < \pi_{sN}$. Therefore, errors in π_{sN} are inflated by a factor $\frac{\zeta}{\zeta-1}$, which is large if $\zeta > 1$ is small.

C.2 Exponentially-spaced grid

One may argue that the poor performance of the truncation method in Table 14 is due to the fact that the truncation point $\bar{w} = 10, 20, 40$ is relatively small compared to the aggregate capital $K = 3.4231$. What if we take $\bar{w}$ much larger, say a million? Then we can no longer use evenly-spaced grids because there will be too few points to cover the bottom of the wealth distribution. Therefore we need to consider an exponentially-spaced grid.

In more general models, the state variable may become negative (e.g., asset holdings), which causes a problem for constructing an exponentially-spaced grid because we cannot take the logarithm of a negative number. Suppose we would like to construct an N-point exponential grid on a given interval (a, b). A natural idea to deal with such a case is as follows.

Constructing the exponential grid.

1. Choose a shift parameter $s > -a$.
2. Construct an N-point evenly-spaced grid on $(\log(a+s), \log(b+s))$.
3. Take the exponential.
4. Subtract s.

The remaining question is how to choose the shift parameter s. Suppose we would like to specify the median grid point as $c \in (a, b)$. Since the median of the evenly-spaced grid on $(\log(a+s), \log(b+s))$ is $\frac{1}{2}(\log(a+s) + \log(b+s))$, we need to take $s > -a$ such that

$$
\begin{aligned}
c &= \exp\left(\frac{1}{2}(\log(a+s) + \log(b+s))\right) - s \\
\iff & c + s = \sqrt{(a+s)(b+s)} \\
\iff & (c+s)^2 = (a+s)(b+s) \\
\iff & c^2 + 2cs + s^2 = ab + (a+b)s + s^2 \\
\iff & s = \frac{c^2 - ab}{a + b - 2c}.
\end{aligned}
$$

Note that in this case

$$s + a = \frac{c^2 - ab}{a + b - 2c} + a = \frac{(c - a)^2}{a + b - 2c},$$

so $s + a$ is positive if and only if $c < \frac{a+b}{2}$. Therefore, for any $c \in \left(a, \frac{a+b}{2}\right)$, it is possible to construct such a grid. The remaining question is how to choose c, but we can use information from the problem we want to solve. Note that by construction, half the grid points will lie on the interval (a, c). Therefore we should choose the number c such that c is a "typical" value for the state variable (e.g., initial capital, aggregate capital in a representative-agent model, etc.).

Returning to our particular Aiyagari model, we choose the shift parameter s such that the median grid point corresponds to the capital in a representative-agent model, which in our case is $K_{\text{RA}} = ((1/(\beta(1 - p)) - 1 + \delta)/(A\alpha))^{\frac{1}{\alpha - 1}} = 4.5577$. Since an exponential grid allows us to use far fewer points than an evenly-spaced grid, we consider $N = 25, 50, 100$. (We have checked that increasing N further also increases the accuracy.) For the truncation point, we consider $\bar{w} = 10^1, 10^2, 10^3, 10^4, 10^5, 10^6$. Table 15 shows the results.

Table 15: Relative error (%) in aggregate capital for the truncation and Pareto extrapolation methods with an exponentially-spaced grid.

Method:	Truncation			Pareto extrapolation		
$\bar{w}$	$N = 25$	50	100	25	50	100
10^1	-41.90	-40.77	-39.92	2.643	0.528	0.276
10^2	-27.54	-23.36	-21.08	-1.697	-0.186	0.588
10^3	-20.85	-15.27	-12.32	-2.423	-0.785	-0.002
10^4	-17.12	-10.78	-7.62	-2.553	-0.905	-0.184
10^5	-14.93	-8.06	-4.93	-2.546	-0.879	-0.222
10^6	-13.19	-6.27	-3.32	-2.445	-0.807	-0.210

Note: N: number of grid points; $\bar{w}$: wealth truncation point.

The accuracy of the truncation method somewhat improves by using the exponential grid with a large truncation point $\bar{w}$ and many points. However, even with a large truncation like a million and $N = 100$ grid points, the error still exceeds 3%. (Increasing N further decreases the error, but only slightly for the truncation method.) Again, the Pareto extrapolation method is overwhelmingly more accurate.

C.3 Simulation

We conduct simulations by recursively computing the wealth of I agents over T periods using the semi-analytical solution for the consumption policies and the risk-free rate. We initialize the wealth distribution at $t = 0$ by setting $w_0 = 1$ for all agents. For every simulation, we compute the cross-sectional mean of the wealth distribution $\widehat{K}$ in the terminal period T, which corresponds to the aggregate capital K in the model.[20] Given that $\widehat{K}$ is a random variable, we compute the relative error for $B = 10{,}000$ independent simulations and take the average to obtain the "mean relative error" defined by $\frac{1}{B} \sum_{b=1}^{B} \left| \widehat{K}_b / K - 1 \right|$, where K is the true aggregate

[20] Since there is a unit continuum of agents in the model, the average wealth is equal to the aggregate wealth.

capital and $\widehat{K}_b$ is the numerical aggregate capital from simulation b. Finally, we compute a measure of the dispersion of results across simulations defined by $\left|\widehat{K}_{p95}/\widehat{K}_{p5} - 1\right|$, where $\widehat{K}_{p5}, \widehat{K}_{p95}$ denote the 5th and 95th percentiles of $\widehat{K}$ across simulations. Table 16 shows the results for different combinations of sample size I and simulation length T.

Table 16: Solution accuracy of the simulation method in the Aiyagari model.

T	I	Relative error (%)	Dispersion (%)	Time (sec)
1,000	10^3	21.93	92.97	0.07
	10^4	10.97	47.99	0.38
	10^5	6.70	27.53	3.13
	10^6	6.69	16.33	41.23
10,000	10^3	23.76	96.68	0.39
	10^4	20.42	50.07	3.76
	10^5	9.69	27.47	31.43
	10^6	6.64	16.07	414.30

Note: T: simulation length in model-years; I: number of simulated agents; Relative error: $\frac{1}{B}\sum_{b=1}^{B}\left|\widehat{K}_b/K - 1\right|$, where $\widehat{K}_b$ is the aggregate capital in simulation b and K is the true value from the analytical solution; Dispersion: $\left|\widehat{K}_{p95}/\widehat{K}_{p5} - 1\right|$, where $\widehat{K}_{p5}, \widehat{K}_{p95}$ are the 5th and 95th percentiles of aggregate capital across $B = 10{,}000$ simulations; Time: computing time of one simulation (one b) in seconds.

A few remarks are in order. First, the simulation method performs poorly on average and the results are very dispersed across simulations. Second, increasing the number of agents I helps reduce both the mean relative error and the dispersion. The gains in accuracy associated with increasing I by an order of magnitude tend to be small, consistent with the fact that the sample mean of a fat-tailed distribution converges very slowly to the population mean. In fact, in our model the Pareto exponent is $\zeta = 1.28$, and the relative errors for sample sizes $I = 10^4, 10^6$ in Table 16, which are about 11% and 7% respectively, are about the same order of magnitude as the error order 11.9% and 4.1% in the column for $\zeta = 1.3$ in Table 1. Third, increasing the simulation length beyond $T = 1{,}000$ does not seem to improve accuracy, suggesting that the simulated wealth distribution has already converged after 1,000 periods.

The last column of Table 16 reports the computing time (without parallelization) associated with producing a *single* simulation using a machine equipped with an `Intel Xeon E3-1245 3.5GHz` processor and 16GB of memory. While increasing the sample size I by a factor of ten is associated with small accuracy gains, it implies a tenfold increase in the computing time. Given the inaccuracy and the large dispersion of results across simulations, we conclude that the simulation method is not a viable option for solving models with fat-tailed wealth distributions.

D Algorithm for Merton-Bewley-Aiyagari model

D.1 Euler and asset pricing equations

First, we derive the Euler and asset pricing equations. Noting that $\tilde{R}_{sj}$ is strictly increasing in j, the borrowing constraint (4.5) can bind only in state $j = 1$. Therefore the Bellman equation

(4.6) is equivalent to

$$\frac{1}{1-1/\varepsilon} v_s(w)^{1-1/\varepsilon} = \max_{c,I\geq 0} \frac{1}{1-1/\varepsilon}\left((1-\beta)c^{1-1/\varepsilon} + \beta\,\mathrm{E}\left[v_{s'}(w')^{1-\gamma} \,\middle|\, s\right]^{\frac{1-1/\varepsilon}{1-\gamma}}\right), \tag{D.1}$$

where

$$w' = \tilde{R}_f(w + (1-\tau_h)\omega h_s - I - c) + \tilde{R}_{s1} I \geq \underline{w}.$$

Let $\mathcal{L}_s(c, I, w)$ be the Lagrangian and $\lambda_s(w), \mu_s(w)$ be the corresponding Lagrange multipliers for the borrowing constraint and the nonnegativity constraint on investment:

$$\begin{aligned}\mathcal{L}_s(c, I, w) = \frac{1}{1-1/\varepsilon}&\left((1-\beta)c^{1-1/\varepsilon} + \beta\,\mathrm{E}\left[v_{s'}(w')^{1-\gamma} \,\middle|\, s\right]^{\frac{1-1/\varepsilon}{1-\gamma}}\right) \\ &+ \lambda_s(w)\left(\tilde{R}_f(w + (1-\tau_h)\omega h_s - I - c) + \tilde{R}_{s1} I - \underline{w}\right) + \mu_s(w) I.\end{aligned}$$

The first-order condition for consumption is given by

$$(1-\beta)c_s(w)^{-1/\varepsilon} = \beta\tilde{R}_f\,\mathrm{E}\left[v_{s'}(w')^{1-\gamma} \,\middle|\, s\right]^{\frac{\gamma-1/\varepsilon}{1-\gamma}} \mathrm{E}\left[v_{s'}(w')^{-\gamma} v'_{s'}(w') \,\middle|\, s\right] + \lambda_s(w)\tilde{R}_f. \tag{D.2}$$

Differentiating both sides of the Bellman equation (D.1) by w, it follows from the Envelope Theorem that

$$v_s(w)^{-1/\varepsilon} v'_s(w) = \beta\tilde{R}_f\,\mathrm{E}\left[v_{s'}(w')^{1-\gamma} \,\middle|\, s\right]^{\frac{\gamma-1/\varepsilon}{1-\gamma}} \mathrm{E}\left[v_{s'}(w')^{-\gamma} v'_{s'}(w') \,\middle|\, s\right] + \lambda_s(w)\tilde{R}_f. \tag{D.3}$$

By (D.2) and (D.3), we obtain the following expression for the derivative of the value function

$$v'_s(w) = (1-\beta)\left(\frac{c_s(w)}{v_s(w)}\right)^{-1/\varepsilon}. \tag{D.4}$$

Substituting (D.4) into (D.2), we obtain the consumption Euler equation

$$c_s(w)^{-1/\varepsilon} = \beta\tilde{R}_f\,\mathrm{E}\left[\left(\frac{v_{s'}(w')}{\mathrm{E}\left[v_{s'}(w')^{1-\gamma} \mid s\right]^{\frac{1}{1-\gamma}}}\right)^{1/\varepsilon-\gamma} c_{s'}(w')^{-1/\varepsilon} \,\middle|\, s\right] + \frac{\lambda_s(w)\tilde{R}_f}{1-\beta}. \tag{D.5}$$

The first-order condition for investment is given by

$$\begin{aligned}\beta\,\mathrm{E}\left[v_{s'}(w')^{1-\gamma} \,\middle|\, s\right]^{\frac{\gamma-1/\varepsilon}{1-\gamma}} &\mathrm{E}\left[v_{s'}(w')^{-\gamma} v'_{s'}(w')(\tilde{R}_{sj} - \tilde{R}_f) \,\middle|\, s\right] \\ &+ \lambda_s(w)(\tilde{R}_{s1} - \tilde{R}_f) + \mu_s(w) = 0.\end{aligned} \tag{D.6}$$

Using the expression for $v'_s(w)$ from (D.4) and rearranging, we obtain the following asset pricing equation

$$\mathrm{E}\left[\left(\frac{v_{s'}(w')}{\mathrm{E}\left[v_{s'}(w')^{1-\gamma}\mid s\right]^{\frac{1}{1-\gamma}}}\right)^{1/\varepsilon-\gamma} c_{s'}(w')^{-1/\varepsilon}(\tilde{R}_{sj}-\tilde{R}_f)\,\middle|\, s\right] = -\frac{\lambda_s(w)(\tilde{R}_{s1}-\tilde{R}_f)+\mu_s(w)}{\beta(1-\beta)}. \tag{D.7}$$

D.2 Policy function and value function iteration

First, we choose an exponential grid as described in Appendix C.2 and replacing the bottom half grid points with an evenly-spaced grid. The grid has 100 points and we set the truncation point to 1,000 times the typical scale of the model (steady-state capital stock in the representative-agent model with the same parametrization). The strategy is start with guesses for the value function $v_s^{\text{old}}(w)$ and policy functions $c_s^{\text{old}}(w)$, $I_s^{\text{old}}(w)$ and update by doing the following steps for each individual state (s,w). Equipped with the asymptotic solution to the individual problem $\{\bar{c}_s,\bar{I}_s,\bar{v}_s\}_{s=1}^S$, we choose the following initial guesses

$$c_s(w)=\bar{c}_s(w-\underline{w}), \quad I_s(w)=\bar{I}_s(w-\underline{w}), \quad v_s(w)=\bar{I}_s(w-\underline{w}).$$

Then, we iterate over steps (1), (2), and (3) until convergence.

1. **Consumption decision**: First, use the old decision rules $c_s^{\text{old}}(w)$, $I_s^{\text{old}}(w)$ to construct next period's wealth $w'_{sj}(w)$ and the upper bound on consumption $c_s^{\text{ub}}(w)$:

   $$\begin{aligned} w'_{sj}(w) &= \tilde{R}_f\left(w+(1-\tau_h)\omega h_s - I_s^{\text{old}}(w) - c_s^{\text{old}}(w)\right)+\tilde{R}_{sj}I_s^{\text{old}}(w),\\ c_s^{\text{ub}}(w) &= w+(1-\tau_h)\omega h_s - \underline{w}/\tilde{R}_f. \end{aligned}$$

   Then, update the consumption using the Euler equation (D.5) as if the borrowing constraint was slack.

   $$c_s^*(w)=\left(\beta\tilde{R}_f\,\mathrm{E}\left[\left(\frac{v_{s'}^{\text{old}}(w'_{sj}(w))}{\mathrm{E}\left[v_{s'}^{\text{old}}(w'_{sj}(w))^{1-\gamma}\,\middle|\, s\right]^{\frac{1}{1-\gamma}}}\right)^{1/\varepsilon-\gamma} c_{s'}^{\text{old}}(w'_{sj}(w))^{-1/\varepsilon}\,\middle|\, s\right]\right)^{-\varepsilon}.$$

   Finally, impose the upper bound on consumption so that $c_s^{\text{new}}(w)=\min\left\{c_s^*(w),c_s^{\text{ub}}(w)\right\}$ and compute the Lagrange multiplier

   $$\lambda_s^{\text{new}}(w)=\frac{1-\beta}{\tilde{R}_f}\left(c_s^{\text{new}}(w)^{-1/\varepsilon}-c_s^*(w)^{-1/\varepsilon}\right).$$

2. **Investment decision**: This step applies only to types s such that $\mathrm{E}\left[\tilde{R}_{sj}\mid s\right]>\tilde{R}_f$, for other

types set $I_s(w) = 0$. Using $c_s^{\text{new}}(w)$, construct the following function

$$\Psi_s(w, I) = \mathrm{E}\left[\left(\frac{v_{s'}^{\text{old}}(w'_{sj}(w,I))}{\mathrm{E}\left[v_{s'}^{\text{old}}(w'_{sj}(w,I))^{1-\gamma}\,\middle|\,s\right]^{\frac{1}{1-\gamma}}}\right)^{1/\varepsilon-\gamma} c_{s'}^{\text{new}}(w'_{sj}(w,I))^{-1/\varepsilon}(\tilde{R}_{sj} - \tilde{R}_f)\,\middle|\,s\right] + \frac{\lambda_s^{\text{new}}(w)(\tilde{R}_{s1} - \tilde{R}_f)}{\beta(1-\beta)},$$

where

$$w'_{sj}(w, I) = \tilde{R}_f\left(w + (1-\tau_h)\omega h_s - I - c_s^{\text{new}}(w)\right) + \tilde{R}_{sj} I.$$

Finally, to solve for the optimal investment, conjecture that investment is interior and find the root I^* that solves $\Psi_s(w, I^*) = 0$ and set $I_s(w) = I^*$. If a root does not exist and $\Psi_s(w, 0) < 0$, set $I_s(w) = 0$.

3. **Value function**: Construct next period's wealth using the new policy functions

$$w'_{sj}(w) = \tilde{R}_f\left(w + (1-\tau_h)\omega h_s - I_s^{\text{new}}(w) - c_s^{\text{new}}(w)\right) + \tilde{R}_{sj} I_s^{\text{new}}(w)$$

and update the value function

$$v_s^{\text{new}}(w) = \begin{cases} \left((1-\beta)c_s^{\text{new}}(w)^{1-1/\varepsilon} + \beta\,\mathrm{E}\left[v_{s'}^{\text{old}}(w'_{sj}(w))^{1-\gamma}\,\middle|\,s\right]^{\frac{1-1/\varepsilon}{1-\gamma}}\right)^{\frac{1}{1-1/\varepsilon}}, & (\varepsilon \neq 1) \\ c_s^{\text{new}}(w)^{1-\beta}\left(\mathrm{E}\left[v_{s'}^{\text{old}}(w'_{sj}(w))^{1-\gamma}\,\middle|\,s\right]^{\frac{1}{1-\gamma}}\right)^{\beta}. & (\varepsilon = 1) \end{cases}$$