Carnegie Mellon University
Tepper School of Business

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
ECONOMICS

Titled

"Essays on Overlapping Generations Models and Social Security"

Presented by

Mauro Moretto

Accepted by

Robert Miller	**5/7/2021**
Co-Chair: Prof. Robert Miller	**Date**
Stephen Spear	**5/10/2021**
Co-Chair: Prof. Stephen Spear	**Date**

Approved by the Dean

Isabelle Bajeux	**5/12/2021**
Dean Isabelle Bajeux	**Date**

Essays on Overlapping Generations Models and Social Security

Mauro Moretto

May 2021

Chapter 1

A General Equilibrium Model with Social Security Endowed with a Trust Fund and Heterogeneous Households

1.1 Introduction

In this chapter, I introduce an Overlapping Generation (OLG) model where the Social Security Security system endowed with a Trust Fund. In economic literature, Social Security has traditionally been approximated as a "pay-as-you-go" ("PAYGO") system, which fails to take into account the fact that, in reality, retirement benefits depend on the lifetime earnings of each retiree and the solvency of the Social Security fund itself. These factors are highly relevant for policy purposes because, based on current projections, Social Security will exhaust its trust fund by 2034, and will no longer be able to pay scheduled benefits to retirees. Therefore, a model that incorporates these factors represents a significant innovation in the literature and could serve as a guide for policy makers with regard to potential policy reforms. In Chapter 1, I successfully develop the proposed model of Social Security, accounting for lifetime earning histories and the potential for bankruptcy of the fund in a stochastic environment.

The proposed model contributes to the literature by introducing a trust fund, whose balances can be used to pay benefits when outlays exceed the revenues collected through payroll taxes. In a PAYGO

system, current benefits paid to the retirees are equal to the payroll taxes levied on the workers, establishing a generational link only between current workers and current retirees. On the one hand, it is able to accumulate financial resources in a designated Trust Fund, which holds non-marketable Treasury bills and bonds, but, on the other hand, the Social Security Administration is not allowed to borrow from either financial markets or other branches of the federal government if the Trust Fund depletes its resources and is not able to meet its obligations. In this event, then, under current law, benefits cannot exceed total revenues, making Social Security an unfunded system in this contingency. The model I present in this chapter takes into account this contingency, as I explicitly model the potential bankruptcy of the Social Security system by recomputing benefits to retirees according to both their employment histories and the total resources available to the Social Security system.

In addition to accounting for the potential for fund bankruptcy, the model proposed in this chapter also accounts for potential sources of heterogeneity across households. Specifically, the model incorporates differences in the life-time earnings profile and retirement age among households, and permits those factors to interact directly with some of the institutional features of the current Social Security system. The model choice allows me to investigate two potential impacts that Social Security has on households welfare. First, the inter-generational reallocation of risk from retirees to workers. Second, the redistribution of resources across households in different groups, which policymakers and the general public have traditionally treated as being important to protecting the welfare of low- and middle-income households during retirement. The literature has traditionally investigated the impact of Social Security on household welfare either by using a homogeneous household framework (see for instance Krueger and Kubler [2006], Hasanhodzic and Kotlikoff [2018] or Harenberg and Ludwig [2019]), or by not accounting for the dependence of benefits on earnings histories in the context of a model with ex-ante heterogeneous agents (see for instance Kim [2018]). My proposed model helps to address this issue by providing a more realistic model of Social Security where benefits depend on life-time earnings histories.

While the models utilized in the literature are able to account for the redistributional features of Social Security across generations, they fail to capture the fact that Social Security redistributes resources within-generations as well. While the marginal tax rate on payrolls is constant[1], retirement benefits vary significantly based on someone's life-time earnings and the rate at which Social Security retirement benefits replace labor income generally decreases as individuals earn more money. For example, at the lowest income

[1]Up to a maximum taxable amount that is determined on a yearly basis. For instance, in 2019, for the retirement portion of Social Security earnings were capped at $132,900. Above that amount, the marginal tax rate is 0%.

levels, Social Security retirement benefits can replace up to 90% of labor income, while at the highest levels, the benefits replace 30% or less than the labor earnings. Ceteris paribus, one may argue that higher income earners are subsidizing lower income retirees, since the return of retirement benefits (calculated on a yearly basis per dollar) based on social security taxes paid is higher for low-income earners than for higher income earners. This observation clearly abstracts from any consideration about differences in retirement age and life-expectancy across different groups. If, for instance, high-income earners receive retirement benefits for a longer period on average as compared to low-income earners, then a lower replacement rate may simply compensate for this factor. We will discuss these matters in greater detail in 3, where we will empirically quantify differences across household groups. At the same time, the estimation performed in 3 allows us to identify the relevant source of heterogeneity that are most relevant for the analysis. In assessing how the current Social Security system and potential policy alternatives impact household welfare, in chapter 3 I consider how the alternative policies interact with the following forms of household heterogeneity:

- Life-time earnings and retirement age, as Social Security retirement benefits directly depend on each worker's earnings history.
- Mortality and life-expectancy, as both impact how long different households will receive benefits on average and the inclusion of these factors will allow me to better characterize the dynamics of Social Security expenditures.

The assumption of the PAYGO system interacts with another feature of Social Security that we believe it is important to capture. In a PAYGO system, the government is forced to balance its budget in every period, but under the current policy regime, benefits paid by Social Security do not depend on the total revenues collected by the system. This is especially relevant in the context of a stochastic economy with aggregate shocks. In this case, the aggregate shocks would determine wages and consequently, the revenues collected through payroll taxes by Social Security. The balanced budget would then transmit the shock directly to retirees' benefits, since Social Security outlays need to match revenues. This does not capture an important characteristic of Social Security retirement benefits, i.e. that benefits are paid as annuities. Figure 1.1 shows the cyclical deviation from the trend component of the real log Social Security expenditures and revenues. As we can see, revenues appear to have a strongly cyclical behavior, with major drops associated with the early 1990s crisis, the dot-com bubble of the early 2000s and the Great Recession of 2008. At the same time, revenues appear to be far more volatile than expenditures, with the standard deviation of the cyclical component of the Social Security revenues being nearly three times that of expenditures. This

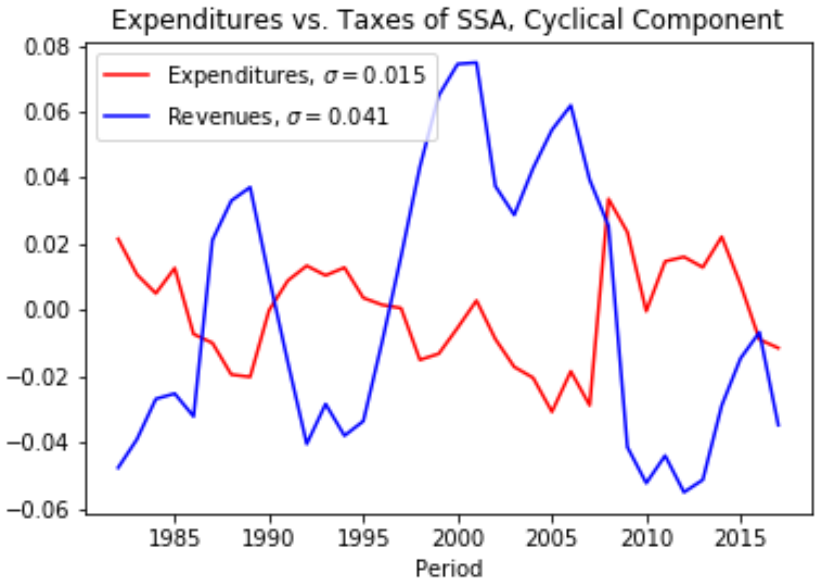

Figure 1.1: Cyclical Component of PCE-Deflated Social Security Revenues vs. Expenditures

Period: 1983-2018.

Source: Social Security Administration and Federal Reserve Bank of St. Louis.

evidence supports the fact that modeling Social Security as a PAYGO system introduces excess volatility in the benefits paid to retirees since aggregate shocks are directly transferred to retirees. This, in turn, could cause some to underestimate the role of Social Security as a form of income insurance at retirement. Incorporating a Trust Fund allows us to better capture the role Social Security in reallocating risk across generations.

There are many computational challenges presented by large-scale OLG models. Indeed, equilibrium prices and allocation will generally depend on the entire distribution of wealth or capital holdings, in addition to the varying Social Security contributions of each type and cohort of households. In general, numerical solutions are characterized by either exploiting the full state space, allowing equilibrium policy functions to depend on the entire distribution of state variables, or by using state-aggregation methods (see for instance Krusell and Smith [1998]). Depending on the numerical technique used, an appropriate equilibrium concept needs to be adopted. For instance, numerical solutions that adopt the full state space rely on fully rational agents, as they optimally choose their consumption and savings while observing the entire distribution of state variables. On the contrary, state space aggregation methods rely on quasi-full rationality, since the policy functions of agents depend only on a subset of the state space. In this Chapter, I

introduce the relevant equilibrium concepts, and in Chapter 2 I develop an algorithm to numerically compute the equilibrium policy functions based on a subset of state variables that depend on the household, utilizing the representational power of deep neural networks.

This chapter is structured as follows. In Section 1.2, I provide an overview of the Social Security Administration, focusing on its role in providing retirement benefits. In Section 1.3, I review the literature, focusing on OLG models and Social Security. In Section 1.4, I introduce the proposed model for Social Security. In Section 1.5, I introduce the relevant equilibrium concepts, and discuss how they interact with the numerical solution proposed in Chapter 2.

1.2 Overview on the Social Security Administration

Social Security is the largest federal government benefit program, distributing approximately 1 trillion dollars to nearly 52 millions of Americans in 2018 alone. Social Security provides three types of insurance: retirement insurance under the Old Age Survivors Insurance (OASI) program, Disability Insurance (DI), and Medicare System Hospitalization Insurance (HI). Benefits to retired workers and their families, in addition to families of deceased workers are paid through the OASI program. Similarly, benefits to disabled workers and their families are paid through the DI program.

Figure 1.2 displays the breakdown of the expenses of the federal government in 2018. The chart clearly shows that the OASI component of the Social Security represents the largest single item in the expenditure side of the federal budget, constituting approximately 22.5% of the total outlays. By way of comparison, the budget for defense only constituted 14.6% of the total expenditures. Social Security is financed through payroll taxes known as FICA and SECA, levied respectively on employed and self-employed individuals. SECA and FICA both contribute to Social Security retirement and disability programs, as well as to Medicare. For the OASI and DI component, employees and employers pay a combined tax rate of 12.4%[2] on employees' labor income, equally partitioned among the employers and employees. The tax applies up to a yearly-adjusted cap, above which the marginal tax rate is 0%. For self-employed individuals, the tax burden fall completely on the individual. Of the 12.4% tax rate, 10.6% contributes to the OASI, while the remaining 1.8% goes to the DI component of Social Security[3]. The FICA and SECA taxes levy an additional

[2] Source: www.ssa.gov/oact/progdata/taxRates.html

[3] The current tax rates were established in 2000, and, except minor modifications in the 2016-2018 period, they have remained unchanged until today.

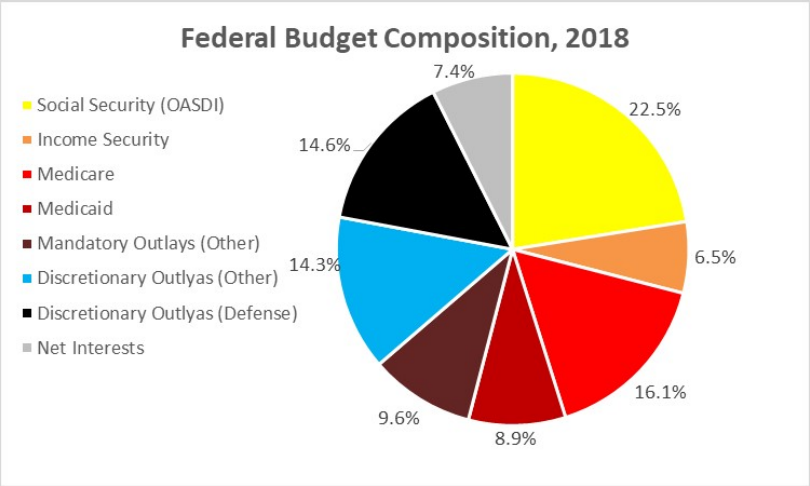

Figure 1.2: Source: Congressional Budget Office.

2.9% tax rate on payrolls for Medicare (HI portion), up to a maximum of $200,000 (or $250,000, depending on the filing status)[4].

The revenues collected are deposited into three separate trust funds, and each type of program (OASI, DI and HI) is financed by the respective trust fund. The three trust funds are separate entities by law. The trust funds, by holding asset reserves, serve a fundamental purpose: these accumulated reserves provide automatic spending authority to pay benefits, since the Social Security Act of 1935 limits trust fund expenditures to benefits and administrative costs and does not allow Social Security to borrow in order to disburse benefits. While the size and the scope of both the DI program and Medicare are very significant, and interact in significant manners with the OASI program, in this paper I focus only on the retirement portion of the OASI program. This choice is motivated by the fact that the the payments of benefits to retired workers represent by far the largest expenditure item in the Social Security budget in 2018. Social Security disburses benefits to four categories of people: retired workers, based on their earning histories, spouses of retired workers, survivors of retired workers and disabled workers. As we can from Figure 1.3, the payment of benefits to retired workers represented 69.4% of the total amount of benefits paid out by Social Security in 2018. This compares to a total of 16.1% distributed to spouses and survivors[5], and 14.5% to disabled workers. Therefore, in this paper we will abstract from any considerations related to either spousal survivor

[4]This provision was added in 2013 through the Affordable Care Act. Above this threshold, employees are responsible for paying an additional 0.9% on payrolls exceeding the cap.

[5]The number of people claiming spousal benefits peaked at 3.1 million in 1992, and was 2.4 millions in 2018, down by 23.1% from the peak. This can be at least partially explained by the increased labor participation of women starting from the 1970s, allowing them to qualify for Social Security benefits trough their earning histories instead of that of their spouses.

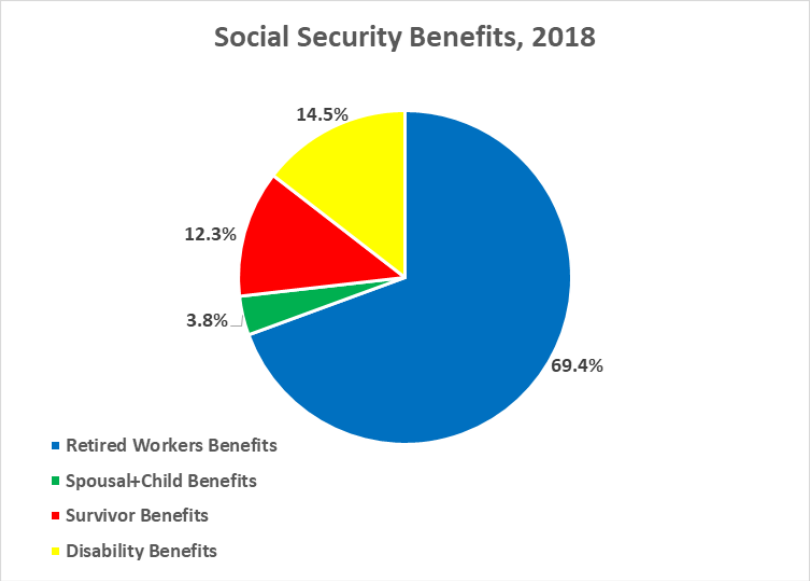

Figure 1.3: Source: Social Security Administration.

benefits or disability and Medicare related benefits.

In order to understand how Social Security interacts with the life-cycle consumption-saving decisions of households, we first need to understand how Social Security retirement benefits are computed. For this reason, we describe how the retirement benefits are computed, and how this affects the modeling choices made later in the model presented in Section 1.4. Every worker with at least 10 quarters of eligible employment history is entitled to receive Social Security retirement benefits based on his or her earning record. For any worker, the computation of the benefits takes into account two main factors: (1) the entire earning history, and (2) when the worker files for benefits. Regarding earning history, Social Security uses the highest 35 years of indexed earnings to compute a worker's benefits. The system uses the Average Wage Indexing Series (AWI) for those purposes. Regarding when a worker decides to file for the benefits, anyone can start collecting retirement benefits as early as 62 years of age, but benefits are adjusted depending on the age at filing. Full Retirement Age (FRA) is defined as the age at which a worker is entitled to receive full benefits. FRA has been changed over the last decades to account for increase in life-expectancy and better health conditions of older workers. The original Social Security Act of 1935 set the FRA at 65. In 1983, lawmakers amended Social Security to create a gradual system of increases to the FRA. Based on these increases, the FRA reached 66 in 2019 and is slated to increase further to 67 by 2027.

Benefits at FRA are called the Principal Insurance Amount (PIA), and are determined through a three-step process. First, a worker's previous earnings are restated in terms of current wages by indexing past

earnings (up to age 60) to wage growth. Second, the highest 35 years' earnings are averaged to a monthly amount called the Average Indexed Monthly Earning (AIME). It is important to keep in mind that nominal earnings are capped by the contribution and benefit base, which is determined by the yearly-set maximum taxable income. The PIA is calculated as a piece-wise linear function of AIME. For instance for someone retiring in 2019, the PIA would be the following function of AIME: 90% of the first $926 of the AIME plus 32% of the AIME between $926 and $5,583 15% of the AIME above $5,583. The maximum AIME in 2019 corresponds to $10,296, as is the result of the cap that is imposed on yearly earnings. While it is clear that PIA is (weakly) increasing with life-time earnings, Figure 1.4 shows the replacement rate of the PIA as a function of AIME. As we can see, the replacement rate is weakly decreasing in the AIME. The downward slope highlights the progressive nature of Social Security benefits: workers with higher life-time earnings have lower share of labor income replaced by Social Security benefits compared to individuals with lower life-time earnings. It is important to notice that, in general, a low AIME can be the result of two factors: (1) a relatively short working history (i.e. the worker worked only a few years), or (2) low annual earnings on average. In this paper, we are focusing on workers who have displayed a significant attachment to the labor force. Therefore, in this case, a low AIME will be the result of low annual earnings on average.

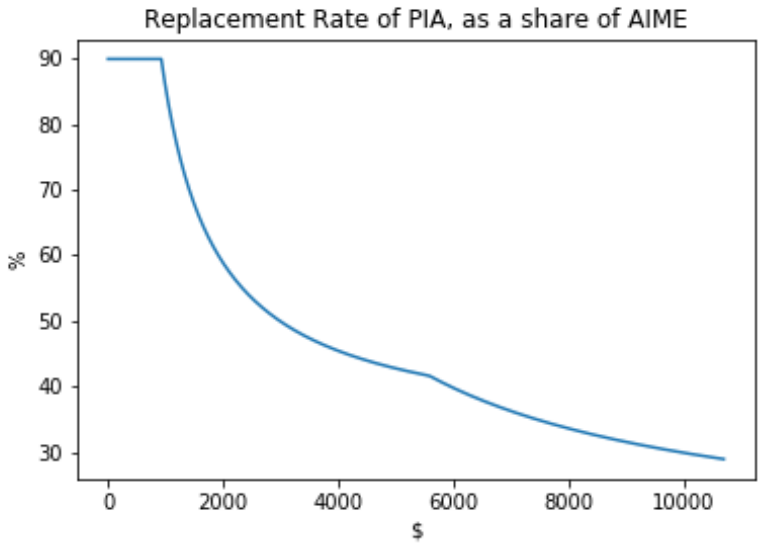

Figure 1.4: Replacement Rate at Full Retirement Age, as of 2019.

In addition to life-time earnings, another important factor that affects the computation of retirement benefits is the age at which a person decides to first apply for Social Security benefits. Benefits are reduced for anyone claiming benefits before FRA, and are increased for anyone who delays collection until after reaching FRA. Once a worker reaches the age of 70, there are no further adjustments, meaning there

are no additional incentives for postponing the collections of benefits beyond that age. In case of early retirement, benefits are reduced by $\frac{5}{9} \times 1\%$ for each month before FRA for up to 36 months; if early retirement anticipates FRA by more than 36 months, then the benefit is further reduced by $\frac{5}{12} \times 1\%$ for each month. Assuming a FRA of 66, this translates to a deduction of 5% per year between the age of 62 and 63, and 6.7% between the age of 63 and 66. If a worker decides to retire after FRA is reached, benefits will accrue by 8% a year, until the age of 70 is reached.[6] Figure 1.5 shows the amount of benefits received relative to FRA as a function of the age of the claimant.

The adjustments made for early retirees, i.e. people retiring before reaching FRA, are computed in such a way to be *actuarially fair* for a person with the average life-expectancy: the expected present value of the total benefits received retiring before FRA and at FRA is the same. On the contrary, benefits are recomputed in an actuarially unfair way for workers deciding to postpone retirement. It is important

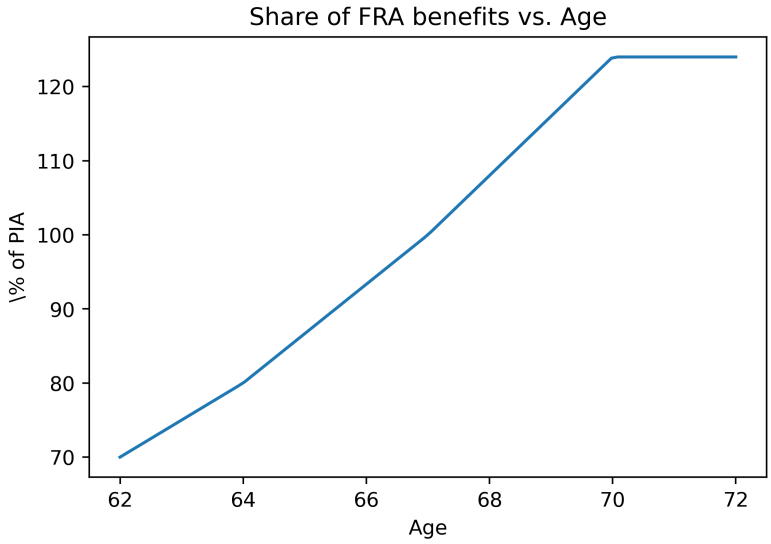

Figure 1.5: Share of Social Security benefits at Full Retirement Age as a function of the age of the claimant, based on a person deciding to retire in 2019.

to consider that the relationship between Social Security benefits and retirement age was designed to be actuarially fair. The option to retire early was first introduced 1956 by Congress for women, and then extended to men in 1961. However, the extent to which these adjustment are still actuarially fair for the average worker depend on the current and future interest rates and on projected life-expectancy. Compared to the time in which these adjustments where made, life-expectancy has increased and interest rates have

[6] Retirement credits do not apply after age 70.

significantly decreased. In addition, the notion of actuarially fair generally applies to the average worker with an average life-expectancy. As documented in the literature, and explored empirically in Chapter 3, life-expectancy positively correlates with measures of socio-economic status, like wealth, life-time earnings, or education. Therefore, actuarial adjustments made in the 1960s are likely to have disparate impacts on workers at different earnings levels today. If life-expectancy and life-time earnings are positively correlated, then part of the redistributional effects of a decreasing replacement rate would be offset by the fact that Social Security benefits do not take into account heterogeneous life expectancy at retirement.

As mentioned above, the OASI benefits are paid by using reserves in the trust fund. These very reserves give the right to the Social Security Administration to pay benefits. Reserves are invested in special non-marketable Treasury bonds that are guaranteed by the full faith of the U.S. Government, which pay a market rate of interest on the held bonds. When Social Security's trust funds run out of reserves and expenditures exceed revenues, Social Security will be able to pay out benefits only up to the total value of revenues. In this case, Social Security will become insolvent, since benefits can only be paid as long as the trust funds actually have assets to draw on to pay them. There are multiple source of income for the Social Security OASI fund: payroll taxes, taxes on the benefits themselves, and interest earned on the Treasury bonds held. In 2019, payroll taxes accounted for the vast majority of Social Security income (87.7%), followed by interest income (8.5%), and taxes levied on the benefits (3.8%). On the cost side, benefit payments account for nearly the totality of the outlays (99.1%).

Since 1983, the Social Security Administration has run an annual budgetary surplus, the peak in the trust funds reserve is expected to be reached sometime in or around 2021. As a result of a nearly thirty-year long budgetary surplus, the OASI trust fund held reservea of approximately 2.9 trillion dollars at the end of 2019, representing 13.5% of nominal US G.D.P. Current Social Security forecasts anticipate that the reserves accumulated in the OASI trust fund will be depleted by 2035.[7] After reserve exhaustion, without policy reform, the Social Security Administration will only be able to pay 73% of the total accrued benefits in the long-run. In addition, Social Security may, in the short term, incur solvency problems even before 2035. Given that the current policy regime is not sustainable, it becomes important to analyze alternative policy scenarios to assess how different policies impact household welfare. A policy intervention in this context is not only advisable, but required in order to keep Social Security alive. We will discuss different policy alternatives in Chapter 3.

[7] *Source:* https://www.ssa.gov/policy/trust-funds-summary.html

1.3 Literature Review

The literature has long focused on the welfare effects of Social Security, and research on this topic goes back to Diamond [1965], with a particular focus on the impact of Social Security on inter-generational risk-sharing. In this context, the welfare implications of Social Security have been analyzed by modeling Social Security as a PAYGO in the context of OLG with homogeneous and heterogeneous households. Diamond [1977] acknowledges that Social Security has both inter-generational and intra-generational redistributional effects. Diamond argues that while Social Security cannot be justified as the sole means of reallocation of resources between low-earners and high-earners, it can still improve household welfare when markets are incomplete (for instance, in the case of imperfect annuities markets or in the absence of risk-free assets). Similarly, Bodie and Shoven [1983] argues that Social Security can improve household welfare when there is another intrinsic market imperfection.

Krueger and Kubler [2004] show that in an economy with incomplete markets,[8] the introduction of a PAYGO Social Security is Pareto-improving, even when the economy is dynamically efficient,[9], if the crowding out-effect on private capital are not taken into account. If, however, crowding-out of private capital is included, then a Social Security system delivers a lower ex-ante expected utility. Their analysis abstracts from redistributional effects across households within the same cohort, since they consider a framework in which households are ex-ante homogeneous. Kim [2018] investigates the impact of a Social Security system modeled as a PAYGO system in the context of an OLG model with ex-ante heterogeneous agents, where households differ in their preferences and life-time incomes. Their analysis shows that the impact of Social Security varies across households, with some households benefiting from Social Security, and others not.

Under certainty, the computation of the numerical approximation of rational expectations equilibrium is relatively straightforward and usually involves the use a combination of fixed-point iteration and backward induction (see for instance Auerbach and Kotlikoff [1987], and, for a specific application to this setting, please refer to A.1). Popular solutions to stochastic general equilibrium models to projection methods are first introduced in the economic literature by Judd [1992]. Projection-based algorithms can be generally divided in two categories: (1) state-space aggregation (see for instance, Krusell and Smith

[8] Market incompleteness is a key assumption because if markets were complete, then Social Security would be a redundant financial instrument, that would have no impact on reallocating resources across generations.

[9] Efficiency follows the definition of Demange [2002] and in the context of incomplete markets refers to the fact that there is no allocation that can be spanned by markets that can improve upon.

[1998]), and (2) full state space (see for instance Krueger and Kubler [2004],Krueger and Kubler [2006]). Aggregation-based methods have been developed to make large-scale problems that suffer the curse of dimensionality tractable, and rely on a quasi-full rationality assumption. Krueger and Kubler [2004] show that the use of the entire state space tends to deliver more precise numerical solutions as compared to state aggregation based methods. They suggest that the performance of aggregation-based methods is inferior in models where agents' propensity to save differs significantly. This is likely to be the case in large-scale OLG models in which agents borrow when they are young. We discuss alternative equilibrium definitions in Section 1.5, and we show in Chapter 2 that a numerical solution developed used a reduced state-space and a quasi-rationality assumption performs well, at par with state-of-the-art full-state space models.

Other techniques have been developed to tackle the curse of dimensionality, and among the most notable ones, we find the use of sparse grids. Krueger and Kubler [2004] introduces the use of Smolyak sparse grids (see Smolyak [1963]) in the context of OLG models, and Judd et al. [2014] shows its numerical properties in a broader context. Hasanhodzic and Kotlikoff [2018] and Reiter [2015] propose solutions of large-scale OLG models, whose projection algorithm rely on the use of sparse grids, and linear polynomials as the functional base. Similarly, Kim [2018] also relies on non-aggregation methods and linear polynomials in the study of the effect of the introduction of a PAYGO Social Security system on heterogeneous welfare households. The solution algorithm developed in Chapter 2 will rely on the use of sparse grids, as well as simulated data, as in Maliar and Maliar [2015] and Azinovic et al. [2019].

1.4 Model

In this Section, I introduce the OLG model with ex-ante heterogeneous agents where Social Security pays retirement benefits and is endowed with a Trust Fund.

1.4.1 Households

It is assumed that the economy is closed. Time is discrete, labeled $t = 1, \cdots, +\infty$. The economy is populated by different types of households. Let $\mathcal{I}$ denote the set of households. Households are heterogeneous at birth, and there are $I = |\mathcal{I}|$ types of households. All households live for a maximum of A periods. Let $\mathcal{A}$ denote the set of cohorts. Each period, a new cohort of households in born for each type. Let $i \in \mathcal{I}$ be the type of

household, and n the age-cohort they belong to. Households value consumption, and the utility flow from consumption $c_{in,t}$ for a household of type i is expressed by:

$$u_i(c_{in,t}) = \frac{c_{in,t}^{1-\sigma_i}}{1-\sigma_i}, \quad c_{in,t}, \geq 0 \tag{1.1}$$

As we can see from Equation (1.1), households have a Constant Relative Risk Aversion (CRRA) utility function, whose coefficient of relative risk-aversion σ_i depends on the household type i. The utility function expressed in (1.1) is strictly increasing, strictly concave, twice continuously differentiable and satisfies the Inada conditions, i.e. $\lim_{c\to+\infty} u'(c) = 0$ and $\lim_{c\to 0} u'(c) = +\infty$.

Households discount future utility geometrically, and are endowed with a type-specific discount factor β_i. For a household of type i who is born at period t, the total expected life-time utility can be expressed as:

$$\mathbb{E}_0\left[\sum_{\tau=1}^{A} \beta_i^{\tau} u_i\left(c_{i\tau,\tau}\right)\right] \tag{1.2}$$

where the conditional expectation is taken with respect to the information set available to the household at time t.

I assume that capital and labor markets are perfectly competitive, so that that households act as price takers. On the capital market side, households have one asset available that can be used to either to borrow or save; the asset earns a stochastic return of r_t. I define as $k_{in,t}$ the amount of the asset held by household of type i of age n at time t. In the presence of an aggregate shock, this implies that market are sequentially incomplete, as the capital asset is used to smooth consumption across time and to insure consumption against aggregate shocks.

Households supply labor inelastically until they reach retirement age $R_i < A$, after which they retire and exit permanently from the labor market. Households are characterized by type and age-specific labor productivities, which are assumed to be a deterministic function of age taking the following form:

$$n_{in} = \begin{cases} f_i(n), & n = 1, \cdots, R_i - 1 \\ 0, & n = R_i, \cdots, A \end{cases} \qquad \forall i \in \mathcal{I} \tag{1.3}$$

In each period, workers earn a real wage rate per efficiency unit w_t, so that the total labor income they earn is $w_{in,t} = w_t n_{in}$. Labor income is taxed at a marginal rate equal to τ. Retired households collect Social Security benefits, which are denoted by $ss_{in,t}$. I assume that households receive inheritances in the form

of accidental bequests, which I define as $b_{in,t}$. During the life-cycle, households face the following set of sequential budget constraints:

$$c_{in,t} + k_{in+1,t+1} \leq w_{in,t}(1-\tau) + b_{in,t} \qquad \text{for } n = 1 \tag{1.4}$$

$$c_{in,t} + k_{in+1,t+1} \leq w_{in,t}(1-\tau) + b_{in,t} + r_t k_{in,t} \qquad \text{for } 2 \leq n \leq R_i - 1 \tag{1.5}$$

$$c_{in,t} + k_{in+1,t+1} \leq ss_{in,t} + b_{in,t} + r_t k_{in,t} \qquad \text{for } R_i \leq n \leq A - 1 \tag{1.6}$$

$$c_{in,t} \leq ss_{in,t} + b_{in,t} + r_t k_{in,t} \qquad \text{for } n = A \tag{1.7}$$

Equation (1.4) shows that each newly born cohort has no capital endowment. In their working year $n = 1, \cdots, R_i - 1$, households receive capital income from their investments k_{in} (Equation (1.5) and from labor. Once households retire at age R_i, they start collecting Social Security benefits ss_{in} (Equation (1.6)) and stop earning labor income. Finally, Equation (1.7) shows the budget constraints of the last period of a household's life-cycle: households consume all their wealth since they do not value bequests.

At any period t, household (i, n) makes a consumption-saving decision based both on individual state variables - average earning profiles and capital holdings - and aggregate states variables - the aggregate shock, the Social Security Trust Fund balance, the distribution of capital holdings and of average earnings of other households. Given the budget constrains described by Equations (1.4)-(1.7), a household's (i, n) problem can be summarized as follows:

$$V_{in}(\mathbf{x}, \mathbf{z}) = \max_{c_{in}, k'_{in+1}} u_{in}(c_{in}) + \beta_i \mathbb{E}\left[V_{in+1}\left(\mathbf{x}', \mathbf{z}'\right) | \left(\mathbf{x}_{in}, \mathbf{x}, \mathbf{z}\right)\right] \quad \text{subject to} \tag{1.8}$$

$$c_{in} + k'_{in+1} = n_{in} w(\mathbf{x}, \mathbf{z})(1-\tau) + k_{in} r(\mathbf{x}, \mathbf{z}) + b_{in,t}(\mathbf{x}, \mathbf{z}) + \mathbf{1}(n \geq R_i) ss_{in}$$

$$\mathbf{z}' \sim f_{\mathbf{z}}(\mathbf{z}'|\mathbf{z})$$

$$\mathbf{x}' = \pi(\mathbf{x}, \mathbf{z}, \mathbf{z}')$$

where the mapping π represents the law of motion of:

- households' capital holdings
- households' cumulative average life-time earnings
- Social Security Trust Fund balance

We can interpret π as a forecast function that agents use to predict future distribution of endogenous variables given the stochastic process of the exogenous shock and optimal choice of households.

1.4.2 Representative Firm

On the production side, I assume that there is a representative firm in the economy. It takes as inputs capital and labor, which it uses to produce a homogeneous final good that can be used for either consumption or investment. The firm is assumed to have a Cobb-Douglas Constant Returns to Scale (CRS) technology, whose output is defined as follows:

$$Y_t = A_t F(K_t, N_t) = A_t K_t^{\alpha} N_t^{1-\alpha} \tag{1.9}$$

In Equation (1.9), A_t represents the aggregate technological shock and δ_t the depreciation shock. Market are perfectly competitive and the firm acts as a price-taker. The representative firm maximizes profits choosing capital and labor, given the price for capital r_t, the price of labor per efficiency unit w_t and the shocks (A_t, δ_t):

$$\max_{K_t, N_t} A_t F(K_t, N_t) + (1 - \delta_t) K_t - r_t K_t - w_t N_t \tag{1.10}$$

The profit maximization problem of the firm results in wage and rental rate of capital being equalized to the marginal product of labor and capital respectively in equilibrium. Under the assumption of a Cobb-Douglas production function, this leads to the following relationships between marginal product of labor and capital and wage (per unit of efficiency) and the interest rate:

$$w_t = A_t F_N(K_t, N_t) = (1 - \alpha)\, A_t K_t^{\alpha} N_t^{-\alpha} \tag{1.11}$$

$$r_t = A_t F_K(K_t, N_t) = \alpha A_t K_t^{\alpha - 1} N_t^{1-\alpha} + (1 - \delta_t) \tag{1.12}$$

1.4.3 Technology

I assume that the aggregate shock A_t and the depreciation shock δ_t jointly follow a first-order discrete Markov process which is exogenously given. I define the vector of aggregate exogenous state variables as $\mathbf{z}_t = (A_t, \delta_t)$. In particular, I assume that:

$$\log A_{t+1} = \rho \log A_t + \varepsilon_{t+1}, \quad \varepsilon_{t+1} \sim \mathcal{N}\left(0, \sigma_{\varepsilon}^2\right) \tag{1.13}$$

$$\delta_t \sim \mathcal{N}\left(\overline{\delta}, \sigma_{\delta}^2\right) \tag{1.14}$$

As we can see from Equation (1.13), the technology A_t shock follows an AR(1) process, with the serial-correlation being ruled by the parameter ρ. From Equation (1.14), the depreciation shock is identically and

independently distributed through time, following a normal distribution with mean $\overline{\delta}$ and standard deviation σ_δ. Following Kim [2018] and Hasanhodzic and Kotlikoff [2019], I assume that the technology and the depreciation shocks are uncorrelated. The introduction of the depreciation shocks allow us to (i) eliminate the perfect correlation between wages and interest rates that would result from the presence of only a technology shock, and (ii) have different volatilities in wages and interest rates. This allows us to isolate more explicitly the sources of risk that characterize the stream of income of workers and retirees: while workers' main source of income are wages which are subject to the technology shock, retirees finance their consumption through the Social Security benefits and their returns on savings, whose interest rate is driven jointly by the technology and the depreciation shock. At the same time, the two sources of shocks have clear impacts on both retirees and workers. For retirees, a negative depreciation shock hurts them as they tend to be the primary savers in the models. As for workers, the depreciation shocks translates into a reduction in the capital stock, leading to a decrease in wages in the short-term.

1.4.4 Demographics

I assume that demographic variables evolve in a deterministic way. In each period t, a new cohort of size $P_{i1,t}$ is born for household of type $i \in \mathcal{I}$. Households are subject to mortality risk, which is both age and type specific. In each period, a fraction μ_{in} of households belonging to type i and cohort n die. It is assumed that the mortality rate does not depend on time. Finally, I assume that once a household reaches the age of A they die with probability one. I define $P_{in,t}$ as the measure of households of type i belonging to cohort n that are alive at time t. Given these assumptions, the population dynamics can be described as follows:

$$\begin{aligned} P_{in,t} &= (1 - \mu_{in-1})\, P_{in,t-1} \quad \forall i, \quad \forall n \\ P_{i1,t} &= \overline{P}_{i1} \end{aligned} \tag{1.15}$$

It is worth noting in Equation (1.15) that the size of the newly born household does not depend on time. Considering that households face mortality risk, they leave accidental bequests to the surviving households. I make the following assumption about the timing at which death occurs, and how wealth of the deceased households is redistributed across the surviving agents.

1. At the beginning of each period t, the aggregate shock $\mathbf{z}_t$ is realized, agents supply labor and capital to the representative firm, earn labor and capital income, and pay payroll taxes to Social Security.

2. A fraction μ_{in} of type i households belonging to cohort n dies; the net assets $r_t k_{in,t} + w_{in,t}(1-\tau)$ of the deceased households become accidental bequests, which is redistributed across the surviving households.

3. Social Security disburses payments to the surviving households in the form of retirement benefits.

4. The surviving households make a consumption/investment decision.

It is assumed that each surviving household (i,n) receives a share η_{in} of the aggregate bequest B_t. Therefore, it is possible to express the bequests received by household i belong to cohort n at time t as follows:

$$B_t = \frac{\sum_{j=1}^{I}\sum_{m=1}^{A}\mu_{jm}P_{jm,t}\left(w_{jm,t}\left(1-\tau\right)+r_t k_{jm,t}\right)}{P_{in,t}(1-\mu_{in})}$$

$$b_{in,t} = \eta_{in}B_t \qquad \forall (i,n) \in \mathcal{I}\times\mathcal{A} \tag{1.16}$$

As we can see from Equation (1.16), I assume that each household bequest is a constant fraction η_{in} of the total bequests, depending on the type and the cohort, that is time independent. For simplicity, I assume that each household receives the same share of aggregate bequests, so that we express η_{in} as follows:

$$\eta_{in} = \frac{1}{\sum_{j\in\mathcal{I}}\sum_{m\in\mathcal{A}}\mu_{jm}P_{jm}} \qquad \forall (i,n) \in \mathcal{I}\times\mathcal{A}$$

1.4.5 Government and Social Security Administration

Households pay payroll taxes on their wages, so that total receipts T_t of the government at time t are:

$$T_t = \sum_{i=1}^{I}\sum_{n=1}^{R_i-1}\tau P_{in,t}w_{in,t} = \tau N_t w_t \tag{1.17}$$

Social Security pays retirement benefits to households once they reach retirement age. The benefits depend on the average life-time earning at retirement age $\overline{e}_{iR_i,t}$ and the retirement age R_i. From a state-space perspective, we need to keep track of the distribution of cumulative average labor income, whose dynamic can be summarized as follows:

$$\overline{e}_{in,t} = \frac{n-1}{n}\overline{e}_{in-1,t-1} + \frac{1}{n}w_{in,t} \qquad \forall t \geq 0,\ \forall n = 1,\cdots,R-1,\ \forall i \in \mathcal{I} \tag{1.18}$$

I assume Social Security benefits $\hat{ss}_{in,t}$ are proportional to the average-indexed lifetime earning through a factor of θ_i, so that we can establish the following relationship between average life-time income, retirement age and Social Security benefits:

$$\hat{ss}_{iR_i,t} = \theta_i \overline{e}_{R_i-1,t-1} \quad \forall i \in \mathcal{I} \tag{1.19}$$

The coefficient θ_i is called replacement rate; it represents the share of the Social Security benefits to the average life-time income, and, as we can see from Equation (1.19), it is an exogenous parameter. It captures two different features of Social Security. First, the amount of benefits received by retirees does not depend linearly on the average life-time labor income. Thus, θ_i decreases with $\overline{e}_{R_i,t-1}$. Second, the benefit amounts also depends on retirement age. The earlier a household starts collecting retirement benefits, the lower the benefits will be. In this model, given that both income profiles and retirement age are type-specific, θ_i will also depend on the household type. A lower θ_i (determined by either a lower retirement age or higher life-time income) will increase private savings. Thus, the Social Security system will directly interact with households' consumption savings decision depending on the size of the replacement rate θ_i. Once households have reached retirement age R_i, they stop paying payroll taxes, and their Social Security benefits will remain unchanged for the rest of their lives:

$$\hat{ss}_{iR_i+\tau,t+\tau} = \hat{ss}_{iR_i,t} \qquad 0 \leq \tau \leq A - R_i, \quad \forall i \in \mathcal{I} \tag{1.20}$$

As previously mentioned, Social Security pays retirement benefits to households after the mortality shock is realized, and it is distributed only to the surviving households. Therefore, the total amount of retirement entitlements S_t can be expressed as follows:

$$\hat{S}_t = \sum_{i=1}^{I} \sum_{n=R_i}^{A} \hat{ss}_{in,t}(1 - \mu_{in}) P_{in,t} \tag{1.21}$$

Social Security balances are invested in the representative firm, and they are paid a return on investment equal to the equilibrium interest rate. This assumption departs from how Social Security invests the resources accumulated in its trust fund. Social Security funds are invested in non-marketable government securities, whose interest rates can be considered risk-free. On the contrary, in this model, the return to investment for Social Security is stochastic, and subject to aggregate uncertainty.

Social Security is able to fulfill its obligation toward retirees only if it is solvent; that is, if Social Security expenses do not exceed the sum of the taxes levied on households' payrolls and the accumulated assets of the trust fund. When that is no longer the case, the Social Security system is considered to be insolvent or in a state of bankruptcy. In light of these assumptions, I define the Social Security expenditures as follows:

$$S_t = \begin{cases} T_t + r_t H_t & \text{if } \hat{S}_t > T_t + r_t H_t \\ \hat{S}_t & \text{if } \hat{S}_t \leq T_t + r_t H_t \end{cases} \tag{1.22}$$

If Social Security is insolvent, then it will pay only a fraction of the benefits. The ratio is determined by the share of (total revenues + asset) to the total entitlements, which can be expressed as follows:

$$ss_{in,t} = \hat{ss}_{in,t} \frac{S_t}{\hat{S}_t} \tag{1.23}$$

This assumption is made to better capture how Social Security currently works. Let H_t be the balance accumulated on the Trust Fund at time t. Then its law of motion can be expressed as follows:

$$H_{t+1} = H_t r_t + T_t - S_t \tag{1.24}$$

As we can see from Equation (1.24), the Social Security Trust Funds earns the prevailing market interest rates on the invested assets.

1.4.6 Markets

At equilibrium, the following market clearing conditions need to hold:

$$N_t = \sum_{i=1}^{I} \sum_{n=1}^{A} n_{in} P_{in,t} \qquad \forall t \geq 0 \tag{1.25}$$

$$K_t = \sum_{i=1}^{I} \sum_{n=1}^{A} k_{in,t} P_{in,t} + H_t \qquad \forall t \geq 0 \tag{1.26}$$

$$C_t = \sum_{i=1}^{I} \sum_{n=1}^{A} c_{in,t}(1 - \mu_{in,t}) P_{in,t} \qquad \forall t \geq 0 \tag{1.27}$$

$$K_{t+1} = \sum_{i=1}^{I} \sum_{n=1}^{A-1} k_{in+1,t+1}(1 - \mu_{in,t}) P_{in,t} + H_{t+1} \qquad \forall t \geq 0 \tag{1.28}$$

$$Y_t = C_t + K_{t+1} - (1 - \delta_t) K_t \qquad \forall t \geq 0 \tag{1.29}$$

Equations have been derived by taking into account the fact that a mortality shock hits each cohort and type of households after they supplied labor and capital to the market, but before the investment-consumption decision is made.

1.5 Equilibrium

We now discuss the equilibrium definition chosen in the context of the problem, and how it interacts with the solution algorithm.

Definition 1 (Recursive Competitive Equilibrium)**.** *A recursive competitive equilibrium is characterized by a forecast function π, a pair of individual functions (V_{in}, C_{in}) for each household, and pricing functions (w, r), such that:*

i) (V_{in}, c_{in}) *solves the consumer problem defined in Equation* (1.8) *for each* $(i, n) \in \mathcal{I} \times \mathcal{A}$

ii) prices (r, w) *are competitive, i.e., determined by the respective marginal productivity*

iii) forecast functions are consistent with the exogenous dynamics and the collection of policy functions

While a formal characterization of the equilibrium definition is important, ultimately our goal is to obtain a numerical approximation of the policy functions, since the stochastic nature of this framework does not allow us to derive a closed-form solution. Therefore, the appropriate choice of the equilibrium definition will be crucial, as it will guide the construction of the numerical algorithm used to approximate the agents' optimal behavior.

When we build a numerical solution in which policy functions depend on the entire state space, we rely on a fully rational expectation equilibrium. Agents that populate the economy are able to observe the entire distribution of state variables, and their optimal choices depend on the whole set of observed state variables. In general, the policy functions are approximated using parametrized functional forms, and the goal of the numerical algorithm is to find the set of parameters that best approximate a set of equilibrium conditions. This equilibrium definition has been introduced by Spear [1988] and Krueger and Kubler [2004]. Many numerical applications have since used it in the context of large scale OLG models, including Krueger and Kubler [2006], and more recently Hasanhodzic and Kotlikoff [2018] and Kim [2018]. In a fully-rational expectation equilibrium, agents make predictions regarding future prices conditional on the entire distribution of endogenous and exogenous state variables. We now formally define the full-rational expectation equilibrium.

Definition 2. *(Functional Rational Expectation Equilibrium) In the context of this problem, we define it as follows:*

- *a hyper-rectangle* $\mathcal{B} \subset \mathcal{X}$*;*
- *a collection of real-valued policy functions* $(c_{in})_{i \in \mathcal{I}, n \in \mathcal{A}}$*, defined as* $c_{in} : \mathcal{B} \times \mathcal{Z} \to \mathbb{R}$ *and belonging to the class of functions* $\mathcal{C}_p$

- *a forecast function:* $\pi_f : \mathcal{B} \times \mathcal{Z} \times \mathcal{Z} \to \mathcal{B}$ *and to the class of function* $\mathcal{C}_f$

with $(c_{in})_{i\in\mathcal{I},n\in\mathcal{A}}$ *and* π_f *satisfying the following conditions:*

- *Optimality of the policy functions given the forecast function*

$$
\begin{aligned}
& u'\left(c_{in}(\mathbf{x},\mathbf{z})\right) = \beta_i \mathbb{E}\left[r(\mathbf{x}',\mathbf{z}')u'\left(c_{in+1}(\mathbf{x}',\mathbf{z}')\right)|\mathbf{x},\mathbf{z}\right] \\
& \mathbf{x}' = \pi_f\left(\mathbf{x},\mathbf{z},\mathbf{z}'\right) \\
& \mathbf{z}' \sim F(\cdot|\mathbf{z}), \quad \forall(\mathbf{x},\mathbf{z}) \in \mathcal{B}\times\mathcal{Z} \\
& \forall(i,n) \in \mathcal{I}\times\mathcal{A}
\end{aligned}
$$

- $\pi_f\left(\mathbf{x},\mathbf{z},\mathbf{z}'\right)$ *is consistent with the policy functions* $(c_{in}(\mathbf{x},\mathbf{z}))_{i\in\mathcal{I},n\in\mathcal{A}}$*, the stochastic process represented by* $F(\cdot|\mathbf{z})$ *and exogenous dynamics of the state variables*

It is important to highlight that in a fully-rational expectation equilibrium, consistency between forecast function π_f and agents' policy functions is guaranteed by the fact that the policy functions depend on the entire state space. In other words, the forecast function defined here are redundant, as they are implied by the policy functions and the exogenous equations ruling the dynamics of the state variables.

Using the full state and thus relying on agents' full rationality involves some numerical challenges, especially in large-scale OLG models. As we increase the dimension of the state space, the computational burden necessary to obtain the numerical solution increases as well, as we are subject to the well-known curse of dimensionality. At the same time, a subset of the state space can be sufficient to derive numerical solutions that are sufficiently precise. This has motivated the development of state-space aggregation methods, which rely on the quasi-fully rational equilibrium concept. First introduced by Grandmont [1977], it became popular thanks to Krusell and Smith [1998]. In this equilibrium framework, agents' policy functions and expectations about future prices depend on a subset of the state space, which are usually agent-specific. The subset includes some agent-specific state variables, namely $\mathbf{x}_{in}$, and some aggregate state variables $\mathbf{x}_a$ that are shared across agents. Applications of this numerical solution strategy can be found in Storesletten et al. [2007] and Harenberg and Ludwig [2019] in the context of OLG models. This framework is not consistent with agents' full rationality, as the information set they have access to when making the optimal decision is restricted. It is in the discretion of the economist to appropriately select the relevant state variables representing the signal, and this level of discretion is the source of one of the major pitfalls of an

equilibrium concept relying on quasi full-rationality. We now proceed by providing a formal definition of the quasi-fully rational expectation equilibrium.

Definition 3. *(Functional Quasi-Fully Rational Expectations) In the context of this problem, we define it as follows:*

- *a collection of hyper-rectangles* $(\mathcal{B}_{in})_{i\in\mathcal{I},n\in\mathcal{A}}$ *for agent-specific state variables;*
- *a hyper-rectangle* $\mathcal{B}_a$ *for the aggregate state variables;*
- *a collection of real-valued policy functions* $(c_{in})_{i\in\mathcal{I},n\in\mathcal{A}}$*, defined as* $c_{in} : \mathcal{B}_{in} \times \mathcal{B}_a \times \mathcal{Z} \to \mathbb{R}$ *belonging to the functional class* $\mathcal{C}_p$
- *a forecast function:* $\pi_f : \mathcal{B}_a \times \mathcal{Z} \times \mathcal{Z} \to \mathcal{B}_a$ *and a collection of agent-specific forecast functions* $(\Gamma_{in})_{i\in\mathcal{I},n\in\mathcal{A}}$*, defined as:* $\Gamma_{in} : \mathcal{B}_{in} \times \mathcal{B}_a \times \mathcal{Z} \to \mathcal{B}_{in}$ *belonging to the functional class* $\mathcal{C}_f$

with the policy functions $(c_{in}, \Gamma_{in})_{i\in\mathcal{I},n\in\mathcal{A}}$ *and the forecast functions* π_f *satisfying the following conditions:*

- *Optimality of the policy functions given the aggregate and household-specific forecast functions:*

$$
\begin{aligned}
& u'\left(c_{in}(\mathbf{x}_{in}, \mathbf{x}_a, \mathbf{z})\right) = \beta_i \mathbb{E}\left[r(\mathbf{x}'_a, \mathbf{z}')u'\left(c_{in+1}(\mathbf{x}'_{in+1}, \mathbf{x}'_a, \mathbf{z}'\right)|\mathbf{x}_{in}, \mathbf{x}_a, \mathbf{z}\right] \\
& \mathbf{x}'_a = \pi_f\left(\mathbf{x}_a, \mathbf{z}, \mathbf{z}'\right), \quad \mathbf{x}'_{in+1} = \Gamma_{in}(\mathbf{x}_{in}, \mathbf{x}_a, \mathbf{z}) \\
& \mathbf{z}' \sim F(\cdot|\mathbf{z}), \quad \forall(\mathbf{x}_{in}, \mathbf{x}_a, \mathbf{z}) \in \mathcal{B}_{in} \times \mathcal{B}_a \times \mathcal{Z} \\
& \forall(i, n) \in \mathcal{I} \times \mathcal{A}
\end{aligned}
$$

- $\pi_f\left(\mathbf{x}_a, \mathbf{z}, \mathbf{z}'\right)$ *is consistent with the policy functions* $\left(c_{in}(\mathbf{x}_{in}, \mathbf{x}_a, \mathbf{z})\right)_{i\in\mathcal{I},n\in\mathcal{A}}$*, the stochastic process represented by* $F(\cdot|\mathbf{z})$ *and exogenous dynamics of the state variables* Γ_{in}*.*

This is different than the fully rational equilibrium, because the consistency between forecast function π_f and agents' policy functions is not guaranteed, but approximated numerically. While consistency between policy and forecast functions is important from a theoretical perspective, it usually comes with significant computational costs in large-scale OLG models. In particular, in some of large-scale models, the use of the entire state space does not allow it to go beyond linear policy functions, (see for instance Hasanhodzic and Kotlikoff [2019] and Kim [2018]), since the state space includes hundreds of variables. The framework proposed in this chapter is no different, since policy functions will generally depend on the

entire distribution of average lifetime incomes, Social Security benefits and capital holdings. This leads to a state space whose size is in the order of $\mathcal{O}(2AI)$. Thus, the choice of the equilibrium concept depends on how it interacts with the complexity of the numerical solution, and it is of particular relevance when we believe that the policy functions display some significant non-linearities.

In Chapter 2, we will show that we can successfully develop a numerical solution that relies on a reduced state space while simultaneously achieving state-of-the-art numerical performance. As the results will show in Chapter 2, policy functions will display some non-negligible non-linearities that capture households' precautionary savings in anticipation of the insolvency of the Social Security Administration.

1.6 Conclusion

In this Chapter, I propose a novel model for Social Security, where I depart from the traditional assumption made in the OLG literature that models Social Security as a PAYGO system. My proposed model incorporates two elements that better capture the current institutional features of Social Security: (1) a Trust Fund; and (2) the dependence of retirement benefits on average-life earnings and when agents retire. In Chapter 2, I propose a novel algorithm to obtain the numerical solution of large-scale OLG models, and I show that it can be successfully used in the context of our problem. The use of neural networks as functional approximators, together with and deep-learning techniques to train them, allow us to solve a complex numerical problem. In Chapter 3, I estimate the relevant structural parameters, and I conduct some counterfactual analysis aimed at evaluating the different policy alternatives on household welfare.

Chapter 2

Policy Approximation with Deep Neural Networks in Large Scale OLG Models

2.1 Introduction

In the previous chapter, I proposed a model of Social Security accounting for lifetime earning histories and the potential for bankruptcy of the fund in a stochastic environment. Upon creating the model, however, an additional challenge was presented by the fact that the complexity of the model made it difficult to solve using traditional economic methods. To address this issue, I turned to the field of machine learning and, specifically, a tool known as neural networks.

Neural networks are a form of real-value mapping, and they have been the workhorse of some of the most important results in artificial intelligence in the last decade. They have been successfully employed for a variety of purposes, from image and text classification to speech recognition, from cell phone apps to self-driving cars. The success of neural networks in empirical applications can be attributed to both theoretical and practical considerations. On the theoretical side, neural networks are considered universal approximators (Hornik et al. [1989],Cybenko [1989] and Hornik et al. [1990]), able to represent irregular, high-dimensional functions. On the practical side, the compositional nature of neural networks combined with gradient-based optimization methods make the fitting process straightforward and suitable for many applications.

While neural networks have been very popular in machine learning settings, their use in economics has, thus far, been very limited (see Maliar et al. [2019], Azinovic et al. [2019] and Duarte [2018]). This can be explained by the fact that training neural networks presents some computational challenges, but recent developments in the Open Source Software community have significantly decreased the cost of entry, making the use of neural networks for empirical and computational purposes significantly more accessible to economists. In particular, companies like Google (Tensorflow) and Facebook (Pytorch) have made Python libraries readily available to those with a working knowledge of programming, thus permitting economists to more easily utilize and apply neural networks to any set of custom data with just a few lines of code.

When evaluating the ways in which I could potentially solve my advanced Social Security model, I determined that neural networks may be applicable as they, and other deep-learning techniques, have proven successful when working with non-linear behavior. This is relevant as my intuition in the context of my model was that agents tend to modify their consumption-saving behavior if they anticipate that the Social Security system will go bankrupt in their lifetime. In particular, a consumption smoothing argument would suggest that agents would tend to save more if they expect lower retirement benefits (triggered, for instance, by insolvency). However, this relationship is unlikely to be linear in nature, as agents' consumption savings behavior will not linearly depend on a measure of distance from bankruptcy. If agents correctly predict that bankruptcy is far away, then they will make little or no adjustment to their savings. On the contrary, if they anticipate that insolvency is on the near horizon, they may decrease their current consumption and save more to compensate for the inevitably lower retirement income from Social Security. Considering the non-linear behavior involved, the application of neural networks seemed a potentially good fit, so I decided to experiment with neural networks to solve my complex model for Socially Security.

The application of neural networks, especially "deep" networks, to my model for Social Security proved highly successful. As such, in this paper I show how economists can take advantage of the approximation power of neural networks in economic applications and, specifically, in the context of large-scale OLG models. Computational issues, especially in models that do not allow for a closed-form solution and in applications where we need to characterize highly dimensional policy or value functions numerically, have always represented one of the major bottlenecks in economic research. This is particularly true with regard to large-scale OLG models, i.e. OLG models populated by a large number of different types of agents or by agents living long lifespans. This type of model is becoming more and more popular, as the development of more detailed microdata has increased the interest in macroeconomic models characterized by a significant source of observed heterogeneity with regard to preferences, income, retirement age, mortality etc. Conse-

quently, the successful application of neural networks to these models could have wide-ranging implications for the future of economic research.

From a methodological perspective, my proposed approach relies on the transformation of an economic problem, characterized by Euler equations, into non-linear regressions, which are suitable to serve as objective functions of a deep-learning framework. More specifically, I plug the policy and forecast functions into the agents' Euler equations, and use a derivative-free projection method to iteratively update the policy functions. The training happens by minimizing the error on a least-square objective function through standard mini-batch stochastic gradient descent, where the use of a global objective function allows me to simultaneously train hundreds of neural networks. The grid of points used is agent-specific and based on Smolyak sparse grids (Smolyak [1963], Krueger and Kubler [2004] and Judd et al. [2014]). The algorithm then proceeds by using the generated policy function to simulate hundreds of economies to train the forecast functions and to update the policy functions based on training points that belong to the ergodic set. In this way, I guarantee that policy and forecast functions are internally consistent numerically, which is a necessary property of any rational expectation general equilibrium problem. In addition, this allows to concentrate the computation of the forecast in the ergodic region to obtain better approximations in regions of the state space that are actually crossed. With regard to this application, deep networks perform better than shallow networks, however, the performance of neural networks characterized by different architectures largely depends on the specific application. Among the different possible neural network architectures, I compare the performance of shallow and deep networks, which are used to approximate policy and forecast functions, and I show that the use of deep networks, together with a reduced state space produce very precise numerical approximations for the policy and forecast functions characterizing the general equilibrium behavior of households. While it is possible to rank the performance of alternative selections of the hyper-parameters characterizing the architectures of policy and forecast functions and the optimization algorithm, ultimately our results are very robust to different specifications. In particular, the results presented in this chapter show that very little fine-tuning is required to achieve a state-of-the-art performance.

As outlined above, in this Chapter, I successfully contribute to the literature by showing how economists can take advantage of the approximation power of neural networks in economic applications and, specifically, in the context of large-scale OLG models. This Chapter is related to the contribution of Maliar et al. [2019] and Azinovic et al. [2019] (discrete time models), and Duarte [2018] (continuous time), who introduce deep-learning in the context of the numerical approximation of policy and value functions. More specifically, this paper shares the use of neural networks as global approximators for policy functions,

and methodologically uses a framework similar to the one proposed by Maliar et al. [2019] based on the use of Euler equations and non-linear regression to iteratively compute the numerical approximation of the policy functions. The application is similar to the deep-learning framework in the context of OLG models proposed by Azinovic et al. [2019], although the solution algorithm developed relies on a different and more parsimonious equilibrium definition.

This chapter is structured as follows. In Section 2.2, I introduce some relevant terminology. In Section 2.3, I present an overview of results using approximation properties of shallow networks 2.3.1, deep networks 2.3.2 and the backpropagation algorithm 2.3.3. In Section 2.4, I introduce the specific deep-learning framework utilized to numerically solve the Social Security model introduced in Chapter 1. In Section 2.6, I describe the algorithm I used to solve the model numerically. In Section 2.7, I discuss the main results. Finally, in Section 2.8, I provide some potential avenues for future research.

2.2 Terminology

Definition 4 (Neural Network)**.** *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be any continuous (non-linear) real-valued function. Let L be the number of hidden layers characterizing the neural network. For each hidden layer l, let H_l denote the number of neurons. Let y denote the output of the neural network. I define neural network as the following mapping $\mathbb{R}^d \to \mathbb{R}$:*

***First Hidden Layer**:*

- *For each neuron in the first hidden layer $k = 1, \cdots, H_1$:*

$$h_{1k} = \sigma\left(\mathbf{w}_{1k}\mathbf{x} + b_{1k}\right), \quad \mathbf{x} \in \mathbb{R}^d$$

***Second to $\mathbf{L^{th}}$ Hidden Layer**:*

- *For each hidden layer $l = 2, \cdots, L$:*
  - *For each neuron $k = 1, \cdots, H_l$ in the l^{th} hidden layer:*

$$h_{lk} = \sigma\left(\mathbf{w}_{lk}\mathbf{h}_{l-1} + b_{lk}\right)$$

***Output Layer**:*

$$y = \mathbf{w}_{L+1}\mathbf{h}_L + b_{L+1}$$

Definition 4 introduces the type of neural networks that I use in this paper. It is worth making two observations. First, considering that I am interested in approximating functions with a uni-dimensional output, I restrict y to be a real-valued scalar. Second, I assume that all of the hidden-layers are characterized by the same activation function σ, while the output layer has a linear activation function.[1]

Definition 5 (Activation Function)**.** *An activation function is a continuous, non-linear, real-valued mapping* $\sigma : \mathbb{R} \to \mathbb{R}$*. The following functions are commonly used as activation functions in the literature:*

- *Sigmoidal activation function:*

$$\sigma(x) = \frac{1}{1 + \exp(-x)}$$

- *Rectified Linear Unit (ReLU):*

$$\sigma(x) = \max(x, 0)$$

- *Hyperbolic Tangent:*

$$\sigma(x) = \frac{\exp(2x - 1)}{\exp(2x + 1)}$$

Definition 5 shows that activation functions can have have different forms: they can be smooth and bounded (sigmoidal and hyperbolic tangent), or they can be unbounded with discontinuous first derivative (ReLU). Their purpose is to introduce non-linearities into the neural network. As I will discuss later, the choice of activation is important, since it is likely to affect the performance of the neural network in the specific task for which it is used. I now turn to the definition of the neurons, which we can think of as the computational units of the neural network.

Definition 6 (Neuron)**.** *Let* $nnet$ *be a neural network as described in Definition 4. A neuron is the output of a real-valued mapping which takes as input the neurons of the previous layer* $\mathbf{h}_{l-1}$*. The real-valued mapping consists of the composition of a linear transformation and a non-linear transformation (activation function), which can be described as follows:*

$$h_{lk} = \sigma\left(\mathbf{w}_{lk}\mathbf{h}_{l-1} + b_{lk}\right)$$

where h_{lk} *is the* k^{th} *neuron in the* l^{th} *hidden layer. The linear transformation takes as input the neurons in the previous layer* $\mathbf{h}_{l-1} \in \mathbb{R}^{N_{l-1}}$ *(or the input of the neural network, if* $l = 1$*), and is parametrized by a*

[1] In general, the output of a neural network can be multi-dimensional.

N_{l-1}-dimensional vector of weights $\mathbf{w}_{lk} \in \mathbb{R}^{N_{l-1}}$ and a bias $b_{lk} \in \mathbb{R}$. The transformation that each neuron operates is parametrized by the following vector:

$$\boldsymbol{\theta}_{lk} := (\mathbf{w}_{lk}, b_{lk}), \quad \boldsymbol{\theta}_{lk} \in \mathbb{R}^{N_{l-1}+1}$$

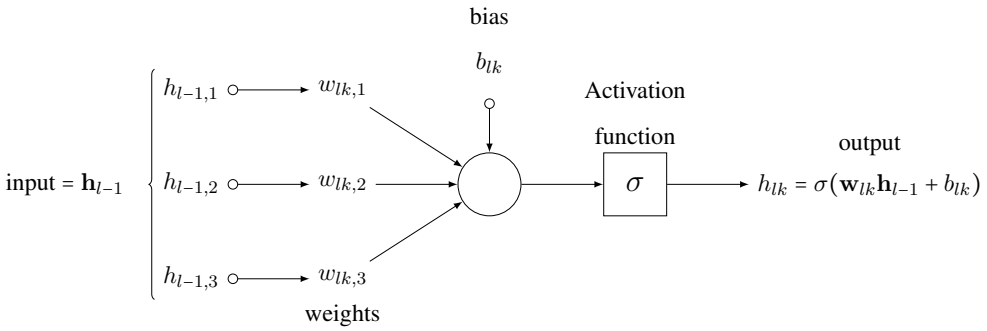

Figure 2.1: Representation of a neuron.

Figure 2.1 shows an example of a neuron. From Definition 6, we can see that the number of parameters characterizing the transformation operated by the neuron depends linearly on the number of neurons in the previous layer. I now proceed by providing definitions aimed at characterizing two classes of neural networks based on the number of hidden layers they have.

Definition 7 (Shallow Neural Network)**.** *A Shallow Neural Network (SNN) is a feed-forward neural network (Definition 4) with one hidden-layer ($L = 1$).*

Definition 8 (Deep Neural Network)**.** *A Deep Neural Network (DNN) is a feed-forward neural network (Definition 4) with at least two hidden-layers ($L \geq 2$).*

2.3 Literature Review

In this Section, I review some important theoretical results related to neural networks. First, I introduce some of the classic results regarding the representational power of shallow and deep neural networks, explaining why neural networks can be used as functional approximators. I then proceed by giving an overview of backpropagation, the workhorse algorithm used in deep-learning to train neural networks. While theoretical

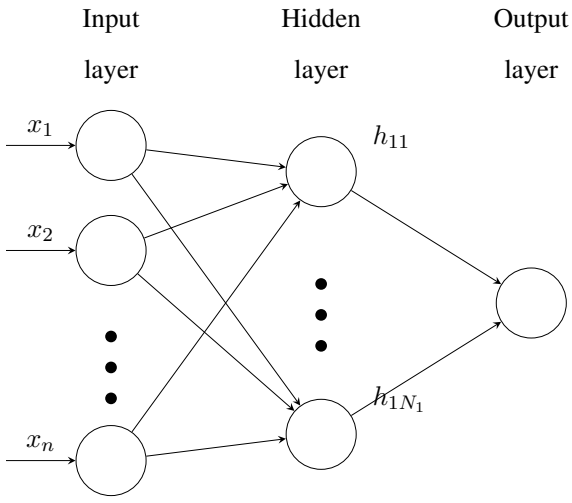

Figure 2.2: Shallow Neural Network.

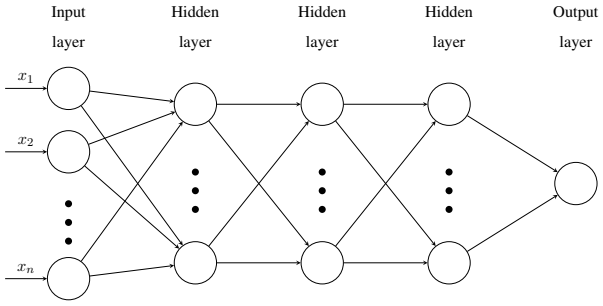

Figure 2.3: Feed-Forward Neural Network

properties are important to justify the consideration of a particular class of functional approximators, they are not, alone, sufficient to explain why neural networks have become successful in many AI applications, and why economists should consider them as functional approximators. Indeed, the development of algorithms tailored to the training of neural networks was an essential component of the success of neural networks. For this reason, in this Section, I describe how the theoretical properties of neural networks interact with the algorithms developed to train them.

2.3.1 Functional Approximation Through Shallow Nets

The main objective of this subsection is to show that neural networks can be chosen as global functional approximators in economic applications. In economics, Chebyshev polynomials are the most widely used

class of functional approximators. Their popularity is justified by the fact that (1) they can achieve any desired degree of accuracy in the limit, and (2) it is possible to compute bounds for their approximation errors with finite degrees. The two theorems presented below show that shallow neural networks can be used as global approximators of any continuous function and that (in the limit) they can achieve any desired level of accuracy. In other words, Theorem 1 and 2 show that shallow neural networks can be used as global approximators, like more traditional Chebyshev polynomials.

Theorem 1 (Superimposition Theorem, Kolmogorov [1957])**.** *Given any continuous function f defined on the n-dimensional compact hyper-cube:*

$$f : \mathbb{I}^n \to \mathbb{R}, \quad f \in \mathcal{C}^0\left(\mathbb{I}^n\right), \quad \mathbb{I}^n = \bigtimes_{i=1}^{n} [0,1]$$

then, it is possible to find collections of continuous functions

$$\chi_p : [0,1] \to \mathbb{R}, \quad \chi_p \in \mathcal{C}^0([0,1]), \quad p = 1, \cdots, n$$
$$\psi_{p,q} : \mathbb{R} \to \mathbb{R}, \quad \psi_{p,q} \in \mathcal{C}^0(\mathbb{R}), \quad p = 1, \cdots, n, \; q = 1, \cdots, 2n,$$

such that:

$$f(\mathbf{x}) = \sum_{q=1}^{2n} \chi_q \left(\sum_{p=1}^{n} \psi_{pq}(x_p) \right) \quad \forall \mathbf{x} \in \mathbb{I}^n \;\; \textit{and}$$

$\psi_{p,q}$ do not depend on f.

Theorem 1 shows that any continuous function defined on a compact subset of $\mathbb{R}^n$ can be exactly represented by using a shallow neural network whose size, measured by the number of neurons used in the first layer, depends linearly on the dimensionality of the input. From Theorem 1 we can see that while the collection of activation functions in the first hidden layer does not depend on the specific function f that we want to represent, the activation functions in the output layer χ_p depend on it. This is not a desirable property, since in a numerical approximation exercise, we would like to be able to pick ex-ante the activation function. Cybenko [1989] and Hornik et al. [1989] show that we can choose activation functions from the class of continuous and bounded functions to approximate any continuous function with a shallow network, and we can achieve any approximation error provided that the neural network is large enough. This result is presented in Theorem 2.

Theorem 2 (Universal Approximation Theorem, Cybenko [1989],Hornik et al. [1989])**.**

Given any continuous function f defined on the $n-$dimensional compact hypercube:

$$f : \mathbb{I}^n \to \mathbb{R}, \quad f \in \mathcal{C}^0\left(\mathbb{I}^n\right), \ \mathbb{I}^n = \bigtimes_{k=1}^{n} [0,1]$$

any approximation error $\varepsilon > 0$, and any continuous, bounded and non-constant activation function $\sigma \in \mathcal{C}^0(\mathbb{R})$, then

$$\exists \ N \in \mathbb{N}_+, \ (\mathbf{w}_{1i}, b_i, w_{2i})_{i=1}^{N}, \quad \mathbf{w}_{1i} \in \mathbb{R}^n, \quad b_i, w_{2i} \in \mathbb{R} \ \ s.t.$$
$$F(\mathbf{x}) = \sum_{i=1}^{N} w_{2i}\sigma\left(\mathbf{w}_{1i}\mathbf{x} + b_i\right), \quad |F(\mathbf{x}) - f(\mathbf{x})| < \varepsilon \quad \forall \mathbf{x} \in \mathbb{I}^n$$

Theorem 2 shows that any continuous function can be approximated by a shallow neural network, where the number of neurons in the hidden layer N depends on the target approximation error ε, the target function itself f and the activation function σ. It is important to highlight the fact that this result holds for any (non-constant) bounded activation function,[2] allowing the researcher to fix ex-ante the choice of the functional form of the activation function. However, Theorem 2 does not provide any intuition as to how to construct weights and the biases of the neural network, nor does it provide any guidance with regard to the choice of the specific activation function or the number of neurons. The literature has tried to address these issues, investigating the relationship between the number of neurons required to approximate a target function with a shallow net, and the regularity and the size of the input of the target function. For instance, Barron [1993] derives bounds on the approximation error for a shallow neural network with a given number of neurons and sigmoidal activation function, and relates it to the regularity of the function (in a Fourier sense) and to the dimensionality of its input. In particular, it shows that a shallow network can achieve a L^2-error in the order of $\mathcal{O}(\frac{1}{N})$, independently of the size of the input. On the contrary, polynomial-based approximations have errors that depend positively on the size of the input, making them more suitable for applications in which the state space is small.

While these results help us understand some of the properties characterizing shallow neural networks, the vast majority of the results in empirical applications rely on the use of deep neural networks (see Lecun et al. [2015]). As noted by Mhaskar and Poggio [2016], both deep and shallow neural networks are universal approximators, and therefore, from a theoretical perspective, both can be used as global functional approximators. From a practical perspective, however, it is important to account for the computational constraints. In the next subsection, I describe how deep neural networks can help us mitigate this problem.

[2] Cybenko [1989] shows the result for shallow networks with a sigmoidal activation function, while Hornik et al. [1989] extends the result to any continuous, bounded activation function.

2.3.2 Functional Approximation through Deep Nets

The deep-learning literature has investigated the relationship between neural network architectures and their representational power. By referring to the architecture of a neural network, I summarize the information about the number of layers, the number of neurons per layer, the activation function in each neuron, how neurons across different layers are connected, etc. This area of research was motivated by two issues arising from empirical applications. First, the choice of the architecture appears to be an important determinant in the success (or failure) of a specific application, potentially undermining the generalization of the results. Second, computational constraints need to be taken into account.

The literature has not reached a general consensus with regard to which architecture to deploy in any given circumstance. That said, the literature does seem to agree on one key point: higher representational power can be more easily achieved through depth than through width.[3] For instance, Cohen et al. [2016] makes this point, but acknowledges that evidence supporting this claim relies either on empirical grounds, or on the construction of functions with pathological behavior. Indeed, while Delalleau and Bengio [2011], Cohen et al. [2016] and Mhaskar et al. [2017] show that deep neural networks can approximate specific classes of functions more parsimoniously[4] than shallow networks, these results only hold for specific classes of functions.

The deep-learning literature has tried to address the question of how to measure the representational power of neural networks by counting the number of regions in the output space that a neural network is able to shatter. In particular, studies have provided estimates of upper and lower bounds on the maximum number of piece-wise linear functions that a neural network is able to generate when ReLU activation functions are used. For instance, Pascanu et al. [2014] show that deep nets separate their input space into exponentially more linear response regions than their shallow counterparts when deep and shallow nets have the same number of neurons. Montúfar et al. [2014], Arora et al. [2018] and Serra et al. [2018] built on these results, providing sharper estimates for the bounds. However, as noted by Serra et al. [2018], the representational power of deep nets is not necessarily superior to those of shallow nets, but in general, it depends on the specific architecture. In addition, they provide an intuition as to how the maximal number of regions is sensitive to the number of neurons in different layers of the neural net. This result gives us some guidance with regard to the best practices for designing the architecture of a deep neural network.

[3] I.e., by increasing the number of layers of the neural network instead of increasing the number of neurons per layer.

[4] The measure of complexity considered here is the total number of neurons.

While the theoretical results are far from conclusive, neural networks have been used extensively to approximate high-dimensional functions. This has led to the use of deep neural networks combined with reinforcement-learning techniques to solve dynamic optimization problems with hundreds of state variables (see for instance Mnih et al. [2013],Mnih et al. [2015]).

2.3.3 Backpropagation

Theoretical properties themselves are not sufficient to justify why neural networks have been successfully used in a wide range of applications in machine learning and AI. When we use neural networks in empirical applications, we are ultimately solving an optimization problem where we want to find the neural network that minimizes the value of some loss function. For what concerns the optimization component, virtually all applications rely on some variant of the stochastic gradient descent (SGD) algorithm. In this context, as stated by Lecun et al. [1998], the success of practical applications of neural networks relies on the ease of the optimization process. Backpropagation is the workhorse algorithm used in the AI literature when we are fitting neural networks. While backpropagation was first introduced in the 1960s-1970s, it became popular thanks to Rumelhart et al. [1986]. The mechanics are simple: it exploits the chain rule to derive an analytical formulation of the gradient thanks to the compositional nature of the neural network. This is used in conjunction with the application of a gradient-descent method. I will provide more details in Section 2.4, where I describe the specific optimization routine I selected for my model.

2.3.4 Deep Learning in Economic Applications

While the use of neural in economic applications dates back to Kelly and Shorish [2000], only very recently new development in deep learning have started to attract the attention of economists. In particular, the literature has started to gain interest in understanding how neural network can be used as functional approximators of policy functions in the context of economic problems characterized by strong non-linearities. The literature has focused on continuous on both discrete time and continuous time applications (for continuous time, see for instance Duarte [2018]. For what concerns discrete time applications, Maliar et al. [2019] show how neural networks can successfully approximate policy functions in the context of large-scale models with infinitely lived agents. On the contrary, Azinovic et al. [2019] focus on applications centered around larger OLG models. Their approach differs from the one proposed in this chapter on multiple dimensions.

Firstly, both Maliar et al. [2019] and Azinovic et al. [2019] rely on policy functions depending on the full state space, while the algorithm proposed in this chapter builds on a reduced and household specific state space. This is motivated by the fact that in this application, I want to characterize each policy function using a single neural network. This issue is automatically taken into account by Maliar et al. [2019], since in their application only one multi-dimensional policy function needs to be approximated; however in Azinovic et al. [2019], the author decide to exploit a neural network with multi-dimensional output to characterize the policy functions of agents in different cohort. In this way, all cohorts share the same parameters in the neural network up to the last hidden-layer. While the authors are able to show the numerical solution achieves a very good performance, our proposed solution is ultimately more parsimonious, as the number of parameters to be computed is significantly lower.

2.4 Deep-Learning Framework

The computation of the equilibrium policy functions characterizing the stochastic economy relies on several steps. In general, convergence is more likely to be achieved if a good initial guess for the policy functions for which we are solving. For this reason, I start from the deterministic version of the model, where there are no aggregate or depreciation shock, and then we proceed by using the policy functions obtained for the deterministic economy as initial guesses of the fixed point algorithm. More details about the initialization are presented in Appendix A.1.

I characterize the policy functions in a neighborhood of the steady state (see A.1.1), by finding the a local linear approximation of the policy function for consumption and transition of the state variables (A.1.2). I then use these functions as the initial guesses for the iterative, derivative-free fixed-point algorithm (see for instance Hasanhodzic and Kotlikoff [2019], or Maliar et al. [2019] for an application in the context of deep-learning). Given the large state space, I decide to reduce the dimension of the input of each policy function a smaller subset of state variable that household-specific, following the approach first proposed Krusell and Smith [1998]. In addition, I use Smolyak sparse grids as proposed by Krueger and Kubler [2004, 2006] and Judd et al. [2014] to mitigate the curse of dimensionality, since even the reduced state space contains seven state variables.

2.4.1 Neural Network Architecture

Households' Consumption Policy Functions

I use feed-forward neural networks to approximate the consumption policy function for household (i, n); it is assumed that the input is a seven-dimensional vector $\mathbf{x}_{in}$, which contains household-specific and aggregate variables:

$$(\mathbf{x}_{in}, \mathbf{z}, \mathbf{x}_a) = (k_{in}, \overline{e}_{in}, A, K, H, \sigma(K), S) \in \mathcal{B}_{in} \times \mathcal{Z} \times \mathcal{B}_a \tag{2.1}$$

I define the parametrized policy function as $C_{in}\left(\mathbf{x}_{in}|\boldsymbol{\theta}_{in}\right)$, where the trainable parameters are defined as:

$$\boldsymbol{\theta}_{in} = \left(\mathbf{W}_l^{i,n}, \mathbf{b}_l^{i,n}\right)_{l=1}^{L}$$

The matrix $\mathbf{W}_l^{i,n}$ represents $n_{l-1} \times n_l$ weights connecting the neurons in the hidden layer $l-1$ to the input activation function in layer l for the policy function of household (i, n). The vector $\mathbf{b}_l^{i,n}$ represents the n_l-dimensional vector of biases for hidden layer l. For each household, the total number of trainable parameters is:

$$\underbrace{\sum_{l=0}^{L-1}(n_l \times n_{l+1} + n_{l+1})}_{\text{Hidden Layers}} + \underbrace{(n_L + 1)}_{\text{Output Layer}}$$

where n_0 denotes the size of the input of the neural network. Figure 2.4 shows an example of a fully connected 2-hidden layer DNN representing the policy function for households for households (i, n).

In each hidden layer I utilize a sigmoidal activation function, as introduced in Definition 5. The output layer is obtained via linear activation. The choice comes from the desirable smoothness that policy functions would inherit from the use of such activation, as the output of the DNN would be the composition of linear combinations of smooth functions.

It is important to keep in mind that, in general, I can use any moment of the distribution of the endogenous state variables as an additional aggregate variable fed in the household's policy function. In the limit, I could resort to the use of the entire distribution of endogenous state variables. The inclusion of the standard deviation of the cross-sectional distribution of households' savings is motivated by performance considerations, as illustrated in Section 2.7.

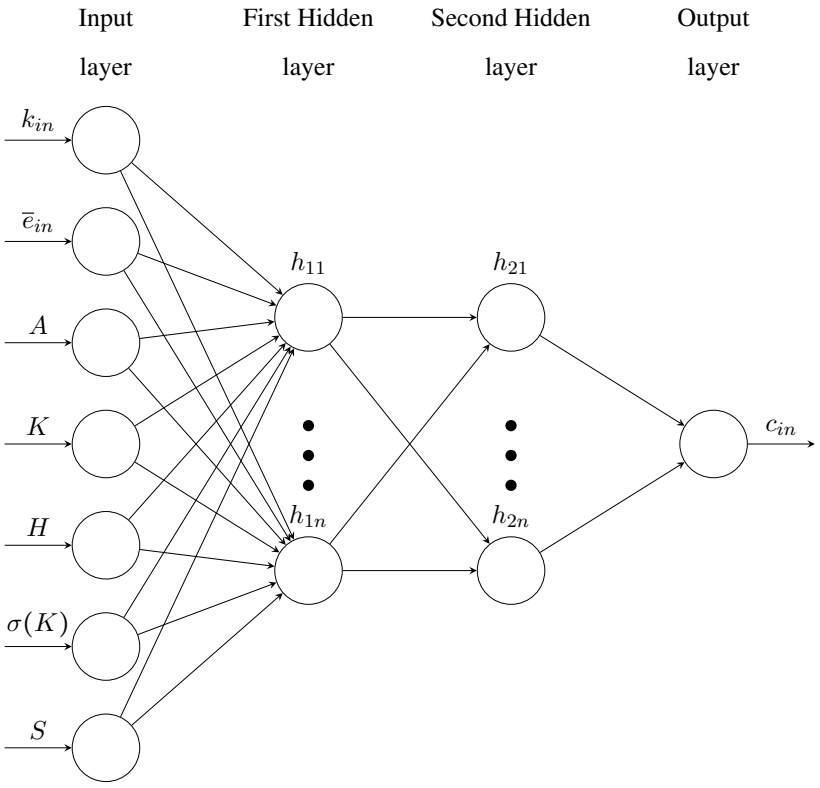

Figure 2.4: Representation of a two-hidden layer neural network for the policy function for household (i, n) consumption.

2.4.2 Forecast Functions

As discussed in Chapter 1, the solution algorithm requires that agents use a forecast function to predict the future value of the endogenous aggregate variables. For this reason, I introduce the additional neural network $\pi_f : \mathbb{R}^5 \to \mathbb{R}^4$, mapping $(K, z, H, S, \sigma(K)') \to (K', H', S', \sigma(K))$. With the values of K', H' and S', the households are able to forecast the amount of benefits they would receive under both a solvent and insolvent regime, while the aggregate state variable $\sigma(K)$ has been added since it increases the fit.

In this model I have introduced mortality, and as a consequence, accidental bequests. Given that, in general, aggregate bequest B left by the deceased households will depend on the distribution of wealth (weighted by the population shares and the mortality rate of each age-type specific household), I also need to forecast the value of aggregate bequests. Given that aggregate bequests are uniformly distributed across the surviving households, agents will be able to use this information to predict the total amount of the bequest $b_{in} = \eta_{in} B$ they receive. For this reason, I introduce the bequest function $\pi_B : \mathcal{B}_a \times \mathcal{Z} \to \mathbb{R}$.

In addition, I assume that all households share the same forecast function π_f to predict the future aggregate state variables and the same bequest function π_B to infer the current value of aggregate bequests. This assumption extends the quasi-fully rational expectation equilibrium approximation proposed by Krusell et al. [2000], where policy and forecast functions were assumed to be linear in the relevant state variables, and the only moment used to describe the distribution of savings is the mean. Here, on the contrary, I assume that the forecast function is a non-fully connected neural network, and I augment the aggregate state space to include the cross-sectional standard deviation of households' savings.

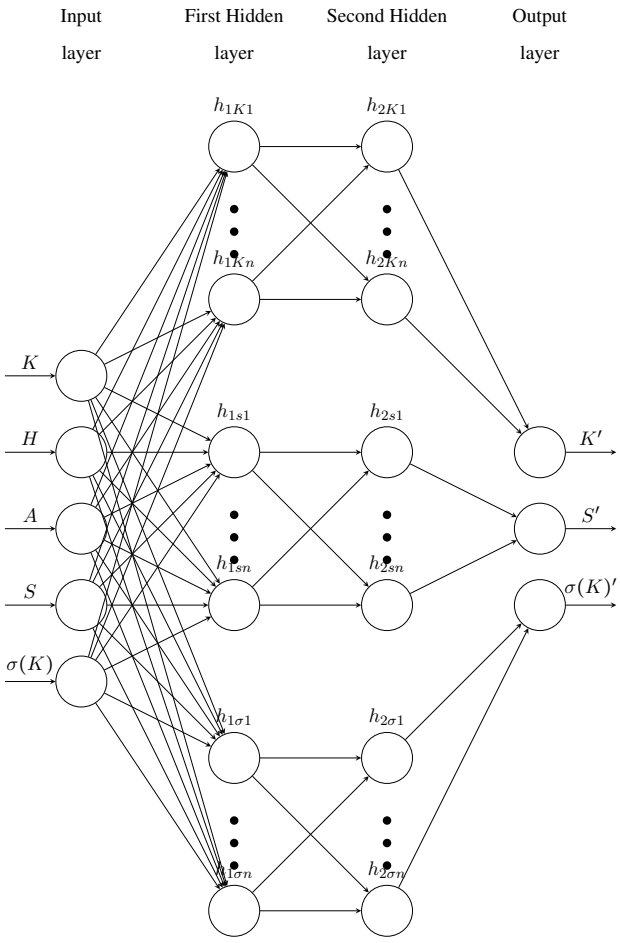

Figure 2.5: Representation of a two-hidden layer neural network forecast function π_f.

As we can see from Figure 2.5, each aggregate variable in the output depends on a subset of weights and biases. Therefore, I can partition the set of parameters according to the variable in the output to which they contribute. In terms of the activation function, I choose ReLUs for both the bequest and the forecast function. In light of these assumptions, I can represent the forecast function depicted in Figure 2.5 as follows:

- The input of the forecast functions is defined as:

$$(\mathbf{x}_a, \mathbf{z}) = (K, H, S, \sigma(K)', A)$$

and includes both exogenous and endogenous aggregate state variables.

- For the first hidden layer:

$$\mathbf{h}_{1H} = \max(0, \mathbf{W}_{1H}(\mathbf{x}_a, \mathbf{z}) + \mathbf{b}_{1H}) \qquad \mathbf{h}_{1K} = \max(0, \mathbf{W}_{1K}(\mathbf{x}_a, \mathbf{z}) + \mathbf{b}_{1K})$$
$$\mathbf{h}_{1S} = \max(0, \mathbf{W}_{1S}(\mathbf{x}_a, \mathbf{z}) + \mathbf{b}_{1S}) \qquad \mathbf{h}_{1\sigma} = \max(0, \mathbf{W}_{1\sigma}(\mathbf{x}_a, \mathbf{z}) + \mathbf{b}_{1\sigma})$$

- For hidden layer $l = 2, \cdots, L$:

$$\mathbf{h}_{lH} = \max(0, \mathbf{W}_{lH}\mathbf{h}_{lH} + \mathbf{b}_{lH}) \qquad \mathbf{h}_{lK} = \max(0, \mathbf{W}_{lK}\mathbf{h}_{lK} + \mathbf{b}_{lK})$$
$$\mathbf{h}_{lS} = \max(0, \mathbf{W}_{lS}\mathbf{h}_{lS} + \mathbf{b}_{lS}) \qquad \mathbf{h}_{l\sigma} = \max(0, \mathbf{W}_{l\sigma}\mathbf{h}_{l\sigma} + \mathbf{b}_{l\sigma})$$

- The output layer can be expressed as follows:

$$H' = \mathbf{W}_{LH}\mathbf{h}_{LH} + b_{LH} \qquad K' = \mathbf{W}_{LK}\mathbf{h}_{LL} + b_{LK}$$
$$S' = \mathbf{W}_{LS}\mathbf{h}_{LS} + b_{LS} \qquad \sigma(K)' = \mathbf{W}_{L\sigma}\mathbf{h}_{L\sigma} + b_{L\sigma}$$

2.4.3 Objective Functions

As the objective function for each household (i, n), I use the sum of the square residuals over the N-dimensional grid of points $\mathbf{X}_{in}$:

$$\mathcal{L}_{in}(\boldsymbol{\theta}_{in}) = \frac{1}{N}\sum_{j=1}^{N}\left(\hat{c}_{in}^{j} - c_{in}(\mathbf{x}_{in}^{j}, \mathbf{x}_{a}^{j}, \mathbf{z}^{j}|\boldsymbol{\theta}_{in})\right)^2 \quad \forall (i,n) \in \mathcal{I} \times \mathcal{A} \tag{2.2}$$

Given the equation for the household-specific policy function (2.2), I now define the global objective function used in the optimization routine:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{|\mathcal{I}||\mathcal{A}|}\sum_{i\in\mathcal{I}}\sum_{n\in\mathcal{A}}\mathcal{L}_{in}(\boldsymbol{\theta}_{in}) \tag{2.3}$$

where $\boldsymbol{\theta}$ denotes the collection of parameters characterizing households policy functions. Similarly, for the forecast function and bequest function I define:

$$\mathcal{L}_a(\boldsymbol{\theta}_f) = \frac{1}{JT}\sum_{j=1}^{J}\sum_{t=0}^{T-1}||\hat{\mathbf{x}}'_{ja,t+1} - \pi_f(\mathbf{x}_{ja,t}|\boldsymbol{\theta}_f)||^2 \tag{2.4}$$

$$\mathcal{L}_b(\boldsymbol{\theta}_b) = \frac{1}{JT}\sum_{j=1}^{J}\sum_{t=0}^{T}\left(\hat{B}_{jb,t} - \pi_b(\mathbf{x}_{ja,t}|\boldsymbol{\theta}_b)\right)^2 \tag{2.5}$$

where $(\mathbf{X}_{jt})_{t=1}^{T}$ represents the j-th sequence of aggregate variables obtained by simulating the economy for T periods using the consumption policy functions. As we can see from equations (2.4) and (2.3), training the neural network involves optimizing a non-linear least regression over the set of choice variables, which are represented by the parameters characterizing the different neural networks.

2.4.4 Optimization and Hyperparameters Tuning

I will now describe the algorithm and the hyperparameters used to train the policy and the forecast functions. One of the main advantages of polynomial-based approaches is that the optima of the loss functions introduced above are characterized by a closed-form solution, usually resulting from linear least-square problem. In addition, the solution is a global optimum within the class of chosen polynomials. When neural networks are used instead, the optimization process differs in two dimensions. First, the global optimum is not unique, as the architecture of neural networks are invariant to permutation of the nodes in the hidden layers. Second, given the composition of neural networks, their output is highly non-linear in the parameters for which we are optimizing. This makes the search for a global optimum not achievable in practice. In addition, the choice of the optimization algorithm and its hyperparameters is crucial. As noted by Orr and Müller [1998], while backpropagation is "conceptually simple, computationally efficient, and often works," training a DNN using backpropagation requires making many seemingly arbitrary assumptions about the architecture and the training such as the number and types of nodes, layers, learning rates, training sets, etc. Clearly, researchers' discretion in the selection of the hyperparameters represents a disadvantage, since, as noted by Orr and Müller [1998], "the choices can be critical, yet there is no foolproof recipe for deciding them because they are largely problem and data dependent." Simply put, what may deliver good results in a specific application, may fail poorly in another one, undermining the generality of the methodology used. While this problem was acknowledged early on in deep-learning literature, it is still an issue in the more recent applications (Lecun et al. [2015]). While the deep-learning literature does not provide definitive answers as to how to structure the training process and select the architecture of the of the neural networks, recent contributions in the literature have tried to establish a set of best practices. Here, I discuss some of the practices that practitioners have agreed on to facilitate the training process, and I describe how their properties interact with the model I proposed.

Parameters Initialization and Input Normalization It is widely accepted in the literature that the initialization of the neural network parameters plays a fundamental role in both the convergence and precision. The reason behind this is very intuitive. For sigmoidal activation functions, the gradients vanish for either very small or very large input values; similarly, for ReLU activation functions, the gradient is 0 when the input is negative. Therefore, it is important for both types of activation functions that the inputs are somewhat centered around 0, as first suggested by Rumelhart et al. [1986]. They propose to initialize the weights according to a distribution $\mathcal{N}(0,1)$. While this initialization works in some applications, it has been shown that it often leads to vanishing gradients, especially in deep neural networks. In order to counteract this issue, Orr and Müller [1998] recommend drawing the weights from the uniform distribution $W_{ij}^l \sim \mathcal{U}[\frac{-\sqrt{1}}{\sqrt{N_l}}, \frac{\sqrt{1}}{\sqrt{N_l}}]$. The proposed initialization depends on the structure of the neural network itself, since the bounds of the uniform distribution depend on the number of neurons in the layer.

Building on this result, Glorot and Bengio [2010] modified the initialization of weights by allowing the bounds to depend on the number of neurons in the previous layer as well: $W_{ij}^l \sim \mathcal{U}[\frac{-4\sqrt{6}}{\sqrt{N_{l-1}+N_l}}, \frac{4\sqrt{6}}{\sqrt{N_{l-1}+N_l}}]$[5]. This initialization ensures that the variance of the input and output of each layer is similar, reducing the risk of incurring a vanishing gradient during the training process. They show that networks initialized in this way perform better in the training process, achieving faster convergence to the local optimum. Glorot initialization (also known as Xavier) has become one of the most common initialization schemes in the deep-learning literature, and in this paper I adopt it for all of the neural networks. In addition, I initialize all biases to be 0.

To further reduce the risk of incurring in vanishing gradients, I follow Lecun et al. [1998]. I normalize the input of the neural networks along each dimension, so that the grid points have mean 0 and variance 1. As shown by Ioffe and Szegedy [2015], this improves the convergence speed, as it decrease the likelihood of remaining in a local minimum. More advanced normalization techniques, like batch-normalization, did not improve our results. I want to emphasize that the lack of normalization of the input of both the policy functions and forecast function will negatively impact performance in a significant way in this particular application, since the scale of the variables used as input varies dramatically, from units (households' savings) to hundreds (aggregate capital).

Optimization Algorithm Given the large number of parameters with which a neural network is endowed, standard second-order optimization methods are unlikely to be successful. This has led to widespread use of

[5] This is the distribution for sigmoidal activation function, the one used for the consumption policy functions.

the backpropagation algorithm discussed in Section 2.3.3. Backpropagation is paired with gradient-descent, and the deep-learning literature has developed a wide range of first-order stochastic optimization algorithms aimed at training neural networks. In this application, I use the momentum-based Adam algorithm proposed by Kingma and Ba [2015], whose adaptive learning rate has been widely used in deep-learning applications. Adam presents two major advantages. First, it is able to characterize a learning rate for different parameters by using estimations of first and second moments of gradient, helping to equalize the learning speeds across parameters. Second, the use of a momentum-based algorithm improves the training speed when the objective function is highly non-spherical. In exploiting the momentum, the Adam gradient-descent algorithm does not only take into account the current gradient, but also use take previous gradients. The weight that previous gradients and momentum have are ruled by two hyperparameters: the momentum decay (set to 0.9) and the scaling decay hyper-parameter (set to 0.999).

Considering that I simultaneously train hundreds of DNNs approximating consumption policy functions using the global objective function described in Equation (2.3), it is important to choose an algorithm that updates learning rates, as each DNN is likely to have a different contribution to the objective function. This decreases the risk that the objective function remains stuck at local optimum during the optimization process. Table 2.1 shows the average absolute error at the initialization step of our algorithm, compared to standard stochastic gradient descent and RMSprop: as we can see, the use of Adam and RMSprop greatly improves on the basic SGD in terms of convergence speed. After 15 epochs, the approximation error of the policy functions using Adam is approximately 0.1% of that obtained using SGD. This suggests that the selection of update rule is fundamental in determining the speed of convergence, and motivates my choice. Also, the results suggest that Adam performs better than RMSprop, supporting the choice of Adam. With regard to the learning rate, i.e. the weight used determine the size gradient update, it is set to be 0.001 (the default value).

N. of Grad. Steps	SGD	RMSprop	ADAM
1	98.9%	6.4%	7.8%
5	81.7%	0.9%	1.2%
10	65.5%	0.8%	0.5%
15	50.5%	0.7%	0.4%

Table 2.1: Average Absolute % Error vs. Number of Gradient steps, Computed at the Initialization Step

Grid I use Smolyak sparse grids to train the policy functions as in Judd et al. [2014]. The choice of the sparse grid is motivated by the fact that, even if the state space for each individual policy function is reduced to include only a subset of the cross-sectional distribution of savings and average life-time earnings, the input of the policy function is still high-dimensional, since it includes seven state variables. It is important to notice that at the beginning of the iterative procedure, it is not possible to identify where the ergodic set lives. For this reason I choose the bounds on the grids based on the values of the state variables at the deterministic steady state for the median productivity value for two scenarios: a Social Security system that is always able to pay out the benefits, and a Social security system that always pay out benefits proportional to the ratio of total revenues to entitlements. During the iterative procedure, I make sure that during simulation step, the values fall within the boundaries of the grid. The choice of the grid is very important. While neural networks do well with approximating functions on the points of the grid, it is not clear whether outside that domain, their approximation can be considered reliable. Therefore, it is important to ensure that the ergodic set is contained within the grid.

2.5 Policy Approximation through Projection Methods

Projection methods have been the workhorse numerical technique used in macroeconomics to find numerical policy approximations via fixed-point iterations since Judd [1992] introduced them to the economics literature. The general idea is to find the function $\hat{\pi}$ within a specified class of functions $\mathcal{F}$ that bests fits a loss function $\mathcal{L}$. The class of functions $\mathcal{F}$ is chosen a priori, and determines the degrees of approximation that we are seeking in our approximation for the policy functions. In general, we can describe the implementation of a fixed-point algorithm in the context of projection methods as follows:

Step 1. Choose a functional class $\mathcal{F}$ to approximate the policy function π, parametrized by the vector $\boldsymbol{\theta} \in \mathbb{R}^d$, and initialize it $\boldsymbol{\theta}_0$. Choose an appropriate finite dimensional grid $\mathbf{X}$, and a loss function $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$.

Step 2. For $k = 0, \cdots, N$:

1. Generate $\mathbf{y}_k$, consistently with $\mathbf{X}$ and $\pi(\cdot|\boldsymbol{\theta}_k)$:

$$\mathbf{y}_k = F\left(\mathbf{X}, \pi(\mathbf{X}|\boldsymbol{\theta}_k)\right)$$

2. Given $\mathbf{y}_k$ and $\mathbf{X}$, and the loss function $\mathcal{L}$, solve optimization problem:

$$\boldsymbol{\theta}_{k+1} \leftarrow \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{L}\left(\pi\left(\mathbf{X}|\boldsymbol{\theta}\right), \mathbf{y}_k\right)$$

3. If $||\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k|| < \varepsilon$, convergence is achieved, terminate; else go back to step 1.

With regard to our numerical solution, we first need to choose:

- the functional class $\mathcal{F}_c$ parametrized by $\boldsymbol{\Theta}_c$ used to approximate the consumption policy functions:

$$c_{in} : \mathcal{B}_{in} \times \mathcal{B}_a \times \mathcal{Z} \to \mathbb{R} \qquad c_{in}(\cdot,\cdot,\cdot|\boldsymbol{\theta}_{in}) \in \mathcal{F}_c$$
$$\forall (i,n) \in I \times A$$

- the functional class $\mathcal{F}_f$ parametrized by $\boldsymbol{\Theta}_f$ for the forecast functions:

$$\pi_f : \mathcal{B}_a \times \mathcal{Z} \times \mathcal{Z} \to \mathcal{B}_a \qquad \pi_f(\cdot,\cdot,\cdot|\boldsymbol{\theta}_f) \in \mathcal{F}_f$$

- the household specific state variables $\mathbf{x}_{in} = (\overline{e}_{in}, k_{in})$, including the average cumulative earnings e_{in} and savings k_{in};

- the aggregate state variables $\mathbf{x}_a = (K, H, S, \sigma(k))$, including the aggregate capital K and Social Security Trust Fund Balance H, the total amount of Social Security benefits due if the system was solvent S, which allows us to pin down the share of benefits Social Security is able to pay, and the cross-sectional standard deviation of savings $\sigma(k)$.

The goal now is to find collection of parameters $(\boldsymbol{\theta}_{in})_{(i,n)\in\mathcal{I}\times\mathcal{A}}$ for the policy functions and $\boldsymbol{\theta}_f$ for the forecast functions that well approximates the following set of conditions:

- $c_{in}(\cdot|\boldsymbol{\theta}_{in})$ optimize given π_f and Γ_{in}, $\forall (i,n) \in \mathcal{I} \times \mathcal{A}$

$$u'\left(c_{in}(\mathbf{x}_{in}, \mathbf{x}_a, \mathbf{z}|\boldsymbol{\theta}_{in})\right) = \beta_i \mathbb{E}\left[r(\mathbf{x}'_a, \mathbf{z}')u'\left(c_{in+1}(\mathbf{x}'_{in}, \mathbf{x}'_a, \mathbf{z}'|\boldsymbol{\theta}_{in+1}\right)|\mathbf{x}_{in}, \mathbf{x}_a, \mathbf{z}\right]$$
$$\mathbf{x}'_a = \pi_f\left(\mathbf{x}_a, \mathbf{z}, \mathbf{z}'\middle|\boldsymbol{\theta}_f\right), \quad \mathbf{x}'_{in+1} = \Gamma_{in}(\mathbf{x}_{in}, \mathbf{x}_a, \mathbf{z})$$
$$\mathbf{z}' \sim F(\cdot|\mathbf{z})$$
$$\forall (\mathbf{x}_{in}, \mathbf{x}_a, \mathbf{z}) \in \mathcal{B}_{in} \times \mathcal{B}_a \times \mathcal{Z},$$
$$\forall (i,n) \in \mathcal{I} \times \mathcal{A}$$

- Consistency between $(c_{in}(\cdot|\boldsymbol{\theta}_{in}), \Gamma_{in})_{(i,n)\in\mathcal{I}\times\mathcal{N}}$ and π_f

2.6 Numerical Algorithm

Here I describe the algorithm I use to numerically compute the equilibrium policy functions of the model. It is composed of two steps. The first step involves the initialization of the policy and forecast functions to fit the deterministic economy. The second step comprises of the iterative procedure in which policy and forecast functions for the stochastic economy are computed.

2.6.1 Initialization

In the initialization step, I choose an architecture (layers, neurons per layer, activation functions) for the policy functions $(c_{in})_{i\in\mathcal{I},n\in\mathcal{A}}$, the forecast function π_f and the bequest function π_b. I then set the boundaries for the ergodic sets of the household-specific state variables $(\mathcal{B}_{in})_{i\in\mathcal{I},n\in\mathcal{A}}$, and aggregate state variables $\mathcal{B}_a$. Based on these boundaries, I build Smolyak sparse grids as Judd et al. [2014] for each household based on the variables in $\mathcal{B}_{in} \times \mathcal{B}_a \times \mathcal{Z}$. The initialization of the household-specific $\mathbf{X}_{in}^0$ grids is based on the values computed for deterministic steady states described below.

Two deterministic steady states are considered. The reason why we are considering two steady states is because we use their linearized local dynamics as educated guesses regarding the policy functions. In the first one, the Social Security system is always solvent, no matter whether the revenues collected and the assets in the trust fund are sufficient to cover the expenditures. In this scenario, we abstract from dynamics of the trust fund, in the sense that the deficits are not accumulated.[6] This economy is constructed to provide an educated guess as to the optimal consumption of agents when the economy is sufficiently far from Social Security insolvency, but, at the same time, the accumulated balances are sufficiently low such that their impact on prices is negligible. In the second steady state, Social Security is assumed to be insolvent, so that the benefits paid out to retirees not only depend on the their earnings histories, but also on the ratio between total expenditures and revenues. I characterize two different steady states in the deterministic dynamics because I want to exploit the log-linearized dynamics as educated guesses for the policy functions. The deterministic dynamics are clearly likely to experience severe non-linearities, especially in the transition, so the choice of linear policy functions in the deterministic setting is likely to poorly capture dynamics of the economy. However, by considering the two economies, we are implicitly

[6] In a sense, we are assuming that if expenditures exceed revenues at the steady state, a transfer of resources from outside of the economy is completed to ensure that the system is solvent.

considering two limit cases.

In terms of the bounds for the values of the grids, we use very conservative estimates. First, we compute the steady states and the log-linearized dynamics for every value of the exogenous shock. Then, for each type of household, I compute the maximum level of capital holdings and average life-time earnings. I use the following value to compute the bounds for the grid points:

$$k_{in}^{L} = \min_{ss} k_{in}^{ss} - \left[\left(\frac{1}{10} + \frac{6}{10}\frac{n}{R_i}\right)\mathbf{1}(n \leq R_i) + \left(\frac{3}{10} + \frac{4}{10}\frac{A-n}{A-R_i}\right)\mathbf{1}(n > R_i)\right] \times \max_{n,ss} k_{in}^{ss}$$
$$k_{in}^{H} = \max_{ss} k_{in}^{ss} + \left[\left(\frac{1}{10} + \frac{6}{10}\frac{n}{R_i}\right)\mathbf{1}(n \leq R_i) + \left(\frac{3}{10} + \frac{4}{10}\frac{A-n}{A-R_i}\right)\mathbf{1}(n > R_i)\right] \times \max_{n,ss} k_{in}^{ss}$$

The bounds are chosen to guarantee that in the sparse grids the ergodic set is included. The bounds for the other variables are chosen so that the lower bound is $\frac{2}{3}$ of the lowest values among the deterministic steady states and the upper bound is $\frac{7}{4}$ the highest values among the deterministic steady states. Finally, the level of the trust fund balance is assumed to range between 0 and $\frac{1}{2}$ the aggregate capital.

I proceed by initializing the parameters of the policy functions, forecast function and bequest function by finding the $(\boldsymbol{\theta}_{in}^{0})_{i\in\mathcal{I},n\in\mathcal{A}}$, $\boldsymbol{\theta}_{0}^{f}$ and $\boldsymbol{\theta}_{0}^{b}$ that best approximates the linear policy and forecast function around the log-linearized steady states (for more details about the derivation of the linear policy functions, refer to the subsections A.1.1 and A.1.2 in the Appendix). This provides an educated guess as to the policy functions in the stochastic steady state. Considering that neither of the two steady states are able to account for the transition between solvency regimes, I make the following assumptions about how policy and forecast functions are related to the two identified deterministic steady states. For strictly positive values of the Social Security trust fund, I assume the policy and forecast functions behave consistent with the regime wherein the Social Security system is always solvent and initialize them accordingly. When Social Security trust fund resources reach zero, I assume that policy and forecast functions behave consistent with the regime wherein the Social Security system is never solvent, and initialize them accordingly.

Finally, the initialization allows me to assess some of the properties of the hyperparameters that I have chosen, and to choose the appropriate optimization algorithm based on the convergence speed. In this case, given that I am using linearized dynamics to characterize policy and forecast functions, I know what to expect from the target function, which allows me to conduct experiments on the hyper-parameters in a relatively controlled environment. The choice of the optimization algorithm used to update parameters was based on the speed of convergence, as reported in Table 2.1.

2.6.2 Iterative Procedure

The iterative procedure is based on three steps, which are repeated until convergence is achieved:

1. Update the parameters of the households' consumption policy functions;
2. Simulation of the economy;
3. Update the parameters of the forecast and bequest functions.

Convergence is reached when the changes in the consumption predicted value are sufficiently small, or when the maximum number of iterations has been achieved. I now describe the algorithm used to compute the policy functions.

Policy Function Update As discussed later, this step is performed only for the first 10 iterations of the algorithm.

- First, I use the forecast functions to obtain the future values of the aggregate state variables and bequests:

$$\mathbf{X}'_a = \pi_f(\mathbf{X}_a, \mathbf{Z}, \mathbf{Z}'|\boldsymbol{\theta}^k_f) \qquad\qquad B = \pi_b(\mathbf{X}_a, \mathbf{Z}|\boldsymbol{\theta}^k_b)$$

- For each household $(i, n) \in \mathcal{I} \times \mathcal{A}$, I compute the target values for the consumption policy functions by using the Euler equations and the transition equation characterized by Γ_{in}:

$$\begin{aligned}
\mathbf{c}_{in} &= c_{in}(\mathbf{X}_{in}, \mathbf{X}_a, \mathbf{Z}|\boldsymbol{\theta}^k_{in}) \\
\mathbf{X}'_{in+1} &= \Gamma_{in}(\mathbf{X}_{in}, \mathbf{X}_a, \mathbf{Z}, \mathbf{c}_{in}, B) \\
\mathbf{c}'_{in+1} &= c_{in+1}(\mathbf{X}'_{in+1}, \mathbf{X}'_a, \mathbf{Z}'|\boldsymbol{\theta}^k_{in+1}) \\
\hat{\mathbf{c}}^k_{in} &= u^{-1}_{in}\left(\beta_i \mathbb{E}\left[u'_{in+1}\left(\mathbf{c}'_{in+1}\right) r(\mathbf{X}'_a, \mathbf{Z}')|\mathcal{I}_{in}\right]\right)
\end{aligned}$$

- I define the global objective function according to Equation (2.3) for the policy functions:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i\in\mathcal{I}} \sum_{n\in\mathcal{A}} ||\hat{\mathbf{c}}^k_{in} - c_{in}\left(\mathbf{X}_{in}, \mathbf{X}_a, \mathbf{Z}|\boldsymbol{\theta}_{in}\right)||^2_2$$

- For each household $(i,n) \in \mathcal{I} \times \mathcal{A}$, I update the coefficients of the policy functions through batch-gradient descent, where the update rule depends on the function F_{in}:

$$\boldsymbol{\theta}_{in}^{k+1} \leftarrow \boldsymbol{\theta}_{in}^{k} - F_{in}^{k}\left(\nabla_{\boldsymbol{\theta}_{in}} \mathcal{L}(\boldsymbol{\theta}_{in})\right) \quad \forall (i,n) \in \mathcal{I} \times \mathcal{A}$$

  It is important to note that the update of the parameters characterizing the policy functions is completed in parallel thanks to the definition of a global objective function.

Forecast and Bequest Function Update

- Using $c_{in}(\cdot|\boldsymbol{\theta}_{in}^{k+1})$, I simulate J economies for T periods, and I collect the values of the simulated aggregate variables $\mathbf{X}_a^k$ and bequest $\mathbf{b}^k$.

- Based on Equations (2.3) and (2.5), I define the objective function:

$$\mathcal{L}_a(\boldsymbol{\theta}_f, \boldsymbol{\theta}_b) = ||\mathbf{X}_a'^{\,k} - \pi_f(\mathbf{X}_a^k, \mathbf{Z}^k, \mathbf{Z}'^k|\boldsymbol{\theta}_f)||_2^2 + ||\mathbf{b}^k - \pi_b(\mathbf{X}_a^k, \mathbf{Z}^k|\boldsymbol{\theta}_b)||_2^2$$

- I update the coefficients of the forecast function through batch-gradient descent based on the the update rules F_b^k and F_f^k:

$$\boldsymbol{\theta}_f^{k+1} \leftarrow \boldsymbol{\theta}_f^k - F_f^k\left(\nabla_{\boldsymbol{\theta}_f} \mathcal{L}_a(\boldsymbol{\theta})\right)$$
$$\boldsymbol{\theta}_b^{k+1} \leftarrow \boldsymbol{\theta}_b^k - F_b^k\left(\nabla_{\boldsymbol{\theta}_b} \mathcal{L}_a(\boldsymbol{\theta})\right)$$

- Based on the simulated values for $\mathbf{X}_a$, $(\mathbf{X}_{in})_{(i,n)\in\mathcal{I}\times\mathcal{A}}$, compute:

$$\mathbf{c}_{in} = c_{in}(\mathbf{X}_{in}^k, \mathbf{X}_a^k, \mathbf{Z}^k|\boldsymbol{\theta}_{in}^{k+1})$$
$$\mathbf{c}_{in+1} = c_{in+1}(\mathbf{X}_{in+1}^{k\,\prime}, \mathbf{X}_a^{k\prime}, \mathbf{Z}^{k\prime}|\boldsymbol{\theta}_{in+1}^{k+1})$$
$$\hat{\mathbf{c}}_{in}^k = u_{in}^{-1}\left(\beta_i \mathbb{E}\left[u_{in+1}'\left(\mathbf{c}_{in+1}'\right) r(\mathbf{X}_a^{k\prime}, \mathbf{Z}')|\mathcal{I}_{in}\right]\right)$$

- I define the global objective function according to Equation (2.3) for the policy functions:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i\in\mathcal{I}} \sum_{n\in\mathcal{A}} ||\hat{\mathbf{c}}_{in}^k - c_{in}\left(\mathbf{X}_{in}^k, \mathbf{X}_a^k, \mathbf{Z}^k|\boldsymbol{\theta}_{in}\right)||_2^2$$

- I update, for each household $(i,n) \in \mathcal{I} \times \mathcal{A}$, the coefficients of the policy functions through batch-gradient descent:

$$\boldsymbol{\theta}_{in}^{k+1} \leftarrow \boldsymbol{\theta}_{in}^{k} - F_{in}^{k}\left(\nabla_{\boldsymbol{\theta}_{in}} \mathcal{L}(\boldsymbol{\theta}_{in})\right) \quad \forall (i,n) \in \mathcal{I} \times \mathcal{A}$$

It is important to note that the policy functions are updated twice. In the first update, the points of the state space are selected ex-ante based on a guess about where the ergodic lives. This allows us to stabilize the algorithm in the first steps of the iterative procedure. The second update, on the contrary, is based on the state space points obtained through simulation. This step is aimed at increasing the precision of the solution for areas of the state space that are relevant, i.e. belong to the ergodic set. This double update method, while inspired by Maliar and Maliar [2015], departs significantly from it, since the created sparse grids are ultimately discarded. It is also worth noting that, in the second step, there is no use of the forecast functions. Consumption policy functions are updated based on the realized values of the future (endogenous) aggregate state variables, rather than the one predicted using the forecast functions. Finally, we perform the update described in the Policy Function Update subsection only for the first 10 iterations. We do so for two reasons. First, the grid of points used to approximate the ergodic set is chosen ex-ante, and while the bounds are chosen very conservatively to minimize the risk of excluding part of the ergodic set, the risk cannot be completely eliminated, since the deterministic dynamics only gives us an educated guess as to where the ergodic dynamics actually live. Second, we find that numerically, the error on the test data stops decreasing after 8-10 iterations when we continue to updated the policy functions based loss functions derived using the sparse grids. This suggests that while the use of a fixed grid of point can help us stabilize the training of the neural networks in the early stages of the training process, it can also represent a detriment in the later stages, since it may contain points that are rarely traversed by the stochastic dynamics of the model.

2.7 Results

In this Section, I present the numerical results related to the policy approximations of my model. I first introduce the estimation procedure I use to obtain the relevant parameters that I feed into the benchmark model. I then discuss the performance of the algorithm in the fitting of the policy functions and the forecast functions.

2.7.1 Policy Functions

In order to assess the quality of the numerical solution, I use the residuals computed using the Euler equations (see Judd [1992], Krueger and Kubler [2004] and Maliar et al. [2019]). I simulate J economies for T periods

of time. For each period $t = 0, 1, \cdots, T$, and for each household $(i, n) \in \mathcal{I} \times \mathcal{A}$, I compute:

$$\hat{\varepsilon}_{in,t} = 1 - \frac{{u'_{in}}^{-1}\left(\mathbb{E}\left[\beta_i u'_{in+1}(\hat{c}_{in+1,t+1}) r_{t+1} | \mathcal{I}_{in,t}\right]\right)}{\hat{c}_{in,t}} \tag{2.6}$$

$$\forall t = 0, \cdots, T \quad \text{and} \quad \forall (i, n) \in \mathcal{I} \times \mathcal{A}$$

As we can see from Equation (2.6), the overall performance of the model is assessed by computing the percentage deviation of the actual predicted consumption versus the consumption implied by the Euler equation, a standard practice in the OLG literature. I then define the following two metrics:

$$\overline{err}_t = \frac{1}{|\mathcal{A}||\mathcal{I}|} \sum_{j \in \mathcal{I}} \sum_{m \in \mathcal{A}} |\hat{\varepsilon}_{jm,t}| \quad \forall t = 0, \cdots, T \tag{2.7}$$

$$\overline{err}_{in} = \frac{1}{T} \sum_{t=1}^{T} |\hat{\varepsilon}_{in,t}| \quad \forall (i, n) \in \mathcal{I} \times \mathcal{A} \tag{2.8}$$

The metric defined in Equation (2.7) aims at measuring the average deviation in the cross section for a specific period of time t. With this measure, I aim at capturing the performance of the numerical solution pre- versus post-insolvency. Differently, Equation (2.8) aims at measuring the average residual over time for each household. With this measure, I aim at assessing whether the approximation is more (or less) accurate for specific groups of households. In general, the insolvency of the Social Security system is more likely to impact retirees' consumption as compared to that of workers'. Table 2.2 displays the hyperparameters used to set the architecture of the policy functions and to train them, and all the results are based on those hyperparameters.

In terms of network architecture, I test different specifications, since there is a trade-off between approximation power and computational costs. Neural networks with more complex architectures are generally endowed with a higher approximation power, but they are more computationally intensive to train. Therefore, in general, it is preferable to choose a more parsimonious architecture, i.e. an architecture with fewer layers and fewer neurons per layer. Figure 2.6 displays the average test error computed at each iteration of the training process for different architectures. For each step of the iterative procedure, I simulate $J = 200$ economies for $T = 250$ periods, and for each economy I compute the absolute error measures as the average cross-section residual. I then average the error across the different economies:

$$\overline{err}_{test} = \frac{1}{|\mathcal{A}||\mathcal{I}|J} \sum_{j=1}^{J} \sum_{i \in \mathcal{I}} \sum_{n \in \mathcal{A}} \sum_{t=1}^{T} |\hat{\varepsilon}^{j}_{in,t}| \tag{2.9}$$

I use this as a metric to assess how the trained neural networks perform on new test data, which allows me to compare the relative performance of different specifications. As we can see from Figure 2.6, the first 50

Hyper-parameters	
N. of hidden layers	(1,2,3)
N. of neurons	-
Activation function	Tanh
Weights initialization	Glorot
Gradient update	ADAM
Epochs/iteration	2
Epochs at initialization	20
Batch size	32
Learning rate	0.001
N. Iterations	200

Table 2.2: Hyper-parameters used to fit the consumption policy functions.

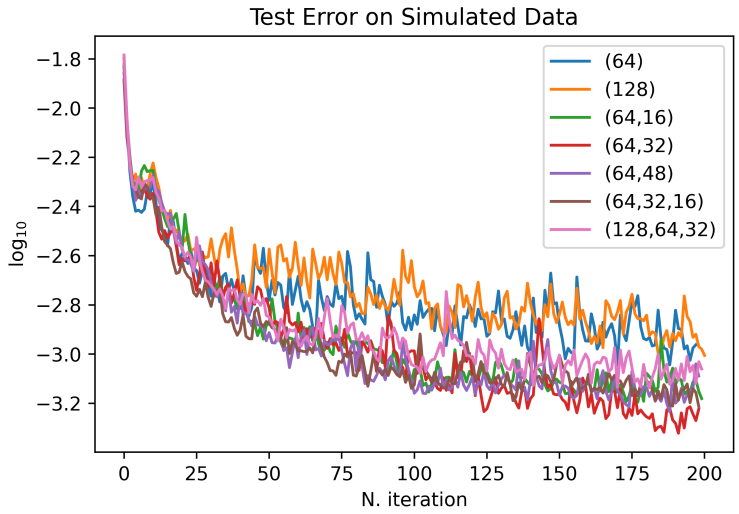

Figure 2.6: Mean absolute relative deviation based on simulated data vs. step of the iterative procedure.

iterations are characterized by fast improvements across the entire architecture, followed by slower changes in the performance over the test data.[7] It is also clear that shallow neural networks, despite having a similar performance over the first 25 iterations, fail to match the performance of deep neural networks once they are trained for a sufficient number of iterations. Shallow neural networks display significantly larger errors on the test data throughout the entire training process, regardless of the number of neurons in the hidden layers.

[7]Test data in this context are generated through simulation.

Architecture	Mean Error	St. dev. Error
(64)	1.14×10^{-3}	1.52×10^{-4}
(128)	1.19×10^{-3}	1.62×10^{-4}
(64,16)	7.83×10^{-4}	1.01×10^{-4}
(64,32)	5.03×10^{-4}	3.7×10^{-5}
(64,48)	7.16×10^{-4}	5.4×10^{-5}
(64,32,16)	7.05×10^{-4}	6.4×10^{-5}
(128,64,32)	8.53×10^{-4}	9.8×10^{-5}

Table 2.3: Mean and standard deviation of the error compute on the test data over the last 20 iterations of the algorithm described in Section 2.6. Tanh is used as activation function.

Table 2.3 shows that between iteration 180 and 200, the average test error is ≈ 1.2 – 1.5 higher for shallow network as compared to deep neural networks, with a higher degree of variability. The results suggest that, in this application, two hidden layer neural networks display very similar performance as compared to three hidden layer neural networks. As we can see, the two hidden layer neural network with $(64, 32)$ neurons has the best performance – in terms of point estimate – but if we factor in the volatility of this measure, this neural network has a very similar performance to a two hidden layer neural networks with 48 neurons in the second layer, and a three hidden layer neural network with 16 neurons in the third hidden layer. Among the deep neural networks, the architecture with the worst performance is the larger three hidden layer architecture – 128 neurons in the first layer, 64 in the second and 32 in the third – followed by the small two-hidden layer network with 64 neurons in the first layer and 16 in the second. Overall, these results suggest that a two hidden layer architecture with 64 neurons in the first hidden layer and 32 neurons in the second achieves the best trade-off in terms of performance and complexity. It is important to notice that while the performance metrics are generally sensitive to the specific architecture choice, the success of the training algorithm is not impacted by the architecture itself. It is important to highlight this point because we do not want our results to be too sensitive to the hyperparameter choice, and, therefore, subject to excessive fine-tuning.

The choice of the other hyperparameters is also very standard. In all of the experiments, the learning rate is set to 0.001, the default set in Tensorflow for the ADAM optimization algorithm. The decay rate for the moments is also set to the Tensorflow default value. The batch size is set to 32, again, the default value. For each step of the iterative process we train the policy functions for two epochs. The choice of the

number of epochs has been made considering the trade-off between achieving a better fit at each step of the iteration (higher number of epochs), and the time required perform a step. I noticed that after two epochs, the decrease in the objective function was relatively small compared to the decrease between the first and the second epoch. Therefore I limited the training of the policy functions to a total of two epochs per each step of the iterative algorithm.

Figure 2.7 shows the average percentage cross-sectional residual implied by households' Euler equations as defined by Equation (2.7). The results are based on the hyperparameters presented in Table 2.2, and on policy functions having two hidden layers, with 64 neurons in the first hidden layer and 32 in the second. The activation function for the hidden layers is the hyperbolic tangent. The results are obtained by simulating an economy for 2000 periods. I use this metric to assess the overall performance of the benchmark specification, i.e. a two hidden layer neural network with 64 neurons in the first layer and 32 in the second. As we can see, the average cross-sectional error are between 0.03% and 0.07%. These results clearly show that our procedure performs very well, as indicated by the results being in line with, if not better than, state-of-the-art methods developed in the literature (see for instance Krueger and Kubler [2004], Hasanhodzic and Kotlikoff [2019], Kim [2018] for traditional polynomial based methods, and Azinovic et al. [2019] for neural network based methods). The spike observed around $t = 20$ represents the moment in which the Social Secu-

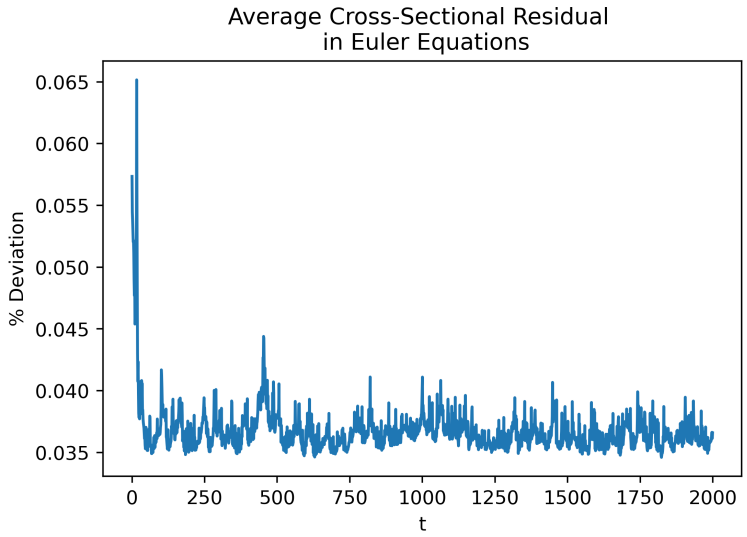

Figure 2.7: Mean absolute relative deviation based on simulated data defined in Equation (2.7) based on policy functions with ReLu activation functions and 64 neurons in the first hidden layer and 32 in the second hidden layer.

rity system becomes insolvent. This suggests that the numerical solution underperforms around the regime

change, as compared to either the pre- or post-bankruptcy regimes. Nevertheless, the maximum average cross-sectional residual is only 0.065%, indicating that the decrease in performance is small in relative terms. As a consequence, these results show that neural networks successfully approximate consumption policy functions in the context of this application. It is worth noting that performance of the numerical solution does not appear to suffer from the use of a reduced state space. This suggests that the use of a full state space, while desirable from a theoretical perspective, may not be necessary. If we were to use the same specification for the neural network for the policy functions specified by Azinovic et al. [2019], then we would need to compute $(634+1)\times1000+(1000+1)\times1000+(1000+1)\times316 = 1,952,316$ weights and biases. In our current specification, the total number of parameters is $312 \times ((8+1)\times 64 + (64+1)\times 32 + (32+1)) = 838,968$, less than half the alternative method.

I now turn to the assessment of the performance of the policy approximations for each household, since I am interested in understanding whether I can identify variation in the quality of the numerical solution across either different types or cohorts of households. Figure 2.8 shows the metric defined in Equation (2.8) based on an economy simulated for 2500 periods. On the x-axis, I distinguish the 4 types of agents. On the y-axis, I represent different cohorts. As we can see from the chart on the left in Figure 2.8, the mean

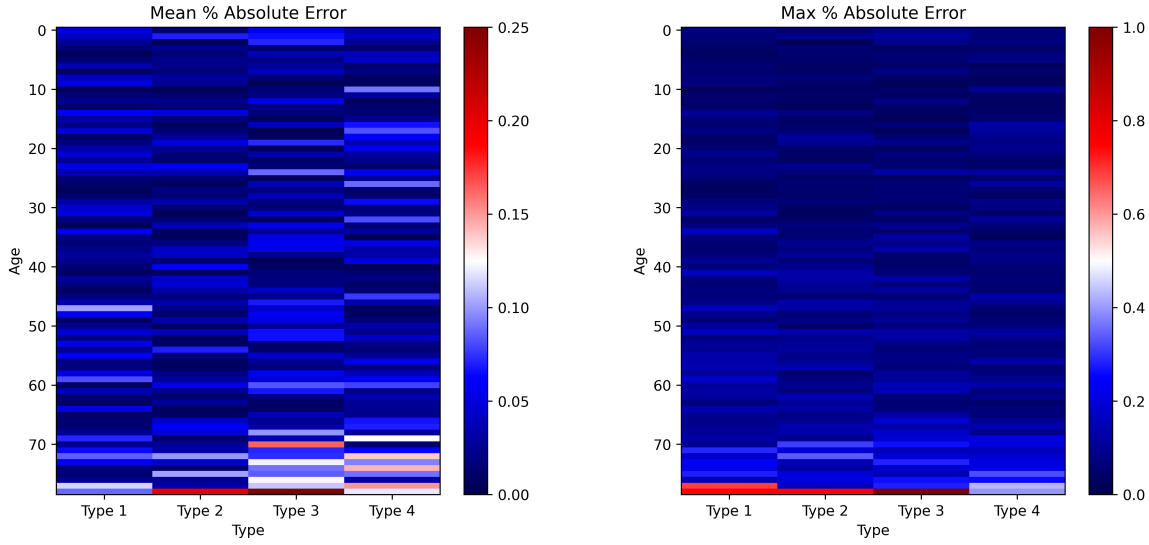

Figure 2.8: max and mean percentage error $\hat{\varepsilon}_{in}$, for each household, based on simulated data. On the x-axis, the type of agent. On the y-axis, ordered from top to bottom, the household cohort.

deviation is comparable across different household types, with older households' (cohort 78 and 79) policy

functions being characterized on average by higher approximation errors. At the same time, we can observe that the average performance varies non-systematically across types or age cohorts. However, the mean absolute error obtained in the test data is bounded above by 0.25%, in line with literature benchmarks in terms of numerical performance (see for instance Krueger and Kubler [2004], Krueger and Kubler [2006], Kim [2018]). Turning to a measure of the lower-bound of the performance, we can see from the chart on the right side in Figure 2.8 that the policy functions of the last two cohorts have the highest maximum absolute residuals, but they remain lower than 1%. Again, no specific pattern is observed in different types of agents. In light of these observations, it is possible to infer that the decrease in the performance of the policy functions of older agents happens around the period of regime transition. I hypothesize that the higher error observed is due to the fact that the consumption of older retirees depends more on Social Security benefits as compared to consumption of younger retirees. Therefore, the higher error rate represents a more imprecise prediction about the bankruptcy state of the economy and is more likely to deteriorate the performance of older retirees' policy functions. Overall, our results indicate that neural networks can be successfully used as numerical approximators in the context of large-scale OLG models.

I now illustrate the importance of using a flexible class of functional approximators like neural networks to represent the policy functions. Figure 2.9 and 2.10 shows how older households' consumption policy function depends on the Social Security trust fund when hyperbolic tangent and ReLU activation are used respectively. In particular, Figure 2.9 and 2.10 displays the consumption policy function belonging respectively to the 76^{th}, 77^{th}, 78^{th} and 79^{th} cohort for each of the agent types. As we can see, for high values of the Social Security trust fund, the policy function is relatively flat: as we move away from bankruptcy, retirees' consumption-saving behavior will be largely unaffected by the trust fund balance, as the funds accumulated will be sufficient to pay out benefits in the near term. It is important to keep in mind that the Social Security trust fund affects agents' decisions through i) prices, as it affects the interest rate that is charged to borrowers/savers and the market wage and ii) the benefit payments (whether the Social Security system is insolvent or not, directly affecting the amount of benefits paid out). As we approach bankruptcy, agents will tend to to decrease consumption and save more, since, as rational agents, they anticipate that Social Security will not be able to fully pay their retirement benefits, and will cut them according to the ratio of total revenues to benefit entitlements.

In addition, it is important to notice from Figure 2.9 that the younger the household, the earlier it will start to decrease consumption, and thus save more, since the bankruptcy (which is an absorbing state in this model) will decrease the benefits they receive for a longer part of their life. The shape of

the policy functions around the transition point changes as we change the the activation function used in the hidden layers. The hyperbolic tangent is a smooth function, and this is reflected in the smooth profile of the implied policy functions. Similarly, the piece-wise linear nature of ReLU activation functions is reflected in the policy functions. One may consider this to be an issue, i.e. the lack of smoothness in the

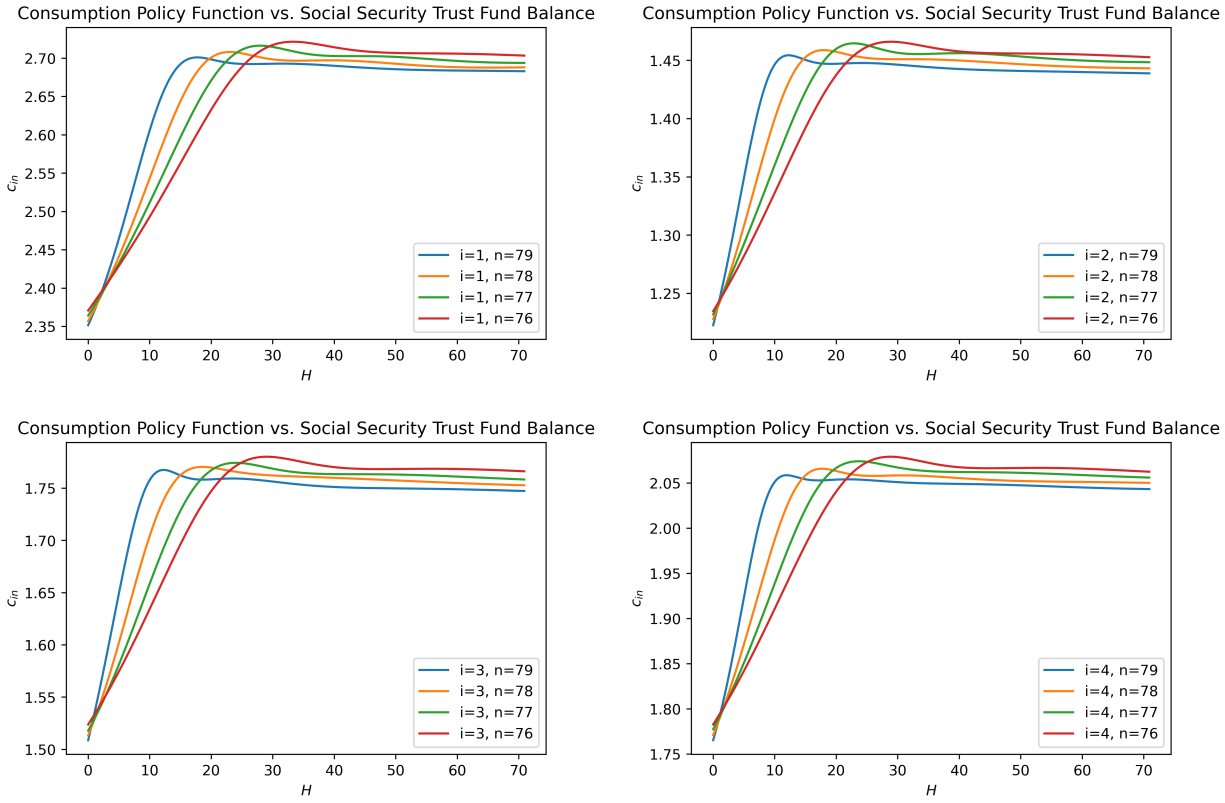

Figure 2.9: Numerical approximation of consumption policy functions for selected groups of retired agents. On the x-axis the Social Security balance H. The neural network has two hidden layers, with 64 neurons in the first hidden layer and 32 in the second. The activation function for all of the hidden layers is the hyperbolic tangent.

ReLU activation function is translated directly into the policy function. However, this particular behavior manifests since that part of the state space is covered only in the transition: once Social Security reaches insolvency, it will remain insolvent forever. Therefore, while training the policy functions on the simulated data, rather than the on grid points, we are implicitly giving higher weight to regions of the state space that belong to the ergodic set, especially if we are simulating the economies for long period of time. In the economy presented, insolvency is reached after 10 to 20 periods, depending on the initial conditions and the sequence of simulated shocks, while in the training process, economies are simulated for 250 periods. This

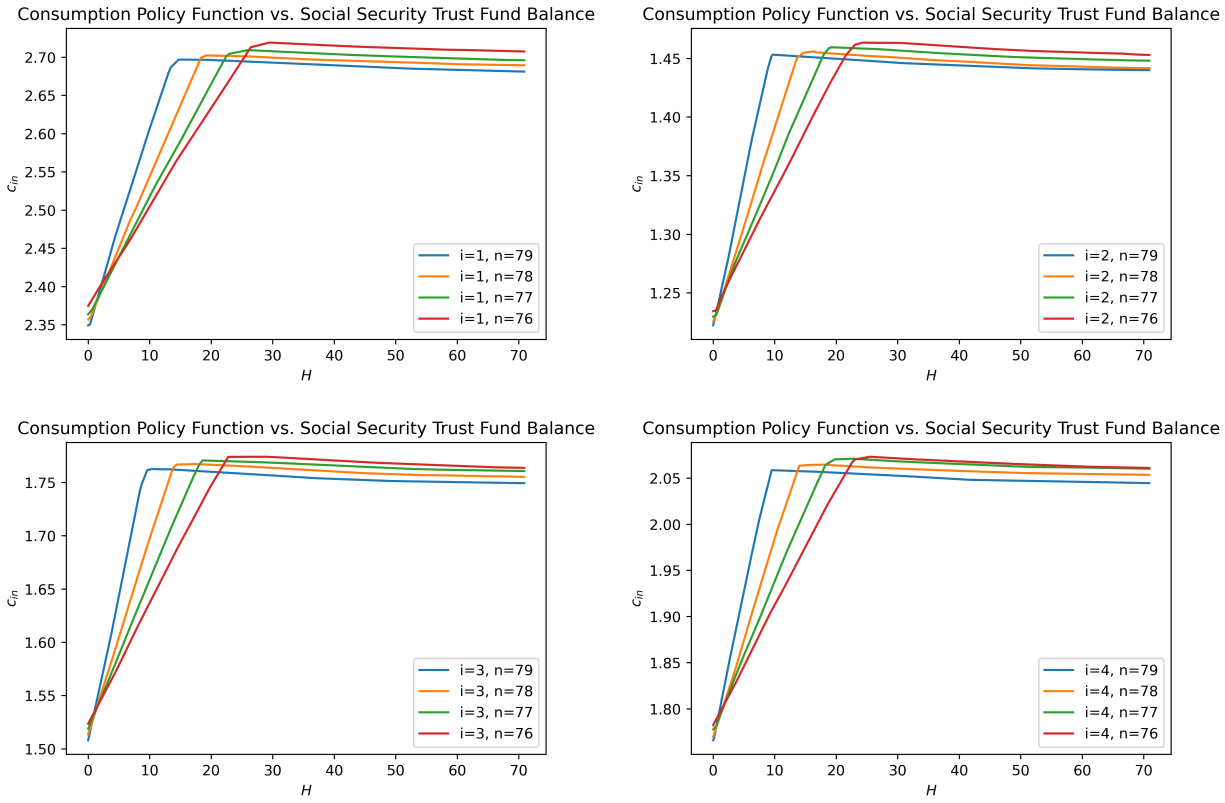

Figure 2.10: Numerical approximation of consumption policy functions for selected groups of retired agents. On the x-axis the Social Security balance H. The neural network has two hidden layers, with 64 neurons in the first hidden layer and 32 in the second. The activation function for all of the hidden layers is the ReLU.

implies that less than 10% of the training points belong to the pre-insolvency regime. From a theoretical perspective,the choice of the activation functions is irrelevant, and generally, it is possible to construct an alternative neural network with a target activation function with the same approximation errors as a corollary of Theorem 2. To further illustrate this point, I compute the numerical solution of the proposed model using the same architecture for the consumption policy functions but different activation functions. In the first experiment, I use the hyperbolic tangent, while, in the second experiment, I ReLU. Figure 2.11 shows the performance over simulated data of the policy functions during the training process. It is clear that while the ReLU activation function has a faster convergence rate in terms of test performance in the earlier stages of the training process, after 150-200 iterations, the performance is virtually the same. These results show that numerically, the choice of the activation function does not impact the performance at convergence.

In Figure 2.12 I plot the consumption policy functions of retirees against both the Social Security trust fund H and the total amount of Social Security entitlements S. The reason why I consider the total

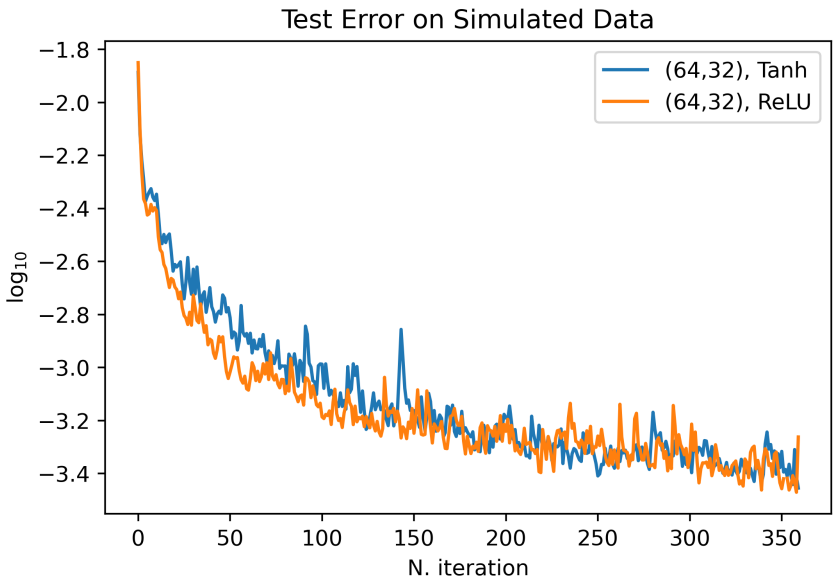

Figure 2.11: Test error on simulated data vs. iteration

amount of Social Security entitlements is because they are a variable that has very low-frequency fluctuations, as, in each period, only a small fraction of retirees die and are replaced by the newly retired cohort of agents. Therefore, a high level of retirement entitlements is likely to be persistent for multiple periods of time. At the same time, in the case of insolvency, the level of entitlement determines the share of benefits that Social Security is able to pay when it is insolvent. Therefore, we expect this variable to play a role when the system is insolvent ($H = 0$) or when it is approximating insolvency. In particular, we expect that consumption of retirees will be negatively impacted by a high level of entitlements, since this will translate into faster insolvency and a lower level of benefits, keeping all other variables fixed. At the same time, consumption of the retirees should be largely unaffected when the economy is far from an insolvent Social Security. As we can see from the plots, a higher level of Social Security entitlements decreases consumption and therefore increases savings of retired agents since it signals that either bankruptcy is more likely to happen, or that the Social Security benefits will be reduced by a larger amount, given the high level of current entitlements. The effect is more pronounced for the older agents, as ceteris paribus have less time to smooth their consumption. At the same time, as expected, when the Social Security trust fund balance is high, we can see that S does not greatly impact consumption. It is interesting to notice how the use of neural networks allows us to approximate functions that display strong non-linearities in a multi-dimensional setting.

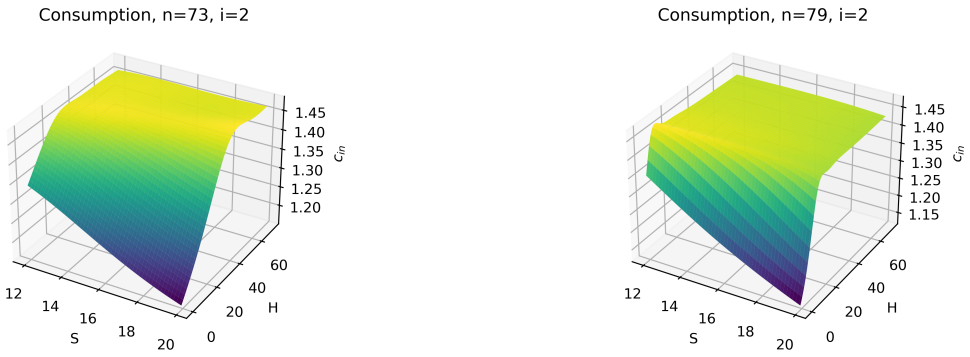

Figure 2.12: Numerical approximation of consumption policy functions for retirees belonging to cohort 73 (left chart) and cohort 79 (right chart). On the x-axis, the total amount of Social Security entitlements S. On the y-axis the Social Security balance H.

2.7.2 Forecast Functions

Table 2.4 shows the hyperparameters used for the neural networks approximating the forecast and bequest functions. As in the case of households' consumption policy functions, I use deep neural networks. Forecast functions are expected to be characterized by significant non-linearities as well, and thus, require the use of flexible functional approximators. As previously discussed, the forecast function can be considered a non-fully connected neural network with a multi-dimensional output. Experimentation in the initialization

Hyperparameters	
N. of hidden layers	3
N. of neurons	(96,48,32)
Activation function	ReLU
Gradient update	ADAM
N. Epochs	2
Batch size	32
Learning rate	0.001
N. Observations	30,000

Table 2.4: Hyperparameters used to train the forecast function π_f and the bequest function π_b.

step indicates that a three-layer deep net performs better than a two-layer neural network or a shallow neural

network. In particular, the pace at which the objective function decreases in the first steps of the initialization process is higher in a three hidden layer configuration as compared to shallower neural network. In addition, forecast functions do not constitute the computational bottleneck of the iterative algorithm. For this reason, I decide to use a more complex neural network for each bequest and policy function. Nevertheless, it is important to highlight that our results are not particularly sensitive to specification of the architecture of either the forecast function or bequest function.

In terms of activation function, I select ReLU as activation function for all the hidden layers, while the output layer has a linear activation function. The learning rate is set to the default rate in Tensorflow for Adam optimization algorithm. In the training process, I divide the training sample into batches of size 32. In addition, for each iteration of the, I train the bequest and forecast functions for 5 epochs. Table 2.5 shows the mean squared error of the forecast and bequest function obtained during the training process. The second row displays the bounds of the interval in which the simulated data live. As we can see, they achieve a satisfactory level of accuracy.

	K	S	$\sigma(K)$	B
MSE	$2.4 \cdot 10^{-1}$	$2.1 \cdot 10^{-3}$	$4.4 \cdot 10^{-6}$	$1.8 \cdot 10^{-1}$
Interval	$[400, 1200]$	$[10, 25]$	$[2, 3]$	$[6, 10]$

Table 2.5: Training performance of the forecast and bequest function.

I now provide further evidence supporting the use of a flexible class of functions to represent the forecast function. Figure 2.13 displays the forecast function for different levels of expenditure S and current trust fund balance H. As we can see, the forecast function is piece-wise linear: for a sufficiently low level of trust fund balance (up to the threshold identified by the kink point), Social Security will not be able to cover the promised benefits to retirees, and therefore will remain in bankruptcy. However, it is important to highlight that the deep network was able to approximate a piece-wise linear function without issue, as it exploited the piece-wise linearity of the ReLU activation functions in the hidden layers. This intuition is reinforced in Figure 2.14.

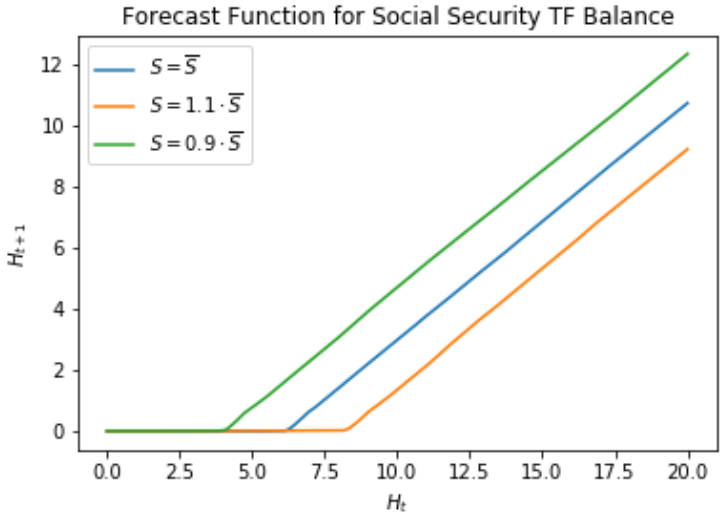

Figure 2.13: Computed forecast function for Social Security trust fund balance. On the x-axis, the current Social Security trust fund balance.

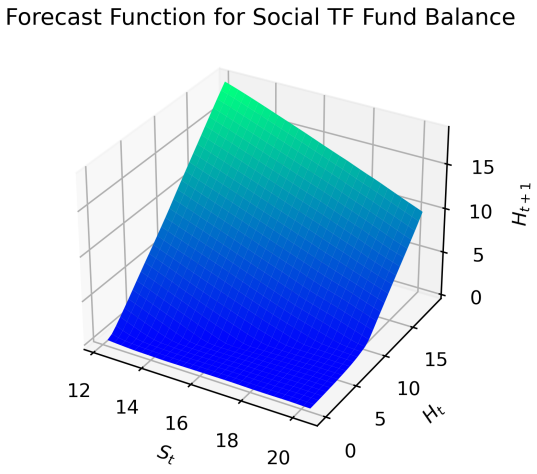

Figure 2.14: Computed forecast function for Social Security trust fund balance. On the x-axis, the current Social Security trust fund balance.

2.8 Conclusion

In this chapter, I show how economists can take advantage of neural networks paired with deep-learning techniques in economic applications and, specifically, in the context of large-scale OLG models. While I use forecast functions and a reduced state space to alleviate the curse of dimensionality, the proposed approach

is able to solve for hundreds of policy functions simultaneously by using a global objective function coupled with an appropriate moment-based gradient-descent algorithm that is able to iteratively adjust the learning rate of each parameter of each individual policy functions. This allows us to easily and automatically parallelize the training of the hundreds of policy functions.

The use of a reduced state space has been criticized in the past, as it has been shown to deliver inferior results as compared to numerical solutions developed using a full state-space approach in certain applications (see for instance Krueger and Kubler [2004]). The results presented in this chapter, however, provide support for the notion that a reduced state space paired with projection-based techniques as introduced by Krusell and Smith [1998] can still be used to decrease the computational burden of the algorithm characterizing the numerical solution, while at the same time achieving similar, if not better numerical performance as compared to state-of-the-art methods relying on the use of the entire state space. In this regard, we can conclude that i) the interaction between the choice of the specific equilibrium concept, the numerical solution, and the relative accuracy depend on the specific problem and the holistic nature of the algorithm developed; and ii) reduced-state space numerical solutions should not be discarded a priori. In particular, I complement the contribution of Maliar et al. [2019] and Azinovic et al. [2019] by showing that neural networks can be easily used even in the context of reduce state-space numerical framework.

In this chapter, we also show that we are able to compute the numerical solution of a complicated OLG model with relatively little fine-tuning of the parameters characterizing the architecture of the neural networks and the training process. We show that in general, two hidden layer deep neural networks perform better than shallow and three hidden layer neural networks in terms of speed of convergence and size of the residuals at convergence. However, the results presented suggest that the specific architecture of the neural network is not a major driver in the success of the application, as long as the architecture chosen is expressive enough. While it is important for economists to take into account the trade-off between computational complexity and the expressiveness of certain architectures, ultimately our results show little sensitivity to the neural network architectures. In addition, the majority of the other hyperparameters is left to the default set in Tensorflow, or follows standard conventions widely shared in deep learning techniques, such as the use of the Glorot initialization for the weights and biases, or the use of ADAM as a gradient-descent based optimization method.

The success on this application shows that it is possible to use neural networks to provide global approximations of non-linear policy functions in the context of large and complex OLG models. Based on

these results, it is clear that the neural networks can be easily extended to other applications. In particular, it would be interesting to use neural networks to numerically approximate policy functions in OLG models where agents face non-linear borrowing costs, or where the non-linearities in the dynamics enter explicitly in the ergodic sets, rather than simply appearing in the transition between stochastic steady states. In Chapter 3, I will use the techniques developed in this chapter to numerically derive the solution under a variety of policy alternatives. As my results will show, the numerical solution developed in this chapter is highly adaptable to different policy scenarios, and requires no additional fine-tuning.

Chapter 3

Estimation and Counterfactual Analysis

3.1 Introduction

In this chapter, I investigate how potential reforms to Social Security would impact the consumption-savings behavior of households in a general equilibrium framework in order to better understand the welfare implications of alternatives policy regimes. The analysis is motivated by the current debate about the future of Social Security; a debate that has only intensified with the recent forecasts that project Social Security will be insolvent by the mid-2030s. Policymakers are called to act, and different policy proposals have been put forward, ranging from the total elimination of Social Security, to long-term reduction in retirement benefits, to an expansion of the scope of the program. However, no systematic analysis has been performed to assess the impact on the alternative policies on households' welfare. As such, to address this void, the goal of this chapter is to assess the differential impact of alternative policy proposals on the welfare of heterogeneous households. As discussed in Chapter 1, the rate at which Social Security replaces labor income during retirement varies greatly, and generally depends on the average life-time earnings and retirement age of the retiree. Therefore, an agent's socio-economic status will determine how their consumption-savings decision interacts with Social Security because retirement benefits paid by Social Security tend to replace a higher percentage of earnings for low-earners as compared to high-earners.

In light of these considerations, it is important to identify and estimate the sources of heterogeneity that are most likely to interact with the consumption-savings decision of households and with the institutional features of the current Social Security system and/or the alternative policy proposals. In section

3.2-3.5, I estimate the sources of agents' heterogeneity described in the OLG model introduced in Chapter 1. In Section 3.2, I estimate the life-time earning profiles, and in Section 3.3, I estimate mortality rates and average life-expectancy based on a semi-parametric Cox proportional hazard model. In Section 3.4, average retirement age is estimated assuming that retirement is a terminal state using a semi-parametric Cos proportional hazard model. In Section 3.5, I use different techniques to pin down parameters characterizing the discount factor and the coefficient of relative risk aversion. I first rely on the use of linear and non-linear GMM routine presented by Alan et al. [2009] to estimate agents' coefficient of relative risk aversion and discount factors using Euler equations as moment conditions. I then propose a novel routine estimation based on the use of the Expectation-Maximization algorithm aimed at uncovering unobserved heterogeneity in the coefficient of relative risk aversion using as moment conditions linearized Euler equations. I find evidence suggesting the presence of heterogeneous coefficient of risk aversion in the population, with the population share of the high risk-averse type being increasing the educational attainment.

Finally, in Section 3.7, I examine three alternative policy scenarios to the current Social Security regime, since, as discussed in Chapter 1, Social Security is projected to become insolvent in the next 15 years. Specifically, I compare the welfare of households under the following three policy alternatives: (1) utilizing an endogenously determined tax rate to prevent insolvency; (2) utilizing a permanent reduction in benefits to prevent insolvency; and (3) embracing insolvency by eliminating Social Security completely. We will discuss the impacts that each alternative policy scenario has on the different groups of households. While the majority of our analysis will rely on the welfare at the stochastic steady states, we will also consider the transition costs that each policy imposes. The analysis will rely on the numerical computation of households' optimal consumption, which we obtain using the algorithm proposed in 2. As we will see, the proposed procedure adapts very well to all of the different scenarios, with no fine-tuning on the hyper-parameters required. This showcases the versatility of the algorithm developed in Chapter 2. In addition, the performance of the algorithm in terms of the precision of the numerical solution it achieves is at least as good as the performance of state-of-the-art methods outlined in the literature. Finally, the analysis proposed in Section 3.7 shows how the use of neural networks scales well and can easily be used in the context of large OLG models, since it is able to numerically handle an economy populated by eight types of households living for eighty periods of time.

3.2 Estimation of the Life-Cycle Earnings Profiles

As discussed in Chapter 1, retirement benefits depend on the life-time earnings of an individual, given that the Principal Insurance Amount (PIA) is computed as the average of the indexed life-time earnings. This, in turn, has an impact on the the replacement rate of Social Security benefits, which is the share of average life-time earnings that Social Security retirement benefits replace. As we have seen in Chapter 1, Figure 1.4 shows that the replacement rate varies significantly across income levels, and ranges from 90% for the lowest income bracket to less that 30% for the top-income level insured by Social Security. In the model retirement coincides with the collection of Social Security benefits. As such, heterogeneous replacement rates will likely have different impacts on the consumption-savings behavior of different households, as consumption at retirement is financed through capital and retirement benefits. For instance, a lower replacement rate may induce higher saving rates during the working life to allow for consumption smoothing. In addition, the profile of the labor earnings itself interacts with the household consumption-saving decision; a steeper profile may lead households to borrow more in the earlier stages of their lives as compared to households with a flatter labor income profile. Therefore, heterogeneity in the life-time income profiles represents an important factor interacting with the institutional details of Social Security retirement benefits.

3.2.1 Data

In this subsection, we describe in greater detail the construction of the sample used. For the estimation of the wage equation, we use the data from the Family-Individual File of the Panel Income of Survey Dynamics (PSID). The Family-Individual database contains a record for each member of all households that are interviewed in a given year. The survey spans the period between 1968 and 2017, and households have been interviewed annually from 1968 to 1997, and biannually from 1999 to 2017. This accounts for a total of 39 years of observations. As of September of 2019, the Family-Individual file had 80,666 records. For each record, the file contains information about socioeconomic variables at the individual and household levels.

The hourly earnings of the individuals were obtained from the Family part of the PSID, since the Individual portion of the dataset does not contain any information about employment or earnings. The Family portion records the hourly earnings of the head of each household, and, if present, the spouse. The data about earnings recorded in a certain year refers to the actual earnings gained the year prior. The

Family part of the survey does not provide any information about the earnings of other members of the household. The earnings of the head of household and the spouse of each household are available for all years except for 1993. From the sample, we remove the earnings observations that have been top-coded. The PSID uses different thresholds over time: 99.99$ before 1993, no top-code between 1994 and 1997, and 999.99$ starting in 1999. In addition, we remove observations of earnings that are lower than a third of the prevailing average federal minimum wage for non farming workers[1] in a given year. In this way, we exclude individuals who report extremely low earnings. We experiment with different thresholds (a half and a quarter of the federal minimum wage), but our results are not affected by this choice. We also remove observations of earnings of individuals aged below 22 and above 66 at the time of the survey. In this way, we want to exclude transitional jobs into full employment for the young or out of employment for those at or nearing retirement age. From the total sample, we remove individuals for which we have less than 15 years of available and (strictly) positive earnings data, of which at least 5 years needs to be consecutive. In this way, we remove from the sample individuals who do not display significant attachment to the labor force. In total, we are left with 7,119 individuals who alone satisfy these criteria, 3,974 males and 3,325 females. We then convert such nominal quantities into real quantities. We use the chain-type price deflator CPI-All Urban Consumers.

We use the Family file to obtain information about the total number of hours worked by individuals. As in the case for hourly earnings, the Family portion of the PSID reports the total number of hours worked in the year preceding the interview for the head of household, and if present, for the spouse. In the survey, a value of 0 indicates that the individual was not working for pay, i.e. was not participating in the labor market.

We use the Individual File to construct our measure of educational attainment. The PSID reports the total number of years of schooling a person has attained, and values range from 1 to 17 for individuals who have actually attended at least a year of school. Information about the number of years of schooling is available for all surveys except for 1969. We code based on the following categories: (1) individuals who have attended less than 12 years of school are coded as less than high-school educated, (2) individuals who have attained 12 years of education are coded as high-school educated, (3) individuals with between 12 and 15 years of schooling are coded as having some college education, and (4) individuals with at least 16 years of schooling are coded as college educated.

[1] https://fred.stlouisfed.org, series: FEDMINNFRWG.

3.2.2 Estimation

In this subsection, we describe the estimation strategy used for the the life-time earnings profile and the replacement rate of Social Security benefits. Consistent with the assumptions made in Chapter 1, it is assumed that markets for labor are perfectly competitive, and wages are equalized to the marginal product of labor. First of all, it is important to start from the earnings equation for individual i in year t:

$$\text{earnings}_{i,t} = w_{i,t} l_{i,t} \tag{3.1}$$

This is comprised of two elements: the hourly wage $w_{in,t}$ and the total number of hours of labor supplied in a year $l_{in,t}$. In the model, it is assumed that labor is supplied inelastically, while the wage rate $w_{in,t}$ is determined by competitive markets. Conditional on individual characteristics $\mathbf{x}_{in,t}$, the wage rate $w_{in,t}$ can be expressed as follows:

$$w_{it} = \omega_t \exp(\boldsymbol{\beta}_w' \mathbf{x}_{in,t} + \varepsilon_{in,t}^w) \tag{3.2}$$

where ε_{it}^w represents a shock unobserved to the econometrician. The wage rate for individuals depends on individual labor market experiences, summarized in $\mathbf{l}_{it}$, and demographic characteristic in $\mathbf{z}_{it}$, so that $\mathbf{x}_{it} = (\mathbf{z}_{it}, \mathbf{l}_{it})$. In Equation (3.2), ω_t denotes the wage rate per efficiency unit of labor, and represents what is determined by the marginal utility of the production function of the representative firm. Our goal is therefore to estimate the average total hours-efficiency supplied by individuals. By taking the logarithm of Equation (3.2), we obtain the following linear equation:

$$\log(w_{it}) = \boldsymbol{\beta}_w' \mathbf{x}_{it} + \varepsilon_{it}^w \tag{3.3}$$

At the same time, we need to estimate the the number of hours of labor supplied by each type and age-cohort. In order to do so, we estimate the following regression:

$$l_{it} = \boldsymbol{\beta}_l \mathbf{x}_{it} + \varepsilon_{i,t}^h \tag{3.4}$$

where $l_{i,t}$ represents the total number of hours worked, and $\mathbf{x}_{i,t}$ represents a vector of explanatory variables, including: a polynomial of age up to the third degree, a categorical variable representing the educational level, and gender. This assumption is consistent with the fact that, in the model, labor is supplied inelastically, and does not respond to any aggregate shocks incorporated in ω_t.

We now turn to the general estimation strategy for the life-time earnings profile: we first estimate via OLS regression Equation (3.3). Based on the estimates of the wage per unit of efficiency, we then

compute, for each of the selected group of individuals (type i-age n), the average value of each of the employment-related explanatory variables in $\hat{x}_{in}$ based on the estimates obtained in (3.4).

$$\hat{n}_{in,t} = \hat{\omega}_t \exp(\hat{\boldsymbol{\beta}}_w \hat{\mathbf{x}}_{in,t}) \hat{l}_{in,t}$$
$$\hat{l}_{in,t} = \hat{\boldsymbol{\beta}}_l \mathbf{x}_{in,t}$$

We now describe the variables used in Equation (1.11). The vector $\mathbf{l}_{it}$ contains variables related to labor-market participation: lagged hours worked (up to 2 lags), lagged labor market participation status, a binary variable taking the value of 1 if an individual i worked at least 200 hours in a year, and 0 otherwise. In this way, we aim at capturing the non-linear relationship between wages and past labor market experience. In the vector $\mathbf{z}_{it}$ we include a polynomial of third degree in age, a categorical variables representing educational attainment and gender, and interaction terms between age and educational attainment.

Table 3.1 shows the estimated coefficients for the demographic variables, and Table 3.2 shows the coefficients for the employment variables. The first column of the two tables displays the coefficients of a richer representation, where interaction terms between age and educational attainment are added. This allows us to potentially incorporate a different return to experience (here proxied by age) based on the educational level. As we can see from Table 3.2, the addition of interaction terms has little impact on the fit of the regression. The regression results obtained are also used to compute the replacement rate of Social Security retirement benefits and are based on the assumption that the individual has always participated in the labor market, so that the dummy indicators in $\mathbf{l}_{it}$ for lagged labor market participation are all set to 1. This assumption is justified by the fact that we are interested at looking at individuals who have Social Security benefits through their working history, and therefore have displayed significant attachment to the labor force throughout their career. Considering that the we have included a time component in the wage regression, and that the Social Security Administration changes the way in which it indexes and caps earnings every year, we make the following additional assumptions to compute the replacement rates. First, we replace the time dummy with a time trend, since the PSID data is reported bi-annually after 1999. Second, we compute the replacement rate for a person who is 19 is 1973, starts working at the age of 22 (compatible with the age at which agents are born in the model), and works until the age of 66. We then compute the life-time earnings and labor supply profile, we index them to make it in real 2019 dollars, we take the highest 35 years of earnings, and we obtain the $\hat{AIME}_{in}$ for each group of individuals. We then compute the Social Security retirement benefits at FRA, namely $\hat{SSB}_{in}$, and we compute the replacement rate as the ratio between the

Table 3.1: Estimated Coefficients for the Wage Equation, Demographic Variables

	Dependent variable:	
	Real Log Earnings	
	(1)	(2)
HS	0.0050	0.2365^{***}
	(0.0666)	(0.0044)
Some College	-0.1835^{**}	0.3993^{***}
	(0.0748)	(0.0050)
College+	-0.5401^{***}	0.7239^{***}
	(0.0776)	(0.0048)
Age	0.0320^{***}	0.0317^{***}
	(0.0069)	(0.0062)
Age_Sq	-0.0005^{***}	−0.0002
	(0.0002)	(0.0001)
Age_Cub	0.000002	−0.000002
	(0.000001)	(0.000001)
Female	-0.3166^{***}	-0.3172^{***}
	(0.0035)	(0.0035)
HS x Age	0.0088^{***}	
	(0.0032)	
Some College x Age	0.0244^{***}	
	(0.0036)	
College+ x Age	0.0527^{***}	
	(0.0037)	
HS x Age_Sq	-0.0001^{**}	
	(0.00004)	
Some College x Age_Sq	-0.0002^{***}	
	(0.00004)	
College+ x Age_Sq	-0.0005^{***}	
	(0.00004)	

Table 3.2: Estimated Coefficients for the Wage Equation, Employment Variables

	Dependent variable:	
	Real Log Earnings	
	(1)	(2)
HRS(t-1)	−0.00001***	−0.00001***
	(0.000003)	(0.000003)
HRS(t-2)	0.00004***	0.00005***
	(0.000003)	(0.000003)
HRS(t-3)	0.00001**	0.00001***
	(0.000003)	(0.000003)
Employed(t-1)	0.1939***	0.1928***
	(0.0103)	(0.0104)
Employed(t-2)	0.1426***	0.1390***
	(0.0093)	(0.0093)
Employed(t-3)	0.1023***	0.0975***
	(0.0093)	(0.0093)
Time Dummy	Yes	Yes
Observations	139,611	139,611
R^2	0.2716	0.2686
Adjusted R^2	0.2715	0.2686
Note:	*p<0.1; **p<0.05; ***p<0.01	

average life-time earnings the the Social Security at FRA:

$$\begin{aligned}\hat{SSB}_{in} =& 0.9 \cdot \mathbf{1}\left[\hat{AIME}_{in}\right] + \\ & 0.32 \cdot \mathbf{1}(\hat{AIME}_{in} \geq 926)\left[\hat{AIME}_{in} - 926\right] + \\ & 0.15 \cdot \mathbf{1}(\hat{AIME}_{in} \geq 5583)\left[\hat{AIME}_{in} - 5583\right] \\ \hat{\theta}_i =& \frac{\hat{SSB}_{in}}{\overline{earnings}_{in}}\end{aligned}$$

Table 3.3 shows the wage index-adjusted (in 2019 dollars) life-time earnings used by Social Security to compute the PIA, by sex and educational attainment. The estimates are based on the average regression results obtained from Equation (3.3). As we can see from Table 3.3, there is a significant gender gap in life-time earnings, even after controlling for educational attainment and labor market participation. Also, as expected, the AIME is increasing with educational attainment.

	<HS	HS	Some College	College+
AIME (Men)	2,980	4,393	5,230	7,356
Replacement Rate at FRA (Men)	0.500	0.442	0.423	0.352
AIME (Women)	1,522	2,335	2,903	4,222
Replacement Rate at FRA (Women)	0.673	0.550	0.505	0.447

Table 3.3: Average AIME, Wage Index Adjusted in 2019 dollars and estimated replacement by gender and educational attainment.

In Table 1.18 we compute the replacement rate at FRA, i.e. the ratio between the PIA and the AIME, which represents the share of income that Social Security benefits replace at full retirement age. As we have seen in Chapter 1, Figure 1.4, the replacement rate is decreasing with income. Male high-school graduates and those with some college education have similar replacement rates (44.2% and 42.3% respectively). For both men and women, estimates show that the replacement rate of college graduates is approximately two-thirds that of their counterparts with less than high-school education. This leads to the conclusion that there is substantial heterogeneity in the replacement rate of individuals across gender and educational attainment groups.

3.3 Mortality Risk and Life Expectancy

Literature results have empirically shown that there exists a positive correlation between life-expectancy and various measures of socio-economic status. For instance, Waldron [2013, 2007] displays that the mortality rates of men exhibit an inverse relationship with different measures of earnings using Social Security records. According to their estimates, men born in 1941 in the top half of the earnings distribution have a life-expectancy higher by 5.8 than men in the bottom half. In a related study, Bound et al. [2015] arrives at a quantitatively similar conclusion, although their sample includes women as well. Interestingly, they document that the life-expectancy gap between top and bottom earners has increased over time, as low- and high-earners of the cohort of 1912 had only an estimated 1.2 years gap in life-expectancy, compared to the cohort of 1932, where the difference in life-expectancy reached 4.7 years. A more recent study by Bosworth et al. [2016], while supporting the results of the previous studies, considers alternative definitions of socio-economic status, including education, income and earnings, and wealth, and investigates whether they could be used as proxies for each other in the context of studies of mortality and life-expectancy. Their results suggest that predicted mid-career earning and educational attainment are strongly predictive for mortality risk, and therefore can both be used as explanatory variables in regressions aimed at capturing the relationship between socio-economic status and life-expectancy. In the context of the model presented, this allows us to characterize the mortality risk of the agents populating the economy by using their educational attainment. This assumption allows us to simplify our framework, since in this way population dynamics, which are governed by mortality, can be considered as ex-ante assigned at the birth of our agents. In this subsection, we estimate the mortality risks of different groups of individuals, grouped by age, sex and educational attainment, where education represents the proxy for socio-economic status. In our framework, the use of income-based measures or education is equivalent, since agents are homogeneous within the group they are born in, and education represents the major driver in difference in life-time earnings profiles. The computation of the mortality risk has two purposes in our model. First, it allows us to better characterize the expenditure side of the Social Security system, letting the total amount benefits disbursed be a decreasing function of the age of the cohort. Second, it allows us to characterize more precisely the weights of each cohort, and this is relevant if we want to use the type-cohort specific weights as welfare weights.

3.3.1 Data

In order to compute the mortality rates, we use the Individual file of the PSID. The Individual file collects, for each individual in the sample, the year at which death has occurred (for deceased individuals). In the Individual file of the PSID, there are 6832 individuals that have died between 1968 and 2017 and have exact year of death recorded, 73544 that have a recorded year of death equal to 0 (corresponding to being currently alive), and 290 individuals who are deceased is not recorded precisely, but belongs to an interval. We remove from the sample individuals for which we do not the exact age at death recorded. Another important variable that we construct is age of the individual. We define as age as the age at death for individuals who have died, and for individuals who have not deceased, their age in 2017. To compute the variable age, we need to infer what the year of birth is. The Individual File of the PSID starts reporting the year of birth starting from 1983. We are able to infer, directly from this part of the data, the year of birth of 68814 individuals in the sample. In total, we can compute the age of death of 4961 individuals who are deceased directly from the year of birth and year of death. Nevertheless, we can infer the age of death of some of the individuals for which the PSID reports the year of death but not the year of birth by inferring the year of birth from the recorded age. For each of such individuals, we compute the year of birth as the average of the difference between their age and the year for which the individual has a recorded age. We take the average as the age variables recorded in the Personal File represents the age at the moment of the survey, and the survey can be administered in different parts of the year. In this way, we are able to extrapolate the year of birth for all but 181 individuals. In particular, our sample contains the age at death of 6756 individuals.

3.3.2 Estimation

In light of these considerations, we estimate the conditional survival probabilities as a function of age and educational attainment and sex. From an estimation perspective, we use as the Cox proportional hazard model, which represent the *risk* of dying at time t conditional on the individual-specific explanatory variables $\mathbf{x}_i$:

$$\lambda(t|\mathbf{x}_i) = \lim_{dt \to 0} \frac{P(t \leq D_i \leq t + dt | D_i \geq 0 | \mathbf{x}_i)}{dt} = \lambda_0(t) \exp\left(\boldsymbol{\beta}_d' \mathbf{x}_i\right) \tag{3.5}$$

where $\mathbf{x}_i = (sex_i, educ_i)$ represents the set of time-invariant demographic variables, and D_i represents the event of death for individual i. The variable $educ_i$ is categorical, and represent four different educational attainments: less than high-school (less than 12 years of education), high-school (12 years of education),

some college (13 to 15 years of education), and college and above (16+ years of education).

Table 3.4: Cox Proportional Hazard Rates, Estimated Coefficients

	Dependent Variable	
	Age Death	
	(1)	(2)
HS	−0.127***	−0.076*
	(0.029)	(0.041)
Some College	−0.316***	−0.228***
	(0.042)	(0.056)
College+	−0.740***	−0.777***
	(0.048)	(0.062)
Female	−0.501***	−0.459***
	(0.025)	(0.035)
Female x HS		−0.101*
		(0.058)
Female x Some College		−0.196**
		(0.085)
Female x College+		0.111
		(0.097)
Note:	*p<0.1; **p<0.05; ***p<0.01	

In Table 3.4, we show the estimates of the coefficients presented in Equation (3.5): in the first specification, we include the dummies for educational attainment and sex; in the second one, we also incorporate an interaction term between sex and educational attainment. As we can see from the estimated coefficients in column (1) and (2) of Table 3.4, educational attainment has a negative impact on the hazard rates, and women have uniformly lower hazard rates than men, as the negative coefficient on Female shows. From column (2) we can observe that overall, the interaction between educational attainment and sex do not appear to be uniformly significant at the 5% level across educational levels. As we can see, the interaction between Female × HS is negative but significant at a 10% level, while the coefficient Female × College is not significant at any standard level. We can conclude that in general, hazard rates seem to depend on sex and the selected proxy (educational attainment) for life-time socio-economic status.

Given Equation 3.5, we can compute the conditional survival probability at age n as:

$$\hat{S}(n|\mathbf{x}_i) = \hat{P}(D_i \geq n|\mathbf{x}_i) = \exp\left(-\int_0^n \hat{\lambda}(u|\mathbf{x}_i)du\right) = \exp\left(-\exp(\hat{\boldsymbol{\beta}}_d'\mathbf{x}_i)\int_0^n \hat{\lambda}_0(u)du\right) \tag{3.6}$$

We use Equation (3.6) to estimate the survival probabilities of each group. Given that we do not find the interaction terms between sex and educational attainment to be significant, we use the coefficients estimated from the more parsimonious specification presented in column (1) of Table (3.6). Figure 3.1 shows the estimated conditional survival rates based on educational group and sex. As we can see from Figure 3.1, men have uniformly lower conditional survival probabilities than women, and conditional on sex, they are monotonic in educational attainment. We can now compute the average life-expectancy of the different

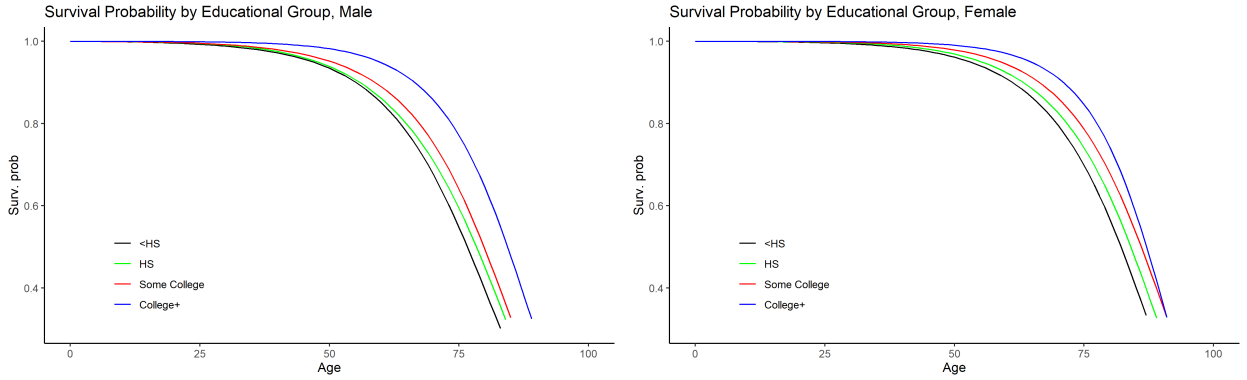

Figure 3.1: Conditional survival probability. Source: Panel Survey of Income Dynamics. Period: 1968-2018.

demographic groups. Given Equation (3.6), we can compute the conditional life-expectancy as follows:

$$\mathbb{E}[D|\mathbf{x}_i] = \int_0^{+\infty} \hat{S}(n|\mathbf{x}_i)dn \tag{3.7}$$

Table 3.5 displays the computed average life-expectancy for each demographic group based on Equation (3.7) and the estimates presented in Table 3.4. Considering that there is not closed-form expression, I numerically integrate using the treapezoid method. Results show that conditional on education attainment, women have higher life-expectancy than men. The smallest gender gap is among the college educated group (2.9 years), while the biggest one is recorded among the group with some college education (6.5 years). Similarly, life-expectancy is strictly monotonic in the educational attainment conditional on sex. Our estimates point to an approximate difference life expectancy of around 7.5 years between high-school dropouts and college educated men. The biggest gap for men is recorded between the group with some college education and college graduates, which is estimated to be of 4.9 years, compared to 1.1 years difference between high-school dropouts and high-school graduates, and 1.5 difference between high-school graduates and the ones

with some college education. For what concerns women, the estimated difference in life-expectancy is 5.2 years between the most- and least-educated groups. The gap is 1.6 years between high-school dropouts and high-school graduates, 2.3 years between high-school graduates and women with some college education, and 1.4 years between women with some college education and college graduates.

		<HS	HS	Some College	College+
Men	Life Expectancy	74.3	75.4	76.9	81.8
	# Years Benefits (Ret. age 66)	8.3	9.5	10.9	15.8
	% rel. to Men HS Dropouts	100%	114%	131%	190%
Women	Life Expectancy	79.5	81.1	83.4	84.7
	# Years Benefits (Ret. age 66)	13.3	16.1	17.4	18.7
	% rel. to Men HS Dropouts	160%	194%	210%	225%

Table 3.5: In the first row average Life Expectancy, calculated using Cox proportional hazard rates, conditional on educational attainment and sex. In the second row, the expected number of retirement benefits received if they were to retired at the current FRA of 66. In the third row, the number of years in benefits relative to male high-school dropouts.

Table 3.5 displays the number of years each demographic group is expected to receive retirement benefits, conditional on retiring at the current FRA of 66. A college-educated man receives benefits for 7.5 more years than a high-school dropout, since the high-school drop-out would collect benefits for an average of 8.3 years, while college-graduate for 15.8 years. This implies that a college graduate men would collect retirement benefits for period that is nearly twice as long as the one of a high-school dropouts. The difference between high-school graduates and dropouts is only 14%. Similar patterns are observed for women. Women in the high-school dropouts category expect to receive benefits for 13.3 years, a 60% increase to their male counterparts, while for women with college education the average collection would last for 18.7 years.

Overall, these computations show differences in life-expectancy are non-negligible across different demographic groups, and they significantly interact with Social Security in the number of expected years of retirement. In particular, while the computation of the benefits is progressive, differences in life-expectancy seem to clearly favor demographic groups in higher socio-economic status. In Section 3.4, we investigate whether differentials in retirement age accentuate or mitigate this factor.

Finally, we use the estimated hazard rates and survival probabilities to characterize the population

dynamics. We define mortality rate at age n as:

$$\hat{\mu}(n|\mathbf{x}_i) = \hat{P}\left(D_i < n+1|\mathbf{x}_i, D_i \geq n\right) = \frac{\hat{P}\left(n \leq D_i < n+1|\mathbf{x}_i\right)}{\hat{P}(D_i \geq n|\mathbf{x}_i)} = \frac{\hat{S}(n|\mathbf{x}_i) - \hat{S}(n+1|\mathbf{x}_i)}{\hat{S}(n|\mathbf{x}_i)} \tag{3.8}$$

where the survival probabilities are computed using Equation (3.6). Equation (3.8) rules the population dynamics described in Section 1.4 of Chapter 1.

To fully characterize the population dynamics, we need now to compute the population share of the initial cohort, namely P_{i1}. In order to do so, I use the individual files from the PSID, specifically from 1972 to 2017. For each year of the survey, I compute the population shares by educational attainment conditional on sex and on age cohort. I then average the yearly shares for each education level to estimate the composition of the initial cohort. Table 3.6 reports the computed separately for two separate waves of surveys: 1972-1997 and 1999-2017, for the cohort group aged 22-30. By doing so, we want to investigate whether there have been significant changes over time in the composition of the initial cohort. As we can

Educational attainment	Men		Women	
	1999-2017	1972-1997	1999-2017	1972-1997
High-School Dropouts	8.16%	14.40%	7.67%	15.70%
High-School Graduates	34.88%	40.01%	29.88%	41.90%
Some College	29.34%	24.13%	29.10%	23.51%
College+	27.06%	21.39%	33.31%	18.89%

Table 3.6: Population shares by educational attainment conditional on sex for the cohorts aged 22-30.

from the two row, the share of high-school dropouts and graduates has significantly decreased for both men and women. High-school dropouts constituted 14.4% (15.70%) of the young men (young women) in the sample in the waves between 1972-1997, while the amounted to only 8.2% (7.7%) of the sample in the 1999-2017 waves. A similar pattern can be observed among high-school graduates: the share decreased by 5.1% for men, and nearly 12% for women. From the third and the fourth row, we can see that share of individuals with at least some college educations has increased over time: for men, then gains amount to approximately 11% while for women are nearly double, 20.0%. These results suggest the following: the composition of the population has changed over time, and on average, has achieved higher educational attainments. These results are largely in line with literature findings (see for instance Krusell et al. [2000]). At the same time, they indicate that while it is not clear whether the educational composition has reached the steady state, for the purposes of our analysis it better to consider only the most recent waves of the PSID survey. For this

reason, in the counterfactual analysis I use the population shares obtained from the 1999-2017 waves.

3.4 Retirement

The effect of differential mortality rates on the redistributional effects of Social Security can be partially mitigated by the negative correlation between retirement age and to measures of economic status. For instance, Bosworth et al. [2016] estimate that a twelve year gap in life expectancy between top and bottom earners translates into only a difference of eight years in the length of the period during which retirement benefits are collected. At the same time, differentials in retirement age affects the actual computation of the benefits received, as they depend on the actual age

The applied microeconomics literature has widely studied the determinants of retirement decisions, and how such decisions affect labor supply. In particular, studies suggest that changes in institutional features characterizing Social Security are likely to affect labor supply of older workers on the extensive margin. For instance, Gustman and Steinmeier [2011] estimate that a combination of the incremental increase in the Full Retirement Age and the elimination of the earning test after Full Retirement Age helps explain the increase participation in the labor market of workers aged recorded in the 1990s and early 2000s. Similarly, French [2005] estimates that permanently reducing retirement benefits by 20% would lead to a postponement of retirement by three months on average. French and Jones [2011] investigate how employer related insurance, Medicare and Social Security jointly impact retirement. They estimate that raising Medicare eligibility by two years (from 65 to 67), would increase the labor supply by 0.074 years of workers aged between 60 and 69. They estimate a similar response in magnitude if the total amount of benefits were reduced by two years. Gustman and Steinmeier [2002] builds a structural model that helps explain how an individual's retirement decision is affected by their spouse's retirement decision; their analysis show that couples with double careers tend to coordinate, with women more likely to anticipate retirement if their spouse decides to retire.

As described in the previous paragraph, results in the literature suggest that labor supply is likely to increase as a response to a cut in benefits, a scenario that we consider in our counterfactual analysis. The question now becomes whether the impact is likely to be meaningful from a macroeconomic perspective, i.e. if alternative counterfactual scenarios would trigger a change in labor supply that affect wages and interest rates. From the evidence presented above, it appears that the overall macroeconomic effects are small for

two reasons. First, the supply of labor only changes for workers in the later stages of their careers as a response to changes in the institutional characteristics of Social Security. Second, the changes are small in absolute terms, in the order of months at most. Therefore, in this subsection, I estimate the retirement age through a reduced form model that does not incorporate as variable of interest institutional characteristic of the Social Security system.

3.4.1 Data

In order to estimate the expected retirement ages across different demographic groups, we use the Health and Retirement Study (HRS), a biennial survey that spans the period between 1992 and 2016. In this particular application, we define as retirement age as the age at which a person first starts to collect Social Security retirement benefits. It is important to highlight that, in the U.S. a person can continue to work while collecting retirement benefits: depending on the age and the earnings, a penalization can be applied by Social Security and retirement benefits may be decreased as a consequence of a continued participation in the labor market. To be more specific, the earning test acts as a temporary disincentive for people who want to continue working after they decide to claim their Social Security benefits before full retirement age (FRA). The penalization depends on the age at which the person decides to claim the benefits while working: after FRA is reached, no penalization is applied. Before then, Social Security withholds 1$ in benefits for every 2$ of earnings in excess of the lower exempt amount (17,640$ in 2019), and 1$ in benefits for every 3$ of earnings in excess of the higher exempt amount (46,920$ in 2019).

In light of these factors, we can make two observations. First, the collection of Social Security does not necessarily coincide with the exit from the labor market. Second, Social Security benefits can change if someone apply for them before FRA but continues working: in general, benefits will increase once someone reaches FRA to account for the withholding of benefits determined by the earning test. These are important considerations to take into account when define what retirement means in this context, and how it is associated to labor market participation and Social Security benefits.

As documented in the literature, a non-negligible share of workers rejoins the labor market after considering themselves as retired. This leads to the possibility of adopting multiple definitions for retirement. In general, retirement definitions can be based on the employment status and the degree to which a person participates to the labor market (based on actual labor market supply, or on measure of earnings); alternatively, can be based on the respondent subjective perception, or on when the respondent has actu-

ally started receiving Social Security benefits from the OASI. Considering that this study focuses on Social Security, we use as retirement age the age at which HRS respondents started to receive Social Security.

Occupation Classification	Group	<HS	HS	Some College	College+
Never work	All	0.635	0.585	0.542	0.479
	Men	0.573	0.527	0.487	0.430
	Women	0.702	0.631	0.593	0.552
Work only PT	All	0.202	0.248	0.272	0.300
	Men	0.223	0.251	0.283	0.300
	Women	0.180	0.246	0.263	0.301
Work FT	All	0.162	0.167	0.185	0.221
	Men	0.204	0.222	0.230	0.270
	Women	0.118	0.123	0.144	0.147
≤1 period of empl.	All	0.772	0.727	0.711	0.636
	Men	0.727	0.671	0.668	0.582
	Women	0.822	0.772	0.750	0.717
>1 period of empl.	All	0.228	0.273	0.289	0.364
	Men	0.273	0.329	0.332	0.418
	Women	0.178	0.228	0.250	0.283
N. obs.	All	2733	4455	2717	2541
	Men	1420	1989	1296	1530
	Women	1313	2466	1421	1011

Table 3.7: Source: Health and Retirement Study. Period: 1992-2016. Share of individuals who never work again, only work part-time, work part-time or full-time by gender or educational attainment.

Table 3.7 shows some descriptive statistics about employment post collection of Social Security for different demographic groups, divided by educational attainment. The employment definition are based on the self-reported status. In particular, the HRS asks to each respondent whether she/he is in the labor force, and if in the labor force, whether she/he worked part-time or full-time. For each respondent, we compute the number of positive answers for each of the questions. The first row shows the share of individuals in the sample that never work after starting collecting Social Security benefits; the second one displays the share of individuals in the sample that only work part time. Finally, one the third row, we present the share

of individuals who continue on working either part time or full time.

As we can see from the first row, for each level of educational attainment, the relative majority of people does participate in the labor force either in full-time or part-time capacity after starting collecting Social Security benefits. The share is decreasing in the educational attainment, with 63.5% of high-school dropouts permanently exiting the labor force after starting collecting Social Security benefits vs. 47.9% of college graduates. If we restrict our attention to full-time employment, only between 16.2% to 20.2% of Social Security recipients either continue to work full-time or go back to full-time employment. In this case, the share of individuals working full-time is decreasing in educational attainment. Similar patterns can be observed for part-time employment: 20.2% of high school dropouts hold a part-time position after starting collecting Social Security benefits, compared to 30.0% of college graduates. This suggests that while retirement as defined by starting receiving Social Security benefits does not determine a permanent exit from the labor market, 77.9% to 83.7% of individuals in the sample neither continue to work full time nor go back to a full-time employment position. It is important to notice the existence of a gap between men and women: on average, women are less likely to be in the labor force after they start collecting Social Security benefits. The gap appear significant, and it is in the order of 10% if we consider the difference in share of women and men who never work again after starting to collect retirement benefits. Minor differences are observed in part-time employment. This suggests that women tend to be less attached to the labor force once they file for Social Security benefits.

The second part of the Table (row 4 and 5) displays the average tenure for different types of employment (part-time or full-time). As we can see from row 4, the majority of workers hold an employment position for at most one year after starting collecting Social Security benefits. The ratio is decreasing in the educational attainment, with 77.2% of high-school dropouts working for at most one year, 72.7% of high-school graduates, 71.7% of individuals with some college education, and 63.6% of college graduates. Similar patterns are observed for mean and women: men on average work a longer period of time compared to women: 27.3% work for more than one period compared to only 22.8% of women. Altogether, the evidence suggest we can proxy the starting collecting Social Security retirement benefits with exit from the labor market and that the exit from the labor market is permanent. This result is in agreement with literature results (see for instance Gustman and Steinmeier [2002],Gustman and Steinmeier [2004]) that show a coordination effect in couples in the joint decision of retirement.

Given the we have established a definition for retirement, i.e. the start of collection of Social

Security benefits, we can now describe the steps taken to compute average retirement age. From the HRS, we can infer the age at which a survey respondent started receiving Social Security benefits. The issue comes from the fact that the survey does not specify whether the first received benefit is related to the disability insurance or the old-age and survivor insurance. A person can start collecting Social Security retirement benefits related to his own working career at the age of 62; therefore, we exclude every from the sample every individual who reports an earlier collection age. This selection rule can be problematic, especially if for instance there is a correlation between the occurrence of disability, early retirement and education attainment or sex. Suppose for instance, that low-educated male are more likely to be employed in jobs that are physically demanding, and this increases their chances to apply for disability benefits before the age of 62; by excluding these individuals from the sample, we would likely obtain an upper biased estimate of the average retirement age of male high-school drop-outs, as we would include in the sample only healthy ones. While this is a reasonable concern, the limitation of the data does not allow us to provide a better solution.

Secondly, we exclude all individuals who have never reported a Social Security age, and are older that 70 in the latest survey in which they participate. Considering that Social Security retirement benefits do not accrue after the age of 70, it is hard to envision why someone would not collect benefits starting the age of 70. In addition, we focus on individuals who have shown a relatively strong attachment to the labor force throughout their life. Therefore, we exclude from the sample everyone who has not reported at least 20 years of professional experience. We experiment with different threshold, but we find that our results do not significantly change based on the different thresholds. In addition, we exclude all individual who have not reported a starting age of collection Social Security benefits, but who Social Security retirement income has been positive at least in one wave. This leaves us with a sample of 12,808 observations, with 10,488 of them classified has having retired at a certain point in time.

3.4.2 Estimation

Using the Health and Retirement Study, we estimate the average retirement age for different educational groups, separately for men and women. By doing so, we achieve two goals: (1) on the one hand, we assess whether differences in life-expectancy are directly translated into differences in the expected of years of benefits received; and (2) on the other hand, we use these estimates to compute the share of Principal Insurance Amount that each demographic group would receive if it were to retire at the estimated age. Our estimates are based on Cox proportional hazard rates defined in Equation (3.5), where retirement is considered to be

the event/terminal status. As explanatory variables, we use dummies for educational attainment and gender as in the survival analysis carried out in the previous subsection.

Table 3.8: Cox Proportional Hazard Rates, Retirement Age

	Dependent Variable	
	Age First SS Retirement Benefits	
	(1)	(2)
HS	−0.033	−0.019
	(0.036)	(0.028)
Some College	−0.192***	−0.245***
	(0.040)	(0.030)
College+	−0.449***	−0.480***
	(0.038)	(0.031)
Female	−0.065	−0.103***
	(0.047)	(0.020)
Female x HS	0.020	
	(0.057)	
Female x Some College	−0.116*	
	(0.062)	
Female x College+	−0.082	
	(0.063)	
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 3.8 shows the estimates for the coefficients of the relevant demographic variables. We run two different specifications: in column 1, we specify a richer model in which the interaction term between the gender dummy and the four educational attainments is introduced; in column 2, we consider a parsimonious version of the model, where no interaction terms are considered. As we can see from specifications, the hazard rate is decreasing in the educational attainment. The coefficients for high-school graduates (HS), female high-school dropouts (Female) and for female high-school dropouts (Female × HS) are not statistically significantly different from 0 at any standard level; this implies that, given that male high-school dropouts represent reference group, that the hazard rate of men and women with at most high-school education is very close. At the same time, from column (1) we can see that the coefficient for men

with some college education and with at least a bachelor degree are negative statistically significant at the 1% level. In particular, we can notice that as we increase the educational level, the size of the relative coefficient becomes more negative. Given the relationship between hazard rate and expected duration of the event, this implies that the average age at which retirement benefits are collected is increasing in educational level. From column (2) of Table 3.8, we can see that the coefficient for the dummy women in negative. Therefore, we can infer that on average women start collecting benefits at older ages. In addition, as column (1) shows, the interaction terms between gender and educational attainment do not appear to be significant at 1% level, suggesting that there is no strong interaction between educational level and gender. Therefore, in estimating the average retirement age, use the coefficients column (2) of Table 3.8.

		<HS	HS	Some College	College+
Men	Avg. Retirement Age	63.7	63.8	64.1	64.7
	Share of PIA	84.7%	85.3%	87.3%	91.3%
Women	Avg. Retirement Age	63.9	63.9	64.5	65.0
	Share of PIA	86.0%	86.0%	90.0%	93.3%

Table 3.9: Average Retirement Age, by Sex and Educational Attainment, implied by Cox proportional hazards model.

Table 3.9 displays the estimated ages at which different groups start collecting Social Security benefits. While clearly there is a positive relationship between age at first collection of Social Security benefits and educational the difference between the least and the most educated groups is approximately one year, for both men and women. Therefore, while college educated individuals are expected to receive retirement benefits for an average of a year less compared to college dropouts based on their retirement age, the difference in number of years of collections is mostly driven by a differential in life-expectation across different groups. As shown in Table 3.5, the average life-expectancy gap between male high-school dropouts and college graduates is 7.5 years, which translates into a 6.5 years different in the average number of retirement years. Our estimate therefore suggest that differentials in retirement age mitigate little the gap in years of retirement benefits stemming from difference in life-expectancy. Our results differ from Bosworth et al. [2016], as their estimate point to a larger retirement gaps among individuals belonging to different socio-economic groups. However, they use life-time earnings to capture socio-economic status, while here we assume it is represented by educational attainment.

Based on the results presented in Table 3.9, we can compute the penalization that would derive

from retiring before FRA. Assuming a full retirement age of 66 years (which is currently applied to the cohorts of 1943-1954), men with less that college degree would receive a Social Security benefit that is approximately 85% of the PIA based on our estimates. For college graduates, with a average retirement age of 12 months higher than non-college graduate, this would translate with a only 9% deduction of the PIA. This implies that the retirement age affects in a non-negligible way the amount of retirement benefits received on a yearly basis. Given the estimates presented in Table 3.9 and in Table 3.3, we can compute the effective replacement rate for the eight identified groups taking into account the penalization that is applied when retiring before full retirement age is reached. Results are displayed in Table 3.10. As we can see,

	<HS	HS	Some College	College+
Men	42.35%	35.83%	34.92%	30.13%
Women	57.88%	47.30%	45.45%	41.71%

Table 3.10: Social Security benefits replacement rate, conditional on educational attainment and sex. Replacement rate is defined as the ratio between the estimated retirement benefits at retirement and the index-adjusted average life-time income. Estimates computed taking into account group-specific retirement ages.

the estimated replacement rate is decreasing in educational attainment, and women have a uniformly higher replacement rate compared to men. Conditional on sex, the gap between the highest and lowest educated groups is significant. For men, the replacement rate of college graduates is 30.13%, while for high-school dropouts is 42.35%, a 29.9% difference in relative terms. High-school graduates men and men with some college experience display very similar replacement rates (35.83% and 34.92% respectively). Similarly, the replacement rate for college educated women is significantly lower than high-school dropouts (respectively 57.88% and 41.71%), corresponding to a relative difference of 27.8%. These computations indeed show how Social Security is redistributive in the way it computes monthly retirement benefits (i.e. abstracting from any consideration about differential life expectancy).

3.5 Structural Estimation of the Coefficient of Relative Risk Aversion and the Discount Factor

In order to understand the welfare implications of potential changes in Social Security, it is important to uncover households preferences for time and risk. The identification and the estimation of preference pa-

rameters has been at the center of the structural estimation literature for more than 30 years. In particular, the use of Euler equations, i.e. the first-order conditions from the dynamic optimization problem of the household, has been widely exploited as moment conditions in the context of linear and non-linear GMM methods.

The imposition of moment conditions allows econometricians to not fully specify the market environment in which households live and the distributions of the underlying stochastic processes. The estimation relies on the parametrization of the utility function that is usually modeled with a iso-elastic form:

$$u\left(c_t^h, \mathbf{x}_t^h\right) = \exp\left(\boldsymbol{\theta}' \mathbf{x}_t^h\right) \frac{\left(c_t^h\right)^{1-\sigma}}{1-\sigma} \tag{3.9}$$

where σ represents the coefficient of relative risk aversion and its inverse $\frac{1}{\sigma}$ denotes the elasticity of intertemporal substitution. As described in Chapter 1, households maximize expected utility, which is intertemporally additive, and discounted at a geometric discount rate β. Under the assumption of rational expectations, agents optimization of the problem presented in Equation (3.9) leads to the following equilibrium conditions:

$$\mathbf{E}_t\left[\beta\left(\frac{c_{t+1}^h}{c_t^h}\right)^{-\sigma} \exp(\boldsymbol{\theta}' \Delta \mathbf{x}_{t,t+1}^h) R_{t+1}\right] = 1 \tag{3.10}$$

where R_{t+1} is the real return of a generic asset that is available for trade to the household. By deriving Equation (3.10) as an optimality condition, we are implicitly assuming that households do not face a liquidity constraint at any period. Although this is a standard assumption in the literature, its validity depends on the sample of households that is considered. As stated by Alan et al. [2018], traditional data sources do not allow to satisfactorily deal with the liquidity constraint, and therefore, the assumption is motivated by empirical feasibility.

We can define η_{t+1} as the expectation error (or the innovation in the discounted value of the marginal utility):

$$\eta_{t+1}^h = \beta\left(\frac{c_{t+1}^h}{c_t^h}\right)^{-\sigma} \exp(\boldsymbol{\theta}' \Delta \mathbf{x}_{t,t+1}^h) R_{t+1} - 1 \tag{3.11}$$

where clearly $\mathbf{E}_t\left[\eta_{t+1}\right] = 0$ by construction. Economic theory implies that the expectation error η_{t+1}^h is orthogonal to any variable that belongs to the information set of household. In general, GMM estimations rely on the orthogonality of the expectation error η_{t+1} expressed in Equation (3.11) with respect to a set of variables that belong to the information set of the households. Therefore, in order to implement the estimation through a GMM approach, we need to find a vector of instruments that are included in the in

the information set of the household. If we assume that the demographic variables are orthogonal to the forecast error, then it would sufficient to use only one instrument to identify the structural parameters (σ, β) in addition to the constant.

3.5.1 Data

As the main source of household-level data, we use the Panel Survey of Income Dynamics (PSID). This survey contains information about food expenditures, and it has been widely exploited in the literature for the purposes of estimating the discount factor and the coefficient of relative risk aversion. As our baseline, we use the food consumption data available from 1974 to 1987. We limit ourselves to this specific window of time since food data are hard to interpret before 1974, and expenditure related questions where suspended until 1999. In addition, starting from 1999, the survey has been conducted only on a biannual basis, and the definition of some of the relevant expenditure categories has changed. For this reason, I decide to only consider the earlier data available. As a measure of food consumption, we use total value of food consumed at home, food away from home, and food stamps. The data is collected on a yearly basis, and the individual level considered is the household. For the purposes of the empirical analysis, I select households that satisfy the following criteria:

- appear in the sample for at least 5 periods;
- whose head is married and does not change over time.

It is assumed that all households face the same interest rate, that is calculated as the yearly average US 3-month T-bill rate deflated by the Consumer Price Index. The sample contains a total of 3382 households, for a total of 31969 observations. We identify the educational level of a household as the educational attainment of its head. The sample contains 1018 high-school dropouts, 1419 high-school graduates, 697 with some college degree and 677 with college education.

Non-Linear GMM Let η_{t+1}^h be the forecast error represented in Equation (3.11) and let $\mathbf{z}_{it}^h$ a vector of instruments belonging to household h information set. Then, the orthogonality condition implied by the Euler equation implies:

$$\mathbb{E}_t \left[\eta_{t+1}^h \mathbf{z}_t^h \right] = 0 \tag{3.12}$$

We estimate a non-linear GMM to recover the structural parameters of interest. In terms of demographic variables, we assume that the vector $\mathbf{x}_t^h$ contains only the number of individuals in the household, as assumed by Alan et al. [2009]. We assume that the vector of instruments contains the constant, the first and second lag of the real risk free interest rate, and under the assumption that household are able to perfectly forecast the future number of family members, the change in the actual number of household members. Under these assumptions, we define the vector of instruments as $\mathbf{z}_t^h = \left(1, R_t, R_{t-1}, \Delta\text{n. people}^h\right)$, so that we can set the empirical moment conditions as follows:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{h\in\mathcal{H}}\mathbf{z}_t\eta_{t+1}^h\left(\sigma,\beta,\theta\right)$$

where $\mathcal{H}$ denotes the set of households in the sample.

As discussed by Alan et al. [2009], problems arise if we assume that consumption is measured with a multiplicative error that takes the following form:

$$c_t^h = c_t^{0h}\mu_t^h \tag{3.13}$$

where c_t^{0h} represents the *true* consumption of the household. In particular, the presence of measurement error would lead to a biased estimate in the discount factor. For this reason, we follow Alan et al. [2009] and we account for the possibility that the consumption expenditure observed by the econometrician is affected by a household specific measurement error. As an identification assumption, we impose that the measurement error is independent from any other variable, either observed or unobserved by the econometrician. In this case, the household's Euler equation would take the following form:

$$\eta_{t+1}^h = \beta\left(\frac{c_{t+1}^h}{c_t^h}\right)^{-\sigma}\left(\frac{\mu_t^h}{\mu_{t+1}^h}\right)^{-\sigma}\exp(\boldsymbol{\theta}'\Delta\mathbf{x}_{t,t+1}^h)R_{t+1} - 1 \tag{3.14}$$

where $\mathbf{E}_t\left[\eta_{t+1}^h\right] = 0$, where the measurement shocks (μ_t^h, μ_t^h) are not observed. In this case, even if η_{t+1}^h and (μ_t^h, μ_{t+1}^h) are independent (as postulated), the moment condition used for the non-linear GMM estimation would depend on the higher-order moments of the measurement error. Following Alan et al. [2009], we assume that i) μ_t is stationary and independent from any other variable and ii) μ_t is log-normally distributed $\mu_t^h \sim \mathcal{N}(\mu,\gamma)$. This allows to the derive the following analytical solution for the expectations of the error terms:

$$\mathbf{E}_t\left[\left(\frac{\mu_{t+1}^h}{\mu_t^h}\right)^{-\sigma}\right] = \mathbf{E}\left[\left(\frac{\mu_{t+1}^h}{\mu_t^h}\right)^{-\sigma}\right] = \frac{\mathbf{E}\left[(\mu_{t+1}^h)^{-\sigma}\right]}{\mathbf{E}\left[(\mu_t^h)^{-\sigma}\right]} = e^{\sigma^2\gamma} \tag{3.15}$$

where we replace the conditional expectation with the unconditional one given, and we obtain a closed-form solution for expected value given the independence assumption and the log-normality of μ_t^h. In a similar

fashion, we can obtain the following analytical formulation between measurement errors at time t and $t+2$:

$$\mathbf{E}_t\left[\left(\frac{\mu^h_{t+2}}{\mu^h_t}\right)^{-\sigma}\right] = e^{\sigma^2\gamma} \tag{3.16}$$

By using Equations (3.15) and (3.16), it is straightforward to derive the following moment conditions:

$$\mathbf{E}_t\left[\beta\left(\frac{c^h_{t+1}}{c^h_t}\right)^{-\sigma} R_{t+1} - e^{\sigma^2\gamma}\right] = 0 \tag{3.17}$$

$$\mathbf{E}_t\left[\beta^2\left(\frac{c^h_{t+2}}{c^h_t}\right)^{-\sigma} R_{t+1}R_{t+2} - e^{\sigma^2\gamma}\right] = 0 \tag{3.18}$$

In terms of instruments, for the first moment condition (3.17), I use the constant, and the values of the real interest rate lagged one time and two times respectively. I define $\mathbf{z}^h_{1t} = \left(1, R_t, R_{t-1}, \Delta\text{N. people}_{t,t+1}\right)$. For the second moment condition (3.18), I choose as instruments $\mathbf{z}^h_{2t} = \left(1, R_t, \Delta\text{N. people}_{t,t+2}\right)$. In total, we estimate 4 structural parameters with 7 moment conditions. In Table 3.11 I report the estimated coefficients

	σ	β	θ	γ
GMM-Measurement Error	1.8018 (0.7100)	0.8414 (0.1049)	0.1069 (0.0398)	0.0030 (0.0062)
GMM	1.3943 (0.4749)	0.9793 (0.0700)	0.4822 (0.27675)	–

Table 3.11: Estimates of the coefficient of relative risk-aversion σ, the discount factor β obtained by solving for the exact non-linear GMM and the non-linear GMM with approximation error. In parentheses, standard errors obtain through bootstrapping.

using non-linear GMMs, when consumption expenditure is measured with and without error. In the first row, I report the estimates for the specification with no measurement error, while the second row shows the parameters for the specification that incorporate the multiplicative error. The reported standard errors are obtain through a bootstrapping procedure. I resample with replacement the same number of households from the original sample, creating an artificial sample from which I then estimate the two different specifications of the non-linear GMM. Standard errors are then obtained as follows:

$$\text{St. }\hat{\text{Dev}}\left(\hat{\sigma}\right) = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\hat{\sigma}_b - \overline{\sigma}^B\right)^2}, \qquad \overline{\sigma}^B = \frac{1}{B}\sum_{b=1}^{B}\hat{\sigma}_b$$

$$\text{St. }\hat{\text{Dev}}\left(\hat{\beta}\right) = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\hat{\beta}_b - \overline{\beta}^B\right)^2}, \qquad \overline{\beta}^B = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_b$$

First, the estimation of the parameters proves to be particularly challenging, since the objective function is non-linear in the parameters of interest. In particular, I observe that the objective function is very flat

in large portions of the parameter space, leading to a significant volatility at the convergence of the of the optimization algorithm. In particular, traditional gradient based optimization methods seem to fail, delivering values for the coefficient of relative risk aversion smaller than 1. For this reason, I resort to the use of Bayesian optimization type of techniques, which prove to be particularly useful when the objective function is not necessarily well behaved, or when the objective function is expensive to evaluate. In particular, I use the algorithm developed by Bergstra et al. [2011] to search for the optimal set of parameters. The use of this algorithm alleviates the difficulty encountered by gradient based methods.

As we can see from the results in Table 3.11, the GMM with no measurement error provides a lower point estimate for the coefficient of relative risk aversion (1.39) compared to the GMM with measurement error (1.80). On the contrary, the estimate for the discount factor in the exact GMM (0.97) is higher than for the GMM with measurement error (0.84), the latter looking implausibly low for a yearly discount factor. In both cases the parameter that accounts for family size is positive, but the magnitude differ quite significantly across specifications. In particular, in the exact GMM case the estimate for θ appears to be imprecise, given the large standard deviation that is obtained via bootstrapping. Overall, our results are not particularly satisfactory, since our moment conditions seem to be able to only weakly identify the parameters of interest in the context of a non-linear optimization problem.

For this reason, I decide to resort to a linear GMM estimation based on a log-linearized version of the Euler equation presented in Equation (3.11). The choice is motivated by two reasons. First, in the case of linear GMM, we are able to obtain estimates in a closed-form fashion. Secondly, Attanasio and Low [2004] show in a Montecarlo study that the log-linearized Euler equations delivers unbiased estimates for the coefficient of relative risk-aversion under a variety of assumptions regarding the underlying data-generating process, and in general, performs better than a non-linear GMM. The only downside of this estimation routine comes from the fact that in general, it is not possible to separately identify the discount factor.

Linear GMM Another common approach in the literature to estimate households preferences starts from the log-linearization of Equation (3.14):

$$\Delta c^h_{t,t+1} = \frac{1}{\sigma}\log\beta + \frac{1}{\sigma}\log R_{t+1} + \boldsymbol{\theta}'\Delta\mathbf{x}_{t,t+1} + \log(1+\eta_{t+1}) + \sigma\left(\log(\mu^h_{t+1}) - \log(\mu^h_t)\right) \tag{3.19}$$

The unobserved component of Equation (3.19) depends on two terms: the expectation error $\log(1+\eta_{t+1})$ and the measurement error $\log\mu^h_{t+1} - \log\mu^h_t$. In general, it is easy to see that unless we are willing to make

some assumptions about the distribution of the expectation error and the measurement error, then:

$$\mathbf{E}_t\left[\log\left(1+\eta_{t+1}\right)+\sigma\left(\log(\mu^h_{t+1})-\log(\mu^h_t)\right)\right]\neq 0 \tag{3.20}$$

For instance, the assumption of stationarity helps us remove the measurement error from (3.19) even if we did not assume that the measurement error follows a 0 mean process or a particular distribution. At the same time, in (3.11) the expectation error enters in a non-linear way. Given that the McLaurin expansion of $\log(1+\eta_{t+1})$ takes the following form:

$$\log(1+\eta^h_{t+1})=\sum_{n=1}^{+\infty}(-1)^n\frac{(\eta^h_{t+1})^n}{n} \tag{3.21}$$

it is easy to see that in general, the expected value $\log(1+\eta_{t+1})$ will depend on the second- and higher-order moments of η_{t+1}. In particular, let define

$$\nu^h_{t+1}=\eta^h_{t+1}-\sum_{n=2}^{+\infty}(-1)^n\frac{\mathbf{E}_t\left[(\eta^h_{t+1})^n\right]}{n} \tag{3.22}$$

given that $\mathbf{E}_t\left[\eta_{t+1}\right]=0$ by construction. Therefore, we can rewrite Equation (3.19) using (3.22), and obtain the following expression:

$$\Delta c^h_{t,t+1}=\underbrace{\sum_{n=2}^{+\infty}\frac{\mathbf{E}_t\left[(\eta^h_{t+1})^n\right]}{n}+\frac{1}{\sigma}\log\beta}_{\text{constant}}+\frac{1}{\sigma}\log R_{t+1}+\boldsymbol{\theta}'\Delta\mathbf{x}_{t,t+1}+\underbrace{\nu^h_{t+1}+\sigma\left(\log(\mu^h_{t+1})-\log(\mu^h_t)\right)}_{\text{error term}} \tag{3.23}$$

From Equation (3.23) we can see that the model-implied constant in Equation (3.23) captures the higher moments of the conditional expectation error of the household; at the same time, under the assumption of stationarity of the measurement error, the model-implied error term has now mean 0. We can see that if we were to run an IV regression based on this relationship, then the constant would capture the discount factor and the means of higher order moments, even if we assumed that the log measurement error had a 0 conditional mean $\mathbf{E}_t\left[\log(\mu^h_{t+1})\right]=0$. It is worth highlighting that without further assumption, in the case of a log-linearized Euler equation as in (3.23) it is not possible to separately identify the discount factor. The residual term would contain the measurement error, and an error deriving from the first-order approximation and including higher moments of consumption growth and interest rates conditional on past information. For estimation purposes, we can rewrite:

$$\begin{aligned}\Delta c^h_{t,t+1}&=\gamma_0+\gamma_1\log R_{t+1}+\theta\Delta\mathbf{x}_{t,t+1}+\epsilon^h_{t+1}\\ \gamma_0&=\sum_{n=2}^{+\infty}\frac{\mathbf{E}_t\left[(\eta^h_{t+1})^n\right]}{n}+\frac{1}{\sigma}\log\beta\\ \gamma_1&=\frac{1}{\sigma}\\ \epsilon_{t+1}&=\nu^h_{t+1}+\sigma\left(\log(\mu^h_{t+1})-\log(\mu^h_t)\right)\end{aligned} \tag{3.24}$$

As first noted by Chamberlain [1984], and discussed by Altug and Miller [1990] and Attanasio and Low [2004], there is no reason to believe that those cross-section means of forecast errors are zero at any point time, especially in the presence of an aggregate shocks (or in Chamberlain [1984] words, "economy-wide innovations"). Therefore, this would translate to a error ϵ_{t+1}^h in equation (3.24) whose average is not 0 in the cross section. Under the assumption of a stationary stochastic process for the expectation error, the household expectation error should be 0 over time, so a sufficiently long panel should be able to deliver consistent estimates. This leads to the following condition moment condition:

$$\lim_{T\to+\infty} \frac{1}{T}\sum_{t=1}^{T} z_t^h \varepsilon_t \xrightarrow{p} 0 \tag{3.25}$$

for any variable z_t^h that belongs to the information set of the household at time t.

A log-linearized Euler equations as in (3.24) produces unbiased estimates of the parameters $(\gamma_0, \gamma_1, \boldsymbol{\theta})$ through standard IV if the instruments used are uncorrelated with the residuals, which are composed by the measurement error, and the the innovation to the conditional second and higher-moment of consumption growth and interest rates. A difficulty in choosing the right set of instruments arises from the fact that the higher moments of consumption growth are determined as part of the optimization problem of the consumer, and therefore, are endogenous. This leads to the problem of determining analytically for which variable this condition holds true, which cannot bet usually solved. In general, lagged values of consumption growth or interest rates represent preferred choices in the literature: two lags represent a preferred choice in many papers (see Alan et al. [2009]), but first lags are allowed under the generally accepted assumption that measurement errors are uncorrelated with any other variable. However, Attanasio and Low [2004] notes that consumption growth can exhibit very persistent heteroskedasticity, and therefore, the use of lagged instruments would potentially introduce bias in the estimates of the parameters. The question then becomes on what is the size of the bias, in particular, in the EIS, the structural parameter that we can identify without further assumption. Attanasio and Low [2004] stress the importance of having data covering long-horizons to estimate structural parameters from Euler equations. In comparing the performance between linear IV estimates based on Equation (3.24), and non-linear GMM estimates based on Equation (3.17) (no measurement error), they show that linear IV-estimates of the EIS outperform non-linear GMM[2]. While they recognize that the discount factor cannot be separately due to the presence of higher order moment of consumption growth and returns in the constant, they recognize that IV estimates are very close to the underlying parameter used in their Montecarlo study.

[2]They note that their GMM estimates appear to be extremely volatile.

Education level	γ_0	γ_1	σ	θ	N_{obs}
Whole Sample	−0.0133 (0.0021)	0.4352 (0.0575)	2.2978 (0.3036)	0.0647 (0.0041)	31,969
High-School Dropouts	−0.0294 (0.0043)	0.4610 (0.1195)	2.1692 (0.5626)	.0494 (0.0071)	8,530
High-School Graduates	−0.0147 (0.0035)	0.4343 (0.0944)	2.3026 (0.5009)	0.0810 0.0105	11,790
Some College	−0.0021 (0.0054)	0.3438 (0.1427)	2.9086 (1.2073)	0.0811 (0.0150)	5,252
College+	0.0016 (0.0043)	0.3740 (0.1154)	2.6738 (0.8247)	0.0783 (0.0091)	6,397

Table 3.12: GMM estimates from log-linearized Euler equation described in Equation (3.24). In parenthesis, estimates standard errors. Estimated mean and standard errors for the σ coefficient are based on the δ-method, where $\sqrt{N}\left(\frac{1}{\hat{\gamma}_1} - \sigma\right) \rightarrow \mathcal{N}\left(0, \frac{1}{(\hat{\gamma}_1)^4}\hat{\sigma}^2(\hat{\gamma}_1)\right)$.

In light of these considerations, I estimate the coefficient of relative risk aversion σ using the log-linearized equation described in Equation (3.24). In the vector $\mathbf{x}_{t,t+1}$, I only include the change in the size of the household. I assume that households have perfect forecast on their future composition, leading to the orthogonality between the residual v_{t+1} and $\Delta\mathbf{x}_{t,t+1}$. As instruments, the value of the log real risk-free interest rate, with one and two lags. The vector of instruments therefore includes $\mathbf{z}_t^h = \left(1, \Delta\text{N. People}^h, \log(R_t), \log(R_{t-1})\right)$. Table 3.12 reports the estimates from the linear-GMM. In the first row, I report the coefficients obtained using the entire sample. Standard errors for the coefficient of relative risk aversion are obtained using the δ method. As we can see, the estimate for coefficient of relative risk aversion is 2.30, which is higher than both estimates based on non-linear GMM. In row (2-5), I estimate the parameters separately for each educational group. The sample is divided into subsamples based on the educational attainment of the head, and I carry out separately an estimation for each of the subsamples. This experiment is motivated by the fact that I am interested in understanding whether we can measure some heterogeneity in preferences using panel data, in particular, in the coefficient of relative risk aversion. Studies have documented a positive relationship between coefficient of relative risk aversion and educational attainment (see for instance Alan et al. [2018]). As we can see from column (3), the point estimates of the coefficient of relative risk aversion for households whose head is less than high-school educated are lower than the other the households who have at least some college education. However, the standard deviations of the estimates are high, and they do not allow to distinguish the coefficients at any statistically

significant level. This suggests that our results do not seem to support the hypothesis of preference heterogeneity across different educational groups. This result however, does not rule out the possibility that households have heterogeneous preferences. In the next subsection, I assume that households can be divided into different sub-populations, characterized by heterogeneous preferences. While the number of groups is set ex-ante, the actual division across groups is the result of the optimization algorithm. Therefore, instead of characterizing the source of heterogeneity based on some arbitrary observable variable (like education in the example provided above), we let the algorithm sort households across different types based on the observed consumption patterns.

3.5.2 Risk Preferences and Unobserved Heterogeneity

In this subsection, I explore how an application of the Expectation-Maximization algorithm (see Do and Batzoglou [2008]) can help us disentangle heterogeneity in preferences for risk exploiting a log-linearized version of the Euler equations in the context of panel data. The estimation routine we exploit here presents a novel way to approach the task of uncovering unobserved heterogeneity in preferences. In fact, the literature has mostly focused on estimating heterogeneity based on some observable variables, such as age, education level, income, etc. On the contrary, in this application, I do not put any restriction as to which type of group an agents belong to, and only ex post I infer whether there is a relationship between the different unobserved types and some relevant observable variables. This exercise is inspired by the work of Arcidiacono and Miller [2011], who propose an estimator aimed at uncovering unobserved heterogeneity in the structural parameters ruling the agents utility function in the context of dynamic discrete choice models.

The question of preference heterogeneity has been explored in the experimental and the structural applied-micro literature. Findings in both fields of research document substantial heterogeneity in individual preferences for both time (discount factor) and risk (coefficients of risk aversion). The experimental literature elicits preferences through the administration of tasks aim at eliciting individual risk-aversion and discount factors (see Andersen et al. [2008] for experiment-based identification of risk-preferences and time-preferences). Evidence of preference heterogeneity is also found in the empirical literature, but it is common to attribute the sources of heterogeneity to some observable variables. For instance, Blundell et al. [2008] estimates risk-preferences and discount factors by dividing the sample into different sub-samples according to certain individual-level observable variables, like their educational attainment, age cohort etc., similarly to what we have performed in the previous analysis. Similarly, Cagetti [2003] performs a simu-

lated method of moments estimation in which he estimates preferences according to three main educational groups, namely high-school drop outs, high-school graduates, and college graduates. Their estimates point to a monotonic relationship between educational attainment, risk-aversion and discount factor, with more educated individuals being on average more risk-averse and having higher discount factors. Alan et al. [2009] clusters individuals by educational attainment into two groups, and estimate the parameters ruling the underlying distribution of the risk-aversion and discount factor structural parameters. They conclude that individuals with higher educational attainment tend to display on average, higher levels of risk aversion and a higher discount factor. Alan et al. [2018] posits a relationship between individual risk-preferences and the stochastic income process; again, they find substantial heterogeneity across households in time and risks preferences.

3.5.3 Model

In this subsection, I introduce the framework at the base of the estimation routine used to identify unobserved heterogeneity in risk-preferences. The data structure we are focusing on is a panel, where there is a set I of individuals, spanning a period of time denotes by T. The panel does not need to be balanced, i.e. each individual i can have a different number of observation available. For notational convenience, I assume that the panel is balanced, but this assumption is not needed.

Let y_{it} represents the endogenous variable, $\mathbf{x}_{it}$ the k-dimensional vector of exogenous variables, $\mathbf{z}_{it}$ and the l-dimensional vector of the instruments. I assume that there are K unobserved types of agents, and that the type of an agent s_i is unobserved to the econometrician. In addition, it is assumed that the type of the agent is time-invariant. The agents differ in relationship between y_{it} and $\mathbf{x}_{it}$, in the sense if individual i belongs to group k, the following relationship between exogenous variables

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta}_k + \varepsilon^k_{it} \tag{3.26}$$

where ε^k_{it} denotes the stochastic shock that is unobserved by the econometrician, whose distribution depends on the specific group. Equation (3.26) shows that heterogeneity across groups is modeled by different parameters $\boldsymbol{\beta}_k$ and distribution of the unobserved shock ε^k_{it} for each group $k \in K$.

Given the panel structure of the data, we define $(\mathbf{X}_i, \mathbf{y}_i) = (\mathbf{x}_{it}, \mathbf{y}_{it})_{t\epsilon T}$. Let $\boldsymbol{\Theta}$ represent the set of parameters to be estimated, i.e. $\boldsymbol{\Theta} = (\boldsymbol{\beta}_k, \sigma_k)_{k=1}^K$. The expected log-likelihood for individual i takes the

following form:

$$\mathbb{E}_{s_i}\left[\log \mathcal{L}\left(\mathbf{y}_i, s_i|\mathbf{X}_i, \mathbf{\Theta}\right)\right] = \sum_{k\in K} \log \mathcal{L}\left(\mathbf{y}_i|s_i = k, \mathbf{X}_i, \boldsymbol{\theta}_k\right) Pr(s_i = k|\mathbf{X}_i, \boldsymbol{\theta}_k) \tag{3.27}$$

where $P(s_i = k|\mathbf{X}_i, \boldsymbol{\theta}_k)$ represents the probability that agent i belongs to group k, and $\mathcal{L}\left(\mathbf{y}_i|\mathbf{X}_i, z_i = k, \boldsymbol{\theta}_k\right)$ represents the type-conditional likelihood, which under the assumption of serially-uncorrelated error terms, can be written as follows:

$$\log \mathcal{L}\left(\mathbf{y}_i|s_i = k, \mathbf{X}_i, \mathbf{\Theta}\right) = \sum_{t\in T} \log \mathcal{L}\left(y_{it}|s_i = k, \boldsymbol{x}_{it}\right)$$

If in addition we assume that the unobserved shocks are uncorrelated across individuals, we can now write the full log-likelihood of the data:

$$\log \mathcal{L}\left(\mathbf{y}|\mathbf{X}, \theta\right) = \prod_{i\in I} \mathcal{L}\left(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta}\right) \tag{3.28}$$

where $(X, y) = (X_i, y_i)_{i\in I}$. Given the expression for the log-likelihood function in (3.28), it follows that:

$$\begin{aligned}\log \mathcal{L}\left(\mathbf{y}|\mathbf{X}, \mathbf{\Theta}\right) &= \sum_{i\in I} \log\left(\sum_{k\in K} \mathcal{L}\left(\mathbf{y}_i|s_i = k, \mathbf{X}_i\right) Pr(s_i = k|\mathbf{X}_i, \boldsymbol{\theta}_k)\right) \\ &= \sum_{i\in I} \log\left[\sum_{k\in K}\left(\prod_{t\in T} \mathcal{L}\left(y_{it}|s_i = k, \boldsymbol{x}_{it}, \boldsymbol{\theta}_k\right)\right) Pr(s_i = k|\boldsymbol{x}_i, \boldsymbol{\theta}_k)\right]\end{aligned} \tag{3.29}$$

Introducing unobserved heterogeneity involves some computational challenges, as the solution of the objective function will not be in a closed form. In addition, the objective function of the optimization problem, i.e. the log-likelihood will possibly be not globally concave, leading to the possibility of reaching local-optima through gradient-based optimization methods. For this reason, we implement the Expectation-Maximization algorithm in the the context of linear regression. In the following sub-sections, we describe the implementation of the two main steps of the algorithm, respectively, expectation and maximization, in the context of linear regression

First Step: Expectation

In the Expectation step of the algorithm, we simplify the problem of computing the conditional likelihood of the data. In order to do this, we start computing the joint conditional likelihood of the endogenous variable (y, s). For individual i, we can derive the following expression:

$$\mathcal{L}\left(\mathbf{y}_i, s_i|\mathbf{x}_i, \mathbf{\Theta}\right) = \prod_{s_i\in K} \left(Pr(s_i = k|\mathbf{X}_i, \mathbf{\Theta})\mathcal{L}\left(y_i|s_i = k, \mathbf{X}_i, \boldsymbol{\theta}_k\right)\right)^{\mathbf{1}(s_i=k)} \tag{3.30}$$

By taking the logarithm of equation (3.30), we can write the following expression for the log-likelihood:

$$\log \mathcal{L}\left(\mathbf{y}_i, z_i|\mathbf{X}_i, \Theta\right) = \sum_{s_i \in K} \mathbf{1}(s_i = k) \left(\log Pr(s_i = k|\mathbf{X}_i, \theta) + \log \mathcal{L}\left(\mathbf{y}_i|s_i = k, \mathbf{X}_i, \boldsymbol{\theta}_k\right)\right)$$

so that the full conditional joint-likelihood of the data is:

$$\log \mathcal{L}\left(\mathbf{y}, s|\mathbf{X}, \theta\right) = \sum_{i \in I} \sum_{s_i \in K} \mathbf{1}(s_i = k) \left(\log Pr(s_i = k|\mathbf{X}_i, \boldsymbol{\theta}_k) + \log \mathcal{L}\left(y_i|s_i = k, \mathbf{X}_i, \boldsymbol{\theta}_k\right)\right) \tag{3.31}$$

We need to remove the uncertainty around the type s_i, since it unobserved to the econometrician. Therefore, we compute the expectation of z, conditional on all observable (x, y). We define η_i^k as the conditional probability of belonging to group k given the observable variables (X_i, y_i) for individual i:

$$\begin{aligned}
\eta\left(s_i = k|\, y_i, X_i, \boldsymbol{\Theta}\right) &= \mathbb{E}\left[\mathbf{1}\left(s_i = k\right)|\mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\theta}_k\right] \\
&= Pr\left(s_i = k|y_i, X_i, \boldsymbol{\theta}_k\right) \\
&= \frac{\mathcal{L}\left(\mathbf{y}_i, s_i = k|\mathbf{X}_i, \boldsymbol{\theta}\right)}{\mathcal{L}\left(\mathbf{y}_i|\mathbf{X}_i, \boldsymbol{\Theta}\right)} \\
&= \frac{\mathcal{L}\left(\boldsymbol{y}_i|s_i = k, \boldsymbol{X}_i, \boldsymbol{\theta}_k\right) Pr(s_i = k|X_i, \boldsymbol{\theta}_k)}{\sum_{k \in K} \mathcal{L}\left(\boldsymbol{y}_i|s_i = k, \boldsymbol{X}_i, \boldsymbol{\theta}_k\right) Pr(s_i = k|\boldsymbol{X}_i, \boldsymbol{\theta}_k)}
\end{aligned} \tag{3.32}$$

where the derivation follows by simply applying Bayes rule. Finally, taking into account what we have derived in (3.32), we can derive the objective function:

$$\mathbb{E}_{\mathbf{s}}\left[\log \mathcal{L}\left(\mathbf{y}, s\right)|\mathbf{X}, \boldsymbol{\Theta}\right] = \sum_{i \in I} \sum_{s_i \in K} \eta(s_i = k|\mathbf{X}_i, \mathbf{y}_i, \boldsymbol{\Theta}) \left[\log \mathcal{L}\left(\boldsymbol{y}_i|s_i = k, \mathbf{X}_i, \boldsymbol{\theta}_k\right) + \log Pr\left(s_i = k|\mathbf{X}_i, \boldsymbol{\theta}_{\boldsymbol{k}}\right)\right] \tag{3.33}$$

As we can see from equation (3.33), the objective function exhibits a non-trivial dependence on the underlying-parameters to be estimated. Therefore, at the $j + 1$ step of the algorithm:

- we first compute the values of $\eta^{j+1}(s_i = k|y_i, X_i, \boldsymbol{\Theta}^j)$, conditional on the the values of the parameters $\boldsymbol{\Theta}^j$ obtained at step j, for each individual i and unobserved group k;
- we use equation (3.33) to derive the objective function for the maximization step by plugging in $\eta^{j+1}(s_i = k|\mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\Theta}^j)$

In this way, we obtain the following objective function for the maximization step at iteration $j + 1$:

$$\sum_{i \in I} \sum_{s_i \in K} \eta^{j+1}(s_i = k|\mathbf{X}_i, \mathbf{y}_i, \boldsymbol{\Theta}^j) \left[\log \mathcal{L}\left(\mathbf{y}_i|s_i = k, \boldsymbol{X}_i, \boldsymbol{\theta}_k\right) + \log Pr\left(s_i = k|\mathbf{X}_i, \boldsymbol{\theta}_{\boldsymbol{k}}\right)\right] \tag{3.34}$$

It is clear from equation (3.34) that we can interpret the $\eta(s_i = k|X_i, y_i, \theta^j)$ individual i has in the part of the objective function associated group k, given that we observed data $(\mathbf{y}_i, \mathbf{X}_i)$, and the parameters $\boldsymbol{\Theta}^j$ was estimated in the previous step.

Second Step: Maximization

We now make that the dependent variable y_{it} is conditionally homeskedastic given the explanatory variable $\mathbf{x}_{it}$ and the type s_i:

$$y_{it}|\mathbf{x}_{it}, s_i = k \sim \mathcal{N}\left(\mathbf{x}'_{it}\boldsymbol{\beta}_k, \sigma_k^2\right), \quad \forall k \in K$$

which leads to the following representation of the conditional likelihood likelihood function for individual i:

$$\mathcal{L}\left(\mathbf{y}_i|s_i = k, \mathbf{X}_i, \boldsymbol{\theta}_k\right) = \prod_{t\in T} \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(y_{it}-\mathbf{x}'_{it}\boldsymbol{\beta}_k)^2}{2\sigma_k^2}} = \frac{1}{\sqrt{2\pi}\sigma_k^T} e^{-\sum_{t\in T}\frac{(y_{it}-\mathbf{x}'_{it}\boldsymbol{\beta}_k)^2}{2\sigma_k^2}} \tag{3.35}$$

Also, we assume that the type s_i is independent from the set of explanatory variables $\mathbf{X}_i$, so that we can define:

$$\pi_k = Pr(z_i = k|\mathbf{X}_i, \boldsymbol{\theta_k}) \ \forall k \in K, \ \forall i \in I \tag{3.36}$$

The assumptions made in equations (3.35) and (3.36) lead to the following definition of parameters: $\boldsymbol{\Theta} = (\boldsymbol{\beta}_k, \sigma_k^2, \pi_k)_{k\in K}$. In order to obtain the values for Θ^{j+1}, we solve the following optimization problem:

$$\max_{\boldsymbol{\Theta}} \sum_{i\in I}\sum_{s_i\in K} \eta^{j+1}(s_i = k|\mathbf{X}_i, \mathbf{y}_i, \boldsymbol{\Theta}^j)\left[\log \mathcal{L}\left(\mathbf{y}_i|s_i = k, \mathbf{X}_i, \boldsymbol{\theta}_k\right) + \log Pr\left(s_i = k|\mathbf{X}_i, \boldsymbol{\theta_k}\right)\right]$$

$$\text{s.t} \sum_{k\in K} \pi_k = 1 \tag{3.37}$$

where the objective function comes from the equation derived in (3.35). Given the optimization problem defined in (3.37), we define the Lagrangian as follows:

$$\Lambda(\boldsymbol{\Theta}, \lambda) = \sum_{i\in I}\sum_{s_i\in Z} \eta(z_i = k|\mathbf{X}_i, \mathbf{y}_i, \boldsymbol{\Theta}^j)\left[\log \mathcal{L}\left(\mathbf{y}_i|s_i = k, \mathbf{X}_i, \boldsymbol{\theta}_k\right) + \log Pr\left(s_i = k|\mathbf{X}_i, \boldsymbol{\theta}_k\right)\right] + \lambda(1 - \sum_{k\in K}\pi_k) \tag{3.38}$$

Given the assumption about the likelihood made in equation (3.35), equation (3.38) becomes:

$$\max_{\boldsymbol{\Theta}} \sum_{i\in I}\sum_{k\in K} \eta^{j+1}(s_i = k|\mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\Theta}^j)\left[\sum_{t\in T}\left(-\frac{1}{2}\log 2\pi\sigma_k^2 - \frac{(y_{it} - \mathbf{x}'_{it}\boldsymbol{\beta}_k)^2}{2\sigma_k^2}\right) + \log \pi_k\right] + \lambda(1 - \sum_{k\in K}\pi_k) \tag{3.39}$$

As it is clear from equation (3.39), once $\eta^{j+1}(s_i = k|\mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\theta}^j)$ has been computed in the Expectation step, we can find analytical solutions for the other variables of the optimization problem. In particular, first order

conditions are sufficient to characterize optimality:

$$\pi_k^{j+1} = \frac{\sum_{i\in I}\eta(s_i = k|\mathbf{y}_i, \mathbf{X}_i, \mathbf{\Theta}^j)}{\sum_{k\in K}\sum_{i\in I}\eta(s_i = k|\mathbf{y}_i, \mathbf{X}_i, \mathbf{\Theta}^j)}$$

$$\boldsymbol{\beta}_k^{j+1} = \left(\sum_{i\in I}\eta(s_i = k|\mathbf{y}_i, \mathbf{X}_i, \mathbf{\Theta}^j)\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\left(\sum_{i\in I}\eta(s_i = k|\mathbf{y}_i, \mathbf{X}_i, \mathbf{\Theta}^j)\mathbf{X}_i'\mathbf{y}_i\right)$$

$$\sigma_k^{j+1} = \sqrt{\frac{\sum_{i\in I}\eta(s_i = k|\mathbf{y}_i, \mathbf{X}_i, \mathbf{\Theta}^j)\left[\sum_{t\in T}(y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta}_k^{j+1})^2\right]}{T\sum_{i\in I}\eta(s_i = k|y_i, \mathbf{X}_i, \mathbf{\Theta}^j)}}$$

for each group $k \in K$. It is interesting to notice that the maximization step can be equivalently expressed as K separate Generalized Least Square (GLS) regressions, where weights are individual specific, represented by the posterior probability of belonging to the group. It is easy to see that by defining the matrix:

$$\boldsymbol{\omega}_{ik}^j = \mathbf{1}_i\eta(s_i = k|\mathbf{y}_i, \mathbf{X}_i, \mathbf{\Theta}^j)$$

$$\mathbf{\Omega}_k^j = \text{diag}\left[\left(\boldsymbol{\omega}_{ik}^j\right)_{i\in I}\right]$$

where $\mathbf{1}_i$ denotes a vector of 1 with length T_i, representing the number of observations that are available for individual i. We can reformulate the coefficients as follows:

$$\boldsymbol{\beta}_k^{j+1} = \left(\mathbf{X}'\mathbf{\Omega}_k^j\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{\Omega}_k^j\mathbf{y}\right), \quad k \in K \tag{3.40}$$

As we can see from Equation (3.40), this formulation allows to have a more clear interpretation of the estimation routine. The application of the Expectation Maximization here proposed can be seen as an iterative continuously updated sequence of K GLS regressions, where the weights are different for each of the groups, and depend on the posterior probabilities to belong to each group computed in the previous steps and summarized in the matrix $\mathbf{\Omega}_k^j$. The higher the probability of a certain individual i to belong to a specific group k, the higher its weight is going to be in the k^{th} GLS regression, reducing at the same time its weight in the other $k-1$ GLS regressions.

3.5.4 Iterative Algorithm

We now describe the iterative algorithm based on the two steps introduced in the previous subsection.

1. Set the number of groups to be $K \in \mathbb{N}_+$.
2. Initialize the values of Θ^0 based on the linear GMM with instrumental variables, and perturb them with a random shock; in this application, I set the random shock to be normally distributed: where

$GMM - lin$ denotes the point estimates obtained using the GMM on the log-linearized moment conditions.

3. For $j = 1, \cdots, N$:

   3.1 Perform the expectation step, and update the values of η_k^i, $i \in I$, $k \in K$

   3.2 Perform the maximization and update the values of Θ

4. If convergence is achieved, end; else go back to step 3.

Convergence is achieved if $||\Theta_{j+1} - \Theta_j||_2^2 < tol$, which I set to be 10^{-4}.

3.5.5 Moment Based Estimator

We refer to the log-linearized Euler equation presented in (3.24). We hypothesize that for each household i is of type $s_i = k$, so that the moment condition given the type can be expressed as follows:

$$\Delta c_{it,t+1} = \gamma_{0k} + \gamma_{1k} \log R_{t+1} + \theta_k \Delta \text{N. People}_{it,t+1} + \epsilon_{it+1,k} \tag{3.41}$$

In this case, $\mathbf{x}_{it} = (\log R_{t+1})$. Given that the error term includes the higher moments of the forecast error, in general it will not be the case that $\mathbb{E}[\varepsilon_{it,t+1}|\mathbf{x}_{it}] = 0$. This automatically violates the assumption about conditional normal distribution of the error term. For this reason, we need to make additional assumptions involving the distribution of the error term and the instruments. Let $\mathbf{z}_{it}$ be a vector of instruments. I assume that:

$$\mathbf{x}_{it} = \boldsymbol{\delta}' \mathbf{z}_{it} + \eta_{it}, \quad \eta_{it}|\mathbf{z}_{it} \sim \mathcal{N}(0, \sigma_x^2)$$
$$\text{cov}_t(\varepsilon_{it,t+1}, \eta_{it}) = 0$$

The first-stage estimates based on maximum likelihood are the following:

$$\hat{\delta}_{1,stage} = \mathbf{Z}'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}\mathbf{X}$$

where $\mathbf{z}_{it} = \left(1, R_t, R_{t-1}, \Delta \text{N. People}_{it,t+1}\right)$, $\mathbf{x}_{it} = (R_{t+1})$, $\mathbf{X} = (\mathbf{x}_{it})_{i\in\mathcal{I}, t\in\mathcal{T}}$ and $\mathbf{Z} = (\mathbf{z}_{it})_{i\in\mathcal{I}, t\in\mathcal{T}}$. Given the distributional assumption made, we have that:

$$\hat{\boldsymbol{\delta}}|\mathbf{Z} \sim \mathcal{N}\left(\delta, \sigma_x^2 \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\right)$$
$$\hat{\mathbf{x}}_{it}|\mathbf{Z} = \hat{\boldsymbol{\delta}}' \mathbf{z}_{\mathbf{it}} \sim \mathcal{N}\left(\boldsymbol{\delta}' \mathbf{z}_{it}, \sigma_x^2 \mathbf{z}_{it}' (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{z}_{it}\right)$$

Under the assumption of no correlation between the error term of the first stage equation and the instrument, it implies that we have an unbiased estimator for δ. This implies that:

$$\hat{\mathbf{x}}_{it} - \mathbf{x}_{it} \sim \mathcal{N}\left(0, \sigma_x^2 \mathbf{z}'_{it} \left(\mathbf{Z}'\mathbf{Z}\right)^{-1} \mathbf{z}_{it}\right)$$

We can therefore rewrite:

$$\Delta c_{it,t+1} = \gamma_{0k} + \gamma_{1k} \hat{\log} R_{t+1} + \theta_k \Delta \text{N. People}_{it,t+1} + \underbrace{\epsilon_{it+1,k} + (\gamma_{1k} \log R_{t+1} - \gamma_{1k} \log \hat{R}_{t+1})}_{\tilde{\varepsilon}_{it}} \tag{3.42}$$

$$\tilde{\varepsilon}_{it} | \mathbf{z}_{it} \sim \mathcal{N}\left(0, \sigma_z^2 + \sigma_x^2 \mathbf{z}'_{it} \left(\mathbf{Z}'\mathbf{Z}\right)^{-1} \mathbf{z}_{it}\right) \tag{3.43}$$

Let define as $\sigma_{it}^2 = \hat{\sigma}_x^2 \mathbf{z}'_{it} \left(\mathbf{Z}'\mathbf{Z}\right) \mathbf{z}_{it}$. In light of these considerations, I modify Equation (3.39) as follows:

$$\max_{\boldsymbol{\Theta}} \sum_{i \in I} \sum_{k \in K} \eta^{j+1}(s_i = k | \mathbf{y}_i, \hat{\mathbf{X}}_i, \boldsymbol{\Theta}^j) \left[\sum_{t \in T} \left(-\frac{1}{2} \log\left(2\pi(\sigma_k^2 + \sigma_{it}^2)\right) - \frac{(y_{it} - \hat{\mathbf{x}}'_{it}\boldsymbol{\beta}_k)^2}{2(\sigma_k^2 + \sigma_{it}^2)} \right) + \log \pi_k \right] + \lambda(1 - \sum_{k \in K} \pi_k) \tag{3.44}$$

Considering that the only variable that is correlated with the error term is $\log R_{t+1}$, as is instrumented using aggregate variables, we can drop the dependence of the first stage error from the $\hat{\sigma}_{it}^2$ from the specific household i, defining with $\sigma_t^2 = \sigma_{it}^2$. In addition, we can redefine $\sigma_{kt}^2 = \sigma_t^2 + \sigma_k^2$, so that we can rewrite the optimization problem of the maximization step as follows:

$$\max_{\boldsymbol{\Theta}} \sum_{i \in I} \sum_{k \in K} \eta^{j+1}(s_i = k | \mathbf{y}_i, \hat{\mathbf{X}}_i, \boldsymbol{\Theta}^j) \left[\sum_{t \in T} \left(-\frac{1}{2} \log\left(2\pi\sigma_{tk}^2\right) - \frac{(y_{it} - \hat{\mathbf{x}}'_{it}\boldsymbol{\beta}_k)^2}{2\sigma_{tk}^2} \right) + \log \pi_k \right] + \lambda(1 - \sum_{k \in K} \pi_k) \tag{3.45}$$

From Equation (3.45), we can see that the two-stage estimation leads to the conditional heteroskedasticity. Therefore, we need to redefine the estimation routine defined in Equation (3.40) to take into account for this factor.

$$\boldsymbol{\beta}_{k,OLS}^{j+1} = \left(\hat{\mathbf{X}}' \boldsymbol{\Omega}_k^j \hat{\mathbf{X}}\right)^{-1} \left(\hat{\mathbf{X}}' \boldsymbol{\Omega}_k^j \mathbf{y}\right) \tag{3.46}$$

$$\hat{\sigma}_{tk,OLS}^{j+1} = \sqrt{\frac{\sum_{t \in T} \sum_{i \in I} \eta(s_i = k | \mathbf{y}_i, \hat{\mathbf{X}}_i, \boldsymbol{\Theta}^j)(y_{it} - \hat{\mathbf{x}}'_{it}\boldsymbol{\beta}_k^{j+1})^2}{\sum_{t \in T} \sum_{i \in I} \eta(s_i = k | y_i, \hat{\mathbf{X}}_i, \boldsymbol{\Theta}^j)}} \tag{3.47}$$

$$\boldsymbol{\omega}_{ik,FGLS}^j = \left(\frac{1}{(\hat{\sigma}_{1k,OLS}^{j+1})^2}, \cdots, \frac{1}{(\hat{\sigma}_{Tk,OLS}^{j+1})^2} \right) \eta(s_i = k | \mathbf{y}_i, \hat{\mathbf{X}}_i, \boldsymbol{\Theta}^j) \tag{3.48}$$

$$\boldsymbol{\Omega}_{k,FGLS}^j = \text{diag}\left[\left(\boldsymbol{\omega}_{ik,FGLS}^j \right)_{i \in I} \right] \tag{3.49}$$

$$\boldsymbol{\beta}_{k,FGLS}^{j+1} = \left(\hat{\mathbf{X}}' \boldsymbol{\Omega}_{k,FGLS}^j \hat{\mathbf{X}}\right)^{-1} \left(\mathbf{X}' \boldsymbol{\Omega}_{k,FGLS}^j \mathbf{y}\right) \tag{3.50}$$

In the first step described in Equation (3.46), I perform a GLS regression using as weights the household-specific group probabilities. In the second step, described in Equation (3.47) based on the estimates obtained

for $\beta^{j+1}_{k,OLS}$, we then compute the residuals, which we use to estimate the time-group specific standard errors $\hat{\sigma}^{j+1}_{tk,OLS}$. In the third step (3.48)-(3.49), I adjust the matrix of weights, taking into account the newly computed standard errors. Finally, in the last step described in Equation (3.50), I estimate the new coefficients based on the updated weights matrix, in what is similar to a Feasible GLS procedure.

3.5.6 Results

In this subsection, I present the empirical results. Table 3.13 displays the estimated coefficients for $K = 1$ (standard instrumental GMM), $K = 2$ and $K = 3$. The standard errors are obtained through non-parametric bootstrapping, by resampling with replacement from the original sample of the households keeping the total number of households fixed (instead of the total number of observations). This choice is motivated by the fact that in this case, the unit of observation that is relevant to attribute the type is the entire set of observations pertaining to a household. We need to also specify how we are comparing estimates obtained using different bootstrapping samples. In general, the order of the groups that the algorithm gives back is random, and is at least in part influenced by the parameters chosen at initialization. If we were to compute the standard deviation of the parameters based on the bootstrapped estimates based on the group assignment generated by the algorithm, the estimates obtained would have little sense, since a particular unobserved group in a bootstrap estimate can very well refer to a different group in another bootstrap estimate. Therefore, I order the groups based on γ_1, i.e. the parameter representing the elasticity of inter-temporal substitution. In the estimates presented below, group $k = 1$ always represent the one that has a higher (lower) elasticity of inter-temporal substitution (coefficient of relative risk aversion).

As we can see from the first column of Table 3.13, the results we obtain are largely in line with the literature (see for instance Alan et al. [2009]), as the implied of risk-aversion is 2.29, and are similar to the ones obtained through non-linear GMM (see Table 3.11). As we can see from the second column, when $K = 2$ unobserved groups are assume, the coefficients of relative risk-aversion among the two groups differ: the first group of households display a coefficient of relative risk aversion of 1.78, while for the second one is 2.84. The two groups appear to be equally numerous among the households in the sample, as the estimates for $\hat{\pi}_1$ ($\hat{\pi}_2$) show. Figure 3.2 shows the distribution for the elasticity of inter-temporal substitution (left panel) and the coefficient of relative risk aversion (right panel). The vertical lines represent the estimated coefficients from the full sample reported in Table 3.13.

Another way to assess whether the identified groups obtained using the proposed estimation rou-

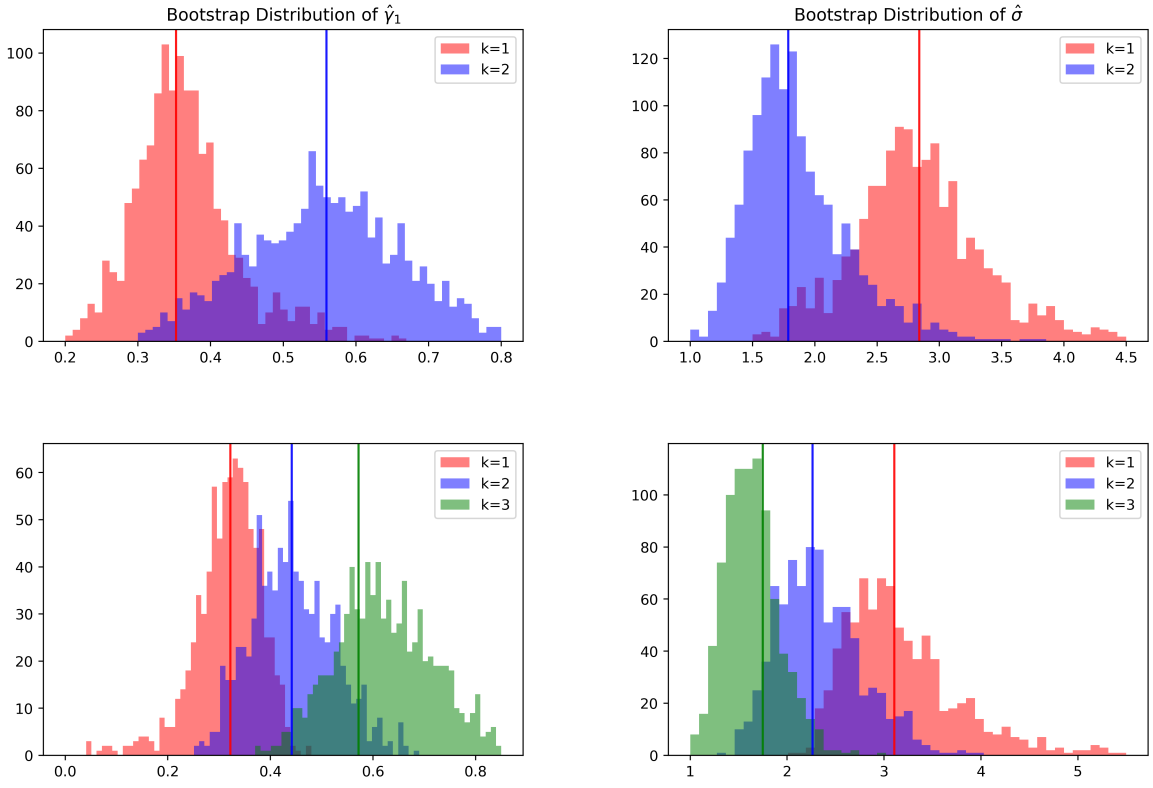

Figure 3.2: Bootstrap distribution of γ_1 (left) and σ (right) over 1000 samples. The top two charts represent the bootstraps distribution when $K = 2$, while the bottom two charts represent the bootstraps distribution when $K = 3$. The vertical lines represent the point estimates presented in Table 3.13 column (2) and (3).

		Estimates		
Type	Parameters	K = 1	K = 2	K = 3
Type 1, Low RRA	$\hat{\gamma}_{11}$	0.4352 (0.0575)	0.5593 (0.1137)	0.5718 (0.1046)
	$\hat{\gamma}_{01}$	-0.0133 (0.0021)	-0.0185 (0.0035)	-0.01988 (0.0046)
	$\hat{\sigma}_1$	2.2978 (0.3036)	1.7876 (0.4095)	1.7487 (0.2772)
	$\hat{\theta}_1$	0.0647 (0.0041)	0.0583 (0.0092)	0.0479 (0.0170)
	$\hat{\pi}_1$	1.000 –	0.4941 (0.0399)	0.3494 (0.0498)
Type 2, Med. RRA	$\hat{\gamma}_{12}$	–	0.3524 (0.0775)	0.4418 (0.0801)
	$\hat{\gamma}_{02}$	–	-0.0079 (0.0023)	-0.0108 (0.0048)
	$\hat{\sigma}_2$	–	2.8378 (1.2372)	2.2637 (0.4324)
	$\hat{\theta}_2$	–	0.0725 (0.0046)	0.07313 (0.0166)
	$\hat{\pi}_2$	–	0.5058 (0.0399)	0.37593 (0.0754)
Type 3, High RRA	$\hat{\gamma}_{13}$	–	–	0.3220 (0.0639)
	$\hat{\gamma}_{03}$	–	–	-0.0079 (0.0030)
	$\hat{\sigma}_3$	–	–	3.1055 (1.5444)
	$\hat{\theta}_3$	–	–	0.0767 (0.0112)
	$\hat{\pi}_3$	–	–	0.2747 (0.0799)

Table 3.13: Estimates based on PSID Food Consumption Data. Period: 1974-1987.

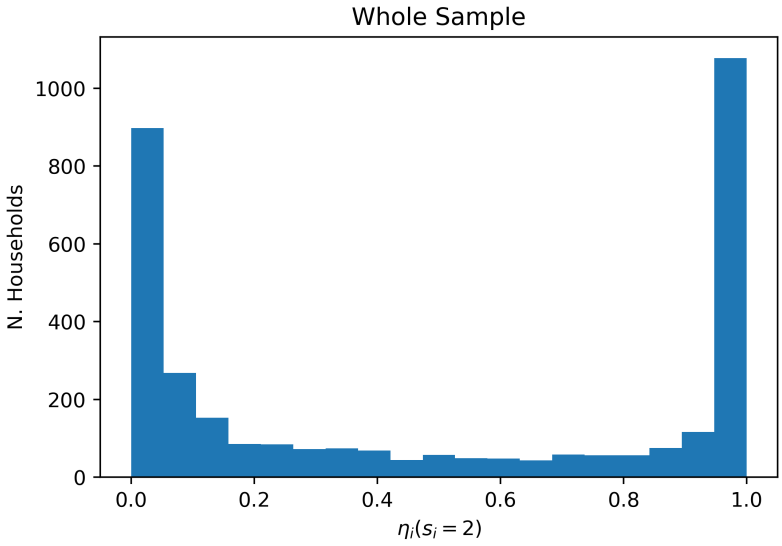

Figure 3.3: Distribution of $\eta_i\left(s_i = k|\mathbf{X}_i, \mathbf{y}_i\right), \boldsymbol{\Theta})$ across households.

tine are sensible is to look at the distribution of $\eta(s_i = k|\mathbf{y}_i, \mathbf{X}_i, \hat{\boldsymbol{\Theta}})$ across households. Ideally, for each household, we would like to assign a high probability of belonging to either one of the two groups, rather than having an equally split probability. If the probability of belonging for instance to group $k = 1$ is close to 0.5, then it would be hard to infer which group that specific households belongs to. We report the results in Figure 3.3. As we can see, there are two peaks around the values of 0 and 1. This implies that for a large group of households, the probability to belong to group $k = 1$ is either 0 or 1, suggesting that the output of the algorithm developed allows us to assign households to either group $k = 1$ or group $k = 2$. We now report the distribution of $\eta(s_i = 1)$ conditional on the educational attainment since we want to understand whether there is any relationship between risk-aversion and educational attainment, as discussed previously. As we can see from Figure 3.4, for all but one educational group, namely college educated households, the peak around 0 is lower than the one around 1. In addition, it appears that the difference across the two peaks is monotonically decreasing in the the educational attainment. In order to reinforce this intuition, we now want to assess what the the relative share of each type of unobserved household among the four observed educational attainments. To do so, we need to classify each household based on the implied probability derived in the Expectation step of the Expectation Maximization algorithm. We define the predicted household group $\hat{s}^i$ as of household i as follows:

$$\hat{s}_i = \arg\max_{k \in K} \eta(s_i = k|\mathbf{y}_i, \mathbf{X}_i, \hat{\Theta}) \tag{3.51}$$

Based on Equation (3.51), we then define the share of the population belonging to group k for each education

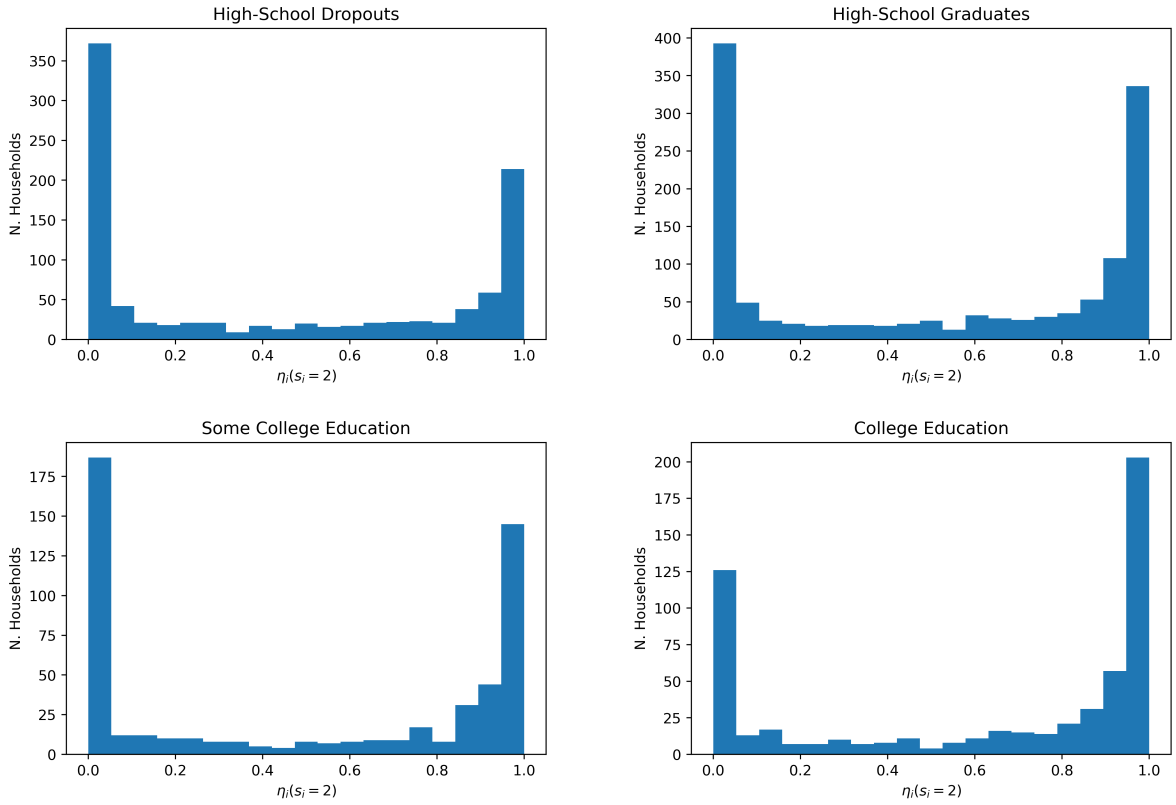

Figure 3.4: Distribution of $\eta\left(s_i = k|\mathbf{X}_i, \mathbf{y}_i, \hat{\mathbf{\Theta}}\right)$ across households, for four different level of educational attainments.

attainment:

$$\hat{\mu}_k(educ) = \frac{\sum_{i \in I} \mathbf{1}(\hat{s}_i = k, educ_i = educ)}{\sum_{i \in I} \mathbf{1}(educ_i = educ)} \quad educ \in \{\text{Lt. HS, HS, Some College, College+}\} \tag{3.52}$$

We report the values of $\tilde{x}_i$ described in Equation (3.52) representing the share of the population that belongs to each unobserved group for the four levels of educational attainments, based on the results of column (2) in Table 3.13. Table 3.14 reports the results when $K = 2$, while in Table 3.15 are displayed the shares for $K = 3$. As we can see from Table 3.14, the share of households belonging to the low-risk averse group ($k = 1$) decreases with the level of educational attainment: the highest share of households in the first group is among high-school dropouts (55.4%), while the lowest one is among college graduates (35.3%). High-school graduates and dropouts display similar conditional distributions across the two groups, with the higher risk-aversion group being relatively more frequent. Results Table 3.15 suggest that by including a third group, our conclusions relating risk-aversion and educational attainment do not change. The share of high-school dropouts in the intermediate ($k = 2$) and high ($k = 3$) risk-aversion group is lower than the ones of college graduates: 34.7% of high-school dropouts are classified in group $k = 2$ vs. 44.0% of college graduates,

and 27.4% are in group $k = 3$ vs. 34.8% of college graduates. Similarly, the conditional distribution across unobserved groups of high-school graduates and households with some college education. The results

Education	$k = 1$	$k = 2$
High-School Dropouts	55.4%	44.6%
High-School Graduates	46.7%	53.3%
Some College	48.0%	52.0%
College+	35.3%	64.7%

Table 3.14: Share of Households by Educational Attainment and Unobserved Type based on the estimates presented in Table 3.13 for column (2), $K = 2$. Groups are ordered in ascending order of relative risk aversion, with group $k = 1$ has the highest (lowest) EIS (CRRA).

Education	$k = 1$	$k = 2$	$k = 3$
High-School Dropouts	38.2%	34.7%	27.4%
High-School Graduates	32.2%	38.8%	30.0%
Some College	34.3%	37.6%	28.2%
College+	21.2%	44.0%	34.8%

Table 3.15: Share of Households by Educational Attainment and Unobserved Type based on the estimates presented in Table 3.13 for column (3), $K = 3$. Groups are ordered in ascending order of relative risk aversion, with group $k = 1$ has the highest (lowest) EIS (CRRA).

presented here support findings in the literature showing a positive correlation between risk aversion and educational attainment. However, the methodology proposed here uncovers this relationship by establishing an indirect relationship the estimated unobserved heterogeneity in preferences and the observed variables of interest.

3.6 Production Technology and Aggregate Shocks

As standard in the literature, I assume that the capital share in the Cobb-Douglas production technology is $\alpha = 0.33$. As discussed in Chapter 1, I assume that the technology shock follows a AR(1) process. The assumption about the independence between the depreciation shock and the technology shock is motivated

empirically. Ambler and Paquet [1994] find that the correlation between the depreciation shock and the technology shock to be -0.057, but not significant at any standard level. In order to estimate the relevant parameters (ρ, σ_e^2), I collect U.S. Total Factor Productivity (TFP)[3] data from the online database of the Federal Reserve Bank of Saint Louis (FRED) from 1970 to 2018. I first detrend the yearly time series assuming a linear trend in time, and then on the the detrended residuals I estimate an AR(1) process. The estimated coefficients are $(\hat{\rho}, \hat{\sigma}_\varepsilon) = (0.851, 0.008)$. The high coefficient for the auto-correlation term suggests a high persistence in the technological process, while the low standard deviation implies the the TFP shocks will be subject to relatively little volatility. Although our we use a much more recent sample compared to Hansen [1985] and Prescott [1986], our estimates are comparable to ones presented in their studies.

In order to compute the average depreciation rate, I use the data on Fixed Assets[4] and Depreciation[5] in current costs from the U.S. Bureau of Economic Analysis from 1970 to 2018. I take the ratio between the depreciation costs and total amount of fixed assets, and I compute the average across time. The estimate is $\hat{\delta} = 5.0\%$ on an annual basis. The estimated volatility of the computed yearly depreciation rate is 0.0035, which is significantly lower than the one estimated by Ambler and Paquet [1994].[6] An underestimation of the volatility of capital depreciation is not surprising, since the Bureau of Economic Analysis constructs capital stock data using an accounting methodology that assumes constants rates of depreciation across various capital categories. At the same time, for our purposes, it is important to try to capture the volatility of the depreciation rate as carefully, as it directly impacts the volatility of the interest rates, and therefore it represents and important source of risk for retirees.

For convenience, I use Tauchen method (Tauchen [1986]) to discretize the state space and to characterize numerically the Markovian distribution of the shocks[7]. For both shocks, I use five nodes, leading to a total of 25 possible values for (A_t, δ_t) and the two 5×5 transition matrices $\mathbf{\Pi}_{\delta'|\delta}$ and $\mathbf{\Pi}_{A'|A}$. The joint transition matrix $\mathbf{\Pi}_{\mathbf{z}'|\mathbf{z}}$ is computed through the Kroeneker product of the two individual transition matrix given the underlying assumption of independence between the two shocks.

[3] The identifier for the time series is RTFPNAUSA632NRUG.

[4] The identifier for the time series in FRED is K1TTOTL1ES000.

[5] The identifier for the time series in FRED is M1TTOTL1ES000.

[6] Ambler and Paquet [1994] estimate that the depreciation rate has a volatility of 0.005245 on a quarterly basis, which translates to 0.0262 on a yearly basis.

[7] An alternative to this method would involve the use of quadrature methods in the computation of integral using appropriate bi-dimensional grids.

3.7 Counterfactual Analysis

In the following analysis, I identify four types of households, namely: high-school dropouts, high-school graduates, some college educated, and college graduates. The identification of the household groups is based on the education level of the head of each household. As described in Chapter 1, in this model I only take into account retirement benefits, and I abstract from any consideration of spousal and survivor benefits. For this reason, consistent with the literature, I decide to identify the four groups as male high-school dropouts, male high-school graduates, male with some college education and male college graduates. This simplification is made for the following reasons. First, the literature and current data suggest that men are more likely than women to qualify for Social Security benefits based on their own earning histories. Second, the introduction of women would significantly complicate the solution of the model. The state space would need to include the entire distribution of women's average life-time earnings, in addition to two additional variables capturing whether husband or wife are alive. In addition, the presence of mortality in this model would significantly increase the number of households in the economy. Suppose that at time t, a household is composed of both husband and wife. In the next period, four possible scenarios can occur: the husband dies, the wife dies, they both die, or neither dies. This leads to the generation of three additional types of households in the next period (if both die, their bequest will be distributed and the household will cease to exist). If, on the contrary, only one of the two spouses is alive, only two scenarios can occur: the spouse either survive or dies. The introduction of mortality combined with joint spouse dynamics makes the problem computationally infeasible. If mortality was not explicitly taken into account, we would only have 4×4 types of ex-ante heterogeneous households in the model. I leave to future studies to explore this avenue.

In the analysis proposed in this chapter, I make some additional assumptions relative to the literature. First of all, in this paper I do not consider a market for risk-free bonds, differently for example, from Kim [2018], Hasanhodzic and Kotlikoff [2018]. As noted in Kim [2018] and Henriksen and Spear [2019], the consumption allocations computed in the presence of one risky asset and one risk-free asset are numerically very similar when markets are sequentially incomplete. This implies the following undesirable property: the implied consumption allocations will vary little as we change the allocations of savings or borrowings in risk-free bonds and risky assets. From a numerical perspective, this implies that while we may obtain good approximations for the consumption policy functions, the implied policy functions for households' portfolio choice may be off by a large margin. This problem has not been explicitly tackled

in the context of large scale OLG models, and while multiple recent papers numerically solve for the optimal portfolio, I decide take a more cautious approach and assume the existence of only one asset. This is motivated by the fact that I am using a new approach to approximate policy functions based on neural networks and deep-learning, whose numerical interaction with the households' portfolio choice is not clear. It is important to note that in the model proposed, it would be ideal to have a bond market, so that the trust fund assets earn a risk-free interest rate. We leave this avenue to future research.

Secondly, I do not assume that there are any borrowing costs, differently from Hasanhodzic and Kotlikoff [2018] and Kim [2018]. The choice is motivated by the fact that the calibration of the parameters ruling the borrowing cost function is achieved by targeting some empirical moments,[8] rather than following a traditional estimation approach based on micro data. To minimize the number of hand selected parameters, I decide to avoid introducing any borrowing costs. It is important to note that it would be straightforward to extend the model presented in Chapter 1 and combine it with the algorithm presented in Chapter 2 to accommodate for borrowing costs. The only modification required would be the introduction of a household specific rate of return, which would account for whether a household is borrowing or saving. The use of neural networks would easily take into account the non-linearities introduced by the borrowing costs, and it would be interesting to use them to showcase the properties in a context in which non-linearities characterize policy functions on the ergodic set. That said, to incorporate borrowing in this way would require me to hand select parameters in a somewhat arbitrary fashion, which would undermine the results to a certain extent.

Finally, as compared to much of the literature, I take a different stance with regard to the policy analysis in this Section. Several papers have focused on understanding the implications of the introduction of a PAYGO Social Security system and compare its welfare implication relative to an economy with no Social Security at all. As discussed at length in 1, however, the current Social Security does not work as a PAYGO system. Further, the tax rate used in many of these studies has been critiqued as unrealistically low (see for instance Krueger and Kubler [2004], Krueger and Kubler [2006] and Kim [2018], where a 2% payroll tax rate is used). In the analysis performed in the next Section, I will consider as the benchmark economy the current status quo, i.e. a Social Security that is solvent in the short-run but can become insolvent in the future. Therefore, the analysis proposed will not discuss what happens when we introduce a Social Security system, but will, instead, focus on understanding the implications of policy reforms on the current Social Security system. In light of these considerations, I conduct counterfactual analyses aimed at evaluating the

[8] For instance, Kim [2018] chooses a parametrization for the cost function in order to achieve a peak in the consumption profile around the age of 30.

impact of various policies on household welfare. Here I utilize the current system as the benchmark and propose three different policy alternatives. The numerical solution for the alternative policy scenarios is derived using the algorithm developed in Chapter 2 and relies on the use of deep-learning techniques in combination with two hidden-layer neural networks as functional approximators for the forecast and policy functions.

Benchmark The benchmark scenario is the current Social Security system as described in Section 1.4 of Chapter 1. The tax rate used in this context is 10.6%, which represents the current payroll tax rate that contributes to the OASI trust fund. In this case, the Social Security system is solvent, but can become insolvent in the future.

Scenario 1 The first policy alternative examined involves increasing the payroll tax rate. The tax rate is set so that at the deterministic steady state, the Social Security annual budget is balanced, with revenues collected through payroll taxes being equal to total outlays. The implied tax rate is computed numerically, using a binary search method. The resulting tax rate is $\tau^b = 15.14\%$.

$$
\begin{aligned}
H_{t+1} &= H_t r_t + T_t - S_t && \forall t \geq 0 \\
S_t &= \sum_{i \epsilon \mathcal{I}} \sum_{n=R_i}^{A} P_{in}(1-\mu_{in}) ss_{in} \\
T_t &= \sum_{i \epsilon \mathcal{I}} \sum_{n=1}^{R_i-1} P_{in} \tau^b w_t n_{in}
\end{aligned}
$$

In this scenario, the trust fund is allowed to borrow and accumulate negative balances. It is assumed that the initial level of assets of the trust fund is around three times the cost of the current retirement program. As the simulation will show, the transition from the initial state to the new stochastic steady state will take some time. Therefore, in the welfare analysis, it will be important to distinguish between the transitory dynamics and the stochastic steady state.

Scenario 2 In the second policy alternative, I reduce the benefits uniformly across households so that at the deterministic steady state, the budget of the Social Security system is balanced, while at the same time I keep the tax rate steady at 10.6%. The implied reduction rate ρ_b in the benefits is computed numerically through binary search, and amounts to 70.0%. The model implied reduction in the benefits is 30%, an

estimate that is very close to the projected cut of 27%[9] in benefits estimated by Social Security necessary to achieve long-term solvency. This policy alternative can be characterized by the following set of equations:

$$H_{t+1} = H_t r_t + T_t - S_t \qquad \forall t \geq 0$$

$$S_t = \sum_{i \in \mathcal{I}} \sum_{n=R_i}^{A} P_{in}(1-\mu_{in})\tilde{ss}_{in}$$

$$T_t = \sum_{i \in \mathcal{I}} \sum_{n=1}^{R_i-1} P_{in}\tau w_t n_{in}$$

$$\tilde{ss}_{iR_i,t} = \rho_b \times \theta_i \bar{e}_{iR_i-1,t-1} \qquad \forall i \in \mathcal{I}$$

Scenario 3 In the third policy scenario, I eliminate Social Security completely: retirement is financed through personal savings, and no taxes are collected from workers nor are retirement benefits distributed to retirees. In this scenario, there is no redistribution of income across households of different groups, as everyone is responsible for their own retirement savings, with no subsidization of retirement benefits from richer to poorer households. This represents a major departure from the current setting, since, as discussed in Chapter 1 and estimated in Table 3.3, the replacement rate is decreasing in the average life-time income, and varies significantly across groups. In this policy alternative, the tax rate is set to $\tau = 0$. The other relevant equations are defined as follows:

$$H_t = S_t = T_t = 0 \qquad \forall t \geq 0$$

$$ss_{in,t} = 0 \qquad \forall t \geq 0, \;\; i \in \mathcal{I}, \;\; n \geq R_i$$

For what concerns the numerical solution of the economy with no Social Security, I use the same techniques discussed in Chapter 2. Considering that in this case, there is no need to track workers' average life-time earnings nor Social Security expenditures, we can use a reduced version of the state space as compared to the economies with Social Security. Therefore, I can set the input of the policy functions to include only the aggregate shocks, the household-specific capital holdings, the aggregate level of private savings and the cross-sectional standard deviation of households' capital holdings. Similarly, for the forecast functions, I only need to predict the future values of the endogenously determined aggregate variables, namely the first and second moment of the of the households' savings distributions.

For the first, second and third policy alternatives, I use the methods discussed in Chapter 2. In particular, I used the same architecture for the neural networks as described in Table 2.2 and Table 2.4 for policy and forecast functions respectively.

[9] Source: Social Security Administration.

The benchmark regime and three policy alternatives described in the previous subsection are evaluated according the impact each has on the welfare of different types of households populating the economy. In terms of welfare measures, similarly to Kim [2018] and Hasanhodzic and Kotlikoff [2019], I use ex-ante utility as expressed in Equation (3.53):

$$\mathrm{W}_{i,standard} = \mathbb{E}_0 \left[\sum_{t=0}^{A-1} \beta_i^t \frac{c_{in+t,t}^{1-\sigma_i}}{1-\sigma_i} \right] \quad \forall i \in \mathcal{I} \tag{3.53}$$

To evaluate welfare, I numerically compute the household's expected utility at the stochastic steady state. By appealing to the ergodic theorem, I consider the simulated data as representing the Markov distribution at the stochastic equilibrium. While the results presented in the next subsection refer to long-term, I will discuss the implications of the different policy alternatives in the transition as well. This will help us better understand whether some costs are paid in the transition by households.

3.7.1 Results

In this subsection, I present the results for two alternative assumptions about agents' risk preferences. In the first case, I assume that there is no unobserved heterogeneity in risk-preferences. I assume that all households have the same coefficient of relative risk aversion, $\sigma_i = 2.30 \quad \forall i \in \mathcal{I}$, as estimated in the first column ($K = 1$) of Table 3.13. The population shares are based on the first column of Table 3.6. In the second case, I assume that the population is divided into two groups, the first one having a lower coefficient of relative risk aversion σ_l, and the second having a higher coefficient of relative risk-aversion σ_h. I set $\sigma_l = 1.78$ and $\sigma_h = 2.84$ based on the estimates presented in the second column ($K = 2$) of Table 3.13. I assume that each household type is divided into the low and high risk-aversion groups according to the population shares derived in Table 3.14. As a result, the low risk aversion group is comprised of 55.4% of high-school dropouts, 46.7% of high-school graduates, 48.0% of individuals with some college education, and 35.3% of college graduates. In contrast, the high risk aversion group is comprised of 44.6% of high-school dropouts, 53.3% of high-school graduates, 52.% of individuals with some college education, and 64.7% of college graduates.

Under these assumptions, the economy is populated by eight types of agents, which we can distinguish by their risk preferences and their level of educational attainment. For the discount factor, I choose the value estimated trough standard non-linear GMM in Table 3.11, setting $\beta = 0.97$. Under either of the assumptions about risk-preferences, the discount factor is always kept fixed at this value. The aim of this

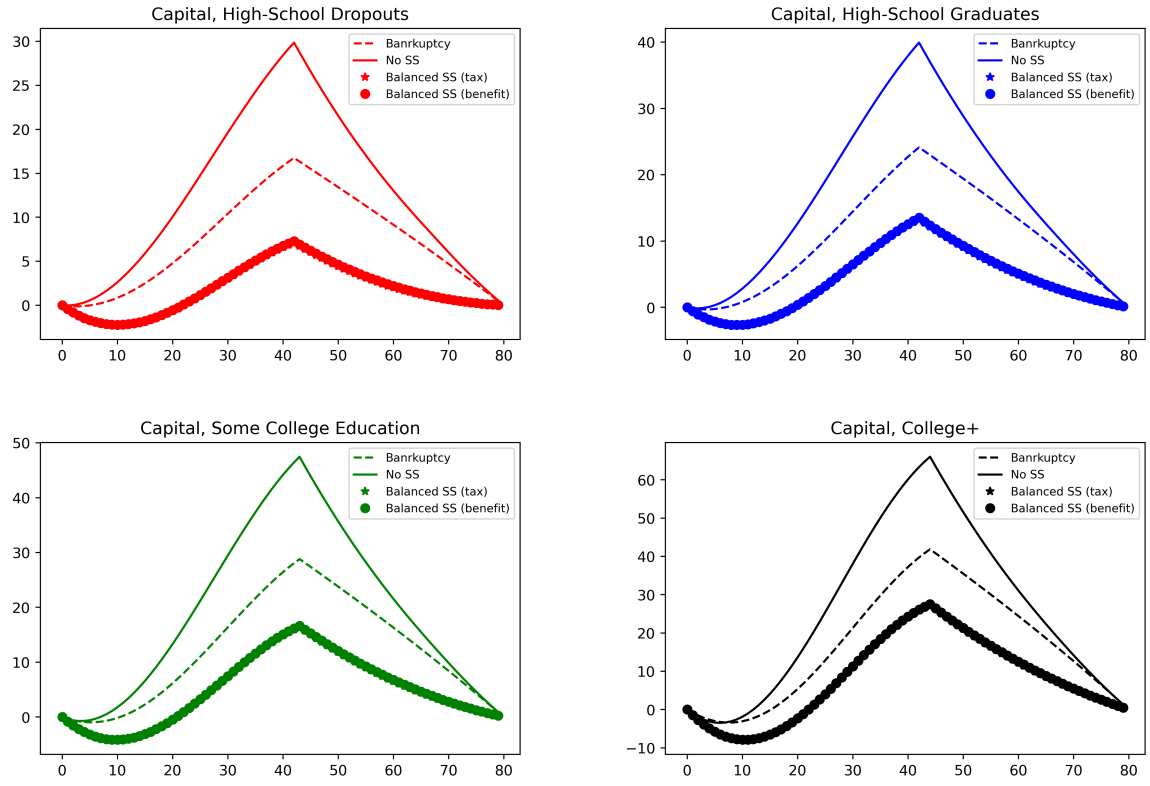

Figure 3.5: Capital holdings by household type in the four alternative policy scenarios.

analysis is to understand whether heterogeneity in risk-preferences potentially impacts our analysis.

Homogeneous Risk Preferences

I present the results based on the assumption of no unobserved heterogeneity, i.e. risk preferences are uniform across the households in the population. Figure 3.5 displays the capital holdings by type of household at the new stochastic steady state under the alternative policy scenarios discussed in the previous subsection. As we can see, the presence of retirement benefits significantly crowds out private investment. In all of the scenarios in which Social Security pays retirement benefits, regardless of the solvency status, private capital holdings are lower than the scenario with no Social Security. In particular, the more generous the retirement benefits, the smaller the incentives are for households to save to finance their retirement. While private investment is crowded out by Social Security, in both of the long-term balanced policy scenarios, the trust fund accumulates a significant amount of assets once it reaches the stochastic steady state. The elimination of Social Security achieves the highest levels of total savings in the economy. As we can see from Table 3.16,

under the first alternative policy scenario, i.e. a tax increase aimed at achieving a balanced Social Security that maintains the current level of benefits, private savings K_p constitute only 25.9% of the total capital in the economy. The remaining 74.9% is represented by public savings accumulated in the Social Security trust fund. Total capital holdings, which includes both private and public savings, is 43.4% higher than the benchmark case. In the second policy alternative, when benefits are reduced but the Social Security system is long-term solvent, households hold a larger share of capital, with private savings accounting for 75.1% of aggregate capital. At the stochastic steady state, the average level of aggregate capital is 35.9% higher than the benchmark case. If we compare the size of the trust fund to the total cost of the Social Security program under the two long-term balanced scenarios, the trust fund accumulates resources that are approximately 20 to 40 times higher than the total expenditures in retirement, depending on how the long-term solvency is achieved. For comparison, as of December 2020, the Social Security trust fund held assets worth 3 times the value of yearly retirement benefits. Therefore, the long-term balanced scenarios presents an undesirable property, that is they both lead to an over-accumulation of savings that is not realistic, which leads to a sizeable decrease in the average rate of return of capital. In a similar fashion, the level of aggregate savings when Social Security is eliminated is significantly higher than the benchmark case, driven by a higher level of private savings used to finance consumption at retirement.

Compared to the benchmark case, the three alternative policy scenarios proposed achieve higher aggregate levels of savings: while more generous retirement benefits decrease the propensity of households to save, the presence of a Social Security trust fund drives up aggregate savings through the accumulated assets. The higher level of aggregate capital drives up the marginal product of labor per unit efficiency, and, consequently, wages. At the same time, the higher level of aggregate capital decrease the interest rate that is paid to households' investments. It is important to note that an increase in wages in the long-run translates to higher Social Security benefits as well, since benefits are keyed to retirees' lifetime income. An increase in the average wage in the first scenario translates to an increase in similar magnitude of retirement benefits, given the linear relationship between the two. At the same, the decrease in the interest rate lowers capital income, which represents the additional source of income that retirees have in addition to Social Security retirement benefits. Overall, the introduction of a balanced Social Security has a mixed effect on retirees' income, as it both increases Social Security benefits while decreasing investment income, since retirees both hold less private savings and earn a lower average interest rate. Relative to the benchmark case, the average interest rate drops by approximately 1.50% in both long-term balanced policy alternatives.

From an insurance perspective, the introduction of a long-term balanced Social Security decreases

Policy	Private Savings K_p	Social Security Trust Fund (H)	Total $(K+H)$	Wage w	Interest rate r
Benchmark	759.3 (–%)	0 (–%)	759.3 (–%)	1.44 (–%)	1.49% (–%)
Scenario 1	282.0 (-62.8%)	806.1 (–%)	1088.2 (+43.3%)	1.63 (13.2%)	-0.01% (-1.50%)
Scenario 2	570.4 (-24.9%)	461.7 (–%)	1032.1 (+35.9%)	1.59 (+10.4%)	+0.02% (-1.47%)
Scenario 3	1293.4 (+70.3%)	0 (–%)	1293.4 (+70.3%)	1.72 (+18.1%)	-0.62% (-2.11%)

Table 3.16: Average private savings $K_p = \sum_{i\in\mathcal{I}}\sum_{n\in\mathcal{A}} k_{in}P_{in}$, assets in the Social Security Trust Fund H, and total capital holdings $K_p + H$ under different policy scenarios at the stochastic steady states. In parenthesis, differences compared to the benchmark. For total capital holdings and wages, we present the relative difference; for the interest rate, the absolute difference.

consumption volatility of retirees. It is important to keep in mind that in the benchmark scenario, retirement benefits are subject to shocks in the aggregate technology A_t, since it impacts wages and consequently the total revenues collected. The mechanism is broken in a long-term balanced Social Security system, since trust fund assets can used to pay benefits in this case. For this reason, we expect consumption of retirees to be less volatile, since it is subject only to the depreciation shock. Figure 3.6 displays the average consumption volatility relative to the benchmark case under the two long term balanced policy alternatives. The chart on the left represents Scenario 1 (long-term balanced with tax increase), while the chart on the right depicts scenario 2 (long-term balanced with benefits cut). The data is obtained by simulating 100 economies for 2000 periods. The first 300 periods for each economy are discarded to make sure that the stochastic steady state has been reached.

As we can see from the top charts in Figure 3.6, both scenarios redistribute consumption risk from younger and middle-aged cohorts to older cohorts. In particular, we can observe the change in consumption volatility relative to the benchmark is a decreasing function of age: the younger the household (independently on the type) the higher its consumption volatility relative to the benchmark. The risk transfer from old to young is more marked for the scenario that has more generous retirement benefits. These results suggest that a long-term balanced Social Security system decreases consumption risk for the retirees at the expense of the young, and the effect is more pronounced in a scenario with more generous benefits.

The bottom chart of Figure 3.6 shows that the elimination of Social Security would, on average, increase consumption volatility relative to the benchmark regardless of the age. There is no clear pattern related to age. This does not come as a surprise, since in a scenario without Social Security, retirees' consumption at retirement depends entirely on their savings, whose return is subject to both the technology shock and the depreciation shock. Additionally, it is interesting to note that younger cohorts experience an increase in the consumption volatility as well. This is driven by their higher level of private capital holdings across all cohorts as displayed in Figure 3.5.

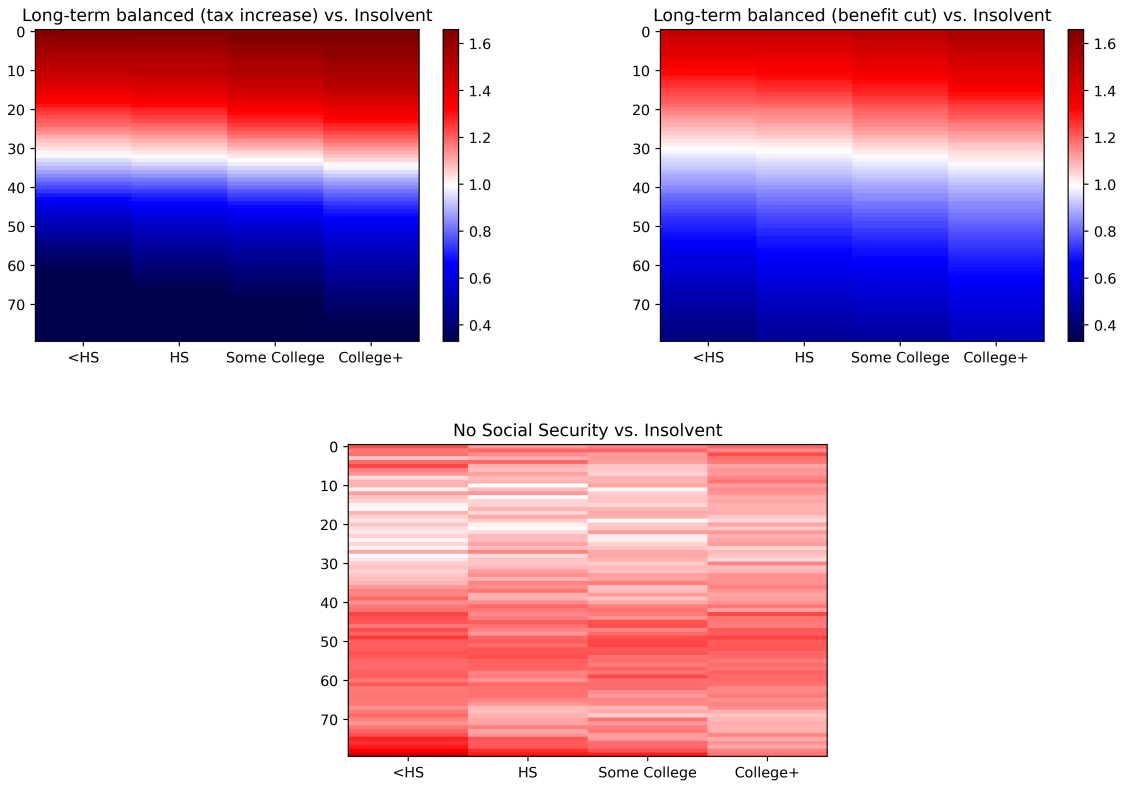

Figure 3.6: Consumption volatility relative to the benchmark scenario by household type (x-axis) and cohort (y-axis). On the top left, Scenario 1 vs. Benchmark. On the top right, Scenario 2 vs. Benchmark. On the bottom, Scenario 3 vs. Benchmark.

I now discuss the impact of different policy alternatives on the measure of ex-ante utility presented in Equation (3.53). The computations of the welfare measures are based on the simulation of 100 economies for 2000 periods. The first 300 periods for each economy are discarded to ensure that the stochastic steady state has been reached. As the results make clear, the elimination of Social Security completely represents the worst policy alternative of the three considered. In this scenario, welfare decreases uniformly across

all types of households, with the absolute magnitude decreasing in the educational attainment. High-school dropouts experience the biggest welfare loss relative to the benchmark (-4.93%), followed by high-school graduates (-4.71%) and the group with some college education (-3.79%). College graduates experience the smallest drop (-1.66%). It is not surprising to observe that the magnitude of the losses are decreasing as the educational attainment increases. As described in Chapter 1 and in Section 3.4 of this chapter, the current Social Security system is redistributive in the way it pays benefits: lower income high-school dropouts have a higher replacement rate than higher income college graduates.[10]. Therefore, the elimination of Social Security will penalize low earners more as the current Social Security system is more generous to low earners with regard to the replacement rate. It is interesting to note that the variation in terms of realized utility when Social Security is eliminated is the largest among all policy alternatives. The standard deviation in realized utility is 1.27, 61% higher relative to the benchmark. This suggests that the presence of retirement benefits, and in general Social Security, increases the degree to which risk is reallocated across generations. Table 3.18 shows the maximum, minimum and the standard deviation of the realized utility in our simulated data. While the lucky cohorts are better off in a world with no Social Security, the unlucky cohorts are significantly worse off compared to all scenarios wherein Social Security exists. This suggests that either version of the Social Security system (solvent or insolvent) is better at reallocating risk across generations.

We now turn to the welfare analysis of the two long-term balanced Social Security systems. As we can see from the second and third rows of Table 3.17, either a tax increase or a benefit reduction decreases the welfare of all types of agents with the exception of college graduates, who experience a 0.62% gain in the case of a tax increase and a 0.74% increase in the case of a reduction of benefits. The losses for the other three groups are minor, in the order of 1.2%-1.4% for high-school dropouts and graduates and 0.10%-0.40% for the group with some college education. Table 3.18 shows that the dispersion in realized utility is uniformly lower for an economy with a long-term balanced Social Security achieved through a tax increase. The highest (lowest) realized utility is lower (higher) than the corresponding one in the benchmark case, suggesting that an increase in taxes (maintaining the same level of benefits) achieves the best inter-generational risk-sharing.

The analysis of the stochastic steady states does not suggest there is a clear policy winner among the benchmark or three policy alternatives. While the benchmark scenario achieves the highest level of ex-ante utility for all but the college graduates, it is not operationally viable in the long-run. As explained in Chapter 1, the assets accumulated in the trust fund give the power to Social Security to pay benefits. Consid-

[10]The replacement rate of male college graduates relative to high-school dropouts is 70%.

ering that payroll taxes are paid on a monthly or bimonthly basis, the share of benefits that Social Security would be able to pay would vary monthly. Looking at the transition between stochastic steady states can be helpful in this context, since every change in the policy regime will take some time to occur, and currently living households will pay the transition costs. And if the transition costs are high enough, the welfare analysis at the stochastic steady state may not be a highly relevant metric to consider for policymakers.

Two considerations are in order. First, it takes 100 to 300 periods to achieve the new steady states, depending on the type of policy implemented and the initial conditions imposed. This corresponds to 100 to 300 years in this model, which is a very long time-horizon for any policy consideration. Second, in the transition, households' ex-ante utility may be significantly different from that of the steady state. Therefore, households living in the transition will pay the policy cost. For instance, an increase in the payroll tax rate will trigger a transition that lasts between 200 and 250 periods, with a simulated realized utility ranging from -19.40 to -19.00 for high-school dropouts, -14.80 and -14.40 for high-school graduates, -12.10 and -11.60 for the group with some college education, and between -8.40 and -8.00 for college graduates. If we compare these values to those presented in Table 3.17, we can see that, in the transition, households' realized utility is significantly lower than that of their counterparts at the stochastic steady state. The difference is only slightly less marked for the second policy alternative that cuts benefits to achieve long-term solvency: utility of high-school dropouts in this case ranges from -19.20 to -18.90, -14.65 and -14.40 for high-school graduates, -11.85 and -11.60 for the group with some college education, and -8.20. and -8.00 for college graduates.

Heterogeneous Risk Preferences

In this subsection, I will present the results based on the assumption of unobserved heterogeneity, i.e. risk preferences are different across households. As explained in the previous subsection, there are two types of agents within each group, one having low risk aversion σ_l and one having high risk aversion σ_h. In total, the economy is populated by eight types of households. The goal of the analysis presented here is to understand the extent to which risk preferences impact the analysis performed in the previous subsection.

The algorithm described in Chapter 2 is used to characterize the numerical solution of the benchmark and each of the three policy alternatives. No fine-tuning in the hyper-parameters is necessary, which displays that it scales very well, since in this application, we are numerically computing the policy functions of 79 = 632 agents simultaneously. This shows that the approach developed is particularly promising in the

Policy	<HS	HS	Some College	College+
Benchmark	-18.67 (–%)	-14.21 (–%)	-11.60 (–%)	-8.05 (–%)
Scenario 1	-18.91 (-1.29%)	-14.41 (-1.41%)	-11.61 (-0.09%)	-8.00 (+0.62%)
Scenario 2	-18.93 (-1.39%)	-14.41 (-1.41%)	-11.64 (-0.35%)	-7.99 (+0.74%)
Scenario 3	-19.60 (-4.93%)	-14.88 (-4.71%)	-12.05 (-3.79%)	-8.15 (-1.24%)

Table 3.17: Average households' ex ante utility obtained under the assumption of no unobserved heterogeneity in risk-preferences across households. The welfare measure used is based on Equation (3.53). Benchmark: Insolvent Social Security. Scenario 1: Social Security, long-term balanced (tax increase). Scenario 2: Social Security, long-term balanced (benefits reduction). Scenario 3: No Social Security.

Policy	Statistics	<HS	HS	Some College	College+
Benchmark	Max	-15.98	-12.21	-10.01	-6.98
	Min	-21.84	-16.55	-13.44	-9.27
	St. dev	0.79	0.58	0.46	0.31
Scenario 1	Max	-16.01	-12.30	-9.94	-6.92
	Min	-21.77	-16.63	-13.38	-9.24
	St. dev	0.73	0.55	0.44	0.30
Scenario 2	Max	-15.88	-12.09	-9.77	-6.71
	Min	-22.26	-16.91	-13.6	-9.28
	St. dev	0.76	0.58	0.46	0.31
Scenario 3	Max	-15.34	-11.67	-9.47	-6.43
	Min	-25.27	-19.12	-15.4	-10.35
	St. dev	1.27	0.94	0.75	0.49

Table 3.18: Maximum, minimum and standard deviation of the realized utility based on simulated data for alternative policy scenarios at the stochastic steady state under the assumption of no unobserved heterogeneity in preferences.

context of large-scale OLG models, since the set of models introduced in this subsection is already at the frontier of the literature in terms of state space size and computational complexity.

Table 3.19 displays the average ex-ante utility defined in Equation (3.53) for the eight identified groups of households. The tax rate and the benefit reduction aimed at obtaining a long-term balanced Social Security are again endogenously determined. The tax rate and the reductions in the benefits are virtually the same as those computed in the economy with homogeneous risk preferences. As we can see, the long-term increase in the payroll tax rate (Scenario 1) leads to the largest welfare gains across the low-risk aversion group if compared with other policy alternatives. Gains increase nearly uniformly as education increases: high-school graduates experience slightly lower gains compared to high-school dropouts (0.40% vs. 0.28%), while individuals with some college and college graduates ex-ante utility is respectively 1.02% and 1.35% higher than in the benchmark case. A similar pattern (but with opposite signs) is observed among the group with high-risk aversion: high-school graduates experience the biggest welfare loss relative to benchmark in this regime (-2.74%), followed by high-school graduates (-2.51%), individuals with some college education (-1.09%) and college graduates (-0.31%). The utility levels achieved in Scenario 2, which involves a cut in benefits, are uniformly lower than those in Scenario 1. Conditional on the risk-aversion level, changes relative to benchmark are monotonic in the educational attainment. The only two groups achieving welfare gains are the low risk-averse agents with some college education and college graduates, who experience gains relative to the benchmark equal to 0.63% and 1.27% respectively. The elimination of Social Security (Scenario 3) is the policy alternative that achieves the lowest level of utility, regardless of the risk-aversion or the educational attainment. Welfare losses relative to the benchmark are particularly significant among the high risk aversion group, ranging from -10.8% for high-school dropouts to -5.21% for college graduates. Similar patterns are observed for the low risk-aversion group, although losses are generally smaller in size. Table 3.20 shows that this policy regime achieves the highest dispersion in realized utility, strengthening the conclusions reached in the analysis with no unobserved heterogeneity.

Similarly to the case with no unobserved heterogeneity, results in Table 3.20 show that an economy with no Social Security is the least effective in transferring risk across different generations. Compared to the alternative policy scenarios incorporating Social Security, the standard deviation of the ex-ante realized utility is significantly higher. Similarly, the policy alternative with no Social Security achieves the highest and the lowest realized levels of ex-ante utility. From a risk-sharing perspective, the best policy is represented by a balanced Social Security achieved with an increase tax rate, like in the case with homogeneous preferences. The dispersion of the realized ex-ante utility is the lowest across the policy options, independent of the level of risk-aversion and the socio-economic group.

Policy	<HS		HS		Some College		College+	
	σ_l	σ_h	σ_l	σ_h	σ_l	σ_h	σ_l	σ_h
Benchmark	-37.95 (–%)	-10.74 (–%)	-32.23 (–%)	-7.3 (–%)	-28.53 (–%)	-5.48 (–%)	-22.91 (–%)	-3.26 (–%)
Scenario 1	-37.81 (0.40%)	-11.01 (-2.51%)	-32.15 (0.28%)	-7.51 (-2.74%)	-28.25 (1.02%)	-5.54 (-1.09%)	-22.6 (1.35%)	-3.27 (-0.31%)
Scenario 2	-37.98 (-0.05%)	-11.13 (-3.63%)	-32.24 (0.00%)	-7.57 (-3.56%)	-28.36 (0.63%)	-5.59 (-2.01%)	-22.62 (1.27%)	-3.28 (-0.61%)
Scenario 3	-38.46 (-1.32%)	-11.9 (-10.8%)	-32.6 (-1.12%)	-8.04 (-9.99%)	-28.71 (-0.6%)	-5.96 (-8.76%)	-22.7 (0.92%)	-3.43 (-5.21%)

Table 3.19: Households welfare obtained using cohorts' weights under the assumption of unobserved heterogeneity in risk-preferences across households, with $K = 2$ unobserved groups. The welfare measure used is based on Equation (3.53). Benchmark: Social Security, insolvent. Scenario 1: Social Security, long-term balanced (tax increase). Scenario 2: Social Security, long-term balanced (benefits reduction). Scenario 3: No Social Security.

3.8 Conclusion

In this chapter, I estimate the parameters necessary to conduct counterfactual analyses aimed at assessing the welfare implications of alternative policy regimes on ex ante heterogeneous households. The source of ex-ante heterogeneity is educational attainment, which is used as a proxy for socio-economic status and interacts significantly with important institutional features of the current Social Security system, including: i) with the rate at which retirement benefits replace labor income after retirement, and ii) mortality rates and life-expectancy. In addition, in this chapter I show how the algorithm developed in Chapter 2, based on the use of neural networks as policy and forecast function approximators, together with deep learning optimization techniques, is successful in computationally solving large-scale OLG models, with four or eight types of households living for eighty periods.

In terms of policy analysis, I consider a policy environment where Social Security reaches insolvency in the short-run as the benchmark scenario, consistent with current forecasts. I also examine three policy alternatives, including i) balancing social security utilizing a tax increase, or ii) balancing social security utilizing a reduction in benefits, and iii) the elimination of Social Security altogether. I analyze the impact of the alternative policies under two alternative assumptions about households preferences. In

Policy	Statistics	<HS		HS		Some College		College+	
		σ_l	σ_h	σ_l	σ_h	σ_l	σ_h	σ_l	σ_h
Benchmark	Max	-34.17	-8.4	-29.13	-5.76	-25.87	-4.35	-20.88	-2.62
	Min	-41.93	-13.77	-35.63	-9.4	-31.48	-7.02	-25.26	-4.18
	St. dev	0.94	0.63	0.77	0.42	0.66	0.3	0.51	0.17
Scenario 1	Max	-34.39	-9.02	-29.42	-6.12	-25.88	-4.51	-20.75	-2.66
	Min	-41.27	-13.5	-35.11	-9.22	-30.85	-6.8	-24.69	-4.0
	St. dev	0.9	0.59	0.74	0.39	0.65	0.29	0.51	0.17
Scenario 2	Max	-34.51	-8.8	-29.32	-5.99	-25.86	-4.44	-20.7	-2.63
	Min	-41.38	-13.74	-35.13	-9.38	-30.88	-6.92	-24.64	-4.07
	St. dev	0.91	0.64	0.77	0.43	0.66	0.32	0.52	0.18
Scenario 3	Max	-33.0	-8.27	-28.04	-5.63	-24.79	-4.21	-19.71	-2.45
	Min	-44.19	-17.02	-37.29	-11.52	-32.68	-8.46	-25.79	-4.83
	St. dev	1.43	1.13	1.18	0.75	1.01	0.54	0.77	0.30

Table 3.20: Maximum, minimum and standard deviation of the realized utility based on simulated data for alternative policy scenarios at the stochastic steady state under the assumption of unobserved heterogeneity in preferences with $K = 2$ groups.

the first, I assume that households in the economy are homogeneous in their risk preferences. In the second, households are treated as having different coefficients of relative risk aversion, which are estimated by applying the Expectation Maximization algorithm to a GMM estimator developed in Section 3.5.

The results presented in this chapter show that the elimination of Social Security would represent the worst policy option, with large welfare losses relative to the benchmark independently of the level of socio-economic status and little inter-generational risk-sharing. The variability in realized utility is significantly higher in this scenario as compared with any policy regime that maintains a Social Security system, regardless of the solvency status. Under both assumptions about risk preferences, the magnitude of the welfare losses relative to the benchmark is decreasing in the educational attainment, which is expected, given the progressive nature in which the current Social Security system computes benefits. When risk preferences are assumed to be homogeneous, welfare losses range from -4.9% for high-school dropouts to -1.2% for college graduates. When heterogeneous preferences are assumed, high risk-averse agents experience larger welfare losses as compared to low risk-averse agents.

With regard to the two policy alternatives that result in a long-term balanced Social Security system, the conclusions reached are sensitive to the underlying assumption made about households preferences. Both policies transfer consumption risk from older to younger cohorts, with the size of the transfer depending on the generosity of the retirement benefits. In addition, a balanced Social Security achieved through an increase in the payroll tax rate leads to the lowest dispersion in realized utility across generations, showing that a more generous retirement benefit system achieves the best inter-generational risk-sharing among the proposed policy options. Under the assumption of homogeneous preferences, the insolvent Social Security appears to be the best policy alternative for all households except for college graduates, who experience welfare gains of similar magnitude in the two long-term balanced policy alternatives. However, as compared to the elimination of Social Security, the welfare losses relative to the benchmark are smaller, ranging from -1.4% for high-school graduates to -0.1% for individuals with some college education. When heterogeneity in risk-preferences is introduced, the low risk aversion group experiences small welfare gains when taxes are raised, ranging from 0.3% for high-school graduates to 1.4% for college graduates. On the contrary, the high risk aversion group experiences losses interdependently on the socio-economic status. Therefore, our results indicate that selection of the coefficients of relative risk aversion plays an important role in assessing the welfare implications of the two long-term balanced scenarios.

While the results presented rely on welfare comparisons at the stochastic steady state, policymakers may need to consider accounting for transition costs in evaluating which alternative policy to implement. In both long-term balanced scenarios with Social Security, the stochastic steady state is reached after an average of 100 to 250 periods, depending on whether taxes are raised or benefits are reduced. In both policy alternatives, the balance of the Social Security trust fund is 20 to 40 times larger than the total retirement entitlements upon reaching the new stochastic steady state. Clearly, this sets an unrealistic level of public savings in the economy. In addition, the analysis presented suggests that welfare losses in the transitional dynamics are particularly marked for the policy alternative that raises taxes to achieve a long-term balanced Social Security system. This poses additional challenges for policymakers in redesigning Social Security, as current generations would need to bear the brunt of the cost of a restructured program.

In terms of future research, it would be interesting to incorporate elastic labor supply into this framework, to examine how the labor supply decision interacts with the consumption-savings decisions of households in a policy environment in which retirement benefits depend on the history of workers' earnings. It is easy to adapt the framework developed in Chapter 2 to incorporate endogenous labor supply. The forecast functions would need to be modified to account for the aggregate level of labor supply, while

policy functions for labor supplies would need to be introduced, but neural networks can easily tackle this numerical challenge. The residuals of the Euler equations linking current labor supply and consumption can then be used to optimize and numerically derive the labor policy functions in an iterative fashion similar to what we do for the consumption function.

Bibliography

Sule Alan, Orazio Attanasio, and Martin Browning. Estimating Euler equations with noisy data: Two exact GMM estimators. *Journal of Applied Econometrics*, 24(2):309–324, 3 2009. ISSN 08837252. doi: 10.1002/jae.1037.

Sule Alan, Martin Browning, and Mette Ejrnæs. Income and consumption: A micro semistructural analysis with pervasive heterogeneity. *Journal of Political Economy*, 126(5):1827–1864, 2018. ISSN 1537534X. doi: 10.1086/699186.

Sumru Altug and Robert A. Miller. Household Choices in Equilibrium. *Econometrica*, 58(3):543, 7 1990. doi: 10.2307/2938190.

Steve Ambler and Alain Paquet. Stochastic Depreciation and the Business Cycle. *International Economic Review*, 35(1):101–116, 1994.

Steffen Andersen, Glenn W Harrison, Morten I Lau, and E Elisabet Rutström. Eliciting Risk and Time Preferences Published by : The Econometric Society Stable URL : https://www.jstor.org/stable/40056458 to Econometrica. 76(3):583–618, 2008.

Peter Arcidiacono and Robert A. Miller. Conditional Choice Probability Estimation of Dynamic Discrete Choice Models With Unobserved Heterogeneity. *Econometrica*, 79(6):1823–1867, 2011. ISSN 0012-9682. doi: 10.3982/ecta7743.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. *35th International Conference on Machine Learning, ICML 2018*, 1: 390–418, 2018.

Orazio P. Attanasio and Hamish Low. Estimating Euler equations. *Review of Economic Dynamics*, 7(2): 406–435, 2004. ISSN 10942025. doi: 10.1016/j.red.2003.09.003.

Alan J. Auerbach and Laurence J. Kotlikoff. Evaluating Fiscal Policy with a Dynamic Simulation Model. *American Economic Review*, 77(2):49–55, 1987.

Marlon Azinovic, Luca Gaegauf, and Simon Scheidegger. Deep Equilibrium Nets. *SSRN Electronic Journal*, pages 1–51, 2019. ISSN 1556-5068. doi: 10.2139/ssrn.3393482.

Andrew R. Barron. Universal Approximation Bounds for Superpositions of a Sigmoidal Function. Technical Report 3, 1993.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, NIPS 2011*, pages 1–9, 2011.

Richard Blundell, Luigi Pistaferri, and Ian Preston. Consumption inequality and partial insurance. *American Economic Review*, 98(5):1887–1891, 2008. ISSN 00028282. doi: 10.1257/aer.98.5.1887.

Zvi Bodie and John Shoven. *Financial aspects of the United States pension system*. University of Chicago Press, 1983. ISBN 0226062813.

Barry Bosworth, Kan Zhang, and Gary Burtless. Later Retirement, Inequality in Old Age, and the Growing Gap in Longevity Between Rich and Poor. *The Brookings Institution*, pages 1–174, 2016.

John Bound, Arline T. Geronimus, Javier Rodriguez, and Timothy Waidmann. The Implications of Differential Trends in Mortality for Social Security Policy. 2015.

Marco Cagetti. Wealth accumulation over the life cycle and precautionary savings. *Journal of Business and Economic Statistics*, 21(3):339–353, 7 2003. ISSN 07350015. doi: 10.1198/073500103288619007.

Gary Chamberlain. Panel Data. In *Handbook of Econometrics*. 1984.

Nadav Cohen, Or Sharir, and Amnon Shashua. On the expressive power of deep learning: A tensor analysis. *Journal of Machine Learning Research*, 49(June):698–728, 2016. ISSN 15337928.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4):303–314, 1989. ISSN 09324194. doi: 10.1007/BF02551274.

Olivier Delalleau and Yoshua Bengio. Shallow vs. deep sum-product networks. *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, NIPS 2011*, pages 1–9, 2011.

Gabrielle Demange. On optimality in intergenerational risk sharing. *Economic Theory*, 20(1):1–27, 8 2002. ISSN 09382259. doi: 10.1007/s001990100199.

Peter A. Diamond. National Debt in a Neoclassical Growth Model: Comment. *American Economic Review*, 59(1):205–10, 1965. doi: 10.2307/1811114.

Peter A. Diamond. A Framework for Social Security Analysis. Technical report, 1977.

Chuong B. Do and Serafim Batzoglou. What is the expectation maximization algorithm? *Nature Biotechnology*, 26(8):897–899, 2008. ISSN 10870156. doi: 10.1038/nbt1406.

Victor Duarte. Machine Learning for Continuous Time Finance. Technical report, 2018.

Eric French. The Effects of Health, Wealth, Wages on Labour Supply and Retirement Behaviour. *Review of Economic Studies*, 2005.

Eric French and John B Jones. The Effects of Health Insurance and Self-Insurance on Retirement Behavior. *Econometrica*, 79(3):693–732, 4 2011. doi: 10.3982/ecta7560.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. *Journal of Machine Learning Research*, 9:249–256, 2010. ISSN 15324435.

Jean-Michel Grandmont. Temporary General Equilibrium Theory Published by : The Econometric Society. *Econometrica*, 45(3):535–572, 1977.

Alan L. Gustman and Thomas L. Steinmeier. Retirement in Dual-Career Families: A Structural Model. *Journal of Labor Economics*, 18(3):503–545, 7 2002. ISSN 0734-306X. doi: 10.1086/209968.

Alan L. Gustman and Thomas L. Steinmeier. Social security, pensions and retirement behaviour within the family. *Journal of Applied Econometrics*, 19(6):723–737, 10 2004. ISSN 08837252. doi: 10.1002/jae.753.

Alan L. Gustman and Thomas L. Steinmeier. How Changes in Social Security Affect Retirement Trends. *SSRN Electronic Journal*, 12 2011. doi: 10.2139/ssrn.1094978.

Gary D. Hansen. Indivisible labor and the business cycle. *Journal of Monetary Economics*, 16(3):309–327, 1985. ISSN 03043932. doi: 10.1016/0304-3932(85)90039-X.

Daniel Harenberg and Alexander Ludwig. Idiosyncratic Risk, Aggregate Risk, and the Welfare Effects of Social Security. *International Economic Review*, 60(2):661–692, 5 2019. ISSN 14682354. doi: 10.1111/iere.12365.

Jasmina Hasanhodzic and Laurence J Kotlikoff. Generational Risk: Is It a Big Deal? Simulating an 80-Period OLG Model with Aggregate Shocks. Technical report, 2018. URL http://www.nber.org/papers/w19179.

Jasmina Hasanhodzic and Laurence J. Kotlikoff. A Study of Generational Risk in Life-Cycle Models . 2019.

Espen Henriksen and Stephen E Spear. Corrigendum to "Endogenous market incompleteness without market frictions: Dynamic suboptimality of competitive equilibrium in multiperiod overlapping generations economies". pages 1–3, 2019.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. ISSN 08936080. doi: 10.1016/0893-6080(89)90020-8.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks*, 3(5):551–560, 1990. ISSN 08936080. doi: 10.1016/0893-6080(90)90005-6.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *32nd International Conference on Machine Learning, ICML 2015*, 1:448–456, 2015.

Kenneth L. Judd. Projection methods for solving aggregate growth models. *Journal of Economic Theory*, 58(2):410–452, 1992. ISSN 10957235. doi: 10.1016/0022-0531(92)90061-L.

Kenneth L. Judd, Lilia Maliar, Serguei Maliar, and Rafael Valero. Smolyak method for solving dynamic economic models: Lagrange interpolation, anisotropic grid and adaptive domain. *Journal of Economic Dynamics and Control*, 2014. ISSN 01651889. doi: 10.1016/j.jedc.2014.03.003.

David L. Kelly and Jamsheed Shorish. Stability of Functional Rational Expectations Equilibria. *Journal of Economic Theory*, 95(2):215–250, 2000. ISSN 00220531. doi: 10.1006/jeth.2000.2690.

Eungsik Kim. Preference Heterogeneity, Aggregate Risk and the Welfare Effects of Social Security. Technical report, 2018.

Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pages 1–15, 2015.

Paul Klein. Using the generalized Schur form to solve a multivariate linear rational expectations model. *Journal of Economic Dynamics & Control*, 2000.

Andrey Kolmogorov. On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. Technical Report 5, 1957.

Dirk Krueger and Felix Kubler. Computing equilibrium in OLG models with stochastic production. *Journal of Economic Dynamics and Control*, 2004. ISSN 01651889. doi: 10.1016/S0165-1889(03)00111-8.

Dirk Krueger and Felix Kubler. Pareto-Improving Social Security Reform when Financial Markets Are Incomplete!? Technical Report 3, 2006.

Per Krusell and Anthony A. Smith, Jr. Income and Wealth Heterogeneity in the Macroeconomy. *Journal of Political Economy*, 106(5):867–896, 10 1998. ISSN 0022-3808. doi: 10.1086/250034. URL https://www.journals.uchicago.edu/doi/10.1086/250034.

Per Krusell, Lee E. Ohanian, José Víctor Ríos-Rull, and Giovanni L. Violante. Capital-skill complementarity and inequality: A macroeconomic analysis. *Econometrica*, 68(5):1029–1053, 2000. ISSN 00129682. doi: 10.1111/1468-0262.00150.

Yann Lecun, Bottou Leon, Genevieve B. Orr, and Klaus-Robert Muller. Efficient BackProp. 1998.

Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. ISSN 14764687. doi: 10.1038/nature14539.

Lilia Maliar and Serguei Maliar. Merging simulation and projection approaches to solve high-dimensional problems with an application to a new Keynesian model. *Quantitative Economics*, 6(1):1–47, 2015. ISSN 1759-7331. doi: 10.3982/qe364.

Lilia Maliar, Serguei Maliar, and Pablo Winant. Will Artificial Intelligence Replace Computational Economists Any Time Soon? *CEPR Discussion Paper Series*, 14024, 2019.

Hrushikesh Mhaskar and Tomaso Poggio. Deep vs. shallow networks : An approximation theory perspective. 8 2016. URL http://arxiv.org/abs/1608.03287.

Hrushikesh Mhaskar, Qianli Liao, and Tomaso Poggio. When and why are deep networks better than shallow ones? *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, pages 2343–2349, 2017.

Volodymyr Mnih, David Silver, and Martin Riedmiller. Playing Atari with Deep Reinforcement Learning. pages 1–9, 2013.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015. ISSN 14764687. doi: 10.1038/nature14236.

Guido Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. *Advances in Neural Information Processing Systems*, 4(January):2924–2932, 2014. ISSN 10495258.

Genevieve B. Orr and Klaus-Robert Müller. *Neural Networks: Tricks of the Trade*. 1998. ISBN 3-540-65311-2. doi: 10.1007/3-540-68339-9{_}34.

Razvan Pascanu, Guido Montúfar, and Yoshua Bengio. On the number of response regions of deep feedforward networks with piecewise linear activations. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, pages 1–17, 2014.

Edward C. Prescott. *Theory ahead of business-cycle measurement*, volume 25. 1986.

Michael Reiter. Solving OLG Models with Many Cohorts, Asset Choice and Large Shocks. 2015.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning Representations by Back-Propagating Errors. *Letters to Nature*, 323(9):533–536, 1986. ISSN 02632241. doi: 10.1016/j.measurement.2017.09.025.

Thiago Serra, Christian Tjandraatmadja, and Srikumar Ramalingam. Bounding and counting linear regions of deep neural networks. *35th International Conference on Machine Learning, ICML 2018*, 10:7243–7261, 2018.

Sergey Smolyak. Quadrature and Interpolation Formulas for Tensor Products of Certain Classes of Functions. *Uspekhi Mat. Nauk*, 163(5):1042–1045, 1963.

Stephen E. Spear. Existence and local uniqueness of functional rational expectations equilibria in dynamic economic models. *Journal of Economic Theory*, 44(1):124–155, 1988. ISSN 10957235. doi: 10.1016/0022-0531(88)90099-3.

Kjetil Storesletten, Christopher I. Telmer, and Amir Yaron. Asset pricing with idiosyncratic risk and overlapping generations. *Review of Economic Dynamics*, 10(4):519–548, 10 2007. ISSN 10942025. doi: 10.1016/j.red.2007.02.004.

George Tauchen. Finite state markov-chain approximations to univariate and vector autoregressions. *Economics Letters*, 20(2):177–181, 1986. ISSN 01651765. doi: 10.1016/0165-1765(86)90168-0.

Hilary Waldron. Trends in Mortality Differentials and Life Expectancy for Male Social Security-Covered Workers, by Average Relative Earnings. Technical report, 2007.

Hilary Waldron. Mortality differentials by lifetime earnings decile: implications for evaluations of proposed Social Security law changes. *Social security bulletin*, 73(1):1–37, 2013. ISSN 0037-7910. URL http://www.ncbi.nlm.nih.gov/pubmed/23687740.

A.1 Numerical Solution of the Deterministic OLG Model

The solution algorithm described below is composed of two steps. First, steady-state equilibrium is computed for the deterministic economy. Second, I use a log-linearized approximation of the dynamics of the steady state to compute the linearized policy functions for an economy in a neighborhood of the steady state.

A.1.1 Steady-State

At the steady state of the deterministic economy, the Euler equation for household belonging to group i and cohort n can be expressed as follows:

$$(c_{in})^{-\sigma_i} = \beta_i (c_{in+1})^{-\sigma_i} r \quad 1 \leq n \leq A-1, \quad \forall i$$

where r represents the steady-state value of the interest rate, determined marginal value of capital for the representative firm. Agents are born with no capital endowment. In the last period of their life, they leave no bequests to future generations, so that their consumption can be expressed by $c_{iA} = w_{iA}(1-\tau) + rk_{iA} + ss_{in}$. It is convenient to define the cash-on-hand of household (i, n) as follows:

$$\omega_{in} = wn_{in}(1-\tau) + rk_{in} + b_{in} + \mathbf{1}(n \geq R_i)ss_{in} \quad \forall 1 \leq n \leq A, \quad \forall i \in \mathcal{I}$$

where w represents the steady state value of the wage per efficiency unit. This implies that I obtain the following set of Euler equations:

$$\begin{aligned}
(\omega_{i1} - k_{i2})^{-\sigma_i} &= \beta_i (\omega_{i2} + k_{i2}r - k_{i3})^{-\sigma_i} r && j = 1 \\
(\omega_{in} + k_{in}r - k_{in+1})^{-\sigma_i} &= \beta_i (\omega_{in+1} + k_{in+1}r - k_{in+2})^{-\sigma_i} r && n = 2, \cdots, A-1 \\
(\omega_{iA-1} + k_{iA-1}r - k_{iA})^{-\sigma_i} &= \beta_i (\omega_{iA} + k_{iA}r)^{-\sigma_i} r && j = A
\end{aligned}$$

The numerical solution adopted to find the steady state relies on fixed-point iteration, in which I iteratively update the values of aggregate capital and aggregate bequest. The algorithm can be described as follows. First, rearranging the Euler equation for the households belonging to cohort $A-1$, and I obtain:

$$k_{iA} = \frac{(\omega_{iA-1} + k_{iA-1}r)(\beta_i r)^{\frac{1}{\sigma_i}} - w_{iA}}{(\beta_i r)^{\frac{1}{\sigma_i}} + r} \tag{54}$$

As we can see from Equation (54), k_{iA} is linear in k_{iA-1}. Therefore, I can rewrite it in more explicit terms as follows:

$$
\begin{aligned}
k_{iA} &= \delta_{iA} + \gamma_{iA} k_{iA-1} \\
\delta_{iA} &= \frac{\omega_{iA-1} \left(\beta_i r\right)^{\frac{1}{\sigma}} - \omega_{iA}}{\left(\beta_i r\right)^{\frac{1}{\sigma_i}} + r} \\
\gamma_{iA} &= \frac{\left(\beta_i r\right)^{\frac{1}{\sigma_i}} r}{\left(\beta_i r\right)^{\frac{1}{\sigma_i}} + r}
\end{aligned}
$$

It is easy to see that it is possible to express the future capital holdings of household (i, n), k_{in+1}, as a linear function of the current capital holdings k_{in}:

$$
k_{in+1} = \delta_{in+1} + \gamma_{in+1} k_{in} \tag{55}
$$

Replacing Equation (55) in the Euler equations, I obtain the following relationship between k_{in+1} and k_{in}:

$$
k_{in+1} = \frac{\left(\omega_{in} + k_{in} r\right) \left(\beta_i r\right)^{\frac{1}{\sigma_i}} - \left(\omega_{in+1} - \delta_{in+2}\right)}{\left(r - \gamma_{in+2}\right) + \left(\beta_i r\right)^{\frac{1}{\sigma_i}}}
$$

In light of this, I can express the parameters ruling the linear relationship between current and future capital in a recursive manner:

$$
\begin{aligned}
\delta_{in+1} &= \frac{\omega_{in} \left(\beta_i r\right)^{\frac{1}{\sigma_i}} - \left(\omega_{in+1} - \delta_{in+2}\right)}{\left(r - \gamma_{in+2}\right) + \left(\beta r\right)^{\frac{1}{\sigma_i}}} \\
\gamma_{in+1} &= \frac{r \left(\beta_i r\right)^{\frac{1}{\sigma_i}}}{\left(r - \gamma_{in+2}\right) + \left(\beta_i r\right)^{\frac{1}{\sigma_i}}}, \quad n = 2, \cdots, A-2
\end{aligned}
$$

Finally, for $n = 1$, we have the following relationship:

$$
\begin{aligned}
\delta_{i2} &= \frac{\omega_{i1} \left(\beta_i r\right)^{\frac{1}{\sigma_i}} - \left(\omega_{i2} - \delta_{i3}\right)}{\left(r - \gamma_{i3}\right) + \left(\beta_i r\right)^{\frac{1}{\sigma_i}}} \\
\gamma_{i2} &= 0
\end{aligned}
$$

Given that the sequence of $(\gamma_{in+1}, \delta_{in+1})$ depend on the steady-state value of the interest rate, wage (both depending on aggregate capital) and aggregate bequests (depending on the entire wealth distribution), the fixed-point algorithm needs to iterate over different values of aggregate capital and bequests. To see this

more clearly, we have that the cash-on-hand variable depends on:

$$\omega_{in} = n_{in}w(z,K)(1-\tau) + k_{in}r(z,K) + \mathbf{1}(n \geq R_i)\theta_i\overline{e}_{in} + b_{in} \quad \text{where}$$
$$\overline{e}_{in} = \frac{1}{R_i - 1}\sum_{j=1}^{R_i-1} n_{in}w(z,K)$$
$$b_{in} = \eta_{in}\sum_{i=1}^{I}\sum_{n=1}^{A}\left(r(z,K)k_{in} + w(z,K)n_{in}(1-\tau)\right)P_{in}(1-\mu_{in})$$

As we can see, households bequests depend on the entire distribution of capital through the heterogeneous mortality rate. Therefore, the algorithm needs to iterate over the value of bequest as well. Given aggregate capital and bequest, we can compute the rental rate of capital, the wage, and therefore obtain the cash-on-hand variable ω_{in} for each household. Once we have that, we can determine the entire distribution of capital allocations, and we can check whether it is consistent with the original guess for aggregate capital. If it is, then the algorithm has found a solution for the steady state; if not, we update the guess for aggregate capital and bequest, and iterate until convergence is achieved. The fixed-point algorithm can be described as follows:

1. Initialize a guess for K^0 and b^0.
2. For $k \geq 0$, given (K^k, b^k), compute r^k, and ω_{in}^0 for each household, and derive the sequence of γ_{in} and δ_{in}.
3. Given the sequence of γ_{in} and δ_{in}, compute the implied capital holdings for each household, k_{in}^{k+1} using Equation (55).
4. Compute the new value of aggregate capital K_{new} and bequest B_{new} based on the new distribution of capital holdings.
5. Check:
   - If $\max\{\frac{|K^k - K_{new}|}{|1+K^k|}, \frac{|B^k - B_{new}|}{|1+B^k|}\} \geq \varepsilon$, then update the values of aggregate variables using the dampening parameter $\lambda = 0.9$:

     $$K^{k+1} = \lambda_K K^k + (1-\lambda_K)K_{new}$$
     $$B^{k+1} = \lambda_b b^k + (1-\lambda_b)B_{new}$$

     and go back to 2.
   - If $\max\{\frac{|K^k - K_{new}|}{|1+K^k|}, \frac{|b^k - b_{new}|}{|1+b^k|}\} < \varepsilon$, then convergence has been achieved. Exit the loop.

A.1.2 Local Linear Dynamics around the Steady State

I use the method proposed by Klein [2000] in order to develop the numerical approximation of the deterministic dynamics of the model around the steady state. As standard in macroeconomics, the method relies on the first-order approximation of the dynamics of the system through a linear system of equations.

By using the $I \cdot (A-1)$ Euler equations, the $I \cdot A$ budget constraints and the $I \cdot (A-1)$ equations ruling the law of motion of the Social Security contribution, the equilibrium conditions can be summarized as follows:

$$\mathbf{H}(\mathbf{c}', \mathbf{k}', \mathbf{s}', \mathbf{c}, \mathbf{k}, \mathbf{s}) = \mathbf{0} \tag{56}$$

I define as $\mathbf{x} = (\mathbf{c}, \mathbf{k}, \mathbf{s})$ as a $AI + 2I \cdot (A-1)$ dimensional vector of households savings, consumption allocations, and average lifetime earnings. By linearizing (56) around the steady state, I obtain the following first order approximation of the dynamic system representing the economy:

$$\begin{aligned} &\mathbf{A}\left(\mathbf{x}' - \mathbf{x}^*\right) = \mathbf{B}\left(\mathbf{x} - \mathbf{x}^*\right) \quad \text{where} \\ &\mathbf{A} = \mathbf{H}_{\mathbf{x}'}\left(\mathbf{x}^*, \mathbf{x}^*\right) \\ &\mathbf{B} = \mathbf{H}_{\mathbf{x}}\left(\mathbf{x}^*, \mathbf{x}^*\right) \\ &\mathbf{A}, \mathbf{B} \in \mathbb{R}^{(3AI-2)\times(3AI-2)} \end{aligned} \tag{57}$$

In order to obtain the linearized consumption policy functions, I perform the generalized Schur decomposition of the matrix pencil $(\mathbf{A}, \mathbf{B})$, as described by Klein [2000]. Given $(\mathbf{A}, \mathbf{B})$, the generalized eigenvalues λ_i can be expressed as follows :

$$\mathbf{B}\mathbf{x} = \lambda_i \mathbf{A}\mathbf{x}, \quad \lambda_i \in \mathbb{C} \tag{58}$$

The complex generalized Schur form a regular matrix pencil guarantees the existence of of a quadruple of matrices $(\mathbf{Q}, \mathbf{Z}, \mathbf{S}, \mathbf{T})$:

$$\begin{aligned} &\mathbf{Q}\mathbf{A}\mathbf{Z} = \mathbf{S} \\ &\mathbf{Q}\mathbf{B}\mathbf{Z} = \mathbf{T} \\ &\lambda(\mathbf{A}, \mathbf{B}) = \left\{ \frac{t_{ii}}{s_{ii}} : \; s_{ii} \neq 0 \right\} \end{aligned}$$

The matrices $\mathbf{S}$ and $\mathbf{T}$ are complex-valued upper-triangular, while $\mathbf{Q}$ and $\mathbf{Z}$ are unitary complex matrices. Given that $\mathbf{A}$ is not invertible, I assign to the missing generalized eigenvalues the value of infinity. The

generalized eigenvalue is table if $|\lambda_i| < 1$, and unstable if $|\lambda_i| > 1$. A value of infinity would correspond to an unstable eigenvalue.

Now, I verify that the number of unstable generalized eigenvalues is equal to the number of consumption allocations of households, in this case IA. Alternatively, I check that the number stable generalized eigenvalues is equal to the number of state variables, in this case $2I(A-1)$. If the number number of stable eigenvalues is equal to the dimension of the state space, then it is possible to characterize uniquely a linearized equilibrium around the steady state, obeying the following equations:

$$\mathbf{c} = -\mathbf{Z}_{21}\mathbf{Z}_{11}^{-1}\begin{bmatrix}\mathbf{k}\\ \mathbf{s}\end{bmatrix}$$

$$\begin{bmatrix}\mathbf{k}'\\ \mathbf{s}'\end{bmatrix} = \mathbf{Z}_{11}\mathbf{S}_{11}^{-1}\mathbf{T}_{11}\mathbf{Z}_{11}^{-1}\begin{bmatrix}\mathbf{k}\\ \mathbf{s}\end{bmatrix}$$

In the following paragraphs, I show how to derive the matrices $\mathbf{B}$ and $\mathbf{A}$.

Euler Equations By taking the logarithm of the Euler equations, I obtain:

$$-\sigma_i \log c_{in} = \log \beta_i + \log r' - \sigma_i \log c'_{in+1}$$
$$\forall i = 1, \cdots, I$$
$$\forall n = 1, \cdots, A-1$$

By taking a first-order Taylor approximation around the steady-state, I can write:

$$-\sigma_i(\log c^*_{in} + \frac{1}{c^*_{in}}(c_{in} - c^*_{in})) = \log \beta_i + \log r + \frac{1}{r}\nabla_{\mathbf{k}} r(\mathbf{k}^*)\cdot(\mathbf{k}-\mathbf{k}^*) - \sigma_i \log c^*_{in} + \frac{1}{c^*_{in}}(c'_{in+1} - c^*_{in+1})$$

At the steady state, we have:

$$-\sigma_i c^*_{in} = \log \beta_i + \log r - \sigma_i c^*_{in+1}$$

Given this relationship, I can simplify:

$$-\frac{\sigma_i}{c^*_{in}}(c_{in} - c^*_{in}) = \frac{1}{r}\nabla_{\mathbf{k}} r'(\mathbf{k}^*)\cdot(\mathbf{k}'-\mathbf{k}^*) - \frac{\sigma_i}{c^*_{in+1}}\left(c'_{in+1} - c_{in+1}\right)$$

Budget Constraints At equilibrium, budget constraints need to hold with equality. Therefore, with the assumption of no-bequest left once age A is reached, we have the following budget constraints:

$$
\begin{aligned}
& c_{in} = w_{in} + k_{in} r - \mathbf{1}(j < A) k'_{in+1} + b_{in} + \theta_i s_{in} \mathbf{1}(n \geq R_i) \\
& \forall i = 1, \cdots, I \\
& \forall j = 1, \cdots, A
\end{aligned}
$$

A first-order Taylor approximations around the steady state implies that:

$$
\begin{aligned}
c_{in} - c^*_{in} = & \left(\nabla_{\mathbf{k}} w_{in}(\mathbf{k}^*) + k^*_{in} \nabla_{\mathbf{k}} r(\mathbf{k}^*) + \nabla_{\mathbf{k}} b_{in}(\mathbf{k}^*)\right) \cdot (\mathbf{k} - \mathbf{k}^*) + \\
& + (k_{in} - k^*_{in}) r - \mathbf{1}(j < A) \left(k'_{in+1} - k^*_{ij+1}\right) + \mathbf{1}(j \geq R_i) \theta_i \left(s_{in} - s^*_{in}\right)
\end{aligned}
$$

We define $\mathbf{p} \in \mathbb{R}_+^{MA}$ as the vector representing the population distribution and $\mathbf{d}\mathbb{R}_+^{M(A-1)}$ as the vector representing the distribution of premature deaths, where deaths are computed as:

$$
d_{in} = P_{in} \mu_{in}
$$

For what concerns the wage, the gradient with respect to households capital holdings is:

$$
\nabla_{\mathbf{k}} w(\mathbf{k}^*) = \frac{\partial w}{\partial K} \nabla_{\mathbf{k}} K = \alpha(1-\alpha) z K^{*\alpha-1} N^{-\alpha} \mathbf{p}
$$

Turning to the interest rate, its gradient with respect to households capital holding is:

$$
\nabla_{\mathbf{k}} r(\mathbf{k}^*) = \frac{\partial r}{\partial K} \nabla_{\mathbf{k}} K = \alpha(\alpha-1) z K^{*\alpha-2} N^{1-\alpha} \mathbf{p}
$$

Finally, for each household (i, n), the gradient of the households bequests with respect to capital holding is:

$$
\begin{aligned}
\nabla_{\mathbf{k}} b_{in}(\mathbf{k}^*) &= \nabla_{\mathbf{k}} \left(\eta_{in} b(\mathbf{k}^*)\right) \\
&= \eta_{in} \left[r\mathbf{d} + \left(\nabla_K r(\mathbf{k}^*) \sum_{i=1}^{I} \sum_{n=1}^{A} d_{in} k^*_{in} + \nabla_K w(\mathbf{k}^*) \sum_{i=1}^{I} \sum_{n=1}^{A} d_{in} n_{in} \right) \mathbf{p} \right]
\end{aligned}
$$

Social Security Contributions Average-indexed lifetime earnings evolve according to the following law of motion:

$$
s'_{ij} = \begin{cases} \frac{j-1}{j} s_{ij-1} + \frac{1}{j} w_{ij} & \forall i,\ 1 < j \leq R_i \\ s_{ij-1} & \forall i,\ R_i < j \leq A \end{cases}
$$

Therefore, we have that:

$$
s'_{ij} - s^*_{ij} = \frac{j-1}{j} \left(s_{ij-1} - s^*_{ij-1}\right) + \frac{1}{j} \nabla_{\mathbf{k}} w_{ij} \cdot (\mathbf{k} - \mathbf{k}^*) \quad 1 < j \leq R_i
$$