

Gouin-Bonenfant, Émilien; Akira Toda, Alexis

Article

Pareto extrapolation: An analytical framework for studying tail inequality

Quantitative Economics

Provided in Cooperation with:

The Econometric Society

Suggested Citation: Gouin-Bonenfant, Émilien; Akira Toda, Alexis (2023) : Pareto extrapolation: An analytical framework for studying tail inequality, Quantitative Economics, ISSN 1759-7331, The Econometric Society, New Haven, CT, Vol. 14, Iss. 1, pp. 201-233, <https://doi.org/10.3982/QE1817>

This Version is available at:

<https://hdl.handle.net/10419/296302>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



<https://creativecommons.org/licenses/by-nc/4.0/>

Pareto extrapolation: An analytical framework for studying tail inequality

ÉMILIE GOUIN-BONENFANT
Department of Economics, Columbia University

ALEXIS AKIRA TODA
Department of Economics, University of California San Diego

We develop an analytical framework designed to solve and analyze heterogeneous-agent models that endogenously generate fat-tailed wealth distributions. We exploit the asymptotic linearity of policy functions and the analytical characterization of the Pareto exponent to augment the conventional solution algorithm with a theory of the tail. Our framework allows for a precise understanding of the very top of the wealth distribution (e.g., analytical expressions for top wealth shares, type distribution in the tail, and transition probabilities in and out of the tail) in addition to delivering improved accuracy and speed.

KEYWORDS. Bewley–Huggett–Aiyagari model, Pareto exponent, power law, solution accuracy.

JEL CLASSIFICATION. C63, D31, D58, E21.

1. INTRODUCTION

What are the economic effects of a wealth tax on billionaires? Providing a quantitative answer to this type of question using the existing toolkit for economic modeling can prove challenging. The reason is that becoming a billionaire is inherently a “tail event”: it is extremely rare yet it has a disproportionate effect on aggregate variables. For example, in the U.S., fewer than 0.0006% of households can claim the title of billionaire, yet they control 3.4% of the capital in the economy.¹ Having a precise quantitative theory of tail

Émilien Gouin-Bonenfant: eg3041@columbia.edu

Alexis Akira Toda: atoda@ucsd.edu

We thank Brendan Beare, Mark Bils, Dan Cao, Chris Carroll, Wouter Den Haan, Edouard Djétem, Simone Galperti, Fedor Iskhahov, Barış Kaymak, Narayana Kocherlakota, Lilia Maliar, Alisdair McKay, Ben Moll, Markus Poschke, John Stachurski, Eric Young, and seminar participants at Australian National University, City University of New York, Claremont McKenna College, Rochester, UCSB, UCSD, University of New South Wales, University of Tokyo, 2019 Computation in Economics and Finance Conference, 2019 Society for the Advancement of Economic Theory Conference, 2019 Society for Economic Dynamics Conference, 2019 Stanford Institute of Theoretical Economics Conference, and 2021 China International Conference in Macroeconomics for comments.

¹According to Forbes (<https://www.forbes.com/billionaires>), there were 724 billionaires U.S. with net worth \$4.4 trillion as of March 2021. According to Federal Reserve’s “Distributional Financial Accounts” (<https://www.federalreserve.gov/releases/z1/dataviz/dfa/>), the aggregate U.S. household net worth in 2021Q1 was \$128 trillion. Therefore with 122 million households in U.S., the fraction of billionaires is $724/(122 \times 10^6) = 0.00059\%$ with a wealth share of $4.4/128 = 3.4\%$.

events—how much wealth is held in the upper tail of the wealth distribution, what type of individuals reach the tail, and how much mobility there is in and out of the tail—is thus key to conducting credible economic analysis on questions such as the effect of a wealth tax.

However, the conventional solution algorithm for heterogeneous-agent models is not well suited to handle tail events since it relies on approximating the wealth distribution by a histogram, which lumps the upper tail into a single bin. Although analytical models of wealth inequality based on the random growth mechanism do not suffer from this issue, they require strong (and often unrealistic) assumptions to obtain closed-form solutions, which makes them a poor device to address policy-relevant quantitative questions.

In this paper, we develop an analytical framework designed to solve and analyze Bewley–Huggett–Aiyagari models (heterogeneous-agent models without aggregate uncertainty) that generate fat-tailed wealth distributions. In a nutshell, we augment the conventional solution algorithm with an analytical “theory of the tail” (e.g., Pareto upper tail of the wealth distribution, mobility in and out of the tail, and policy functions in the tail). Our contribution is to produce a number of new theoretical results including: an algorithm to compute the wealth distribution (*Pareto extrapolation*), correction terms for aggregate quantities to account for the contribution of agents in the tail, and analytical expressions for tail event moments such as top wealth shares, type distribution in the tail, and the transition probabilities in and out of the tail.

Our framework builds on the conventional solution algorithm and extends it with two additional steps: (i) the “asymptotic analysis” of the individual optimization problem to approximate the behavior of agents in the tail and analytically compute the Pareto exponent of the wealth distribution, and (ii) the “Pareto extrapolation” of the wealth distribution outside the grid to accurately compute the equilibrium. Our analytical framework allows researchers to easily and accurately analyze rich heterogeneous-agent models with features such as persistent earnings and investment risk, borrowing constraint, portfolio choices, etc. and compute tail event moments, which existing methods have difficulty (or are unable) to compute despite their importance.

In the “asymptotic analysis” step, we solve a simplified, or “asymptotic” individual optimization problem semianalytically. Roughly speaking, this problem abstracts from additive elements and focuses on proportional elements.² The benefit of studying the asymptotic problem is that its solution determines the behavior of wealthy agents, which governs the tail property of the model such as the Pareto exponent of the wealth distribution and the transition probabilities in and out of the tail. Since asymptotic policy functions (e.g., consumption, investment) are linear in wealth, we are able to derive analytical expressions for the contribution of agents in the tail to aggregates that depend only on asymptotic slopes (for instance, the asymptotic marginal propensity to consume) and the Pareto exponent.

²For 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 labor income (for agents in the tail, labor income is negligible compared to capital income), which can be solved analytically as in Merton (1969) and Samuelson (1969).

In the “Pareto extrapolation” step, we approximate the upper tail of the wealth distribution using the theoretical Pareto exponent computed in the asymptotic analysis step. Recall that the conventional solution algorithm approximates the wealth distribution by assigning probabilities to a finite wealth grid. To do so, one must simply compute the transition probability matrix for wealth and its eigenvector associated with the dominant eigenvalue (see [Young \(2010\)](#)). Yet, this approach is problematic because in many situations including heterogeneous-agent models with return heterogeneity, mortality risk, or random discount factors, the wealth distribution is naturally unbounded and fat-tailed. Our approach is thus to approximate the bulk of the wealth distribution using the conventional approach but to “append” an (unbounded) Pareto distribution above the largest grid point. The main challenge is to construct a transition probability matrix such that the values associated with the largest wealth grid point reflect the transition probabilities in and out of the upper tail. We provide an algorithm to do so whose key inputs are the asymptotic policy functions and the Pareto exponent computed in the asymptotic analysis step.

The new steps that we propose are based on closed-form formulas, and thus do not generate additional computational cost. In fact, our framework tends to be much faster than the conventional solution algorithm. In principle, one could approximate a fat-tailed wealth distribution using a very large (but finite) grid. In contrast, our approach is to solve the “non-Pareto” part of the wealth distribution using a short but dense grid and extrapolate the upper tail using the theoretical Pareto exponent, where all calculations are done analytically. Since most of the computational cost scales with the number of grid points, our approach can deliver improved accuracy *and* a higher speed.

To assess the accuracy of our framework, we use a simple heterogeneous-agent model that admits a closed-form solution as a laboratory. We find that our method is extremely accurate and robust to the grid choice, even when looking at moments such as the wealth share of the top 0.0006% of households (billionaires). Unsurprisingly, the common practice of truncating the upper tail of the wealth distribution generates a severe downward bias in top wealth shares, even for extremely large truncation points.

For the benefit of the users, we have provided `MATLAB` files for implementing the Pareto extrapolation algorithm at <https://github.com/alexisakira/Pareto-extrapolation>. Due to space limitations, this paper omits a quantitative application of Pareto extrapolation. An application to the wealth tax in a calibrated general equilibrium model is discussed in Sections 5 and 6 of the working paper version ([Gouin-Bonenfant and Toda \(2018\)](#)).

Related literature Our paper is related to a large literature that spans across many disciplines, including quantitative macroeconomics, economic theory on consumption-portfolio choices and general equilibrium, 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 \(2019, 2020\)](#) 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 (which empirically has a thinner tail than the wealth distribution³). 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. Our paper contributes to the quantitative macroeconomics literature by providing a general solution algorithm to solve and analyze such models.

As mentioned in the Introduction, since existing 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. 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).⁴ Our paper bridges this literature on power law in economics and quantitative macroeconomics by providing an analytical framework that combines a theory of the tail to the conventional numerical solution algorithm. Our framework is general enough to be applied to state-of-the-art quantitative models of wealth inequality (Kaymak and Poschke (2016), Hubmer, Krusell, and Smith (2020), Guvenen, Kambourov, Kuruscu, Ocampo-Diaz, and Chen (2019)).

The asymptotic analysis step of our approach exploits the asymptotic linearity of policy functions in models with homothetic utility as well as the analytical characterization of the Pareto exponent as the solution to an eigenproblem. Although the asymptotic linearity of policy functions with homothetic preferences is intuitive, a rigorous proof with an exact analytical characterization of asymptotic slopes was obtained only recently by Ma and Toda (2021, 2022) after the working paper version of this paper (Gouin-Bonenfant and Toda (2018)) was circulated. To analytically characterize the Pareto exponent of the wealth distribution in a general Markovian environment, we apply the recent results from Beare and Toda (2022), who characterize the Pareto exponent as a solution to an eigenproblem. Toda (2019) pointed out the usefulness of the asymptotic problem for computing the Pareto exponent in general models that admit no closed-form solutions. However, he neither considers the solution algorithm for general equilibrium models with fat-tailed wealth distributions nor discusses the implications for analytically characterizing tail event moments, which is the heart of our analysis. We build

³de Vries and Toda (2021) estimate capital and labor income Pareto exponents across 475 country-year observations and document that capital income (hence wealth) inequality is higher than labor income inequality (median Pareto exponents 1.46 and 3.35, respectively) and the two inequalities are uncorrelated.

⁴Other recent applications include firm dynamics (Daron 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)), entrepreneurship (Jones and Kim (2018)), and the spread of a new infectious disease (Beare and Toda (2020)). For reviews of generative mechanisms of Pareto tails used in these papers, see Gabaix (2009).

on these results to produce our new results, which include: an algorithm to compute the wealth distribution (*Pareto extrapolation*), correction terms for aggregate quantities to account for the contribution of agents in the tail, and analytical expressions for tail event moments such as top wealth shares, type distribution in the tail, and the transition probabilities in and out of the tail. 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.

Our paper is also related to the literature on solution methods for heterogeneous-agent models. 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. We use [Young \(2010\)](#)'s nonstochastic simulation to compute the rest of the wealth distribution from the transition probability matrix implied by the law of motion, though Pareto extrapolation is likely applicable to other approaches such as updating the CDF.

Finally, our paper is close in spirit to [Achdou, Han, Lasry, Lions, and Moll \(2022\)](#). They study a continuous-time version of the Bewley–Huggett–Aiyagari model, 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 apply finite-difference methods and propose a fast solution algorithm that can be applied to general heterogeneous-agent models in continuous time. While our paper is different—we focus on the complications arising from fat-tailed wealth distributions—we share the same goal of bridging the gap between theoretical and quantitative work in macroeconomics.

2. ISSUES WITH THE CONVENTIONAL SOLUTION ALGORITHM

Before introducing our framework, we briefly discuss issues with the conventional solution algorithm. Suppose that we want to solve a Bewley–Huggett–Aiyagari model numerically when the wealth distribution could be unbounded, and in particular, fat-tailed.⁵ The conventional solution algorithm for heterogeneous-agent models (henceforth “the truncation method”) combines dynamic programming over a finite grid ([Blackwell \(1965\)](#), [Coleman \(1990\)](#)) with nonstochastic simulation ([Young \(2010\)](#)). The algorithm would be roughly as follows:

⁵There exist models in which the stationary equilibrium wealth distribution has a compact support. For instance, [Stachurski and Toda \(2019, Theorem 8\)](#) show the boundedness of wealth distribution in any Bewley–Huggett–Aiyagari model in which agents are infinitely lived, have constant discount factors, have bounded relative risk aversion utility, have bounded income, and save only in a risk-free asset. Similarly, there exist models in which the stationary equilibrium wealth distribution has a Pareto upper tail: see footnote 4 and the surrounding discussion.

- (i) 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 $w_1 < \dots < w_N$. 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.
- (ii) Having solved the individual optimization problem and obtained the law of motion for individual wealth, the researcher constructs the $SN \times SN$ joint transition probability matrix Q of exogenous state and wealth. The stationary distribution $\pi \in \mathbb{R}_+^{SN}$ is obtained by solving $Q'\pi = \pi$ (so π is an eigenvector of Q' corresponding to the eigenvalue 1).
- (iii) 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 algorithm 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 half-line $w \in [w_N, \infty)$. 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

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

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.⁶ As we will show, this second problem is particularly severe when the object of interest is a top wealth share.

3. THE PARETO EXTRAPOLATION ALGORITHM

Our new solution algorithm, which we call the ‘‘Pareto extrapolation’’ method, builds on the conventional solution algorithm described in Section 2 but differs at several steps,

⁶This point is important as Kubler and Schmedders (2005) show that approximate equilibria can be very far from exact equilibria.

most importantly when computing the stationary distribution and when aggregating individual behavior to evaluate the market clearing condition. It can be applied to solve for the stationary equilibrium of any heterogeneous-agent model, although the novel steps are needed only when the model generates a Pareto-tailed wealth distribution. In general, the stationary wealth distribution has a Pareto upper tail in models that combine homothetic preferences (e.g., additive CRRA, Epstein–Zin, etc.) with either random discount factors, stochastic returns on wealth, and/or birth and death. A byproduct of our algorithm is that it will tell the user whether the model generates a Pareto-tailed wealth distribution, and if so, provides an analytical characterization of the Pareto exponent.

We now provide a step-by-step description of the Pareto extrapolation method. As a leading example, we focus on a simplified (no aggregate shocks) version of the classical [Krusell and Smith \(1998\)](#) (henceforth KS) model with random discount factors, which is known to generate a Pareto-tailed wealth distribution ([Toda \(2019\)](#), [Ma, Stachurski, and Toda \(2020\)](#)). Since the KS model is well known, we only briefly describe the Bellman equation for the value function of the household:

$$v_s(w) = \max_{c \geq 0} \left\{ \frac{c^{1-\gamma}}{1-\gamma} + \beta_s(1-p) E[v_{s'}(w') | s] \right\}, \tag{3.1a}$$

$$w' = R(w - c) + y_{s'}, \tag{3.1b}$$

$$w' \geq \underline{w}. \tag{3.1c}$$

Here, $s = 1, \dots, S$ denotes Markov states that evolve over time according to a transition probability matrix $P = (p_{ss'})$, $\beta_s > 0$ is the discount factor in state s , $p \in [0, 1)$ is the birth/death probability (the infinitely-lived case corresponds to $p = 0$), $\gamma > 0$ is the relative risk aversion, c is consumption, y_s is income in state s , w is the beginning of period wealth including current labor income,⁷ R is the gross interest rate, and \underline{w} is the minimum wealth (borrowing limit).


The Pareto extrapolation algorithm.

- (i) Asymptotic analysis
 - (a) Compute the “asymptotic” policy/value functions semi-analytically
 - (b) Compute the theoretical Pareto exponent
- (ii) Dynamic programming
 - (a) Initialize value/policy functions using a guess
 - (b) Update the value/policy functions over a finite grid using the Bellman/Euler equations

⁷Sometimes researchers use the wealth excluding current labor income as the state variable. Such variations in the timing convention are unimportant for applying the algorithm.


(iii) Transition probabilities

(a) Construct the joint transition probability matrix for exogenous state and wealth over a finite grid using nonstochastic simulation

(b)  To account for transitions in and out of the grid, extrapolate the model using the asymptotic policy functions and the Pareto exponent

(iv) Aggregation

(a) Aggregate individual behavior of agents inside the grid

(b)  To account for the contribution of agents outside the grid, extrapolate the model using the asymptotic policy functions and the Pareto exponent

Below, we explain each step in more detail and pay particular attention to the “new” steps, namely (iii)b and (iv)b.

3.1 Asymptotic analysis

The first step of the algorithm consists of characterizing the “asymptotic” properties of the model. We use two (related) features of models with homothetic preferences. First, the control variables (consumption, investment, etc.) are approximately linear in wealth for wealthy agents. Second, the endogenously determined wealth distribution has a Pareto upper tail.

3.1.1 Computing the asymptotic policy functions Given that labor income enters additively into the budget constraint, whereas capital income is proportional to wealth, the former becomes negligible as the wealth of an agent tends to infinity.⁸ To characterize the behavior of wealthy agents, we consider a simplified problem where labor income is set to zero. Assuming that agents have (asymptotically) homothetic preferences (e.g., CRRA, HARA, 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 semianalytically and the decision rules become linear in wealth.⁹

⁸Note that it is the *individual* labor income that becomes negligible relative to individual capital income. At the aggregate level, labor income generally comprises a substantial portion of aggregate income.

⁹Toda (2014, Theorem 5) discusses the analytical solution to homogeneous problems in a Markovian (non-IID) environment. The usefulness of asymptotic analysis to compute the theoretical Pareto exponent was pointed out by Toda (2019). Appendix A of the Online Supplementary Material (Gouin-Bonenfant and Toda (2023)) formally defines the asymptotic problem and heuristically discusses the asymptotic linearity of policy functions in an abstract dynamic programming setting. For a rigorous proof of asymptotic linearity as well as an analytical characterization of asymptotic slopes, see Ma and Toda (2021, 2022).

For concreteness, consider the KS model. In this case, income y_s and the borrowing limit \underline{w} are negligible asymptotically, so we replace the budget constraint (3.1b) and the borrowing constraint (3.1c) by

$$w' = R(w - c), \tag{3.2a}$$

$$w' \geq 0, \tag{3.2b}$$

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 (3.2) semianalytically quite efficiently, as explained in Toda (2014) in detail. In the case of the KS model, the asymptotic consumption rule can be computed as follows.

The Euler equation in the KS model is

$$c_t^{-\gamma} = \beta_s(1 - p)RE[c_{t+1}^{-\gamma} | s].$$

Let $w_t = w$ and conjecture a solution of the form $c_s(w) = \bar{c}_s w$, where $\{\bar{c}_s\}_{s=1}^S$ can be interpreted as the asymptotic marginal propensity to consume (MPC) out of wealth in each patience state. Using the asymptotic budget constraint (3.2a), we obtain

$$w' = w_{t+1} = R(w_t - c_t) = R(1 - \bar{c}_s)w.$$

Noting that $c_{t+1} = \bar{c}_{s'} w'$ and combining the above equations, we obtain

$$\bar{c}_s^{-\gamma} = (1 - p)R^{1-\gamma}\beta_s \sum_{s'=1}^S p_{ss'} [(1 - \bar{c}_s)\bar{c}_{s'}]^{-\gamma}. \tag{3.3}$$

The asymptotic consumption rules in the KS model can be solved semi-analytically as the solution to the asymptotic Euler equation (3.3), which admits a (necessarily unique) solution if, and only if

$$(1 - p)R^{1-\gamma}\rho(DP) < 1, \tag{3.4}$$

where $D = \text{diag}(\beta_1, \dots, \beta_S)$ is the diagonal matrix of discount factors and $\rho(A)$ denotes the spectral radius (largest absolute value of all eigenvalues) of the matrix A .¹⁰

¹⁰See Appendix B in the Online Supplementary Material for a proof. For the intuition and technical details on condition (3.4), see the discussion around equation (2.13) in Ma and Toda (2021). When the spectral condition (3.4) is violated, the asymptotic MPCs are zero by the results in Ma and Toda (2021, 2022). If needed, the asymptotic value functions can be obtained by conjecturing that $v_s(w) = \bar{v}_s \frac{w^{1-\gamma}}{1-\gamma}$ and using the Bellman equation (3.1a) combined with the asymptotic MPCs $\{\bar{c}_s\}_{s=1}^S$:

$$\bar{v}_s = \bar{c}_s^{1-\gamma} + (1 - p)R^{1-\gamma}\beta_s \sum_{s'} p_{ss'} \bar{v}_{s'} (1 - \bar{c}_s)^{1-\gamma}.$$

3.1.2 *Computing the theoretical Pareto exponent* Equipped with the solution to the asymptotic problem, we can now compute the Pareto exponent of the wealth distribution. We use the insight from Toda (2019), who argues that the tail property of the wealth distribution depends only on the behavior of wealthy agents. Substituting the asymptotic consumption rule $c_s(w) = \bar{c}_s w$ into the asymptotic budget constraint (3.2a), we obtain the asymptotic law of motion for wealth in the KS model, which is linear:

$$w' = R(1 - \bar{c}_s)w. \tag{3.5}$$

The key insight here is that, since the patience state s evolves randomly over time (and so does the asymptotic MPC \bar{c}_s), the wealth accumulation process of wealthy agents becomes a “random growth model.”

More generally, suppose that the law of motion of the asymptotic problem is

$$w_{t+1} = G_{t+1}w_t, \tag{3.6}$$

where $G_{t+1} > 0$ is the gross growth rate of wealth between time t and $t + 1$ and w_t 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 (2022) show that under mild conditions the stationary wealth distribution has a Pareto upper tail and characterize the Pareto exponent ζ , as follows. For $z \in \mathbb{R}$, let

$$M_{ss'}(z) = E[e^{z \log G_{t+1}} \mid s_t = s, s_{t+1} = s'] = E[G_{t+1}^z \mid s_t = s, s_{t+1} = s'] \tag{3.7}$$

be the moment generating function of the log growth rate $\log G_{t+1}$ conditional on transitioning from state s to s' , and

$$M(z) = (M_{ss'}(z)) \tag{3.8}$$

be the $S \times S$ matrix of conditional moment generating functions (3.7). (In the case of the KS model, we simply have $M_{ss'}(z) = [R(1 - \bar{c}_s)]^z$ using (3.5) and (3.7).) Then under mild conditions Beare and Toda (2022) show that the equation

$$(1 - p)\rho(P \odot M(z)) = 1, \tag{3.9}$$

where $P \odot M(z)$ denotes the Hadamard (entrywise) product of P and $M(z)$, has a unique positive solution $z = \zeta > 0$, and that the stationary wealth distribution has a Pareto upper tail with exponent ζ . The following proposition gives a simple test for the solvability of (3.9).

PROPOSITION 3.1. *If $G_{t+1} \leq 1$ always, then (3.9) does not have a solution $z > 0$. If $M(z)$ is finite for all $z > 0$, P is irreducible, and*

$$p_{ss} \Pr(G_{t+1} > 1 \mid s_t = s_{t+1} = s) > 0 \tag{3.10}$$

for some s , then (3.9) has a unique solution $z = \zeta > 0$.

For some parametrization, the model might generate a bounded wealth distribution. Intuitively, $G_{t+1} \leq 1$ means that wealth shrinks (or stays the same), so there is no random *growth*. In this case, the wealth distribution does not have a Pareto tail, and the Pareto extrapolation algorithm reduces to the conventional one (because $\zeta = \infty$ makes the extra steps discussed in Sections 3.2 and 3.3 redundant). If P is irreducible, every state is visited eventually. The condition (3.10) says that in state s , with positive probability the agent’s wealth grows *and* the agent can remain in that state. Because there is random growth, the wealth distribution has a Pareto tail. 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 (3.9) to compute the theoretical Pareto exponent.

3.2 Transition probabilities

The conventional solution algorithm approximates the wealth distribution over a finite grid. The idea is to first compute the joint transition probability matrix Q over the exogenous state s and wealth w and then compute the stationary distribution π by solving $Q'\pi = \pi$. The key challenge when the wealth distribution is fat-tailed is that for any truncation point w_N , some agents will “escape” the grid (i.e., transition from $w \leq w_N$ to $w' > w_N$) and similarly some will “enter” the grid ($w > w_N$ and $w' \leq w$). Equipped with the policy functions, their asymptotic counterpart, and the theoretical Pareto exponent ζ , we now provide a simple algorithm that computes the transition probability matrix Q while accounting for transitions in and out of the grid.

We first introduce some notation. Let $\mathcal{W}_N = \{w_n\}_{n=1}^N$ be the grid for wealth, $I_n = [w_n, w_{n+1})$ be the half-open interval with endpoints w_n and w_{n+1} , and $I_N = [w_N, \infty)$. Let $w' = g_{ss'}(w)$ be the law of motion for wealth (conditional on transitioning from state s to s') implied by the policy functions obtained in the dynamic programming step. Finally, let $Q = (q_{sn,s'n'})$ be the $SN \times SN$ joint transition probability matrix of exogenous state s and wealth $\{w_n\}_{n=1}^N$ and $\pi = (\pi_{sn})$ be the $SN \times 1$ stationary distribution of wealth.

To compute the elements of Q , we proceed in two steps. For agents “inside the grid,” we use nonstochastic simulation exactly as in Young (2010). For agents “outside the grid,” we extrapolate the model beyond the largest grid point and obtain analytical expressions for the transition probabilities. Then the stationary distribution π can be computed as the (unique) eigenvector of Q' corresponding to the eigenvalue 1.

We now describe in detail how to compute the transition probabilities $q_{sn,s'n'}$. For simplicity, we focus on the case when there is no death ($p = 0$). The case $p > 0$ is similar and it is a matter of taking the weighted average of transition probabilities conditional on survival and death, weighting by $1 - p$ and p .

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

$$g_{ss'}(w_n) = (1 - \theta)w_k + \theta w_{k+1} \iff \theta = \frac{g_{ss'}(w_n) - w_k}{w_{k+1} - w_k}. \tag{3.11}$$

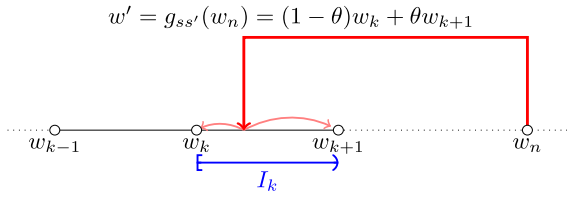


FIGURE 1. Construction of transition probabilities from a grid point.

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 $g_{ss'}(w_n) < w_1$ or $g_{ss'}(w_n) \geq w_N$, then just assign probability 1 to state 1 or N .

Thus for $n < N$ we construct the transition probability as

$$q_{sn,s'n'} = p_{ss'} \times \begin{cases} 1 & \text{if } g_{ss'}(w_n) < w_1 \text{ and } n' = 1, \\ 1 - \theta & \text{if } g_{ss'}(w_n) \in I_k \text{ and } n' = k, \\ \theta & \text{if } g_{ss'}(w_n) \in I_k \text{ and } n' = k + 1, \\ 1 & \text{if } g_{ss'}(w_n) \geq w_N \text{ and } n' = N, \\ 0 & \text{otherwise,} \end{cases} \quad (3.12)$$

where θ is defined by (3.11).

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 r_n . By definition, we have $\sum_{n=N}^\infty r_n = 1$. Now for each $n \geq N$, we can do precisely as in the previous case, and add probabilities $(1 - \theta)r_n$ and θr_n (where θ is defined by (3.11)) to the grid points w_k, w_{k+1} whenever $w' = g_{ss'}(w_n) \in I_k$ for $k < N$ (Figure 2). If $g_{ss'}(w_n) < w_1$ or $g_{ss'}(w_n) \geq w_N$, then just add probability r_n to the transition to state 1 or N . The nice thing is that for large enough n , the next period's state $w' = g_{ss'}(w_n)$ is likely large (contained in I_N), so we only need to compute θ for finitely many n (say $n = N, \dots, N'$, where $w_{N'}$ is the smallest grid point such that the next period's wealth always exceeds w_N). For the probability r_n , because the theoretical density is Pareto with exponent ζ , we can simply set $r_n \propto w_n^{-\zeta-1}$ if the grid spacing $w_{n+1} - w_n$ is constant for $n \geq N$.

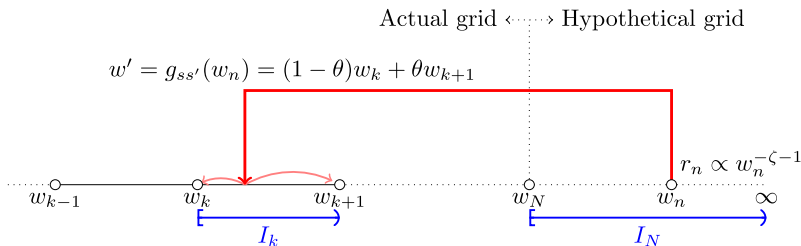


FIGURE 2. Construction of transition probabilities from a hypothetical grid point.

More formally, we do as follows. First, let $h = w_N - w_{N-1} > 0$ be the grid spacing of the hypothetical grid points $\{w_n\}_{n=N+1}^\infty$ taken to be the distance between the largest two actual grid points.¹¹ Define the untruncated grid $\mathcal{W}_\infty = \{w_n\}_{n=1}^\infty$ by $w_n = w_N + (n - N)h$ for $n > N$. Compute the smallest index $N' \geq N$ such that $g_{ss'}(w_{N'}) > w_N$ for all s, s' :

$$N' = \min\{n \geq N \mid \forall s, s', g_{ss'}(w_N + (n - N)h) > w_N\}. \tag{3.13}$$

To evaluate the law of motion outside the grid, we can simply linearly extrapolate the law of motion for $w \geq w_N$ as

$$g_{ss'}(w) = g_{ss'}(w_N) + G_{ss'}(w - w_N), \tag{3.14}$$

where $G_{ss'} > 0$ is the theoretical slope obtained in the asymptotic analysis step. Combining (3.13) and (3.14), after some algebra we obtain

$$N' = N + \max_{s,s'} \max\left\{\left\lceil \frac{w_N - g_{ss'}(w_N)}{G_{ss'}h} \right\rceil, 0\right\}, \tag{3.15}$$

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

To compute the conditional probability r_n , because theoretically the stationary distribution has a Pareto upper tail with exponent $\zeta > 1$ (if $\zeta \leq 1$, then the mean is infinite, which is impossible in equilibrium), using the density (conditional on $w \geq w_N$) $f(x) = \zeta w_N^\zeta x^{-\zeta-1}$, we set

$$r_n \approx \zeta w_N^\zeta (w_N + (n - N)h)^{-\zeta-1} h = \zeta \frac{h}{w_N} \left(1 + (n - N) \frac{h}{w_N}\right)^{-\zeta-1}$$

for $n \geq N$. Since for $n \geq N'$ the next state will always be N ($w' = g_{ss'}(w_n) \in I_N$), there is no need to compute r_n individually. Using the theoretical Pareto density, we obtain

$$\begin{aligned} \sum_{n=N'}^\infty r_n &\approx \frac{1}{2} r_{N'} + \int_{w_{N'}}^\infty \zeta w_N^\zeta x^{-\zeta-1} dx = \frac{1}{2} r_{N'} + (w_{N'}/w_N)^{-\zeta} \\ &= \left(1 + (N' - N) \frac{h}{w_N}\right)^{-\zeta-1} \left(1 + \left(N' - N + \frac{1}{2}\right) \frac{h}{w_N}\right). \end{aligned} \tag{3.16}$$

(We explain the mysterious term $r_{N'}/2$ in Section 4.4.) Therefore we set

$$\begin{cases} r_n := C \zeta \frac{h}{w_N} \left(1 + (n - N) \frac{h}{w_N}\right)^{-\zeta-1} & \text{for } N \leq n < N', \\ r_{N'} := C \left(1 + (N' - N) \frac{h}{w_N}\right)^{-\zeta-1} \left(1 + \left(N' - N + \frac{1}{2}\right) \frac{h}{w_N}\right), \end{cases} \tag{3.17}$$

where the constant of proportionality C is chosen to satisfy $\sum_{n=N}^{N'} r_n = 1$.

¹¹This choice is motivated by the trapezoidal formula for quadrature and we have numerically confirmed to be optimal.

Now for each s and extra grid point $n = N, \dots, N'$, define the transition probability $\tilde{q}_{sn,s'n'}$ exactly as in (3.12). The remaining elements of the joint transition probability matrix $Q = (q_{sn,s'n'})$ can be computed as

$$q_{sN,s'n'} = \sum_{n=N}^{N'} r_n \tilde{q}_{sn,s'n'}. \tag{3.18}$$

A few remarks are in order. First, the algorithm for constructing Q has essentially zero additional computational cost relative to the existing method, 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 conventional solution algorithm needs to compute anyway, the Pareto extrapolation algorithm requires only $S \times 1 = S$ additional operations, which is negligible. In our numerical example in Section 5, 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 Q is sparse, meaning that Q has few nonzero elements. To see this, let us evaluate the number of nonzero elements of Q . For each s, s' and $n < N$, there are at most two states the next wealth can take because nonstochastic simulation assigns positive probabilities to two neighboring grid points (Figure 1). For $n = N$, in principle the next wealth state can be anything. Therefore the number of nonzero elements of Q is at most

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

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

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

as $N \rightarrow \infty$, so Q is sparse.¹² Therefore, computing the stationary distribution π is feasible, despite the fact that Q is in practice a very large matrix.

3.3 Aggregation

When computing the equilibrium of a heterogeneous-agent model, we need to impose market clearing conditions in some way or another. To do so, we need to aggregate individual behavior. Let us first focus on computing the aggregate capital supply in the KS model. Given that the capital supply of an agent in state s with wealth w is $w - c_s(w)$, the

¹²Achdou et al. (2022) 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 advantages over discrete-time problems (e.g., partial differential equations versus nonlinear difference equations), discrete-time problems also do possess sparsity if appropriately solved. One key advantage of continuous-time models arises when the exogenous state variables obey a diffusion process. In that case, the transition probability matrix is also sparse in the number of possible exogenous states (S using our notation).

aggregate capital supply is $K = \int (w - c_s(w)) d\Gamma(w, s)$, where $\Gamma(w, s)$ is the theoretical joint distribution of wealth and exogenous state. The truncation method approximates K with

$$K^{\text{trunc}} = \sum_{s=1}^S \sum_{n=1}^N \pi_{sn}(w_n - c_{sn}), \tag{3.19}$$

where c_{sn} and π_{sn} are consumption and unconditional (stationary) probability at state (s, n) .

The only caveat is that for the wealth state N , the probability is not concentrated on the grid point w_N but in principle distributed over the half-line $I_N = [w_N, \infty)$. We can easily overcome this problem by (i) extrapolating the policy functions outside the grid using the asymptotic policy functions, and (ii) extrapolating the wealth distribution outside the grid using the theoretical Pareto exponent ζ . First, notice that individual capital supply is asymptotically equivalent to $(1 - \bar{c}_s)w$. We can therefore approximate the capital supply of agents with $w \geq w_N$ by linearly extrapolating as

$$w - c_s(w) \approx w_N - c_{sN} + (1 - \bar{c}_s)(w - w_N).$$

Second, notice that the density of wealth conditional on $w \geq w_N$ is approximately a Pareto distribution with exponent ζ and minimum size w_N , which has density $f(x) = \zeta w_N^\zeta x^{-\zeta-1}$. Therefore

$$E[w \mid w \geq w_N] \approx \int_{w_N}^{\infty} \zeta w_N^\zeta x^{-\zeta} dx = \frac{\zeta}{\zeta - 1} w_N.$$

Combining both observations, we can approximate the capital supply of an agent in state (s, N) by

$$E[w - c_s(w) \mid w \geq w_N] \approx w_N - c_{sN} + \frac{1}{\zeta - 1} (1 - \bar{c}_s) w_N.$$

Therefore, using Pareto extrapolation, the correct approximation of the aggregate capital supply K is

$$K^{\text{PE}} = \underbrace{\sum_{s=1}^S \sum_{n=1}^N \pi_{sn}(w_n - c_{sn})}_{=K^{\text{trunc}}} + \underbrace{\frac{1}{\zeta - 1} \sum_{s=1}^S \pi_{sN}(1 - \bar{c}_s) w_N}_{\text{correction term}}. \tag{3.20}$$

Comparing (3.19) to (3.20), we can see that the truncation method introduces an error because the correction term $\frac{1}{\zeta - 1} \sum_{s=1}^S \pi_{sN}(1 - \bar{c}_s) w_N$ is absent. If ζ is close to 1 (Zipf's law), then failing to account for this term will introduce significant error. The formula (3.20) highlights the fact that choosing a large grid point w_N (so that the fraction of agents at the largest grid point $\sum_{s=1}^S \pi_{sN}$ is small) does not necessarily solve the problem. The reason is that w_N and $\sum_{s=1}^S \pi_{sN}$ enter multiplicatively in the correction term. For example, if $\sum_{s=1}^S \pi_{sN} = 10^{-6}$ and $w_N = 10^6$, then the correction term is $\frac{1}{\zeta - 1} \sum_{s=1}^S (1 - \bar{c}_s)$, which can be substantial as a fraction of K .

The insight carries well beyond the KS model. In general, we can approximate the integral of any policy function $x_s(w)$ that is asymptotically equivalent to $\bar{x}_s w$ against the stationary distribution using

$$X \approx \sum_{s=1}^S \sum_{n=1}^N \pi_{sn} x_{sn} + \frac{1}{\zeta - 1} \sum_{s=1}^S \pi_{sN} \bar{x}_s w_N, \tag{3.21}$$

where $x_{sn} = x_s(w_n)$. A special case is the computation of aggregate wealth, which is simply

$$E[w] \approx \sum_{s=1}^S \sum_{n=1}^N \pi_{sn} w_n + \frac{1}{\zeta - 1} \sum_{s=1}^S \pi_{sN} w_N. \tag{3.22}$$

4. ADDITIONAL CONSIDERATIONS FOR PARETO EXTRAPOLATION

In this section, we discuss many additional details and considerations for the Pareto extrapolation algorithm.

4.1 Top wealth shares

Oftentimes, the object of interest is the wealth distribution itself. One common way to summarize the concentration of wealth (other than to report the Pareto exponent) is to compute “top wealth shares” (i.e., the share of wealth owned by the top fraction $p \in (0, 1)$ agents). When using the Pareto extrapolation method, we suggest computing the 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. Dividing that number by aggregate wealth (3.22) 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 suggest using 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 . The density for $x \geq w_N$ is then $f(x) = \pi_N \zeta w_N^\zeta x^{-\zeta-1}$. Using this, the tail probability is

$$\Pr(X \geq x) = \int_x^\infty \pi_N \zeta w_N^\zeta x^{-\zeta-1} dx = \pi_N w_N^\zeta x^{-\zeta}. \tag{4.1}$$

On the other hand, the total wealth held by wealthy agents is

$$E[X; X \geq x] = \int_x^\infty \pi_N \zeta w_N^\zeta x^{-\zeta} dx = \frac{\zeta}{\zeta - 1} \pi_N w_N^\zeta x^{-\zeta+1}. \tag{4.2}$$

Therefore, letting W be the aggregate wealth, setting $p = \Pr(X \geq x)$, and eliminating x , the wealth share $s(p)$ of the wealthiest fraction $p \in (0, 1)$ of agents is

$$s(p) = \frac{\zeta}{\zeta - 1} \pi_N^{1/\zeta} \frac{w_N}{W} p^{1-1/\zeta}. \tag{4.3}$$

4.2 Top tail type and exit probabilities

For particular applications, it is of interest to compute the distribution of types (states) $s \in S$ in the top tail and the probability that an agent with a particular type exits the top tail. The following proposition, which is new, provides an answer.

PROPOSITION 4.1. *Suppose the law of motion for asymptotic agents is the random growth model (3.6). Let $M(z)$ be the matrix of conditional moment generating functions in (3.8) and $\zeta > 0$ be the Pareto exponent that solves (3.9). Let $\bar{\pi}$ be the left Perron vector of $P \odot M(\zeta)$ normalized such that $\sum_{s=1}^S \bar{\pi}_s = 1$. Then $\bar{\pi}$ is the top tail type distribution:*

$$\lim_{w \rightarrow \infty} \Pr(s_t = s \mid w_t > w) = \bar{\pi}_s. \tag{4.4}$$

Furthermore, the conditional top tail exit probability is given by

$$\lim_{w \rightarrow \infty} \Pr(w_{t+1} \leq w \mid w_t > w, s_t = s) = 1 - (1 - p) E[\min\{1, G_{t+1}^\zeta\} \mid s_t = s]. \tag{4.5}$$

The formulas in Proposition 4.1 are highly nontrivial yet quite useful for calibrating and analyzing models. For example, (4.4) and (4.5) can be used to answer questions such as “what is the probability that a billionaire is an entrepreneur?” and “what is the probability that an individual in the Forbes 400 list drops out?” (or “what is the probability that the wealthiest individual is no longer wealthiest?”) within the model. These types of questions cannot be satisfactorily answered using existing approaches because it either requires a very large and fine grid or a prohibitively large scale simulation.

4.3 Dynamic programming

Dynamic programming methods such as value function iteration (Blackwell (1965)) and policy function iteration (Coleman (1990)) consist of numerically solving the individual optimization problem over a finite grid for wealth. Many different algorithms exist, but they all share the same structure: the researcher starts with a guess for the policy/value functions and uses optimality conditions such as the Euler/Bellman equations to update those guesses until convergence.

Dynamic programming is by far the most computationally intensive step when solving a heterogeneous-agent model. Compared to the asymptotic problem, which reduces to solving for an $S \times 1$ object, dynamic programming needs to solve for an $S \times N$ object. However, we can use the solution to the asymptotic problem to construct a “good” initial guess, which helps speed up convergence. For example, when solving the KS model using policy function iteration, we suggest using the initial guess

$$c_s^{(0)}(w) = \bar{c}_s(w - \underline{c}),$$

where \bar{c}_s is the asymptotic MPC and $\underline{c} < w_1$ is an arbitrary number that ensures that consumption is positive. Intuitively, the reason why this guess speeds up convergence is that the distance between the true solution $c_s(w)$ and the guess $c_s^{(0)}(w)$ is already small, especially at the upper end of the grid.¹³

¹³Ma and Toda (2022, Figure 5) study this point in detail and find that using the good initial guess speeds up convergence by about 25%.

4.4 The term $r_{N'}/2$ in (3.16)

Consider the Pareto distribution with exponent $\zeta > 1$ and minimum size $w_N > 0$, which has density $f(x) = \zeta w_N^\zeta x^{-\zeta-1}$ for $x \geq w_N$. Let $w_n = w_N + (n - N)h$ for $n > N$ and consider the probability $P = \Pr(w \geq w_{N'})$, where $N' \geq N$. On the one hand, this probability can be analytically computed as

$$P = \int_{w_{N'}}^{\infty} \zeta w_N^\zeta x^{-\zeta-1} = (w_{N'}/w_N)^{-\zeta}.$$

On the other hand, using the trapezoidal formula for quadrature, we obtain

$$P \approx \frac{1}{2}hf(w_{N'}) + \sum_{n>N'} hf(w_n) = \frac{1}{2}r_{N'} + \sum_{n>N'} r_n,$$

where

$$r_n = hf(w_n) = \zeta w_N^\zeta (w_N + (n - N)h)^{-\zeta-1} h = \zeta \frac{h}{w_N} \left(1 + (n - N) \frac{h}{w_N}\right)^{-\zeta-1}.$$

Therefore

$$\sum_{n=N'}^{\infty} r_n = \frac{1}{2}r_{N'} + \frac{1}{2}r_{N'} + \sum_{n>N'} r_n \approx \frac{1}{2}r_{N'} + P = \frac{1}{2}r_{N'} + (w_{N'}/w_N)^{-\zeta},$$

which explains (3.16).

4.5 Aggregating nonlinear functions

Suppose we want 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 the power ν is below the theoretical Pareto exponent ζ , the conditional expectation of the upper tail is

$$E[w^\nu \mid w \geq w_N] = \int_{w_N}^{\infty} \zeta w_N^\zeta x^{\nu-\zeta-1} dx = \frac{\zeta}{\zeta - \nu} w_N^\nu.$$

Therefore the analog of (3.22) is

$$E[w^\nu] \approx \sum_{s=1}^S \sum_{n=1}^N \pi_{sn} w_n^\nu + \frac{\nu}{\zeta - \nu} \sum_{s=1}^S \pi_{sN} w_N^\nu. \tag{4.6}$$

Similarly, noting that

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

setting $\nu = 0$ we obtain

$$E[\log w] \approx \sum_{s=1}^S \sum_{n=1}^N \pi_{sn} \log w_n + \frac{1}{\zeta} \sum_{s=1}^S \pi_{sN}. \tag{4.7}$$

In the KS model, the value function is asymptotically equivalent to $\bar{v}_s \frac{w^{1-\gamma}}{1-\gamma}$ (see footnote 10). To approximate the welfare function \mathcal{W} , which is defined as the integral of the value function against the stationary distribution, we can apply (4.6) to $\nu = 1 - \gamma$ and dividing by $1 - \gamma$. Hence the welfare in consumption equivalent is

$$\mathcal{W} \approx \left(\sum_{s=1}^S \sum_{n=1}^N \pi_{sn} v_{sn} + \frac{1}{\zeta - 1 + \gamma} \sum_{s=1}^S \pi_{sN} \bar{v}_s w_N^{1-\gamma} \right)^{\frac{1}{1-\gamma}}. \tag{4.8}$$

4.6 Choosing the grid and truncation point

The description of our algorithm in Section 3 implicitly assumes that the researcher has already chosen the grid $\mathcal{W} = \{w_n\}_{n=1}^N$ and, in particular, the truncation point w_N . Here, we provide a practical guidance on how to choose the grid and the truncation point.

Exponential grid In common applied settings, the researcher would like to use a grid that covers a large part of the state space without too many points. A natural idea is to use a grid such that w_n exponentially grows with n . Because we are not aware of a systematic approach for constructing an exponential grid, we propose a simple solution.

In many models, the state variable may become negative (e.g., asset holdings), which causes a problem for constructing an exponential 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.

- (i) Choose a shift parameter $s > -a$.
- (ii) Construct an N -point evenly-spaced grid on $[\log(a + s), \log(b + s)]$.
- (iii) Take the exponential and subtract s .

The remaining question is how to choose the shift parameter s . The following proposition shows that the shift parameter s is automatically determined by the median grid point (corresponding to $n = N/2$).

PROPOSITION 4.2. *Let $a < c < \frac{a+b}{2}$. Then the exponential grid with shift parameter $s = \frac{c^2-ab}{a+b-2c}$ has median grid point c .*

To choose the median grid point c , we can use information from the problem we want to solve. Note that by construction, half of 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, for instance, the aggregate capital in a representative-agent model.

Affine-exponential grid Although the exponential grid covers a large part of the state space, according to our experience it is not always ideal because the grid spacing $w_n - w_{n-1}$ tends to be large in the bulk of the state space. An alternative is to simply set the grid spacing $h = w_n - w_{n-1}$ to be constant, but that necessarily makes the number of points N quite large when the truncation point w_N is large and the grid spacing h is small.

As a compromise, we suggest using the hybrid (affine-exponential) grid: construct the exponentially-spaced grid as discussed above, but replace the bottom (say below the median grid point) 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.

Through many numerical experiments, we have confirmed that the affine-exponential grid outperforms both the evenly- and exponentially-spaced grids. All of our results in Section 5 are based on this affine-exponential grid.

Truncation point Because our algorithm relies on the asymptotic linearity of policy functions, we suggest choosing a truncation point that implies a small difference between the average propensity to consume (APC) at the largest grid point, c_{sN}/w_N , and the asymptotic MPC \bar{c}_s determined by solving (3.3). Therefore the researcher should choose a truncation point w_N that implies a small maximum APC relative error

$$\max_s \left| \frac{1}{\bar{c}_s} \frac{c_{sN}}{w_N} - 1 \right|, \quad (4.9)$$

where w_N is the largest grid point (truncation point) and c_{sN} denotes the policy function in state s at that point. The idea is that, if the APC at the largest grid point is close to its asymptotic value, then the consumption function should already be approximately linear, making all of the above approximations accurate.

4.7 General equilibrium

So far, we have considered evaluating the market clearing condition for a guess of equilibrium prices. To solve for equilibrium prices, we can apply the Pareto extrapolation method for successive guesses of equilibrium prices and update the guesses using the excess supply computed in the aggregation step. Here, the asymptotic analysis step can be used to narrow down the set of prices consistent with an equilibrium.

Notice that we can rule out any prices such that the theoretical Pareto exponent ζ is below or equal one.¹⁴ When $\zeta \leq 1$, aggregate wealth is infinite, which is inconsistent with market clearing. In the KS model, narrowing down the set of interest rates R consistent with an equilibrium amounts to evaluating (3.4) (if one would like to impose positive MPC) and (3.9) for a range of values of R and computing a bound (\underline{R}, \bar{R}) . Since

¹⁴This condition is only necessary for equilibrium existence. Establishing sufficiency is beyond the scope of this paper. Cao (2020) establishes the existence of equilibrium in the Krusell and Smith (1998) model.

the asymptotic analysis step is not computationally intensive, we can make substantial efficiency gains by avoiding the dynamic programming step as much as possible.

5. EVALUATING SOLUTION ACCURACY

As in any new numerical method, the first order of business is to evaluate the solution accuracy. For this purpose, we present a simple (minimal) heterogeneous-agent model that admits a semianalytical solution and Pareto-tailed wealth distribution, which we use as a benchmark for evaluating numerical solutions.

5.1 Model

We consider a dynamic general equilibrium model similar in spirit to [Krusell and Smith \(1998\)](#) but with the following features: (i) mortality risk and heterogeneous discount factors (for obtaining a Pareto-tailed wealth distribution), (ii) no aggregate or income risk (for analytical tractability), (iii) subsistence consumption (for mimicking the borrowing constraint). Time is discrete and denoted by $t = 0, 1, \dots$

Agents The economy is populated by a mass 1 continuum of S types of agents indexed by $s = 1, \dots, S$. An agent's type evolves over time according to a Markov chain with irreducible transition probability matrix $P = (p_{ss'})$. Agents die with probability $p \in (0, 1)$ each period and is replaced by newborn agents. A newborn agent becomes type s with probability $\pi_s > 0$, where $\pi = (\pi_1, \dots, \pi_S)'$ is the stationary distribution (left Perron vector) of P . A type s agent has discount factor β_s . All agents supply one unit of labor inelastically. A typical agent has utility function

$$E_0 \sum_{t=0}^{\infty} \left(\prod_{i=0}^{t-1} \beta_{s_i} (1 - p) \right) \log(c_t - \underline{c}_t), \tag{5.1}$$

where $\underline{c}_t \geq 0$ is the minimum (subsistence) consumption of agents at time t . We assume the minimum consumption is a constant fraction of the current labor income, so

$$\underline{c}_t = \phi \omega_t, \tag{5.2}$$

where $\phi \in [0, 1)$ is the constant of proportionality and ω_t is the wage at time t .

Technology Technology is represented by a representative firm with a Cobb–Douglas production function $AF(K, L) = AK^\alpha L^{1-\alpha}$, where A is Total Factor Productivity (TFP) and $\alpha \in (0, 1)$ is the capital share. Capital depreciates at rate $\delta \in [0, 1]$. Therefore the firm's problem at time t is

$$\max_{K, L \geq 0} \left[-K + \frac{1}{R_t} (A_t F(K, L) - \omega_t L + (1 - \delta)K) \right], \tag{5.3}$$

where R_t is the gross risk-free rate from time $t - 1$ to t . That is, the firm buys capital K at the end of time $t - 1$, hires labor to produce at the beginning of time t , and pays the profit and depreciated capital to shareholders (who discount using the risk-free rate since there is no aggregate risk).

Budget constraint There are perfectly competitive life insurance companies that provide annuities and life insurances. Letting R_t be the gross risk-free rate, due to mortality risk, the effective gross risk-free rate is $\tilde{R}_t = R_t/(1 - p)$. Letting a_t be the financial wealth of a typical agent at the beginning of time t excluding current labor income, the budget constraint is therefore

$$a_{t+1} = \tilde{R}_{t+1}(a_t - c_t + \omega_t). \tag{5.4}$$

The agents face the natural borrowing constraint (see Appendix C in the Online Supplementary Material for details).

Equilibrium We consider the stationary equilibrium.

DEFINITION 5.1 (Stationary equilibrium). A stationary equilibrium consists of a gross risk-free rate R , a wage ω , aggregate capital K , aggregate labor L , optimal consumption rules $\{c_s(a)\}_{s=1}^S$, and a stationary distribution $\Gamma = \Gamma(a, s)$ such that:

- (i) given R and ω , aggregate capital K and aggregate labor L solve the profit maximization problem (5.3),
- (ii) given R and ω , for each s the optimal consumption rule $c_s(a)$ maximizes the utility (5.1) subject to the budget constraint (5.4) and the natural borrowing constraint,
- (iii) the capital market clears, so

$$RK = \sum_{s=1}^S \int a \Gamma(da, s), \tag{5.5}$$

- (iv) the labor market clears, so

$$L = 1, \tag{5.6}$$

- (v) Γ is the stationary distribution of the law of motion

$$(a, s) \mapsto \begin{cases} (\tilde{R}(a - c_s(a) + \omega), s') & \text{with probability } (1 - p)p_{ss'}, \\ (0, s') & \text{with probability } p\pi_{s'}, \end{cases} \tag{5.7}$$

where $\tilde{R} = R/(1 - p)$.

The reason why the left-hand side of (5.5) is RK is as follows. Capital is installed at the end of time $t - 1$ and pays aggregate dividend RK to shareholders at the beginning of time t . This quantity must equal aggregate asset holdings at the beginning of period, which is the right-hand side of (5.5).

Appendix C in the Online Supplementary Material proves the existence of a stationary equilibrium.

TABLE 1. Exogenously set parameters.

Description	Symbol	Value
Birth/death probability	p	0.025
Discount factor	β_s	(0.9, 0.95, 1)
TFP	A	1
Capital share	α	0.38
Capital depreciation rate	δ	0.08

5.2 Calibration

We use a numerical example to evaluate the solution accuracy. We first exogenously set some parameters as in Table 1. The birth/death probability $p = 0.025$ implies that one generation is economically active on average for $1/p = 40$ years. We assume there are three patience types with average discount factor 0.95. Without loss of generality, we set the TFP to $A = 1$. Finally, we set the capital share and depreciation rate to standard values.

We parsimoniously model the discount factor process $\{\beta_t\}$ and assume that the transition probability matrix is

$$P = \begin{bmatrix} 1 - q & q & 0 \\ q/2 & 1 - q & q/2 \\ 0 & q & 1 - q \end{bmatrix},$$

where $q \in (0, 1)$ is the probability of transitioning to another state. We calibrate the remaining parameters q, ϕ by targeting the stationary wealth distribution. Namely, we target the top 1%, 10%, and 50% wealth shares and the wealth Pareto exponent in U.S. and obtain $q = 0.0927$ and $\phi = 0.7286$.¹⁵ This calibration implies that the patience state changes on average every 11 years and nondiscretionary consumption is 73% of average labor income. Table 2 shows that the stationary wealth distribution in the model closely matches the empirical counterpart.

TABLE 2. Targeted moments.

Moment	Data	Model
Top 1% wealth share (%)	32.38	32.90
Top 10% wealth share (%)	69.95	68.72
Top 50% wealth share (%)	98.71	98.11
Pareto exponent	1.52	1.55

¹⁵To compute the wealth shares, we average the 2017–2019 quarterly household wealth shares data from Federal Reserve’s “Distributional Financial Accounts” (footnote 1). The U.S. wealth Pareto exponent is taken from Vermeulen (2018, Table 8).

5.3 Solution accuracy

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. To implement our algorithm, we use an N -point affine-exponential grid supported between the natural borrowing limit and largest grid point \bar{w} as discussed in Section 4.6, where the median grid point is the aggregate capital in a corresponding representative-agent model, which is

$$K_{RA} = ((1/(\bar{\beta}(1 - p)) - 1 + \delta)/(A\alpha))^{\frac{1}{\alpha-1}} = 4.0510. \tag{5.8}$$

(Here, $\bar{\beta}$ is the average discount factor.)

Aggregate wealth 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 semianalytical solution to compute the stationary distribution on the wealth grid, and then compute the implied aggregate financial wealth using (3.22) (with or without the correction term) for the Pareto extrapolation and truncation methods, respectively. For this exercise, our primary interest is the relative error $\widehat{W}/W - 1$, where W and \widehat{W} are the aggregate financial wealth from the semianalytical and numerical solutions, respectively.

Table 3 shows the relative error $\widehat{W}/W - 1$ in the aggregate wealth using this grid for various truncation point \bar{w} and number of points N , both for the truncation and Pareto extrapolation methods.

The message from Table 3 is clear: there is no accuracy-efficiency trade-off with the Pareto extrapolation method, while the conventional truncation method is subject to this trade-off. Pareto extrapolation is quite accurate, with relative error robustly in the range 0.001–0.4% when $\bar{w}/K_{RA} \geq 100$ regardless of the grid specification. On the other hand, truncation can achieve acceptable accuracy only with a large, fine grid.

Another way to see the issue with the truncation method is to simply calculate the fraction of wealth (wealth share) held by agents outside the grid. For any truncation

TABLE 3. Relative error (%) in aggregate wealth for the truncation and Pareto extrapolation methods.

Method: \bar{w}/K_{RA}	Truncation			Pareto extrapolation		
	$N = 10$	100	1000	$N = 10$	100	1000
10^1	-36.920	-26.850	-26.530	-3.195	-1.882	-2.260
10^2	-21.520	-8.690	-7.860	-0.437	0.062	-0.031
10^3	-15.640	-2.950	-2.240	-0.261	0.036	0.017
10^4	-12.930	-1.080	-0.640	-0.234	0.011	0.008
10^5	-11.510	-0.420	-0.180	-0.221	0.003	0.002
10^6	-10.720	-0.180	-0.050	-0.212	0.001	0.001

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

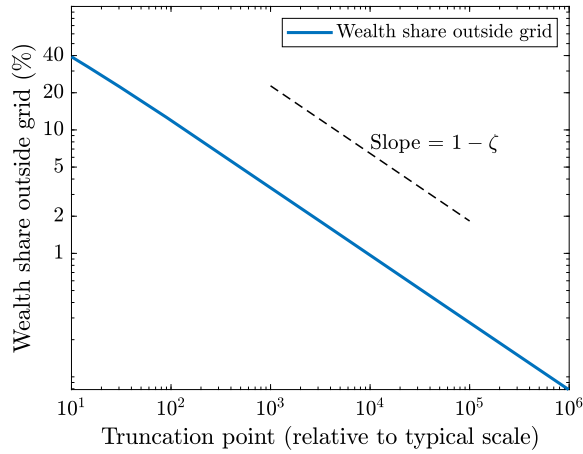


FIGURE 3. Wealth share of agents outside the grid.

point \bar{w} , define the top wealth share by $s(\bar{w}) = \int_{\bar{w}}^{\infty} w d\Gamma(w) / \int_0^{\infty} w d\Gamma(w)$, where $\Gamma(w)$ is the theoretical CDF of the wealth distribution. According to (4.2), if Γ has a Pareto tail with exponent ζ , then $s(\bar{w}) \sim \bar{w}^{1-\zeta}$ for large \bar{w} . Figure 3 plots this wealth share for truncation points $\bar{w}/K_{RA} \in [10, 10^6]$ in log-log scale.¹⁶ With $\bar{w} = 10 \times K_{RA}$, about 40% of wealth is left outside the grid. This is why it is important to correct for the truncation error using the Pareto extrapolation algorithm.

Top wealth shares Although in Table 3 we evaluated the solution accuracy using the aggregate wealth, this quantity is usually not of interest beyond evaluating the market clearing condition. One may be interested in other quantities, such as the top 1% wealth share. To address this point, we compute top wealth shares using the Pareto extrapolation method as described in Section 4.1. 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. Because in general top wealth shares need to be computed only once after solving for the equilibrium, for both truncation and Pareto extrapolation methods, we use a finer grid with $N = 1000$ points. Figure 4 plots the top wealth shares against the truncation point for $\bar{w}/K_{RA} = 10^1, \dots, 10^6$. We consider the wealthiest 0.0006% (fraction of billionaires, see footnote 1) as well as the top 0.01%, 0.1%, and 1%. We can see that the truncation method vastly underestimates top wealth shares when the truncation point \bar{w} is small, as expected. To reasonably match the wealth share of billionaires, the truncation method requires a huge truncation point

¹⁶Technically, for the semianalytical 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 $\bar{w} = 10^6 \times K_{RA}$, which we take as the truth.

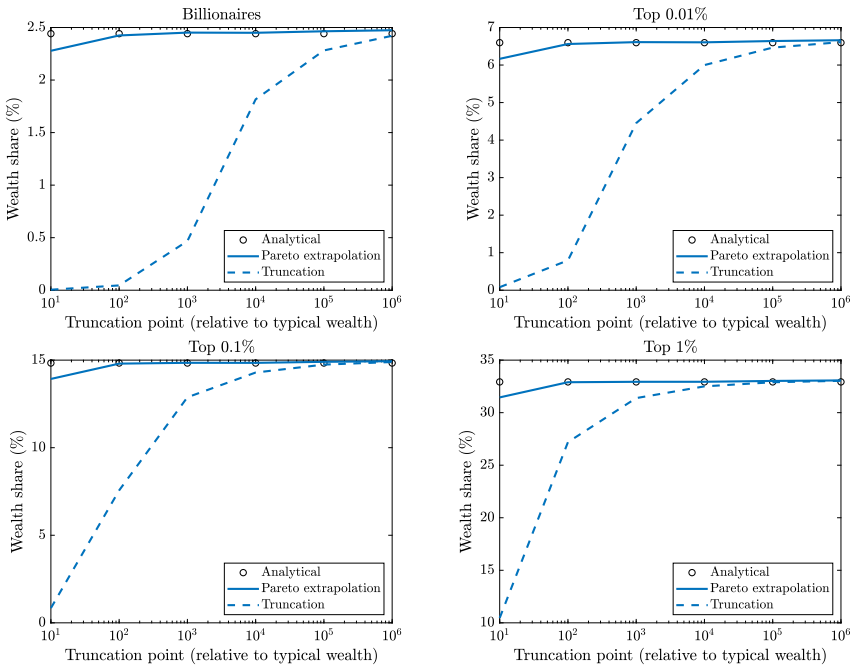


FIGURE 4. Top wealth shares in the analytical model. *Note:* “Analytical,” “Pareto extrapolation,” and “Truncation” refer to the semianalytical solution and the numerical solutions using the Pareto extrapolation and truncation methods, respectively.

such as $\bar{w}/K_{RA} = 10^4$.¹⁷ On the other hand, the Pareto extrapolation method is indistinguishable from the truth regardless of the truncation point.

Evaluation with fixed grid density Although Figure 4 verifies the accuracy of the Pareto extrapolation algorithm, it could be somewhat misleading. For instance, the accuracy of the billionaire wealth share (top left) slightly worsens as we increase the truncation point beyond 10^4 . This is because in Figure 4, the number of grid points is fixed at $N = 1000$, so the grid *density* decreases as we increase the truncation point, which could worsen the accuracy. To isolate the effect of truncation on accuracy, we now vary the truncation point, fixing the grid density. Namely, we solve the model using 300 grid points per one order of magnitude in wealth, so for example with $\bar{w}/K_{RA} = 10^4$, there are $4 \times 300 = 1200$ grid points.

¹⁷Note that a truncation point of 10^4 is extremely large. The average net worth of U.S. households was \$1.05 million in 2021Q1 (see footnote 1). Multiplying this number by 10^4 , we obtain \$10.5 billion. According to Forbes, there were only 60 households with net worth above this threshold in March 2021. A common philosophical criticism for modeling wealth with an unbounded (Pareto) distribution is that in the data there will always be a wealthiest household, and hence the wealth distribution is bounded. This criticism is misguided for two reasons. First, the largest wealth is an order statistic (random variable) and we cannot in general select an a priori upper bound. Second, what practically matters for computation is whether the wealth distribution has a large probability mass in the tail (so truncation error becomes an issue) and not necessarily whether it is bounded or not.

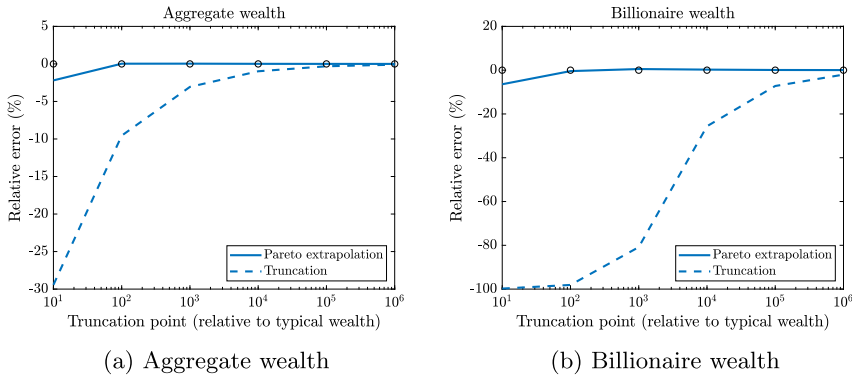


FIGURE 5. Relative error (%) of analytical KS model. *Note:* the grid is an affine-exponential grid discussed in Section 4.6. The horizontal axis shows the truncation point \bar{w} relative to the typical scale K_{RA} defined by (5.8). For each specification, the number of grid points is 300 per one order of magnitude in wealth.

Figure 5 shows the relative error in aggregate wealth and billionaire (top 0.0006%) wealth when we solve the model using Pareto extrapolation and truncation. The relative error with Pareto extrapolation is essentially none and the performance is robust across the grid specification. On the other hand, to calculate the aggregate wealth accurately (which is necessary for solving the market clearing condition), the truncation method requires a very large truncation point such as 10^4 times the “typical scale.” If an intermediate truncation point such as 10^2 is used, the truncation method predicts that billionaire wealth is zero, where in fact they hold 2.5% of aggregate wealth in the model. Even with a large truncation point such as 10^4 , the truncation method still suffers from 26% error for billionaire wealth.

APC error In Section 4.6, we argued that the APC error (4.9) can be used to guide the choice of the truncation point w_N . To see this point, Figure 6 plots (in a log-log scale) the relative errors in aggregate and billionaire wealth computed in Figure 5 as well as the

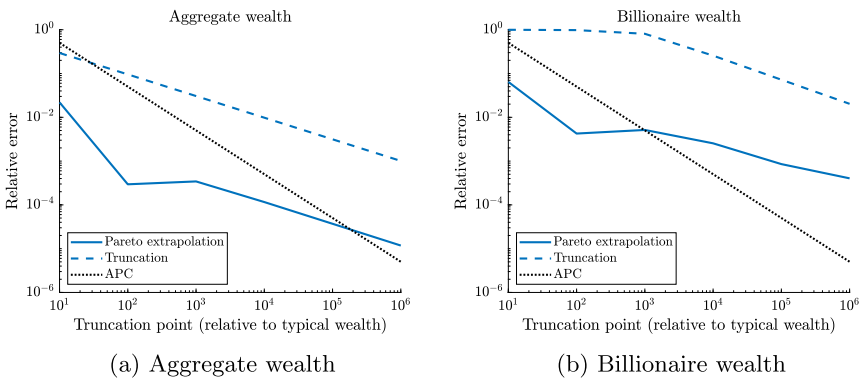


FIGURE 6. APC and solution accuracy of analytical KS model.

APC error (4.9). We can see from these graphs that when the APC error is less than 10^{-2} (1%), the relative error in the aggregate and billionaire wealth is less than 10^{-3} (0.1%), which is more than enough for practical purposes.

Which truncation error matters more? In Section 2, we argued that the truncation method suffers from two kinds of truncation errors, one when computing the transition probabilities and the other when aggregating. The Pareto extrapolation method corrects both errors as discussed in Sections 3.2 and 3.3, respectively. A natural question is which correction matters more.

To address this issue, we consider two intermediate Pareto extrapolation methods, one that only corrects the transition probabilities (as in Section 3.2) and the other that only corrects the aggregate wealth (as in Section 3.3). Figure 7 shows the relative errors in the aggregate wealth with the grid used in Figure 5.

According to Figure 7, correcting the transition probability only has a negligible impact on the solution accuracy, whereas correcting the aggregate wealth improves accuracy by an order of magnitude. However, combining both increases the solution accuracy dramatically. The intuition for this (surprising) result is as follows. In the correction term $\frac{1}{\zeta-1} \sum_{s=1}^S \pi_{sN} (1 - \bar{c}_s) w_N$ in (3.20), 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 multiplicatively. Therefore correcting only one error need not improve the accuracy.

Discussion With the truncation method, in principle one could increase the truncation point further to 10^6 to minimize the truncation error. However, that is inefficient because it increases the number of grid points, which substantially increases computing time (especially when the number of possible individual states S is large). Our approach thus allows for an accurate approximation of the full wealth distribution with few grid points. The key idea is that we approximate the right tail of the wealth distribution with

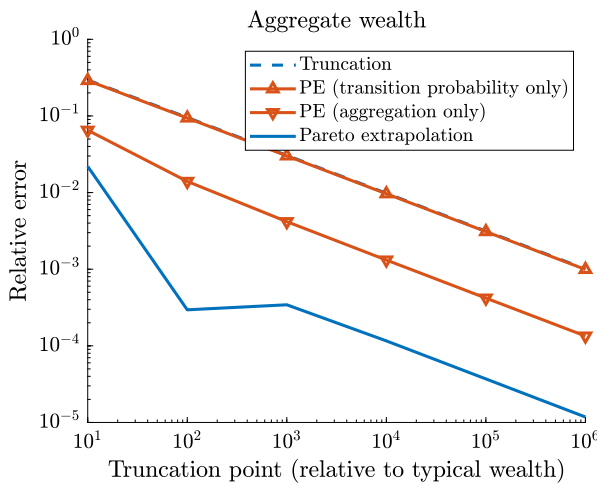


FIGURE 7. Decomposition of Pareto extrapolation method.

a single parameter (i.e., the Pareto exponent ζ), which allows us to substantially reduce the computational cost associated with solving the policy functions and wealth distribution, hence making the algorithm faster.

As is clear from the aggregate capital formula (3.20), the Pareto extrapolation method reduces to the conventional truncation method when $\zeta = \infty$. Therefore whether the truncation method is enough for practical purposes depends on the magnitude of the Pareto exponent ζ . The numerical example in Figure 5 has Pareto exponent $\zeta = 1.5$, which is the value in U.S. When we apply our algorithm to solve the [Krusell and Smith \(1998\)](#) model (without aggregate risk) with their original calibration, we find that the Pareto exponent is $\zeta = 3.23$, which is much larger. Thus the truncation error in their model is less of an issue. However, because Pareto extrapolation improves both accuracy and speed and the computation of the Pareto exponent is not possible without Pareto extrapolation, we strongly recommend using Pareto extrapolation for solving any Bewley–Huggett–Aiyagari model that features the random growth mechanism.

6. CONCLUDING REMARKS

This paper proposes a simple, systematic approach—Pareto extrapolation—to analyze and solve heterogeneous-agent models that endogenously generate fat-tailed wealth distributions. The core insight that we take advantage of is due to Pareto, who noticed that household wealth 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.

We are now much closer to understanding the economic forces that generate such a concentration of wealth. Yet, the conventional solution algorithm for heterogeneous-agent models is not well suited to handle such wealth distributions, since it relies on approximating the wealth distribution by a histogram, lumping the upper tail into a single bin. Our paper fills this gap by combining new theoretical and numerical results.

There are still many open questions that are not addressed in this paper. First, we focus on the solution algorithm for the stationary equilibrium, but many applied works consider the transition dynamics along a deterministic path. The challenge is that the upper tail of the wealth distribution is not exactly Pareto during the transition. We plan to address this issue in a separate paper. Second, we focus on models without aggregate risk. At present, the mathematical theory of Pareto tails such as [Beare and Toda \(2022\)](#) only allow for idiosyncratic risk, and extending it to the case with aggregate uncertainty remains beyond the frontier. Finally, the Pareto exponent formula (3.9) is proved only for Markov multiplicative processes (where Gibrat’s law holds exactly). We conjecture

that the formula is valid in a large class of processes that are “asymptotically linear” (see Mirek (2011) for the case with IID shocks), but establishing that seems challenging.

REFERENCES

- Achdou, Yves, Jiequn Han, Jean-Michel Lasry, Pierre-Louis Lions, and Benjamin Moll (2022), “Income and wealth distribution in macroeconomics: A continuous-time approach.” *Review of Economic Studies*, 89 (1), 45–86. doi:10.1093/restud/rdab002. [205, 214]
- Algan, Yann, Olivier Allais, and Wouter J. Den Haan (2008), “Solving heterogeneous-agent models with parameterized cross-sectional distributions.” *Journal of Economic Dynamics and Control*, 32 (3), 875–908. doi:10.1016/j.jedc.2007.03.007. [205]
- Aoki, Shuhei and Makoto Nirei (2017), “Zipf’s law, Pareto’s law, and the evolution of top incomes in the United States.” *American Economic Journal: Macroeconomics*, 9 (3), 36–71. doi:10.1257/mac.20150051. [204]
- Arkolakis, Costas (2016), “A unified theory of firm selection and growth.” *Quarterly Journal of Economics*, 131 (1), 89–155. doi:10.1093/qje/qjv039. [204]
- Beare, Brendan K. and Alexis Akira Toda (2020), “On the emergence of a power law in the distribution of COVID-19 cases.” *Physica D: Nonlinear Phenomena*, 412, 132649. doi:10.1016/j.physd.2020.132649. [204]
- Beare, Brendan K. and Alexis Akira Toda (2022), “Determination of Pareto exponents in economic models driven by Markov multiplicative processes.” *Econometrica* (forthcoming). doi:10.3982/ECTA17984. [204, 210, 229]
- Benhabib, Jess, Alberto Bisin, and Shenghao Zhu (2011), “The distribution of wealth and fiscal policy in economies with finitely lived agents.” *Econometrica*, 79 (1), 123–157. doi:10.3982/ECTA8416. [204]
- Benhabib, Jess, Alberto Bisin, and Shenghao Zhu (2016), “The distribution of wealth in the Blanchard–Yaari model.” *Macroeconomic Dynamics*, 20, 466–481. doi:10.1017/S1365100514000066. [204]
- Blackwell, David (1965), “Discounted dynamic programming.” *Annals of Mathematical Statistics*, 36 (1), 226–235. doi:10.1214/aoms/1177700285. [205, 217]
- Cagetti, Marco and Mariacristina De Nardi (2006), “Entrepreneurship, frictions, and wealth.” *Journal of Political Economy*, 114 (5), 835–870. doi:10.1086/508032. [204]
- Cao, Dan (1998), “Recursive equilibrium in Krusell and Smith (1998).” *Journal of Economic Theory*, 186, 104978. doi:10.1016/j.jet.2019.104978. [220]
- Cao, Dan and Wenlan Luo (2017), “Persistent heterogeneous returns and top end wealth inequality.” *Review of Economic Dynamics*, 26, 301–326. doi:10.1016/j.red.2017.10.001. [204]

Carroll, Christopher D., Jiri Slacalek, Kiichi Tokunaka, and Matthew N. White (2017), “The distribution of wealth and the marginal propensity to consume.” *Quantitative Economics*, 8 (3), 977–1020. doi:10.3982/QE694. [204]

Coleman II, Wilbur John (1990), “Solving the stochastic growth model by policy-function iteration.” *Journal of Business and Economic Statistics*, 8 (1), 27–29. doi:10.1080/07350015.1990.10509769. [205, 217]

Daron, Acemoglu and Dan Cao (2015), “Innovation by entrants and incumbents.” *Journal of Economic Theory*, 157, 255–294. doi:10.1016/j.jet.2015.01.001. [204]

de Vries, Tjeerd and Alexis Akira Toda (2021), “Capital and labor income Pareto exponents across time and space.” *Review of Income and Wealth*. doi:10.1111/roiw.12556. [204]

Gabaix, Xavier (2009), “Power laws in economics and finance.” *Annual Review of Economics*, 1, 255–293. doi:10.1146/annurev.economics.050708.142940. [204]

Gabaix, Xavier, Jean-Michel Lasry, Pierre-Louis Lions, and Benjamin Moll (2016), “The dynamics of inequality.” *Econometrica*, 84 (6), 2071–2111. doi:10.3982/ECTA13569. [204]

Gibrat, Robert (1931), *Les Inégalités Économiques*. Librairie du Recueil Sirey, Paris. [210]

Gouin-Bonenfant, Émilien and Alexis Akira Toda (2018), “Pareto extrapolation: An analytical framework for studying tail inequality.” <https://ssrn.com/abstract=3260899>. [203, 204]

Gouin-Bonenfant, Émilien and Alexis Akira Toda (2023), “Supplement to ‘Pareto extrapolation: An analytical framework for studying tail inequality.’” *Quantitative Economics Supplemental Material*, 14, <https://doi.org/10.3982/QE1817>. [208]

Güvenen, Fatih, Gueorgui Kambourov, Burhan Kuruscu, Sergio Ocampo-Díaz, and Daphne Chen (2019), “Use it or lose it: Efficiency gains from wealth taxation.” NBER Working Paper 26284. <https://www.nber.org/papers/w26284>. [204]

Hubmer, Joachim, Per Krusell, and Anthony A. Jr. Smith (2020), “Sources of US wealth inequality: Past, present, and future.” In *NBER Macroeconomics Annual*, Vol. 35, Chapter 6 (Martin Eichenbaum and Erik Hurst, eds.). University of Chicago Press, Chicago. [204]

Jones, Charles I. and Jihee Kim (2018), “A Schumpeterian model of top income inequality.” *Journal of Political Economy*, 126 (5), 1785–1826. doi:10.1086/699190. [204]

Kasa, Kenneth and Xiaowen Lei (2018), “Risk, uncertainty, and the dynamics of inequality.” *Journal of Monetary Economics*, 94, 60–78. doi:10.1016/j.jmoneco.2017.11.008. [204]

Kaymak, Barış and Markus Poschke (2016), “The evolution of wealth inequality over half a century: The role of taxes, transfers and technology.” *Journal of Monetary Economics*, 77, 1–25. doi:10.1016/j.jmoneco.2015.10.004. [204]

Kesten, Harry (1973), “Random difference equations and renewal theory for products of random matrices.” *Acta Mathematica*, 131 (1), 207–248. doi:10.1007/BF02392040. [204]

Krueger, Dirk, Kurt Mitman, and Fabrizio Perri (2016), “Macroeconomics and household heterogeneity.” In *Handbook of Macroeconomics*, Vol. 2, Chapter 11 (John B. Taylor and Harald Uhlig, eds.), 843–921, Elsevier, Amsterdam. doi:10.1016/bs.hesmac.2016.04.003. [203]

Krusell, Per and Anthony A. Jr. Smith (1998), “Income and wealth heterogeneity in the macroeconomy.” *Journal of Political Economy*, 106 (5), 867–896. doi:10.1086/250034. [204, 207, 220, 221, 229]

Kubler, Felix and Karl Schmedders (2005), “Approximate versus exact equilibria in dynamic economies.” *Econometrica*, 73 (4), 1205–1235. doi:10.1111/j.1468-0262.2005.00614.x. [206]

Ma, Qingyin, John Stachurski, and Alexis Akira Toda (2020) “The income fluctuation problem and the evolution of wealth.” *Journal of Economic Theory*, 187, 105003. doi:10.1016/j.jet.2020.105003. [207]

Ma, Qingyin and Alexis Akira Toda (2021), “A theory of the saving rate of the rich.” *Journal of Economic Theory*, 192, 105193. doi:10.1016/j.jet.2021.105193. [204, 208, 209]

Ma, Qingyin and Alexis Akira Toda (2022), “Asymptotic linearity of consumption functions and computational efficiency.” *Journal of Mathematical Economics*, 98, 102562. doi:10.1016/j.jmateco.2021.102562. [204, 208, 209, 217]

McKay, Alisdair (2017), “Time-varying idiosyncratic risk and aggregate consumption dynamics.” *Journal of Monetary Economics*, 88, 1–14. doi:10.1016/j.jmoneco.2017.05.002. [204]

Merton, Robert C. (1969), “Lifetime portfolio selection under uncertainty: The continuous-time case.” *Review of Economics and Statistics*, 51 (3), 247–257. doi:10.2307/1926560. [202]

Mirek, Mariusz (2011), “Heavy tail phenomenon and convergence to stable laws for iterated Lipschitz maps.” *Probability Theory and Related Fields*, 151 (3–4), 705–734. doi:10.1007/s00440-010-0312-9. [230]

Nirei, Makoto and Shuhei Aoki (2016), “Pareto distribution of income in neo-classical growth models.” *Review of Economic Dynamics*, 20, 25–42. doi:10.1016/j.red.2015.11.002. [204]

Nirei, Makoto and Wataru Souma (2007), “A two factor model of income distribution dynamics.” *Review of Income and Wealth*, 53 (3), 440–459. doi:10.1111/j.1475-4991.2007.00242.x. [204]

Pareto, Vilfredo (1897), *Cours d'Économie Politique*, Vol. 2. F. Rouge, Lausanne. [229]

Quadrini, Vincenzo (2000), “Entrepreneurship, saving, and social mobility.” *Review of Economic Dynamics*, 3 (1), 1–40. doi:10.1006/redo.1999.0077. [204]

Reed, William J. (2001), “The Pareto, Zipf and other power laws.” *Economics Letters*, 74 (1), 15–19. doi:10.1016/S0165-1765(01)00524-9. [204]

- Samuelson, Paul A. (1969), “Lifetime portfolio selection by dynamic stochastic programming.” *Review of Economics and Statistics*, 51 (3), 239–246. doi:10.2307/1926559. [202]
- Stachurski, John and Alexis Akira Toda (2019), “An impossibility theorem for wealth in heterogeneous-agent models with limited heterogeneity.” *Journal of Economic Theory*, 182, 1–24. doi:10.1016/j.jet.2019.04.001. [203, 205]
- Stachurski, John and Alexis Akira Toda (2020), “Corrigendum to ‘An impossibility theorem for wealth in heterogeneous-agent models with limited heterogeneity’ [Journal of Economic Theory 182 (2019) 1–24].” *Journal of Economic Theory*, 188, 105066. doi:10.1016/j.jet.2020.105066. [203]
- Toda, Alexis Akira (2014), “Incomplete market dynamics and cross-sectional distributions.” *Journal of Economic Theory*, 154, 310–348. doi:10.1016/j.jet.2014.09.015. [204, 208, 209]
- Toda, Alexis Akira (2019), “Wealth distribution with random discount factors.” *Journal of Monetary Economics*, 104, 101–113. doi:10.1016/j.jmoneco.2018.09.006. [204, 207, 208, 210]
- Toda, Alexis Akira and Kieran Walsh (2015), “The double power law in consumption and implications for testing Euler equations.” *Journal of Political Economy*, 123 (5), 1177–1200. doi:10.1086/682729. [204]
- Toda, Alexis Akira and Kieran James Walsh (2017), “Fat tails and spurious estimation of consumption-based asset pricing models.” *Journal of Applied Econometrics*, 32 (6), 1156–1177. doi:10.1002/jae.2564. [204]
- Vermeulen, Philip (2018), “How fat is the top tail of the wealth distribution?” *Review of Income and Wealth*, 64 (2), 357–387. doi:10.1111/roiw.12279. [223]
- Winberry, Thomas (2018), “A method for solving and estimating heterogeneous agent macro models.” *Quantitative Economics*, 9 (3), 1123–1151. doi:10.3982/QE740. [205]
- Young, Eric R. (2010), “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. doi:10.1016/j.jedc.2008.11.010. [203, 205, 211]

Co-editor Kjetil Storesletten handled this manuscript.

Manuscript received 19 January, 2021; final version accepted 21 December, 2021; available online 13 January, 2022.