Flexible Retirement and Optimal Taxation

Abdoulaye Ndiaye*

This version: November 3, 2017

Job Market Paper

Please find the latest version at:

https://www.dropbox.com/s/2a1n5qvczi303u1/Ndiaye_2017_FR-OT.pdf

Abstract

This paper studies optimal income taxes and retirement benefits in a life-cycle model with an intensive margin of labor supply and an endogenous retirement age. The government insures and redistributes resources across individuals who privately observe persistent shocks to their productivity. In this environment, the optimal labor tax is hump-shaped in age, unlike in existing models with no endogenous retirement choice, in which the optimal tax is everywhere increasing. Because of the retirement margin, the total Frisch elasticity of labor supply increases with age. This elasticity effect flattens the labor tax for old workers relative to the model without an extensive margin. In addition, as high-productivity workers retire later than low-productivity workers, the distribution of productivity in the labor force features, over time, a higher mean and lower variance than in the general population. This novel composition effect pushes for a labor tax that declines for old workers. Optimal policy balances these effects with the insurance benefits of taxation, yielding the hump-shape in tax rates. In numerical simulations, the optimum achieves sizable welfare gains that approximately optimal age-dependent taxes fail to capture under the current US Social Security system. Yet, an optimal combination of age-dependent linear taxes with increasing-in-age delayed retirement credits generates welfare gains that are close to those from the optimum.

JEL classification: H21, H55, J26

Keywords: Retirement, Optimal Taxation, Social Security, Continuous-Time, Optimal Stopping

*Department of Economics, Northwestern University. email: ndiaye@u.northwestern.edu. I am grateful to Guido Lorenzoni, Alessandro Pavan, Larry Christiano and Mariacristina de Nardi for their invaluable advice and guidance. I would like to thank Gideon Bornstein, Gaby Cugat, Mike Golosov, Jean-Baptiste Michau, Paul Mohnen, Jordan Norris, Giorgio Primiceri, Stefanie Stantcheva, Bruno Strulovici, Yuta Takahashi and Northwestern Macro Lunch Seminar and Grad Students Seminar participants for numerous discussions and suggestions.

1 Introduction

Planning for retirement and choosing when to retire are important decisions for most people. Workers pay payroll taxes on their labor income to be eligible for retirement benefits, save and invest in retirement pensions, and choose whether to retire early or delay retirement.

There is strong evidence that the tax and Social Security (SS) system affects retirement behavior. Labor income taxes affect people's labor supply, which adjusts both through the number of daily hours worked—an intensive margin—and through the timing of retirement—an extensive margin. Capital income taxes on retirement savings and the value of retirement pensions determine income after retirement. In turn, retirement behavior affects the income distribution and the duration of retirement, which are key inputs of the tax and retirement benefits system.

The goal of this paper is to investigate the consequences of flexible retirement for the optimal design of income taxes and retirement benefits over a person's life-cycle. Since Mirrlees (1971) and Diamond and Mirrlees (1978), the vast majority of optimal tax theory assumes that retirement occurs at an exogenous date instead of being an endogenous decision. Recent literature analyzes the consequences of endogenous retirement for optimal tax and pension systems. Until now, the analysis has been restricted to economies in which agents experience a permanent shock at birth in a static setting (cf. Michau (2014) and Shourideh and Troshkin (2015)). In realistic life-cycle settings where earnings risk is gradually resolved over time, the implications of flexible retirement for the pattern of optimal income taxes and retirement benefits are yet to be understood.

The main question my paper addresses is the following: How is the optimal tax and retirement benefit system altered when we acknowledge that people choose when to retire? I derive an analytical characterization of optimal history-dependent policies and describe the economic forces that shape their patterns over the life-cycle. Finally, I calibrate the model to the US economy and ask how the welfare gains can be achieved by simple policies. I study two such policy experiments: a reform of the tax system, and a joint reform of the tax system and SS system.

I jointly determine optimal tax and retirement benefits and the resulting retirement decisions in a dynamic life-cycle model in which workers adjust their labor supply through working hours and the timing of retirement. Individuals live for T years, work, consume, and choose when to retire. During their working years, labor income is the product of intensive labor supply and productivity, which evolves as a persistent Markov process. A fixed utility cost of staying in the labor market

creates a non-convexity in the disutility of labor. This fixed cost has important implications for the retirement decision. First, workers adjust their working hours continuously until they irreversibly exit the labor force, then their hours of work discretely drop to zero. Second, when productivity is public information, low-productivity agents efficiently retire earlier than high-productivity agents. Third, there is an option value of waiting for higher earnings before retirement. As a consequence, at the retirement age, the marginal utility of staying in the labor market is lower than the marginal utility from not working. This option value decreases over time as the value of waiting for higher earnings vanishes at T.

The government chooses consumption, output and retirement age in order to maximize social welfare. As in the standard Mirrlees (1971) model, individual productivity and labor effort are privately observed by the workers. Therefore, the government's goal is to design a dynamic mechanism that is incentive compatible. I describe the distortions to the optimal retirement decision and analyze constrained efficient allocations by studying the wedges, or implicit marginal taxes, and consumption after retirement.

Methodologically, because of a large number of incentive constraints, the government's problem cannot be treated using a direct approach. A First-Order Approach (FOA) provides analytic tractability in characterizing allocations. This approach relaxes the problem by imposing only a subset of incentive constraints. I follow the implementation of this approach by Farhi and Werning (2013) in the context of optimal taxation. I formulate the model in continuous-time for a sharper characterization of the retirement decision.

In the analytical part of the paper, I determine how optimal policies evolve over time and I provide some intuition for the numerical results. I show through two effects, that optimal policies imply a labor tax that is hump-shaped in age under flexible retirement, while it is increasing in age under exogenous retirement. First, despite the intensive Frisch elasticity of labor supply being constant, the total Frisch elasticity increases with age. This is because of the retirement margin and the decreasing option value of waiting for higher earnings before retirement. This elasticity effect implies that the optimal labor tax is flatter in age relative to the model without an extensive margin. Second, in the constrained optimum, agents with a history of high productivity shocks are provided with higher retirement consumption and are incentivized to retire later than agents with a history of low productivity shocks. Therefore, through selection, the labor force becomes increasingly more productive than the pool of all potential workers of the same age. When the forces

of this composition effect are stronger than the increase of variance in productivity that occurs with age, the older labor force is also more equal in productivity than the general population. Setting a decreasing-in-age labor tax for old workers increases the efficiency of the intensive labor supply of these high-productivity workers. These two effects, balanced with the government's motive for increasing the level of insurance with age in the standard dynamic Mirrlees model, imply that the optimal labor tax is hump-shaped in age. I consider two reference economies that allow me to decompose the effects of elasticity and composition of the labor force on the pattern of the labor wedge: one without an extensive margin of labor supply and one where endogenous retirement is accounted for, by the planner, but is distributed uniformly accross workers at each age.

In the quantitative part of the paper, I exploit a recursive formulation of the FOA to numerically illustrate these effects. I calibrate the model to the US economy under several specifications of the fixed utility cost of staying in the labor market. I find that the average implicit labor tax is hump-shaped in age under flexible retirement, while it is increasing with age under exogenous retirement. I compute the welfare gains from maintaining the current SS system and moving from the existing US tax code to the linear age-dependent labor and capital taxes that mimic optimal policies. I find that this reform brings modest welfare gains under flexible retirement, while it achieves the bulk of welfare gains when retirement is exogenous. The modest welfare gains are because the current SS system does not provide appropriate incentives for delayed retirement like the optimal system does. I find that this tax reform, when coupled with a simple SS reform that increases the delayed retirement credits, can generate sizeable welfare gains. Because of the fixed cost that is incurred both by low-productivity workers and high-productivity workers, most of the agents who delay retirement as a result of the SS reform are highly productive. The decreasing-in-age labor tax for old workers increases the efficiency of the intensive labor supply of these high-productivity agents and delivers welfare gains from the age-dependent linear labor tax that is hump-shaped in age. These calibrations suggest that when the endogeneity of retirement is accounted for, introducing age-dependency into the tax code alone is not enough, and one needs to reform the SS system as well in order to capture the bulk of welfare gains from optimal policies.

Related Literature A large empirical literature documents the relationship between retirement behavior and tax and SS systems around the world. Gruber and Wise (1998), Gruber and Wise (2002), and their accompanying volumes of comparative studies document that over much of the second half of the 20th century, disincentives to continue working have created a trend towards early

retirement. This trend has shown signs of reversing in the mid-2000s because of a combination of factors including longevity, gender composition, social norms, tax provision and SS reforms.

This paper builds on the insights of the early non-linear income taxation literature. Mirrlees (1971) develops the theory and optimal tax formulas that Saez (2001) links to estimated elasticities. Albanesi and Sleet (2006) develop a dynamic Mirrleesian model and focus on the implementation of the optimal allocations with a restricted set of instruments. The subsequent literature develops the dynamic Mirrleesian model with persistent productivity shocks (Farhi and Werning (2013)) and focuses on the evolution of implicit labor taxes. Golosov *et al.* (2016) disentangle the motives of insurance and redistribution. Stantcheva (2017) incorporates endogenous human capital acquisition. Makris and Pavan (2017) investigate the effects of learning-by-doing on optimal taxes. A comprehensive survey of the dynamic Mirrleesian literature can be found in Golosov *et al.* (2006) and in Golosov and Tsyvinski (2015). All these papers assume an exogenous retirement age and find that, as inequality in hourly wages increases with age, the average labor tax should increase with age.

The model considered in this paper is similar to the one in Farhi and Werning (2013), augmented with an endogenous retirement age. I find that accounting for an endogenous retirement age, the average labor tax should be hump-shaped in age. I also find that introducing age-dependency into the tax code alone is not enough and delayed retirement needs to be incentivized through a reform of the SS system.

The first analysis of retirement and taxation comes from Diamond and Mirrlees (1978). In their framework, workers are subject to disability shocks. All able workers choose the same retirement age and, at any given age, they all share the same productivity. Hence, their retirement decisions do not generate the composition effect, which is at the heart of my analysis. Also, Diamond and Mirrlees (1978) do not allow for an intensive margin of labor supply. Other papers study optimal taxation with an extensive margin of labor supply in a static framework (Saez (2002), Jacquet *et al.* (2013),Gomes *et al.* (2017), Rothschild and Scheuer (2013)).

Recent literature has analyzed optimal tax and retirement benefits and the timing of retirement. Michau (2014), Choné and Laroque (2014), Cremer *et al.* (2004) and Shourideh and Troshkin (2015) introduce the retirement margin in the analysis of optimal tax and retirement benefit systems. In these papers, there is a permanent shock that deterministically pins down the whole history of productivity, as in a static setting. My paper analyzes a setting in which earnings risk is gradually

resolved over time and is therefore able to describe the lifetime evolution of the optimal labor income tax.

This paper is also connected to the literature on age-dependent taxation. In the Ramsey tradition, Erosa and Gervais (2002) focus on linear taxes in an economy without uncertainty within a cohort, and find that when the intensive elasticity of labor supply varies over an individual's lifetime, optimal tax rates are age-dependent. Conesa *et al.* (2009) postulate a specification of preferences that are isoelastic in leisure instead of labor. As a consequence, the elasticity of labor supply is high when labor is low. In their model, a low labor supply corresponds to the labor supply of older workers; therefore, they find decreasing the labor tax with age to be optimal. Assuming preferences that feature an increasing intensive elasticity parameter, Karabarbounis (2016) finds that the optimal labor, within the class of the Heathcote *et al.* (2014) tax function, is hump-shaped in age. The result of my paper does not rely on these particular specifications of preferences. I keep the intensive Frisch elasticity fixed, so that the information structure and increasing inequality in earnings are responsible for the increasing profile of the labor tax at the beginning of work life, and the retirement margin and the selection of the labor force are responsible for their decreasing profile in old age. In a recent contribution, Heathcote *et al.* (2017) analyze the optimal degree of progressivity of age-dependent tax systems. Considering a productivity process that is on average increasing in age and has increasing variance in age, they find that the optimal degree of progressivity in the tax system is U-shaped in age. In the Mirrlees approach, Weinzierl (2011) justifies the rising age profile of wages as a reason to increase the labor tax with age but limits his sample to the ages 30 to 59. Farhi and Werning (2013) find that a rising variance of wages justifies increasing the linear labor tax with age and that such an age-dependent tax achieves nearly the entirety of welfare gains from the second-best. When one accounts for flexible retirement, the labor tax should be on average hump-shaped in age and age-dependent taxes alone do not achieve significant welfare gains unless they are complemented by SS reform.

As for the methodological approach, this paper builds on the dynamic mechanism design and optimal contracting literatures. Pavan *et al.* (2014) develop the First Order Approach that simplifies the dynamic mechanism design problem. Bergemann and Strack (2015) adapt the theory in continuous-time. Strack and Kruse (2013) studies pure stopping problems under private information. My paper analyzes the design of optimal mechanisms for optimal stopping problems with stochastic controls.

Outline The remainder of the paper is structured as follows. Section 2 sets up the framework of the model and defines the planning problem. Section 3 solves the first-best planning problem and highlights features of the retirement decision within the full information benchmark. Section 4 develops a recursive formulation of the second-best planning problem, solves for the optimal policies and retirement decision, and discusses the parameters and forces that shape them. Section 5 presents the numerical and welfare analyses. Section 6 presents three extensions of the model: (i) non-separable preferences in consumption and labor, (ii) workers with uncertain lifetimes, and (iii) productivity-dependent fixed costs of staying in the labor market. Section 7 concludes.

2 Model Setup

In this section, I describe an economy in which workers are ex-ante heterogeneous in productivity, experience idiosyncratic productivity shocks over their lifetime, and adjust their labor supply through flexible working hours and the timing of their retirement.

Productivity, Technology, and Preferences Consider a continuous-time economy populated by a continuum of agents who live until age T. At each time t, each agent privately observes the realization of his[1] current labor productivity $\theta_t \in (0, +\infty)$. Agents provide $l_t \geq 0$ units of labor at time t at a wage rate equal to their productivity and earn gross income $y_t = \theta_t l_t$.

At time $t = 0$, initial productivity $\theta_0 \in (0, +\infty)$ is drawn from a distribution F with density f. A standard Brownian Motion $B = \{B_t, \mathcal{F}_t; 0 \leq t \leq T\}$ on $(\Omega, \mathcal{F}, \mathcal{P})$ drives the productivity shocks in future periods. A history of productivities $(\theta^t) = \{\theta_s\}_{s\in[0,t]}$ is a sequence of realizations of the productivity process that evolves according to the law of motion

$$\frac{d\theta_t}{\theta_t} = \mu_t dt + \sigma_t dB_t. \tag{1}$$

By Ito's lemma, the real constants $\mu_t - \frac{1}{2}\sigma_t^2$ and σ_t are respectively the drift and volatility of log-productivity. When the drift and volatility are independent of time, productivity is a Geometric Brownian Motion (GBM) and log-productivity is the continuous-time limit of a random walk.

Agents have time-separable preferences over consumption $\{c_t\}_{0\leq t\leq T}$ and labor $\{l_t\}_{0\leq t\leq T}$ pro-

[1] Throughout the text, I use the male pronoun for an agent and the female pronoun for the planner.

cesses that are progressively measurable with respect to the filtration $\mathcal{F}_t$.[2] When an agent is working, $(l_t > 0)$, he incurs a flow utility cost of staying in the labor market denoted by a deterministic function of age $\phi(t)$; and his current period utility is $u(c_t, l_t) - \phi(t)$, where u is increasing in consumption, decreasing in labor, twice continuously differentiable, and concave.

Utility along the intensive margin is separable in consumption and labor and isoelastic in labor:

$$u(c_t, l_t) = u(c_t) - \kappa \frac{l_t^{1+\frac{1}{\varepsilon}}}{1+\frac{1}{\varepsilon}}$$

where $\varepsilon > 0$ is the intensive Frisch elasticity of labor supply. In Section 6, I extend the analysis to preferences that are non-separable in consumption and labor.

The fixed utility cost of staying in the labor market can be thought of as the utility cost of commuting time, work-related consumption costs, or taste for leisure. I write it in units of utils for tractability. This fixed cost creates a non-convexity in the disutility of work as agents prefer no work to a few hours of work. As in French (2005) and Rogerson and Wallenius (2013), these non-convexities trigger retirement at some point in the worker's life. In Section 6, I extend the analysis to fixed utility costs that depend both on age and current productivity $\phi_t(\theta_t)$.

Retirement, $l_t = 0$, is an irreversible decision. Define a stopping time $\mathscr{T}_R \in \mathcal{T}$,[3] after which a retired agent provides zero labor effort and does not incur the fixed utility cost. After retirement, an agent's utility in each period is $u(c_t, 0)$. I define the retirement age as the age at which an individual chooses to exit the labor force forever[4]—which the model allows to differ from the age at which an individual chooses to start claiming the Old-Age, Survivors and Disability Insurance

[2] Consumption $c_t(\theta^t)$ and labor $l_t(\theta^t)$ depend on the whole history of productivities until time t. In the text, I drop the realisations θ^t when referring to $\mathcal{F}_t$.-measurable processes $\{c_t, y_t\}$ to simplify the notation.

[3] A random variable $\mathscr{T}_R$ is a stopping time if $\{\mathscr{T}_R \leq t\} \in \mathcal{F}_t, \forall t \geq 0$. Intuitively, this definition means that at any time t, one must know whether retirement has occurred or not.

[4] The irreversible retirement assumption is motivated by empirical and theoretical reasons. Rogerson and Wallenius (2013) find empirical evidence in the Current Population Survey data that retirement occurs as abrupt transitions from full-time to little or no work in the US. By age 70, the age by which individuals should start claiming SS benefits, 75% of men report working zero hours. In addition, this assumption can actually be easily relaxed. The main predictions of the model remain unchanged if this paper allows for retirees to return to the labor market at a lower wage. A more involved theoretical reason is in Grochulski and Zhang (2016). In a setting similar to Sannikov (2008) that allows for agents to put in zero labor effort temporarily, they find that when the utility of consumption is unbounded below, workers almost surely provide positive labor efforts as the planner can threaten to provide arbitrarily low utility to shirking agents with zero consumption. I use a logarithmic utility of consumption in most of my analysis and it satisfies these assumptions. In my setting the fixed cost of staying in the labor market has to be paid even if labor effort was allowed to be suspended temporarily; therefore, retirement would be triggered. This utility function, coupled with the fixed cost, is another justification for an interior labor effort $l_t > 0$ before irreversible retirement.

(OASDI) benefits.[5]

Planning Problem Preferences over consumption and labor $\{c_t, l_t\}$ and retirement decisions $\{\mathscr{T}_R\}$ are summarized by an agent's indirect utility at time zero:

$$v(\theta_0) \equiv \mathrm{E}\Big\{ \int_0^{\mathscr{T}_R} e^{-\rho t}[u(c_t, l_t) - \phi(t)]dt + \int_{\mathscr{T}_R}^{T} e^{-\rho t} u(c_t, 0)dt \;\Big|\; \theta_0 \Big\} \tag{2}$$

in which ρ is the rate of time preference. A utilitarian planner chooses incentive compatible (IC) allocations to maximize social welfare:

$$\max_{\{c_t, l_t, v(\theta_0), \mathscr{T}_R\}} \int_0^{\infty} v(\theta_0) dF(\theta_0) \tag{3}$$

subject to the law of motion of productivity (1), the definition of indirect utility (2) and an intertemporal resource constraint. For simplicity, I work in partial equilibrium and the planner can save aggregate resources in a small open economy and borrow at a net rate of return r. I study the planner's problem for a single cohort in isolation and abstract from intergenerational redistribution issues.[6] The planner's resource constraint is therefore:

$$\mathrm{E}\Big\{ \int_0^{T} e^{-rt} c_t dt \Big\} + G \leq \mathrm{E}\Big\{ \int_0^{\mathscr{T}_R} e^{-rt} \theta_t l_t dt \Big\}. \tag{4}$$

The left-hand side includes exogenous government spending G[7] and the cost of providing lifetime consumption to agents. The right-hand side is the sum of the net present value of income y_t generated by workers until they retire. Because of a law of large numbers, the aggregate resource constraint is the expectation over the histories of productivities (θ^t).

3 The First-Best Planning Problem

This section solves the planning problem with full information. I highlight features of the optimal retirement decision that are absent in existing models with no endogenous retirement choice, but

[5] In a decentralized economy, workers can actually claim SS benefits whenever they want and their optimal retirement benefits system are computed according to the history of their earnings. Because I work with allocations directly in this primal approach, the SS benefits are implicit in the model.

[6] Given that I study insurance and redistribution across one cohort, time is equivalent to age for my cohort.

[7] G can capture many sources of exogenous government revenues and expenses as well as intergenerational transfers to or from another cohort etc.

have important implications for optimal policy.

Let the rate of time preference equal the rate of return of government savings, $\rho = r$. From the intertemporal Euler equation, there is perfect insurance against productivity shocks and consumption is the same across all histories: $u'(c_t(\theta^t)) = \lambda$, with λ the multiplier on the planner's resource constraint (4). When it is optimal to work, the marginal rate of transformation of labor into consumption is the wage rate θ_t. Therefore labor supply satisfies $\kappa l_t^{\frac{1}{\varepsilon}} = \lambda\theta_t$. With full information, consumption is smoothed and more productive agents work more hours and produce more output. To maximize social welfare, the planner maximizes total resources available in the economy and makes high-productivity workers retire later than low-productivity workers, as long as the fixed cost of staying in the labor market for high-productivity workers is not too high compared to that of low-productivity workers. The following propositon confirms that it is indeed the case.

Proposition 1. *(First-best retirement decision) There exists a time-dependent deterministic productivity threshold $\theta_R^{fb}(t)$ such that retirement occurs if and only if productivity falls below it: $\mathscr{T}_R^{fb} = \inf\{t; \theta_t \leq \theta_R^{fb}(t)\}$.*

The proof is in Appendix A. This proposition means that the planner balances the need to induce the highly productive (high earning) agents to continue working with the need to avoid the fixed utility cost for less productive (low earning) workers. In the first-best, it is therefore optimal to set productivity cut-offs below which retirement occurs. To understand the determinants and lifetime evolution of these cut-offs, I consider the case in which agents are risk neutral.

The Risk-Neutral Case To qualify results further, I now consider agents who are risk neutral in consumption, so that $u(c_t) = c_t$. Consumption is not pinned down by the Euler equation. I eliminate consumption from the planner's problem by replacing the resource constraint into the planner's social welfare function:

$$w \equiv \max_{\mathscr{T}_R} \mathrm{E}\Big\{ \int_0^{\mathscr{T}_\mathscr{R}} e^{-\rho t}[\theta_t l_t^{fb} - \kappa \frac{(l_t^{fb})^{1+\frac{1}{\varepsilon}}}{1+\frac{1}{\varepsilon}} - \phi(t)]dt \Big\} - G \tag{5}$$

subject to the law of motion of productivity (1). Normalizing government spending to zero, $G = 0$, and replacing the first-best labor allocations using the optimality condition $\kappa(l_t^{fb})^{\frac{1}{\varepsilon}} = \theta_t$, the social

welfare function $w(\theta_t, t)$ satisfies the following Hamilton-Jacobi-Bellman (HJB) equation:

$$0 = \max\left\{-w(\theta,t),\ -\rho w(\theta,t)+\frac{\theta^{1+\varepsilon}}{\kappa^{\varepsilon}(1+\varepsilon)}-\phi(t)+(\mu_t\theta)\partial_\theta w(\theta,t)+\frac{\sigma_t^2\theta^2}{2}\partial_{\theta\theta}w(\theta,t)+\partial_t w(\theta,t)\right\}. \quad (6)$$

The terms to the right of $-\rho w(\theta, t)$ consist of the marginal social value of labor minus the fixed cost and derivatives of social welfare with respect to time and productivity.

Now consider the case of productivity that evolves according to a GBM i.e μ_t and σ_t are respectively constants μ and σ. I show that even when the fixed cost is a constant $\phi(t) = \phi$, there is an option value of waiting for higher productivity shocks before retirement. In addition, this option value decreases over time. Therefore, even when the fixed cost is constant over time, the elasticity over the retirement margin increases over time. Hence the total Frisch elasticity increases over time, despite the intensive Frisch elasticity and the fixed cost being constant. The following corollary summarizes this result in terms of the retirement thresholds $\theta_R^{fb}(t)$.

Corollary 1. *(Option value of continued work vs retirement)*

1. *Suppose ϕ is constant and productivity is a GBM. Denote θ^* the unique level of productivity below which the marginal value of labor is less than the fixed utility cost of work, that is, $\theta^* l^{fb}(\theta^*) - \kappa\frac{(l^{fb}(\theta^*))^{1+\frac{1}{\varepsilon}}}{1+\frac{1}{\varepsilon}} = \phi$. Then for all $t < T$, $\theta_R^{fb}(t) \leq \theta^*$ and the marginal social value of continued work is negative, i.e, $\theta_R^{fb}(t) l^{fb}(\theta_R^{fb}(t)) - \kappa\frac{(l^{fb}(\theta_R^{fb}(t)))^{1+\frac{1}{\varepsilon}}}{1+\frac{1}{\varepsilon}} - \phi \leq 0$.*
2. *The retirement threshold function $\theta_R^{fb}(t)$ is increasing in t. In addition, $\lim_{t\to T}\theta_R^{fb}(t) = \theta^*$.*

Point 1 of the corollary states that retirement occurs below a productivity level at which it would be efficient not to work in a static environment. This creates an option value of waiting for higher productivity shocks and higher earnings before retirement that is not present in models with permanent productivity shocks like Michau (2014) or Shourideh and Troshkin (2015). Working today instead of retiring preserves the option of retiring later at a higher wage, hence the term "option value" of work. Indeed, when there is no uncertainty on future earnings, the marginal value of labor is equal to the fixed utility cost of work at retirement, and the option value is zero. In practice, this option value is negative at retirement. Rust (1989), Lazear and Moore (1988) and Stock and Wise (1988) estimate structural models of retirement with uncertain earnings and find that people continue to work at any age as long as the expected present utility value of continuing work is greater or equal to the expected present value of immediate retirement.

Point 2 of the corollary states that the option value of continued work decreases over time as the horizon shortens. Therefore, the total Frisch elasticity increases over time, despite the intensive Frisch elasticity and the fixed cost being constant. The option value of continued work vanishes at the end of the horizon, and only then is the irreversible retirement decision similar to a static participation decision and the marginal value of labor equal to the fixed utility cost of work.

To develop some intuition, let us consider the infinite horizon limit $T \to \infty$. In this case, the HJB equation is time-homogeneous and the retirement threshold is independent of time, θ_R^{fb}. The proof in Appendix A proceeds similarly to Leland (1994) by decomposing the value of social welfare into

$$w(\theta) = \underbrace{A\theta^{1+\varepsilon} - \frac{\phi}{\rho}}_{\text{social value of working forever (SVWF)}} - \underbrace{(\frac{\theta_R^{fb}}{\theta})^x}_{\text{discounting at retirement } \mathrm{E}[e^{-\rho \mathscr{T}_R^{fb}}|\theta]} \Big[\underbrace{A(\theta_R^{fb})^{1+\varepsilon} - \frac{\phi}{\rho}}_{\text{SVWF starting at the retirement threshold}} \Big] \tag{7}$$

where

$$A = \frac{1}{\kappa^{\varepsilon}(1+\varepsilon)[\rho - (1+\varepsilon)(\mu + \frac{\sigma^2}{2}\varepsilon)]} \tag{8}$$

and $x(\rho, \mu, \sigma)$ is a positive constant defined in the Appendix A. The value of social welfare $w(\theta)$ is given by the value of lifetime utility of output if the agent were to work forever minus the value of lifetime utility of output if he were to work forever at the optimal retirement threshold discounted by the expected value of the discount factor at retirement. This value is zero at retirement. From a smooth pasting argument as in Dixit (1993), the value of its derivative is also zero at retirement. This gives an explicit value of the threshold:

$$\theta_R^{fb} = \Big(\frac{\phi}{\rho} \frac{x}{A(1+\varepsilon+x)}\Big)^{\frac{1}{\varepsilon}}. \tag{9}$$

Now, θ_R^{fb} is increasing in the fixed cost ϕ.[8] Agents retire earlier when their fixed cost is large. In addition, $\theta^* = (\phi\kappa^{\varepsilon}(1+\varepsilon))^{\frac{1}{\varepsilon}}$. It can be deduced that $\theta_R^{fb} < \theta^*$ since $\frac{\rho-(1+\varepsilon)(\mu+\frac{\sigma^2}{2}\varepsilon)}{\rho} < 1$ and $\frac{(x)}{(1+\varepsilon+x)} < 1$. The marginal social value of continued work is negative at retirement.

In summary, the solution of the first-best planning problem generates the following insights about the implications of optimal retirement: Low-productivity agents retire earlier than high-productivity agents. There is an option value of waiting for higher earnings before retiring. Lastly, the total Frisch elasticity increases over time, despite the intensive Frisch elasticity and the fixed

[8] For convergence of net present values, I assume that $\rho > \mu > \sigma^2\varepsilon/2$ in the proof in the Appendix A.

cost being constant. When the planner cannot observe productivity, this allocation is not achievable with constant consumption as any agent would be better off retiring immediately.

4 The Second-Best Planning Problem

This section studies the second-best problem in which productivity and its evolution is private information to the planner. I start by setting up the planning problem with full IC constraints. Then, I relax the incentive problem using the First Order Approach (FOA) procedure developed in Farhi and Werning (2013) and I incorporate the retirement decision. Finally, through a redefinition of the state space, I write a recursive formulation of the FOA.

4.1 Incentive Compatibility

In the second-best problem, both the agents and the planner observe consumption $\{c_t\}$, retirement status $\mathscr{T}_R$ and income from work $\{y_t\}$. However, the planner does not observe $\{\theta_t\}$, and therefore does not observe labor $\{l_t = y_t/\theta_t\}$ either. As a result, the planner needs to incentivize the agents with dynamic contracts.

A contract is a consumption process $\{c_t\}$ and a stochastic retirement time $\mathscr{T}_R$ adapted to the filtration generated by $\{y_t\}$.[9] By the revelation principle, a contract is a mapping from any reported process of productivities $\{\tilde{\theta}_t\}$ to a triplet $\{\tilde{c}_t, \tilde{y}_t, \tilde{\mathscr{T}}_R\}$ of processes adapted to the filtration generated by $\{\tilde{\theta}_t\}$. It specifies how much consumption is given to the agents, how much output the agents should produce, and whether they should retire or continue to work at any time. An allocation is IC if it is the outcome of a contract in which it is optimal for the agent to truthfully reveal his true productivity process $\{\theta_t\}$. In other words, for all reporting strategy $\{\sigma_t\{\theta_s\}_{s\in[0,t]}\}$, $\mathrm{E}\{v(\theta_0)\} \geq \mathrm{E}^{\sigma}\{v(\sigma(\theta_0))\}$, where E^{σ} is the expectation over the paths generated by reports. The planner commits to a contract at time zero. In particular, the contract is not renegotiable.

After retirement, the incentive problem stops since the agent does not need to be incentivized to work. Therefore, the planner does not need to distort consumption decisions after retirement.

Lemma 1. *Suppose $r = \rho$ and u is strictly concave in consumption. For any allocation that solves the planner's second-best problem, consumption is constant after retirement.*

[9] The planner's objective is concave and the optimal contract cannot be strictly improved by randomization over allocations and stopping times.

The result is intuitive: since output is zero after retirement, there is no information for the planner to learn about the agent's real productivity after retirement. Since there is no incentive constraint after retirement, the problem is one of full insurance. The Euler equation holds intertemporally, and the marginal utility of consumption at $l = 0$ is equalized cross-sectionally. Since u_c is strictly decreasing, it follows that consumption is constant after retirement.

This lemma implies that consumption after retirement only depends on the history of productivities until retirement. However, it also allows for a jump in consumption "at" retirement. Because of this possibility, I denote by "$c_{\mathscr{T}_R^+}$" consumption after retirement.[10]

Following this lemma, allocations before retirement and the retirement decision pin down retirement consumption through the resource constraint. In order to characterize allocations before retirement, I now relax the planner's incentive constraints.

4.2 Recursive Formulation of the Planning Problem

Given constant consumption after retirement, an agent's ex-ante indirect utility, or promised utility, given consumption and labor $\{c_t, l_t = y_t/\theta_t\}$ and retirement $\{\mathscr{T}_R\}$ is

$$v_0 = \mathrm{E}\Big\{ \int_0^{\mathscr{T}_R} e^{-\rho t}[u(c_t, \frac{y_t}{\theta_t}) - \phi(t)]dt + \int_{\mathscr{T}_R}^{T} e^{-\rho t} u(c_{\mathscr{T}_R^+}, 0)dt \Big\}. \tag{10}$$

Denote $g(t) \equiv \int_t^T e^{-\rho(s-t)}ds = \frac{1}{\rho}(1 - e^{-\rho(T-t)})$ a shorthand to represent by how much the utility of constant consumption is discounted at retirement at t. Promised utility at time t before retirement is then

$$v_t = \mathrm{E}\Big\{ \int_t^{\mathscr{T}_R} e^{-\rho(s-t)}[u(c_s, \frac{y_s}{\theta_s}) - \phi(s)]ds + e^{-\rho(\mathscr{T}_R - t)} u(c_{\mathscr{T}_R^+}, 0) g(\mathscr{T}_R) \Big| \mathcal{F}_t \Big\} \tag{11}$$

and the feasibility constraint is

$$\mathrm{E}\Big\{ \int_0^{\mathscr{T}_R} e^{-rt} c_t dt + e^{-\rho \mathscr{T}_R} c_{\mathscr{T}_R^+} g(\mathscr{T}_R) \Big\} \leq \mathrm{E}\Big\{ \int_0^{\mathscr{T}_R} e^{-rt} y_t dt \Big\}. \tag{12}$$

[10] The fact that consumption is constant after retirement in this setting is linked to different forces than those in Sannikov (2014). In that model, agents have their consumption distorted and compensations optimally delayed after retirement. The planner continues to observe positive post-termination output, which itself depends on the persistent labor effort of the agent; in my model, output is zero after retirement.

By duality, it is equivalent for the planner to maximize ex-ante promised utility (10) than it is for her to minimize the cost of providing allocations:

$$K_0(v) = \min_{\{c,y,\mathscr{T}_R\}} \mathrm{E}\Big\{ \int_0^{\mathscr{T}_R} e^{-\rho t}(c_t - y_t)dt + e^{-\rho \mathscr{T}_R} c_{\mathscr{T}_R^+} g(\mathscr{T}_R) \Big\} \tag{13}$$

subject a minimum promised utility $v_0 \geq v$, full incentive compatibility and the law of motion of productivity (1).

The First Order Approach (FOA) relaxes the IC constraints by restricting attention to local deviations. An IC mechanism must be immune to such deviations. As a result, the sensitivity of promised utility with respect to reports, denoted by $\Delta_t \equiv \partial_\theta v_t$, satisfies an envelope condition on the agent's optimal reporting problem. I discuss the optimal reporting problem in detail in Appendix A.

The FOA has been implemented by Kapička (2013), Farhi and Werning (2013) and Golosov *et al.* (2016) in the context of optimal taxation and by Williams (2011) and Sannikov (2014) in the context of optimal contracting in continuous-time. It is a necessary condition for an allocation to be IC. In general, it is not a sufficient condition to characterize full incentive compatibilty.[11] In the numerical analysis, I verify ex-post that the allocations obtained from the FOA satisfy full incentive compatibility using a method developed by Farhi and Werning (2013) that does not require solving for the fully incentive compatible mechanism. I continue the recursive formulation of the problem and reparametrize the state space in a simpler form. The lemma below derives the law of motion of promised utility and its sensitivity and allows me to solve the problem recursively.

Lemma 2. *(Law of motion of promised utility and the sensitivity process)*

1. *The law of motion of promised utility is*

$$dv_t = (\rho v_t - u(c_t, \frac{y_t}{\theta_t}) + \phi(t))dt + \theta_t \Delta_t \sigma_t dB_t \tag{14}$$

with the boundary condition

$$v_o = v.$$

[11] Nevertheless, it gives a lower bound on the cost of providing a given promised utility to the agents.

2. (FOA) The law of motion of the sensitivity process $\Delta_t \equiv \partial_\theta v_t$ is

$$d\Delta_t = \Big[\Big(\rho - \mu_t\Big)\Delta_t - u_\theta(c_t, \frac{y_t}{\theta_t}) - \sigma_{\Delta,t}\sigma_t\Big]dt + \sigma_{\Delta,t}\sigma_t dB_t \tag{15}$$

with the boundary condition

$$\Delta_0 = \arg\min_{\Delta} K_0(v, \Delta).$$

Point 1 of this lemma states that the drift of promised utility is the discounted flow utility. More importantly it highlights that the volatility of promised utility is controlled by the sensitivity process. The boundary condition is the promise-keeping constraint. Point 2 of the lemma characterizes how the sensitivity with respect to reports is linked to allocations in an incentive compatible mechanism, i.e. the evolution of informational rents.[12] The term u_θ constitutes the rent in the static Mirrlees model, while the term $\sigma_{\Delta,s}\sigma_t$ is a dynamic rent that summarizes an agent's advance information about his future productivity profile. The term $\mu\Delta_s$ captures how a misreport today affects the planner's perceived distribution of productivities in the future. The boundary condition ensures that the initial sensitivity is chosen to minimize the ex-ante cost of providing promised utility v. The proof is in Appendix A.

These recursive formulations allow me to analyze the relaxed planning problem. Promised utility v_t, its sensitivity with respect to reports Δ_t, time t and current productivity θ_t can be used as state variables of the recursive formulation. At retirement, promised utility $v_{\mathscr{T}_R}$ is provided with a constant consumption $c_{\mathscr{T}_R^+}$, so $v_{\mathscr{T}_R} = g(\mathscr{T}_R)u(c_{\mathscr{T}_R^+}, 0)$ and $c_{\mathscr{T}_R^+} = u_{l=0}^{-1}\Big(\frac{v_{\mathscr{T}_R}}{g(\mathscr{T}_R)}\Big)$ in which $u_{l=0}^{-1}$ is the inverse function of $u(c, 0)$. At each time t the planner's problem is to minimize the cost:

$$K(v, \Delta, \theta, t) = \min_{\{c\},\{y\},\sigma_\Delta,,\mathscr{T}_R} \mathrm{E}\Big\{\int_t^{\mathscr{T}_R} e^{-\rho(s-t)}(c_s - y_s)ds + e^{-\rho(\mathscr{T}_R - t)}g(\mathscr{T}_R)u_{l=0}^{-1}\Big(\frac{v_{\mathscr{T}_R}}{g(\mathscr{T}_R)}\Big)\Big\} \tag{16}$$

subject to the law of motion of productivity (1), the law of motion of promised utility (14) and the law of motion of the sensitivity process (15).

In what follows, I work for tractability with dual variables of (v_t, Δ_t) which are derivatives of the cost function with respect to these state variables. I introduce the co-states $\lambda_t = K_v$ and $\gamma_t = K_\Delta$. The economic intuition behind these variables is that they represent the marginal change in the cost of providing allocations when promised utility v_t or respectively its sensitivity Δ_t are

[12] Informational rents are rents the high-productivity agents derive from having information on their types that is not available to the planner.

marginally increased.[13]

4.3 Optimal Policies

4.3.1 Wedges, Retirement Consumption, and Distortions of The Retirement Decision

The approach to solving the planner's problem by finding the allocations that maximize her objective is called the primal approach.[14] To characterize the planner's optimum it is useful to define some wedges that capture distortions in the constrained optimal allocation relative to the first best.

Definition 1. The labor wedge (or intratemporal wedge) τ^L on workers is the gap between the marginal rate of substitution and the marginal rate of transformation between consumption and labor before retirement.

$$\tau_t^L \equiv 1 + \frac{\frac{1}{\theta_t} u_l(c_t, \frac{y_t}{\theta_t})}{u_c(c_t, \frac{y_t}{\theta_t})} \tag{17}$$

The capital wedge (or intertemporal wedge) at time t and horizon s is the difference between the expected marginal rate of intertemporal substitution between time t and time $t+s$ and the return on savings.

$$\tau_{t,s}^K \equiv 1 - e^{-(\rho - r)s} \frac{u_c(c_t, \frac{y_t}{\theta_t})}{\mathrm{E_t}\Big\{ u_c(c_{t+s}, \frac{y_{t+s}}{\theta_{t+s}}) \Big| \mathcal{F}_t \Big\}} \tag{18}$$

The intertemporal wedge at time t is the marginal intertemporal wedge between t and $t+dt$ i.e. $\tau_t^K = \frac{d\tau_{t,s}^K}{ds}\Big|_{s=0}$.

A positive wedge on labor means that labor is distorted downwards. The capital wedge represents the deviation from the Euler equation. These wedges have been the focus of the dynamic taxation literature. In addition to these wedges, I am interested in consumption after retirement and its net present value:

$$\{c_t\}_{\{\mathscr{T}_R < t \leq T\}} \qquad \mathrm{E}\Big\{ \int_{\mathscr{T}_R}^{T} e^{-r(t-\mathscr{T}_R)} c_t dt \Big| \mathcal{F}_{\mathscr{T}_R} \Big\} \tag{19}$$

and the percentage change, if any, in consumption before and after retirement, which I denote $\frac{\Delta c_{\mathscr{T}_R^+}}{c_{\mathscr{T}_R^-}}$ with an abuse of notation. Finally, I compare the second-best retirement rule $\mathscr{T}_R^{sb}$ to the

[13] Because of the Pontryagin Maximum Principle, (See Bismut (1973)) this method of working directly with the Lagrangians of the problem makes the problem tractable.

[14] As is well known, there can be several policies that implement the planner's optimal allocations.

first-best retirement rule $\mathscr{T}_R^{fb}$, which is summarized by the threshold function $\theta_R^{fb}(t)$ in the separable utility case, to analyze the distortions to the retirement decision.[15]

4.3.2 Optimal Retirement Policy

Since after retirement consumption is constant and labor effort is zero, promised utility is not sensitive to the reports after retirement. The endogenous retirement boundary is $\mathscr{T}_R^{sb} = \inf\{t;\ \Delta(\theta^t) = 0\}$. For incentive compatibility, given the same past history of productivity, promised utility is higher for higher reports, so $\partial_\theta v = \Delta \geq 0$. The sensitivity process starts at a positive value defined by $\Delta_0 = \arg\min_\Delta K_0(v, \Delta)$, follows the law of motion (15) until it hits zero at which point retirement is triggered.

The second-best retirement decision is more complex than the first-best one and depends on the whole history of productivities through $\Delta(\theta^t)$.[16] In Section 6, I show in a risk-neutral case with a progressive redistributive motive for the government, that the second-best retirement rule is determined by thresholds as in the first-best. In that case, low-productivity agents retire earlier than high-productivity agents.

To build intuition on why agents with a history of low productivity shocks retire earlier than agents with a history of high productivity shocks in the risk-averse case, consider first order conditions for consumption c_t, output y_t and the variance of sensitivity $\sigma_{\Delta,t}$ in the planner's problem:

$$[c_t]:\ \lambda_t = \frac{1}{u'(c_t)}, \quad [y_t]:\ \frac{\tau_t^L}{1-\tau_t^L} = -(1+\frac{1}{\varepsilon})\frac{\gamma_t}{\lambda_t}\frac{1}{\theta_t},\ \text{and}\ [\sigma_{\Delta,t}]:\ \sigma_{\Delta,t} = \frac{\gamma_t\sigma^{-1} - K_{v\Delta}\theta_t\Delta_t - K_{\Delta\theta}\theta_t}{K_{\Delta\Delta}}.$$

These first order conditions determine consumption, output and wedges as a function of the Lagrangians λ_t, γ_t. In particular, λ_t is the inverse of marginal utility of consumption and γ_t links the marginal utility of consumption to the labor wedge, the intensive Frisch elasticity, and current

[15] At each age, the planner compares the expected value of continued work against the expected value of retiring today taking incentives into account. Corollary 1 implies that there is no simple "marginal benefit = marginal cost" equation that holds in the first-best and defines a retirement wedge because of the option value of continued work. This is unlike in Michau (2014) or Shourideh and Troshkin (2015) who define the retirement margin as the deviation from the retirement age that equalizes the marginal value of labor with the fixed utility cost of work. In my setting, the marginal value of labor is optimally lower than the fixed utility cost of work at retirement in the first-best. One can define a retirement margin using equation (7) and the *value matching* and *smooth pasting* conditions that define the retirement thresholds. However one needs to assume first that the optimal retirement rule is a cut-off rule at the second-best, which in general is not the case. Even if so, the resulting expression does not provide more intuition than directly comparing the optimal retirement rules.

[16] The retirement boundary is the optimal exercise boundary of an exotic American option with stochastic dividends. Its derivation in the space of states (θ^t) is non-trivial.

productivity. By definition, γ_t is the marginal change in the planner's cost of providing allocations with respect to the sensitivity Δ. When Δ is larger, it hits the retirement boundary $\{\Delta = 0\}$ later, and the expected retirement age is delayed. Agents work longer and the cost of providing allocations for a given promised utility is lower. This means that γ starts at zero ($\gamma_0 = 0$) and takes negative values for all $t > 0$, $\gamma_t \leq 0$. The process $\gamma(v_t, \Delta_t, \theta_t, t)$ is defined for Δ_t non-zero. Denote by the same symbol γ_t, the extension of the process to the whole space of productivity histories. Since labor effort jumps discontinuously from a positive value to zero at retirement because of the fixed cost ϕ_t, γ_t jumps from a negative value to zero at retirement. In other words, because of the fixed cost, the *super contact condition*[17] does not hold at the retirement boundary.

Replacing the first order condition on c_t in the law of motion of Δ yields

$$d\Delta_t = \Big[\Big(\rho - \mu_t\Big)\Delta_t - \Big(\frac{1-\tau_t^L}{\lambda_t}\Big)^{1+\varepsilon}\Big(\frac{\theta_t}{\kappa}\Big)^{\varepsilon} - \sigma_{\Delta,t}\sigma_t\Big]dt + \sigma_{\Delta,t}\sigma_t dB_t. \tag{20}$$

Consider two agents with histories $\{\hat{\theta}^t\} \neq \{\theta^t\}$ such that $\hat{\theta}_t = \theta_t$ and $\hat{\tau}_t^L = \tau_t^L$. From the first order condition for y_t, these states are those for which the corresponding (γ, λ) are on a given line at time t, $\gamma_t = \alpha\lambda_t$. Assuming the volatility of the sensitivity σ_Δ is small, from (20) one can see that the agent for whom λ_t is lower has a more negative drift for Δ_t, and therefore has an earlier expected retirement age. Equivalently, this is the agent for whom consumption today is lower by the FOC on c_t. Respectively, the agent for whom γ_t is lower has an earlier expected retirement age. This is reminescent of the retirement rule in the first best in Proposition 1 but in a setting in which past history matters through the level of promised marginal utility of consumption. Agents who have a history of low productivity shocks, and who have lower consumption, retire earlier than the agents who have a history of high productivity shocks. In the second-best, current productivity alone does not determine the retirement decision; but in fact, the whole history as summarized by the endogenous state variable $(\lambda_t, \gamma_t, \theta_t, t)$, does.[18]

[17] The super contact condition means that the value function is twice differentiable.

[18] Further, replacing the first order condition on y_t in the law of motion (20) yields

$$d\Delta_t = \Big[\Big(\rho - \mu_t\Big)\Delta_t - \Big(\frac{\theta_t\tau_t^L}{(-\gamma_t)(1+1/\varepsilon)}\Big)^{1+\varepsilon}\Big(\frac{\theta_t}{\kappa}\Big)^{\varepsilon} - \sigma_{\Delta,t}\sigma_t\Big]dt + \sigma_{\Delta,t}\sigma_t dB_t. \tag{21}$$

Labor effort is continuous before retirement. At retirement, the labor effort jumps to zero and γ_t jumps to zero. Therefore, from (21), at retirement $\Delta = 0$, the drift of Δ_t jumps to $-\infty$ and the volatility $\sigma_{\Delta,t}$ jumps to zero. These jumps but have important implications for the optimal wedges.

4.3.3 Optimal Capital Wedge

Under separable utility, a standard Inverse Euler Equation of optimal contracting and dynamic moral hazard models holds.

Proposition 2. *(Capital wedge)*

1. There exists a process $\sigma_{c,t}$ such that

$$d\Big(\frac{1}{u'(c_t)}\Big) = \frac{1}{u'(c_t)}\sigma_{c,t}\sigma_t dB_t \qquad \text{(Inverse Euler Equation)} \tag{22}$$

2. The intertemporal wedge between t and $t+s$ is positive and satisfies

$$\tau^K_{t,s} = \int_t^{t+s} \sigma^2_{c,t'}\sigma^2_t dt'$$

and the intertemporal wedge at time t is $\tau^K_t = \sigma^2_{c,t}\sigma^2_t$.

The proof is in Appendix A. Point 1 states that the standard Inverse Euler Equation extends to the case with endogenous retirement. The inverse of marginal utility of consumption is a martingale. A direct consequence of this is that the intertemporal wedge is positive since Jensen's inequality applies to the inverse function that is concave.

Point 2 highlights that the intertemporal wedge τ^K_t is linked to the volatility of the inverse of the marginal utility of consumption. This volatility is a control for how much the changes in productivity translate into changes in consumption. It is, therefore, a measure of risk exposure. A high volatility of the inverse of marginal utility of consumption implies that the planner exposes the agents to risk to provide incentives at the expense of insurance.

This risk exposure stops at retirement and the volatility process $\sigma_{c,t}$ goes to zero.[19] Because of the fixed cost, labor effort jumps discontinuously at retirement. This translates into the volatility process $\sigma_{c,t}$, and therefore the intertemporal wedge, jumping from a positive value $\sigma_{c,\mathscr{T}_R^-}$ to zero at retirement. This jump has consequences on the range of values taken by the intertemporal wedge. Section (5) investigates the magnitude and range of the volatility process and intertemporal wedge numerically.

[19] In Sannikov (2014), risk exposure does not go to zero at retirement. Instead, it builds up to target, starts falling at an age before retirement, and goes to zero at the end of the horizon. The difference is due to the fact that, in my setting, there is no output after retirement, and therefore there is no need for the agent to be exposed to risk after retirement.

4.3.4 Optimal Labor Wedge

The evolution of the labor wedge is obtained from the evolution of γ_t:

Proposition 3. *(Labor wedge)*

The law of motion of γ_t, is

$$d\gamma_t = \Big[-\theta_t \lambda_t \sigma_{c,t} \sigma_t^2 + \mu_t \gamma_t \Big] dt + \gamma_t \sigma_t dB_t, \quad \gamma_0 = 0.$$

In addition, the labor wedge satisfies

$$d\Big(\frac{\tau_t^L}{1-\tau_t^L}\Big) = [(1+\frac{1}{\varepsilon})\sigma_{c,t} + \frac{\tau_t^L}{1-\tau_t^L}\sigma_{c,t}^2]\sigma_t^2 dt - \frac{\tau_t^L}{1-\tau_t^L}\sigma_{c,t}\sigma_t dB_t.$$

The proof is in Appendix A. The first order condition on y_t, coupled with the law of motion of γ_t, implies that

$$d\Big(\lambda_t \frac{\tau_t^L}{1-\tau_t^L}\Big) = [(1+\frac{1}{\varepsilon})\lambda_t \sigma_{c,t}\sigma_t^2]dt. \tag{23}$$

This expression states that the process $\lambda_t \frac{\tau_t^L}{1-\tau_t^L}$ has zero instantaneous volatility. This means that for insurance its paths are less dispersed than the paths of productivity. Applying Ito's lemma to (23) yields:

$$d\Big(\frac{\tau_t^L}{1-\tau_t^L}\Big) = [(1+\frac{1}{\varepsilon})\sigma_{c,t}]\sigma_t^2 dt + \frac{\tau_t^L}{1-\tau_t^L}\lambda_t d\Big(u'(c_t)\Big). \tag{24}$$

In order to understand the implications of an endogenous retirement age for the patterns of the labor wedge, define the model with flexible retirement age $\mathscr{T}_R^{sb}$, the "Flexible Retirement model", and the model with fixed retirement age $\mathrm{E}[\mathscr{T}_R^{sb}]$, the "Fixed Retirement model". In the Fixed Retirement model, the planner imposes on all workers to retire exogenously at the same age and sets the retirement age at $\mathrm{E}[\mathscr{T}_R^{sb}]$ before computing allocations. Note that the Fixed Retirement model generate the counterfactual that agents do not work after $\mathrm{E}[\mathscr{T}_R^{sb}]$, and the average labor wedge is undefined for the ages $\mathrm{E}[\mathscr{T}_R^{sb}]$ to T.[20]

I compare my results to two useful reference economies and motivate their definition. The first economy is one in which the planner expects all agents to retire at the end of the horizon, when $\mathscr{T}_R^{planner} = T$ and computes optimal policies accordingly. Because agents work longer in this economy, they produce more output than in the Fixed Retirement model. To make the pair

[20] The extension of the labor wedge to those ages would be a 100% tax rate in order that all agents retire at $\mathrm{E}[\mathscr{T}_R^{sb}]$.

comparable, I assume in this economy that after the planner has computed optimal policies, agents retire exogenously in a way that: (i) the retirement distribution is uniformly distributed at each age across workers and (ii) matches the the labor force participation rate of the Flexible Retirement model at each age. Formally, there exists an exogenous process $\mathscr{T}_R^{exog}$ and, at each age t, a binary random variable $z_t \in \{0, 1\}$, such that (i) $\mathscr{T}_R^{exog} = \inf\{t;\ z_t = 1\}$ and (ii) $P(\mathscr{T}_R^{exog} = t) = P(\mathscr{T}_R^{sb} = t)$ for all periods t. This process determines the work status of an agent $\mathscr{T}_R^{agent} = \mathscr{T}_R^{exog}$. I call this model the "Exogenous Retirement model". In this model, retirement is unanticipated by the planner but occurs for an exogenous reason.[21] The second economy is one in which the planner expects all agents to retire optimally as in the Flexible Retirement model, $\mathscr{T}_R^{planner} = \mathscr{T}_R^{fb}$ and computes optimal policies accordingly. Then agents retire with uniform probability across workers at each age $\mathscr{T}_R^{agent} = \mathscr{T}_R^{exog}$, as in the first economy. This exercise is useful because it captures the effect of the different labor supply elasticity due to the planner's anticipation of the endogenous retirement decision. In addition, it mutes the selection of the labor force by keeping the distribution of productivity in labor force the same as in the general population. I call this model the "Uniform Retirement model". In this model, retirement is anticipated by the planner but occurs for an exogenous reason.

In the Flexible Retirement model, retirement is anticipated by the planner and is endogenously determined. Comparing the three regimes (Flexible Retirement, Uniform Retirement, and Exogenous Retirement) is useful for understanding the forces at work. The Uniform Retirement model eliminates the selection of the labor force that occurs in the Flexible Retirement model by making workers retire ex-post uniformly at random. As a result, I decompose the effects of elasticity and composition on the patterns of the labor wedge using this intermediate economy: I define the composition effect as the gap in the labor wedge between the Flexible Retirement model and the Uniform Retirement model. In addition, I define the elasticity effect as the gap in the labor wedge between the Uniform Retirement model and the Dynamic Mirrless model. I make these comparisons in the following paragraphs. But first, the Exogenous Retirement model highlights the forces that generate an increasing-in-age average labor wedge in the standard model.

The Exogenous Retirement model The labor wedge formula (24) applies to all productivity histories for which agents work. Consider the time periods for which all agents work, which corre-

[21] Because the fixed cost enters in the planner's problem additively, this model is equivalent to the model in Farhi and Werning (2013) with a fixed cost of staying in the labor market equal to zero.

spond to the Exogenous Retirement model. On one hand, the first term of (24) is the instantaneous covariance between log-productivity and the inverse of marginal utility of consumption scaled by the inverse of the intensive Frisch elasticity of labor supply. When the instantaneous variance of log-productivity is non-zero, this drift is positive and gives a positive slope to the labor wedge. The covariance of consumption growth and log-productivity captures the benefits of added insurance since it depends on the variability of consumption and the degree of risk aversion. Insurance comes at the cost of decreased incentives for work. The more elastic is the labor supply, the stronger is the effect, explaining the role of the intensive Frisch elasticity. In addition, the second term is autoregressive and is scaled by the change in the marginal utility of consumption. Since the inverse of the marginal utility of consumption is a martingale, the marginal utility of consumption is a submartingale and its paths trend upwards. Therefore, the labor wedge is increasing at a young age when all agents are working. These are the standard forces in the Exogenous Retirement model with a fixed retirement age in Farhi and Werning (2013).

The law of motion of the labor wedge also captures two effects that are present once one accounts for a flexible retirement age. I compare the Exogenous Retirement model with the Uniform Retirement model to explain the elasticity effect.

The Elasticity Effect The elasticity over the retirement margin is captured by the jump in $\sigma_{c,t}$ at retirement from a positive value $\sigma_{c,\mathscr{T}_R^-}$ to zero. When age t inches closer to retirement, the volatility process decreases on average and the elasticity over the retirement margin increases. However, in contrast with the Exogenous Retirement model, $\sigma_{c,t}$ goes to zero earlier in the Uniform Retirement model because in the latter model the planner anticipates retirement. Therefore, the slope of $\frac{\tau_t^L}{1-\tau_t^L}$ is flatter in old age in the Uniform Retirement model. This is a manifestation of the elasticity effect. A higher total Frisch elasticity of old workers calls for a flatter labor wedge for old workers. Nevertheless, this does not necessarily translate into the profile of average labor wedges being hump-shaped. With the elasticity effect alone, a planner in the Exogenous Retirement model, who then anticipates retirement in the Uniform Retirement model, adjusts the profile of the increasing average labor distortions in the Exogenous Retirement model by decreasing its slope for old workers and increasing its slope at a young age while collecting the same revenue.

I compare the Uniform Retirement model with the Flexible Retirement model to explain the composition effect.

The Composition Effect The second term in the labor wedge formula is autoregressive and shows that innovations in the labor wedge must follow innovations in the marginal utility of consumption. When productivity increases, consumption increases and the marginal utility of consumption decreases. Therefore, over short horizons, the labor wedge must decrease when productivity increases. Farhi and Werning (2013) call this a form of "short-run regressivity". Here I highlight the implications of the negative covariance between consumption and the labor wedge when retirement is endogenous, as in the Flexible Retirement model, compared to when retirement is exogenous but is anticipated, as in the Uniform Retirement model. When t is closer to retirement, from Section 4.3.2 retirement occurs earlier for agents with a history of low productivity shocks compared to agents with a history of high productivity shocks in the Flexible Retirement model. Therefore, by selection, the labor force becomes more productive than the general population in old age in the Flexible Retirement model, while such selection does not occur in the Uniform Retirement model. In the short-run, this calls for lower labor wedges for this more productive sub-population. This is the novel composition effect. To draw out implications more clearly, apply Ito's lemma to the Inverse Euler equation and replace $d(u'(c_t)) = u'(c_t)\sigma_{c,t}^2\sigma_t^2 dt - u'(c_t)\sigma_{c,t}\sigma_t dB_t$ in (24) to obtain the formula of the labor wedge in the proposition:

$$d\Big(\frac{\tau_t^L}{1-\tau_t^L}\Big) = \Big[(1+\frac{1}{\varepsilon})\sigma_{c,t} + \frac{\tau_t^L}{1-\tau_t^L}\sigma_{c,t}^2\Big]\sigma_t^2 dt - \frac{\tau_t^L}{1-\tau_t^L}\sigma_{c,t}\sigma_t dB_t. \tag{25}$$

The full composition effect is captured by the last two terms on the right. As the labor force becomes increasingly productive when agents retire, over infinitesimal periods the remaining workforce has on average positive productivity shocks, $\sigma_t\theta_t dB_t > 0$. The last term on the right-hand side of the equation above $-\frac{\tau_t^L}{1-\tau_t^L}\sigma_{c,t}\sigma_t dB_t < 0$ captures that the labor force in old age becomes more productive and must have lower labor wedges in short-run. However, the term $\frac{\tau_t^L}{1-\tau_t^L}\sigma_{c,t}^2\sigma_t^2 dt$ captures the volatility of consumption growth and the increase in volatility of log-productivity over a longer infinitesimal horizon dt and calls for higher labor distortions. Therefore, there is a race between selection and rising inequality in productivity and consumption. If the force of selection into a more productive labor force is stronger than the increase in volatility of log-productivity, the composition effect yields a decreasing-in-age average labor wedge for old workers.

4.3.5 Optimal Retirement Consumption

Proposition 4. *Consumption after retirement is constant. In addition, consumption after retirement is equal to the final period consumption:* $c_{\mathscr{T}_R^+} = c_{\mathscr{T}_R^-}$.

The fact that consumption after retirement is equal to consumption at retirement is a consequence of the smooth pasting condition of optimal stopping. It implies that the marginal change in the cost of providing an infinitesimal promised utility before and after retirement are equal. In the separable utility case, it implies that there is no jump in consumption at retirement, i.e $\frac{1}{u'(c_{\mathscr{T}_R^-})} = K_v^- = K_v^+ = \frac{1}{u'\left(\frac{v_{\mathscr{T}_R}}{g(\mathscr{T}_R)}\right)} = \frac{1}{u'(c_{\mathscr{T}_R^+})}$.

To minimize distortions, agents are given their last period consumption at retirement in the separable utility case. Agents with a history of high productivity shocks are offered correspondingly higher retirement consumption than agents with a history of low productivity shocks, in order to induce them to retire later. In addition, the net present value of retirement consumption only needs to depend on their remaining life expectancy $T - \mathscr{T}_R$ and last period consumption (which in turn depends on the whole history until retirement).[22]

5 Numerical Analysis

This section highlights the quantitative implications of the model for the evolution of optimal wedges over the life-cycle and the welfare gains from optimal policies and simple tax and SS reforms. Subsection 5.1 calibrates the model parameters in a baseline US economy. Then, Subsection 5.2 presents optimal policies for those calibrated parameters. Finally, Subsection 5.3 quantifies welfare gains from optimal policies and those from simple tax and SS reforms.

5.1 Calibration

In order to provide some background, I start by discussing the empirical evidence on the fixed cost of staying in the labor market, a crucial parameter in the model.

[22] As in Proposition 4, consumption after retirement depends on the state $(v, \Delta, \theta, \mathscr{T}_R)$ that is a sufficient summary of the whole history until retirement for the purpose of computing allocations. Since at retirement $\Delta = 0$, one can infer that retirement consumption depends only on $(v, \theta, \mathscr{T}_R)$, or equivalently $(v, C(v, 0, \theta, \mathscr{T}_R^-), \mathscr{T}_R)$ in which C is the function that maps state variables to consumption before retirement. Now, the smooth pasting condition links the promised utility before retirement v and consumption before retirement $C(v, 0, \theta, \mathscr{T}_R^-)$. Therefore, the net present value of retirement consumption depends on the history of productivities only through $(c_{\mathscr{T}_R^-}, \mathscr{T}_R)$.

Estimates of the Fixed Cost in Dynamic Models French (2005), Rogerson and Wallenius (2013), Prescott *et al.* (2009), and Chang *et al.* (2014) estimate life-cycle models with endogenous retirement. They consider non-convexities in the labor supply decision due to fixed time costs that match the hours worked and labor force participation of old workers. They find that, one needs large fixed time costs, around 5 to 6 hours a day, to match the retirement data. In their estimations of extensive margin elasticities, Chetty *et al.* (2012) find, in a model similar to Rogerson and Wallenius (2013), that extensive margin labor supply responses ought to be very large to explain the gap between the micro and macro Frisch elasticities. In addition, Banks *et al.* (1998) and Aguila *et al.* (2011) posit that there are sizable fixed consumption costs related to work. In my analysis, I calibrate the fixed utility cost of staying in the labor market and compare its time value and consumption value to the time costs and consumption costs estimated in the literature.

Parametrisation I perform the numerical simulation in a discrete time version of the model, in which agents live for $T = 55$ periods, with each period corresponding to a year between the ages of 25 to 79.[23] I set the discount factor to $e^{-\rho} = 0.95$ and the interest rate $r = \rho$. Since Deaton and Paxson (1994), there is evidence that inequality in consumption and income increases with age within a cohort. Consistent with these findings, I assume that productivity is a geometric random walk as in Farhi and Werning (2013) or Stantcheva (2017):

$$\log(\theta_t) = \log(\theta_{t-1}) + \epsilon_t$$

where $\epsilon_t \sim \mathcal{N}(-\frac{\sigma^2}{2}, \sigma^2)$.

Storesletten *et al.* (2004) have found a high estimate of the volatility $\sigma_H^2 = 0.0161$ and Heathcote *et al.* (2010), a low estimate of $\sigma_L^2 = 0.00625$. In this simulation, I choose an intermediate value of $\sigma_M^2 = 0.0095$, in line with Heathcote *et al.* (2005)'s estimate of a medium volatility, and I perform robustness checks with the low and high volatility estimates in Appendix B.

[23] The theoretical analysis performed in continuous-time allowed for a simple representation of forces shaping the dynamics of the wedges. Additionally, the continuous-time analysis allowed for explicit analytic results in special cases that are not available in discrete time. I choose to perform a numerical simulation of the discrete time model presented in Appendix B rather than solving the HJB equation, using the Markov Chain Approximation Method as in Kushner and Dupuis (2013). By using balanced growth preferences, I reduce the dimensionality of the problem in discrete time with one less state variable.

Preferences during working years are

$$\log(c_t) - \frac{\kappa}{1+\frac{1}{\varepsilon}}\left(\frac{y_t}{\theta_t}\right)^{1+\frac{1}{\varepsilon}} - \phi(t)$$

with $\varepsilon = 0.5$ and $\kappa = 1$, consistent with the estimate of Chetty (2012). During retirement, per period utility is simply $\log(c_t)$. While many parameters are readily estimated from the literature, the fixed cost function $\phi(t)$ is an important parameter to calibrate in my model. I endogenously calibrate the fixed costs in a baseline US economy.

Calibration to US Data in the Baseline Economy The baseline economy is the income fluctuation model in which agents, who experience idiosyncratic productivity shocks, can freely borrow, save in a risk-free asset, and choose their consumption, hours worked, and retirement age. I assume that the agents start claiming retirement benefits when they exit the labor force. The tax system is set to mimic the US tax system. I follow Heathcote *et al.* (2014) and set the labor income tax equal to the approximation function

$$T(y_t) = y_t - \lambda y_t^{1-\tau}$$

where the value of the progressivity parameter τ is 0.181. The capital tax is set to a flat tax rate equal to 20% of capital gains, equivalently 1% of the capital stock.

The SS benefits system features three specific ages that are important for the availability and value of retirement benefits. The Normal Retirement Age (NRA) is the age at which a worker can claim the full amount of retirement benefits, the Primary Insurance Amount (PIA). I set the NRA to 66 for the present cohort. The PIA is a function of the Average Indexed Monthly Earnings (AIME) which is the average monthly earnings of the 35 highest earning years. In the calibration, I set the replacement rate to the (earnings weighted) average replacement of 40%, i.e PIA=40%AIME.[24] The Early Retirement Age (ERA=62) is the age at which an agent can start claiming retirement benefits. For each year between the ERA and the NRA, an individual who starts claiming benefits at that age loses 6.67% points of the PIA per early year (the Actuarial Reduction Factor, ARF). For

[24]In the US SS system, the PIA is a non-linear function of the AIME. The first bracket gives a PIA with a replacement rate of 90% of the AIME until the AIME reaches \$885. The second bracket gives a replacement rate of 32% until it reaches \$5,336. Finally, the third bracket replaces 15% of the AIMEs over \$5,336 and below \$127,200. Munnell and Soto (2005) report a median replacement of SS benefits of 42% in 2001. I set an (earnings weighted) average replacement of 40% in my benchmark calibration.

instance, someone who retires at age 63 gets 80% of his PIA. The End of Eligibility Age (EEA=70) is the age at which an individual should start claiming benefits that would otherwise be lost. For each year between the NRA and the EEA, an individual who starts claiming benefits at that age gains 8% points of the PIA per year delayed (the Delayed Retirement Credit, DRC). For instance, someone who retires at age 70, gets 132% of his PIA. These "actuarial"[25] adjustments to benefits stop at the EEA and are capped at 132% of the PIA. The SS benefits system of my calibration features these adjustments.

In this baseline economy, I calibrate different specifications of $\phi(t)$. One target I match is the labor force participation rate for ages 65 to 69 in the US population. In Toossi (2015), the Bureau of Labor and Statistics reports a labor force participation rate of individuals between ages 65 to 69 of 31.6% in 2014. For the specification with a constant fixed cost $\phi(t) = \phi$ over the life-cycle, I match the labor force participation rate for ages 65 to 69 with a fixed cost that is the utility equivalent of 6.8 hours per day. To compute the time value of fixed utility costs, I follow Shourideh and Troshkin (2015) and use parameters from Chang *et al.* (2014) who estimates a model similar to this paper's baseline economy. I take the estimates of $\hat{\kappa} = 82.70$ from Table 1 of Chang *et al.* (2014) for $\varepsilon = 0.5$ and the lowest variance σ_x which (annualized) is closest to the variance σ_M in my model. I link the estimate of the fixed utility cost $\hat{\phi}$ to its time cost $\hat{l}$ by solving $\hat{\kappa}\frac{\hat{l}^{1+1/\varepsilon}}{1+1/\varepsilon} = \hat{\phi}$.

For this specification, the average retirement age in the baseline economy is 63.73 years old. In the Flexible Retirement model, the optimal average retirement age is large and equal to 76 years old. I find, similarly to Rogerson and Wallenius (2013) that with a constant fixed cost one needs a large ϕ to generate retirement.[26] The intuition is that with a constant fixed cost, the only force for an extensive Frisch elasticity of labor supply that increases with age is the decreasing option value of staying in the labor market. With the medium instantaneous variance of productivity of $\sigma_M^2 = 0.0095$, this option value is low.

To obtain a more realistic optimal average retirement age, I calibrate a specification of fixed costs that increase in age. On top of the labor force participation rate for ages 65-69, I target a measure of dynamic total elasticity of labor supply, as in French (2005), at age 65.[27] In this

[25]The standard term used for these adjustments does not necessarily imply that they are actuarially fair.

[26]The average retirement age in the baseline economy is significantly lower than the optimal average retirement age when ϕ is constant. The SS sytem in the baseline economy is such that most agents retire at the ERA or at the NRA. In the optimum however, there are no retirement spikes at the ERA or the NRA in the optimal distribution of retirement ages induces by the retirement benefits.

[27]I define the dynamic extensive elasticity of labor supply by computing the ratio of a 1% unexpected increase in income at age 65 on the percentage change in the average retirement age. The total elasticity is obtained by adding

specification, the fixed cost is constant until age 55 - when the first point of entry into retirement through the Social Security's disability program occurs in the US - then increases linearly until age 79 as $\phi(t) = a + b(t-55)^+$. I calibrate a and b targeting the labor force participation rate for ages 65-69 and a total elasticity of labor supply at age 65 of 1.3 in the range of values estimated in French (2005).[28]

The qualitative result of the average labor wedge eventually declining with age only needs $\phi(t)$ not to decrease too fast with age given the evolution of the option value. A constant $\phi(t) = \phi$, in particular, can generate an average labor wedge that eventually declines with age as I show in Appendix B. However, to make the decline quantitatively significant, it is useful to have an increasing $\phi(t)$. The calibration controls for the speed by which the fixed cost increases with age given the persistence of the productivity process. This section presents results for this specification.

concept	functional form	values	source/target
Exogenously parametrized			
productivity	$\log\theta_t = \log\theta_{t-1} + \varepsilon_t$	$\sigma_L^2 = 0.00625$	Heathcote *et al.* (2010)
	$\varepsilon \sim N(-\frac{\sigma^2}{2}, \sigma^2)$	$\sigma_M^2 = 0.0095$	Heathcote *et al.* (2005)
		$\sigma_H^2 = 0.0161$	Storesletten *et al.* (2004)
utility	$\log c - \frac{\kappa}{1+\frac{1}{\varepsilon}}(\frac{y}{\theta})^{1+\frac{1}{\varepsilon}}$	$\kappa = 1, \varepsilon = 0.5$	Chetty (2012)
Endogenously calibrated in baseline US economy			
fixed cost	$\phi(t) = a + b(t-55)^+$	$a = 0.32$ (5.42h/day)	lfp65-69=31.6% (BLS)
		$b = 0.03$ (+7mn/day/year)	$\varepsilon_{\text{total\_60}} = 1.3$ French (2005)

Table 1: Calibration

Table 1 summarizes the calibrated values. I obtain a fixed cost equivalent to 5.42 hours per day in terms of time cost at age 55 that increases by 7 minutes each year until attaining 8.28 hours per day at age 79. These estimates are within the range of estimates in Chang *et al.* (2014). Although the qualitative features of the model are unaffected for a wide range of parameters, the quantitative results are. Therefore, I perform several alternative calibrations in Appendix B including robustness checks with respect to the variance of productivity, the specification of a constant fixed cost ϕ, other specifications of $\phi(t)$ and other targets for the value of the dynamic elasticity.

the intensive Frisch elasticity $\varepsilon = 0.5$ with the dynamic extensive elasticity.

[28] Alpert and Powell (2013) report extensive elasticities with respect to after-tax labor income equal to 0.76 for women and 0.55 for men at age 65. These measures are slightly lower than, but of the same magnitude as, my target of a dynamic extensive elasticity of 0.8 as in French (2005).

I compute the policy functions in the Flexible Retirement model for the calibrated values above. From these policy functions, I perform a Monte Carlo simulation with N=1,000,000 draws. I set the initial states and the rate of taxation λ in the labor income tax function $T(y_t)$ to yield a zero present value resource cost for the allocations, $G = 0$. This provides comparable allocations across simulations.

5.2 Optimal Policies

In this section, I describe the properties of the optimal policies obtained from the simulations of optimal allocations in the Flexible Retirement Model.

Optimal Labor Force Participation Rate The left panel of Figure 1 shows the optimal labor force participation rate as a function of age. The labor force participation rate decreases until age 75 after which it is non-zero at each age but less than 0.1%. The Average Retirement Age (ARA) is 65.84 and the labor force participation rate at age 65 is 53.63%. These are larger than in the baseline economy in which the ARA is 63.66 and the labor force participation rate at age 65 is the target of 31.6%. This is consistent with the fact there are still considerable implicit disincentives to continued work between ages 62 and 65 in the US tax and SS system as documented by Gruber and Wise (1998).

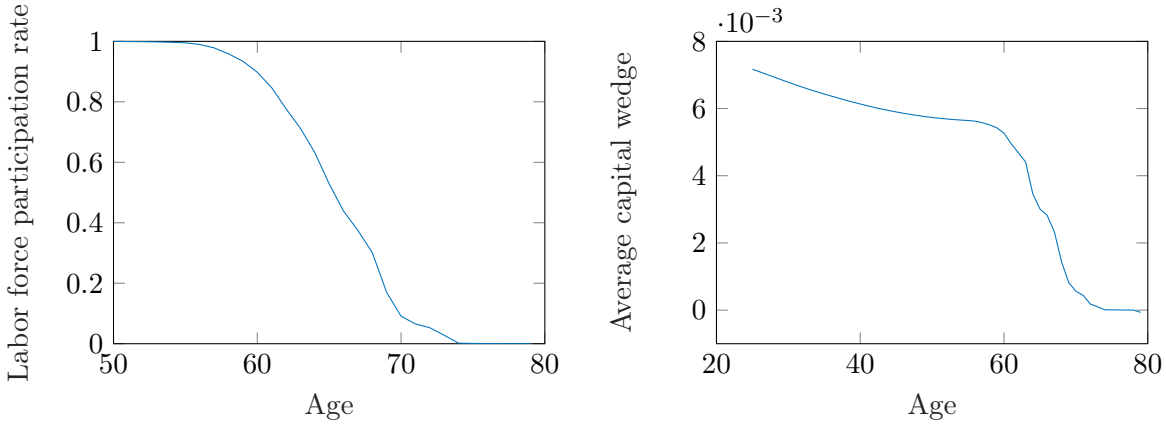

Figure 1: Left: Labor force participation rate as a function of age. Right: Average capital wedge as a function of age.

Optimal Capital Wedge The right panel of Figure 1 shows the cross-sectional average of the capital wedge as a function of age. Distortions on savings decline on average with age. As shown in

Section 4.3, the capital wedge is directly linked to the variance of consumption growth $\tau_t^K = \sigma_{c,t}^2 \sigma_t^2$. This variance decreases over time until it equals zero at retirement when consumption is constant. Because of the discontinuity in labor effort, $\sigma_{c,t}$ jumps to zero one period before retirement. Since in discrete time there is always a positive mass of agents retiring at each time, when the variance jumps to zero at retirement before T, the average capital wedge sharply declines.[29] In addition, the average capital wedge is small in magnitude, going from a tax rate of 0.7% of the capital stock or equivalently 14% of capital gains to zero percent at age 79.

Optimal Labor Wedge Figure 2 displays the average labor wedge as a function of age. The profile of the average labor wedge is hump-shaped in age. The blue curve (solid line) represents the average labor wedge over the whole population. The red curve (double dashed line) and yellow curve (dashed-dotted line) represent the average labor wedge in a population of agents with a history of low and high productivity shocks respectively.[30] For incentive compatibility, the average over the population with low productivity shocks is higher than the average in the population with high productivity shocks. Since the red (double dashed) curve is above the yellow (dashed-dotted) curve, once low-productivity agents start retiring, the blue (solid) curve comes closer to the yellow (dashed-dotted) one. This is a manifestation of the composition effect. At age 73, the blue (solid) curve and the yellow (dashed-dotted) curve are indistinguishable as the remaining labor force is mainly composed of highly productive workers. Overall, the average labor wedge increases from 2.16% at age 25 to 42.5% at age 64 then decreases up to 32% at age 79.

As a reminder from the analytic section, I consider two other reference economies. The Exogenous Retirement model is the model where retirement is unanticipated by the planner but occurs for an exogenous reason. The Uniform Retirement model is the model where retirement is anticipated by the planner but occurs for an exogenous reason. In the Flexible Retirement model, retirement is anticipated by the planner and is endogenously determined. Comparing the three regimes (Flexible Retirement, Uniform Retirement, and Exogenous Retirement) is useful for understanding the forces at work. The Uniform Retirement model, eliminates the selection of the labor force that occurs in the Flexible Retirement model by making workers retire ex-post uniformly at random. I describe,

[29] The inflection points in the capital wedge curve are therefore a result of the bunching of many agents at the same retirement age in discrete time rather than computational imprecision from state variable grids. The downside of my approach is that with a time step of 1 year a significant mass of agents retires at each period when old.

[30] I define the population with a history of low productivity shocks as agents who receive at each period a shock lower than the mean shock plus the quarter of the standard deviation of instantaneous shocks, such that $\exp(\varepsilon_t^L) \leq 1 + \sigma/4$ and the population with a history of high productivity shocks as agents who receive at each period a shock higher than the mean shock minus the quarter of the standard deviation of instantaneous shocks, with $\exp(\varepsilon_t^H) \geq 1 - \sigma/4$.

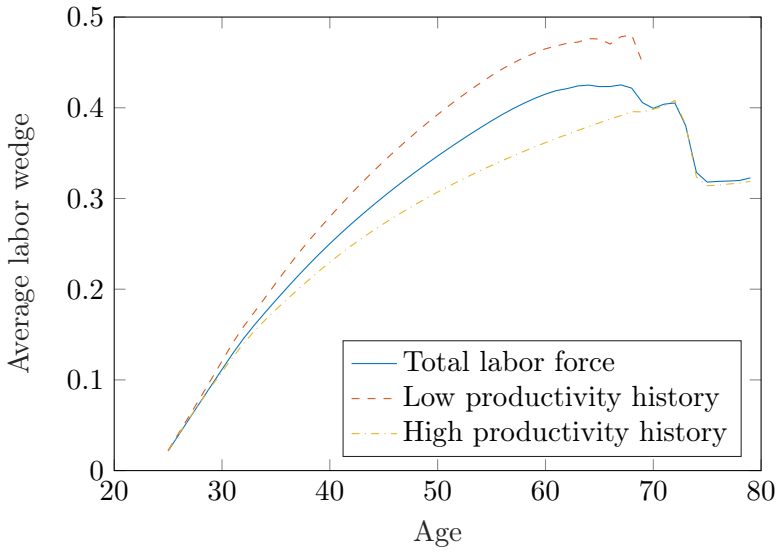

Figure 2: Labor wedge in the Flexible Retirement model

in what follows, the gap in the labor wedge between the Flexible Retirement model and the Uniform Retirement model—the composition effect—and between the Uniform Retirement model and the Dynamic Mirrless model—the elasticity effect.

The Elasticity Effect The left panel of Figure 3 plots, as a function of age, the average capital wedge of the Uniform Retirement model in red (dashed) and the average capital wedge of the Exogenous Retirement model in blue (solid). Suppose the planner starts from the solid curve and realizes that the total Frisch elasticity is larger for old workers than in the Exogenous Retirement model. The planner can lower distortions on old agents and raise the same revenue as in the solid curve by decreasing the average capital wedge for the old and increasing it for the young. This is what the red curve accomplishes.

Similarly, the right panel of Figure 3 plots, as a function of age, the average labor wedge of the Uniform Retirement model in red (dashed) and the average labor wedge of the Exogenous Retirement model in blue (solid). Suppose again that the planner starts from the solid curve and wants to lower distortions on the old while raising the same revenue. The planner makes the average labor wedge curve flatter for old agents and steeper for the young; this is what the red curve accomplishes. Since the slope of the average labor wedge is in part determined by the variance of consumption and therefore the capital wedge, this is consistent with the left panel of Figure 3. Therefore, the elasticity effect calls for an average labor wedge curve for old workers that is flatter

in age. The largest average labor wedge is 51.58% for the Exogenous Retirement model and 48.32% for the Uniform Retirement model, which is a -3.26% decrease in the average labor wedge due to the elasticity effect.

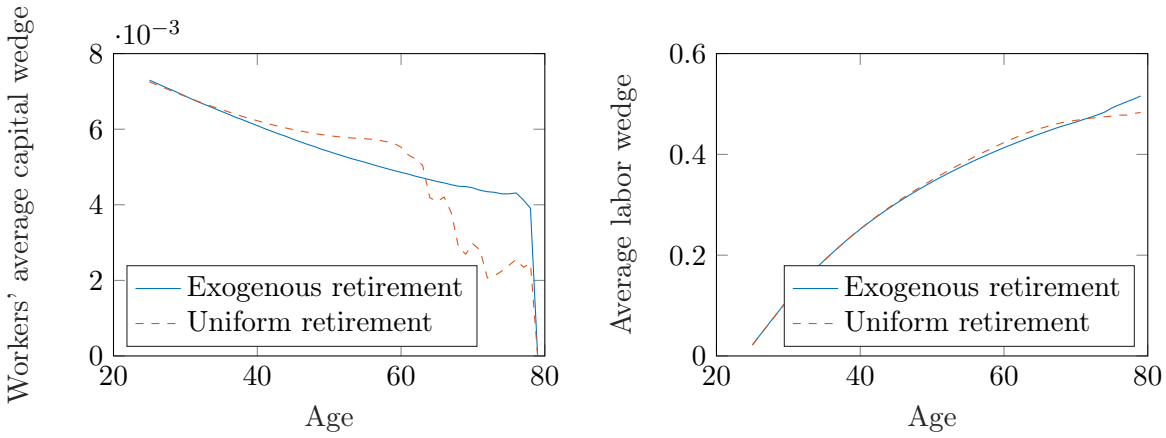

Figure 3: Elasticity effect. Left: Capital wedge. Right: Labor wedge.

The Composition effect The left panel of Figure 4 plots, as a function of age, the average labor wedge of the Uniform Retirement model in red (dashed) and the average labor wedge of the Flexible Retirement model. When the composition effect is strong enough as illustrated in the model with endogenous retirement in Figure 2, the average labor wedge is hump-shaped in age. Despite the fact that $\phi(t)$ increases with age in the simulation, the average labor wedge of the Uniform Retirement model increases in age. This further justifies that the hump-shaped profile of the average labor wedge in the Flexible Retirement model is not mainly driven by extensive margin elasticity through $\phi(t)$ but is a result of the composition effect. The average labor wedge for the Flexible Retirement model is equal to 31.91% in the final period, which compared to the Uniform Retirement model (48.32%) implies a -16.41% reduction due to the composition effect.

The right panel of Figure 4 shows the allocations over the working population in the model with endogenous retirement. Average output is in blue (solid) and average consumption in red (dashed). Until age 58 almost all agents work. Average consumption is constant and the average output is decreasing. After 58, agents with a history of low productivity shocks start retiring. The labor force becomes increasingly selected towards more productive agents. As a result, average output among workers increases and average consumption increases. This justifies that there is a selection of the labor force as a result of the nature of the retirement decision that creates the composition

effect in the pattern of the average labor wedge.

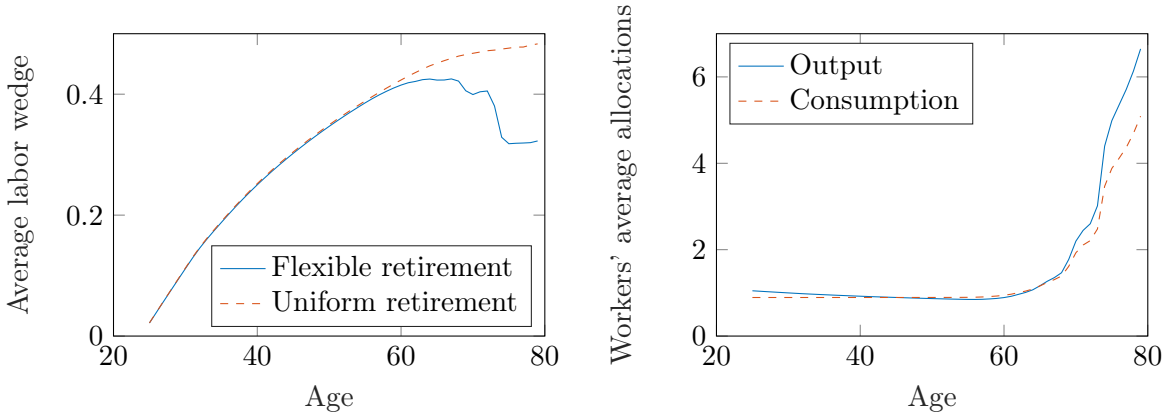

Figure 4: Left: Composition effect, Right: Mean allocations over working population

Retirement Consumption and Allocations The left panel of Figure 5 displays the allocations over the whole population. With log utility, the Inverse Euler equation implies that consumption is a martingale. Therefore, average consumption (dashed) is constant both before retirement and after retirement, while average output (solid) decreases slowly when most agents are working and decreases sharply once agents start retiring. The right panel of Figure 5 plots the mean consumption of retired agents. Over time, agents with higher consumption retire, which increases the average consumption of the retirees. The average consumption of retirees is increasing until almost all agents have retired, at which point it equals the average consumption of the whole population.

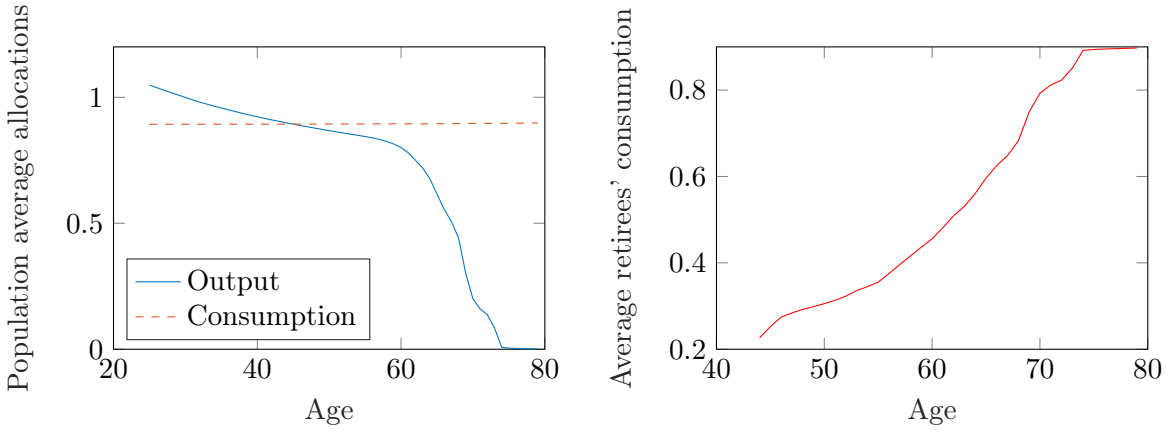

Figure 5: Left: Mean allocations over the whole population, Right: Mean consumption among retirees

5.3 Welfare Analysis

5.3.1 Welfare Gains and Simple Policies

The first line of Table 2 summarizes key features of the baseline economy, with a target labor force participation rate at 65 of 31.6% and an Average Retirement Age (ARA) of 63.66. The second line of the Table 2 shows the welfare gains from the second-best (fully optimal) system compared to the baseline economy.[31] In the second-best, agents retire later on average compared to the baseline economy, with an ARA of 65.84 and a labor force participation rate at 65 of 53.63%. The second-best improves welfare as an equivalent increase in consumption at all histories and periods of +3.75%. These welfare gains are large and correspond to an upper bound on welfare gains from jointly reforming the US tax system and SS system when productivity is unobservable.

From the optimal policies found above, I conduct several experiments. In all these experiments I am interested in the welfare gain relative to the current US tax code in the baseline economy. The simple policies I study are history independent but age-dependent. I consider linear taxes (marginal tax that are flat in income) equal to the cross-sectional average of taxes from the simulations. I compare their welfare gains or losses with respect to the US tax and SS system as a percentage of the welfare gains from the second-best. These experiments are motivated by the fact that they yield the bulk of the welfare gains in the optimal taxation literature that assumes a fixed retirement age (Farhi and Werning (2013), Golosov *et al.* (2016)).[32] I qualify their results accounting for a flexible retirement age.

Average Wedges from the Flexible Retirement model In this reform, the labor tax and capital tax in the baseline economy are replaced by the linear (flat in income marginal tax) taxes equal to the average labor wedge (hump-shaped in age) over the working population and the average capital wedge (small and decreasing in age) over the whole population in the model with a flexible retirement age. The goal of this experiment is to measure the welfare gains from a reform of the tax code alone. The third line of Table (2) shows the welfare gains achieved by this reform.

[31] The literature has usually compared the welfare from the second-best to the welfare achieved in a laissez-faire economy with no taxes or subsidies. Because of the importance of the SS benefits system, here the relevant economy to compare the second-best with is the baseline US economy. In addition, such a direct comparison with a parametrization of the US tax code allows me to measure welfare gains of tax reforms.

[32] Optimizing over age-dependent taxes in this dynamic economy is computationally heavy because of the number of tax functions, one for each period, and the non-negligible time it takes for the income fluctuation algorithm to run for one set of parameter values.

Replacing the tax code with the hump-shaped one brings modest welfare gains, with an increase in consumption at all histories and periods equivalent to 47.20% to that of the second-best. These welfare gains in consumption of +1.77% are slightly higher than the welfare gains in Farhi and Werning (2013) in absolute terms but significantly lower than those under the second-best policies. While Farhi and Werning (2013) find that, with a fixed retirement age, simple age-dependent taxes achieve 95% of welfare gains from the second-best, I find that with flexible retirement those welfare gains shrink to less than half the gains from the second-best. A reform of the tax code alone, despite the increasing then decreasing-in-age labor tax, induces agents to retire even earlier than in the baseline economy with an ARA of 59.11; and almost all agents retire by 65. The current SS system does not provide as much incentives for delayed retirement as the optimal system does and the increasing-in-age linear labor tax before age 64 is, for most ages, above the current levels of taxation, and induces agents to retire even earlier than in the baseline economy.

Table 2: Welfare gains from reforming the tax system while maintaining the current SS system.

reforms	welfare gains	l. f. p. 65-69	ARA
baseline economy age-independent tax system	N/A	31.6%	63.66
history-dependent optimal policies (sb)	+3.75% c_t	53.63%	65.84
$\tau^K(t)$ and hump-shaped in age $\tau^L(t)$ from Flexible R. m.	47.20% of sb	$< 0.1\%$	59.11
$\tau^K(t)$ and increasing-in-age $\tau^L(t)$ from Fixed R. m.	45.87% of sb	$< 0.1\%$	59.12
$\tau^K(t)$ and increasing-in-age $\tau^L(t)$ from Exogenous R. m.	46.67% of sb	$< 0.1\%$	59.09
$\tau^K(t)$ and increasing-in-age $\tau^L(t)$ from Regression m.	44.27% of sb	$<0.1\%$	59.92

Line 1 summarizes the labor force participation rate for ages 65 to 69 and the Average Retirement Age (ARA) in the baseline economy. Line 2 reports welfare gains from the second-best policies. Line 3 reports welfare gains from replacing the tax system with the hump-shaped in age linear labor tax over the cross-section of workers from the optimal policies. Line 4 to 6 gives welfare gains from alternative tax reforms with an increasing-in-age linear labor tax. Under the current SS system, a hump-shaped in age linear labor tax and an increasing-in-age linear labor tax achieve comparable welfare gains that are modest compared to those from second-best policies.

To capture more welfare gains, I augment this tax reform with a reform of the SS system that aims to match the optimal labor force participation for ages 65-69 in the optimum which is 53.63% from the second line of Table 2. Because I assumed the retirement age and benefits claiming age are the same, changing the replacement rate of SS benefits from 40% has little effect on labor force participation. Instead, I focus on reforms that increase the absolute value of the "actuarial" adjustment rates of SS benefits (the Actuarial Reduction Factor before the NRA and the Delayed Retirement Credit after the NRA) that affect at the margin workers' decision to retire before or after

the NRA. These rates are equalized in absolute value and increased until the retirement distribution matches its counterpart in the second-best. This reform requires setting a large uniform adjustment rate of 16%.[33] The goal of this experiment is to measure the welfare gains from a joint reform of the tax code and SS system that simply tries to mimic the optimal history-dependent system in the second-best.

The second line of Table 3 summarizes welfare achieved by this reform. Tilting the retirement distribution to match the retirement distribution of the second-best achieves sizable welfare gains with an increase in consumption at all histories and periods equivalent to 81.33% to that of the second-best. Like French (2005) found, exit from the labor force between ages 62-65 is mostly determined by the tax structure of SS benefits. By increasing "actuarial" adjustment rates of SS benefits in absolute value, one can induce agents to retire later. Because of the fixed cost that is incurred both by low-productivity workers and high-productivity workers, most of the agents who delay retirement are highly productive. The decreasing-in-age labor tax after age 64 increases the efficiency of the intensive labor supply of these high-productivity agents and we obtain the bulk of welfare gains from the age-dependent linear tax that is hump-shaped in age. These two experiments suggest that, when accounting for flexible retirement, reforming the tax code alone is not enough and the SS system also needs to be reformed to correct for the retirement distribution.

I also investigate the benefits of reforming the SS system alone while keeping the tax system as in the status quo system in the baseline economy in the first line of Table 3. The optimal retirement distribution is matched by an increase in absolute value of the Actuarial Reduction Factor from -6.67% to -7%. The welfare gains from a reform of the SS system alone are a small fraction, 5%, of the welfare gains from the second-best, which is equivalent to a 0.19% increase in consumption in all histories and all periods. The joint reform of the tax system and SS system achieves more welfare gains than the sum of each individual reform alone.

5.3.2 Alternative Reforms

I consider a series of alternative reforms in order to determine: (i) how important the hump-shaped profile of the labor tax is in terms of welfare compared to the family of increasing-in-age labor

[33] Another reform would be to increase the NRA. The actuarial adjustment I found, that is the double of the current Delayed Retirement Credit and Actuarial Reduction Factor, is equivalent to a 14 months increase of the NRA for a 65-year-old worker and 28 months increases of the NRA for a 64-year-old worker and so on.

Table 3: Welfare gains from joint reforms of the tax system and the SS system

reforms	welfare gains	ARA
SS reform alone	5.01% of sb	65.12
SS reform + $\tau^K(t)$ and hump-shaped in age $\tau^L(t)$ from Flexible R. model	81.33% of sb	65.84
SS reform + $\tau^K(t)$ and increasing-in-age $\tau^L(t)$ from Fixed R. model	69.60% of sb	63.47
SS reform + $\tau^K(t)$ and increasing-in-age $\tau^L(t)$ from Exogenous R. model	72.53% of sb	63.52
SS reform + $\tau^K(t)$ and increasing-in-age $\tau^L(t)$ from Regression model	67.47% of sb	64.29

SS reforms set a uniform Actuarial Reduction Factor (ARF) and Delayed Retirement Credit (DRC) and the resulting actuarial adjustment rate is increased to match the labor force participation rate for ages 65-69 in the optimum equal to 53.63%. Line 1 reports welfare gains from a reform of the SS alone. Line 2 gives welfare gains from jointly choosing an age-dependent labor tax that is hump-shaped in age from the optimum and reforming the SS system. Lines 3 to 5 reports welfare gains from from alternative tax reforms with an increasing-in-age linear labor tax coupled with SS reforms. With a SS reform, through an increase in Delayed Retirement Credits, a hump-shaped in age linear labor tax achieves a large fraction of those from the second-best and significantly more welfare gains than those from an increasing-in-age linear labor tax or a reform of the SS system alone.

taxes and (ii) how important a joint reform of the tax system and SS system is in terms of welfare compared to a reform of the tax code alone.

Average Wedges from the Fixed Retirement model The labor tax and capital tax in the baseline economy are replaced by linear labor taxes respectively equal to the average labor wedge (increasing in age) and the average capital wedge (small and decreasing in age) in the Fixed Retirement model at age $\mathrm{E}[\mathscr{T}_R^{sb}] = 66$. The goal of this experiment is to see whether the standard result of an average labor wedge that is increasing in age achieves significant welfare once I account for a flexible retirement age.

The fourth line of Table 2 shows welfare achieved by such a reform. Changing the tax code with a "wrong" model and increasing-in-age linear labor tax actually achieves welfare comparable to the hump-shaped in age labor tax (45.87% of the second-best for the former and 47.20% for the latter) but by significantly less than the second-best optimum. This is consistent with the finding in Farhi and Werning (2013) or Golosov *et al.* (2016) that with a fixed retirement age, an increasing-in-age linear labor tax is close to optimal. With a flexible retirement age, the increasing-in-age linear labor tax achieves similar welfare as the hump-shaped in age labor tax because they both induce agents to retire even earlier than in the baseline economy with an ARA of 59 so that almost all workers retire by 65. Therefore, the welfare gains from the decreasing portion of the hump-shaped in age

labor tax after age 64 are not present without a reform of SS system. Line 3 of Table 3 shows the welfare gains from a joint reform of the tax and SS system, 69.6% of the second-best. Compared to the first line, 81.33% of the second-best, the welfare gains from the increasing-in-age linear labor tax are significantly lower than the welfare gains from the hump-shaped in age labor tax once the SS system is reformed to correct for the retirement distribution. The increasing-in-age linear labor tax achieves a lower ARA (63.47) than the hump-shaped one that surprisingly achieves the same ARA as the second-best of 65.84 (while only the labor force participation rate for ages 65-69 of the second-best is targeted by the SS reform). This suggests that a decreasing-in-age the labor tax for old workers and reforming the SS system can be welfare improving by allowing high-productivity agents to work longer (over the extensive margin) and more efficiently (over the intensive margin).

Average Wedges from the Exogenous Retirement model One might worry that the above comparison does not give justice to the Fixed Retirement model because this model implicitly sets a linear labor tax of 100% after age 66. Therefore, I investigate the reforms that use the average labor wedges from the Exogenous Retirement model defined earlier, in which retirement is unanticipated by the planner but occurs for an exogenous reason.

The fifth line of Table 2 gives the welfare gains from replacing the tax system with the linear labor tax equal to the average wedges from the Exogenous Retirement model. As expected, this reform achieves similar welfare gains (46.67% of the second-best) as the reform with the average wedges from the Flexible Retirement model when the SS system is unchanged. However, once the SS system is reformed as well, fourth line of Table 3, it achieves fewer welfare gains (72.53% of the second-best) than the hump-shaped in age linear labor tax (81.33% of the second-best) and its ARA (63.52) is significantly lower than that of the second-best. Therefore, this reform suggests that the hump-shaped in age linear labor tax from the Flexible Retirement model achieves at least 8.8% more welfare as a fraction of the second-best welfare gains, or +0.33% consumption at all histories at all times compared to the planner using the "wrong" Exogenous Retirement model in a context where retirement is flexible.

Average Wedges from the Regression Model However, one can argue that the increasing-in-age linear labor tax from the Exogenous Retirement model in a context where retirement is flexible is expected to underperform compared to the hump-shaped one from the Flexible Retirement model because the model of the economy is the wrong in the former model. To make the point about

comparing the performance of hump-shaped in age linear labor taxes against the class of increasing-in-age linear labor taxes, I run a linear regression of the average labor wedges from the Flexible Retirement model as a function of age.

The best linear (in age) approximation of the hump-shaped in age average labor wedge from the Flexible Retirement model gives a flat labor tax of 13.51% at age 25 that increases linearly in age until age 79 when it becomes 45.37%. The last line from Table 2 shows that replacing the tax system with the increasing-in-age labor tax from the Regression model and the capital tax with the average capital wedge of the Flexible Retirement model brings 44.27% of the welfare gains from the second-best, this sits in the low end of range of welfare gains from other reforms. In addition, since the labor tax from the Regression model is lower than that of the Flexible Retirement model at most ages before 64 (because the former is a linear-in-age approximation of the latter that is hump-shaped in age), the ARA in the former reform (59.92) is higher than the ARA in the latter (59.11). Once the SS system is reformed as well, as shown on the last line of Table 3, the Regression model only achieves 67.47% of the welfare gains from the second-best. These reforms suggest that a hump-shaped in age linear labor tax improves welfare more compared to an increasing-in-age linear labor tax even when both taxes are set to mimic the optimal wedges of the "right" model with a flexible retirement age.

6 Extensions

In this section, I present the extensions of my results to the case of non-separable utility in consumption an labor, agents with stochastic lifetimes and productivity-dependent fixed costs.

6.1 Non-Separable Utility

In this section, I relax the assumption of separable intensive preferences in consumption and labor. In particular, I allow for non-separabilities between consumption and leisure. Saez (2002) argues that this non-separability is important to study optimal income taxation. Non-separability between consumption and leisure brings difficulties in that the Inverse Euler equation does not hold. It is well known that with nonseparable preferences, the no capital tax result of Atkinson and Stiglitz (1976) does not hold. The reason is that income and productivity now directly affect the intertemporal

rate of substitution for consumption. Intertemporal distortions allow to separate types and relax incentive constraints.

Denote the consumption function $C(y, u, \theta)$ the inverse of $u(\cdot, \frac{y}{\theta})$. Define

$$\eta(y, u, \theta) \equiv \frac{-\theta C_{y\theta}(y, u, \theta)}{C_y(y, u, \theta)}.$$

By differentiation of the implicit function C, $C_y = -u_y/u_c = |MRS_t| = 1 - \tau_t^L$ is the marginal rate of substitution between consumption and leisure. Therefore η represents the elasticity $-\frac{d\log|MRS_t|}{d\log\theta_t}$ and plays an important role in this section. In the separable isoelastic utility case above, this elasticity is $\eta(y, u, \theta) = 1 + \frac{1}{\varepsilon}$. Define the co-state $\lambda_t = K_v$ as in the separable utility case. With non-separable utility, λ is still a martingale $d\lambda_t = \sigma_{\lambda,t}\sigma_t dB_t$ but is not the inverse of the marginal utility of consumption since the Inverse Euler equation does not hold. The labor wedge satisfies

$$d\Big(\frac{1}{u_c}\frac{1}{\eta}\frac{\tau_t^L}{1-\tau_t^L}\Big) = [\lambda_t\sigma_{\lambda,t}\sigma_t^2]dt. \tag{26}$$

The no-volatility result generalizes: the stochastic process $\frac{1}{u_c}\frac{1}{\eta}\frac{\tau_t^L}{1-\tau_t^L}$ has zero instantaneous volatility so that its realized paths vary much less than those for productivity, in the sense that they are of bounded variation. To qualify the wedges further, I consider the Greenwood *et al.* (1988) preferences

$$u(c, l) = \frac{1}{1-\nu}\Big(c - \frac{l^{1+\frac{1}{\varepsilon}}}{1+\frac{1}{\varepsilon}}\Big)^{1-\nu} \tag{27}$$

for $\nu > 0$. Then $\eta = 1 + \frac{1}{\varepsilon}$ and the labor wedge satisfies

$$d\Big(\frac{\tau_t^L}{1-\tau_t^L}\frac{1}{u_c}\Big) = [(1+\frac{1}{\varepsilon})\lambda_t\sigma_{\lambda,t}]\sigma_t^2 dt.$$

as well as

$$d\Big(\frac{\tau_t^L}{1-\tau_t^L}\Big) = [(1+\frac{1}{\varepsilon})(\lambda_t u_c)\sigma_{\lambda,t}]\sigma_t^2 dt + \frac{\tau_t^L}{1-\tau_t^L}\frac{1}{u_c}d(u_c). \tag{28}$$

The dynamics of the labor wedge depend on the covariance between growth in λ and log-productivity, the inverse intensive Frisch elasticity of labor supply, $\lambda_t u_c$ (which is one in the separable utility case) and the innovations in marginal of consumption. The first term of labor wedge is positive and pushes the labor wedge up as in the Exogenous Retirement model. The term that mirrors

the marginal utility of consumption is responsible for the composition effect. Therefore as long as high-productivity agents retire earlier than low-productivity agents, the composition effect is active and the average labor wedge is hump-shaped in age. The following lemma shows that it is the case in the first-best problem.

Lemma 3. *Suppose u is a Greenwood et al. (1988)-type utility function. The optimal retirement rule in the first-best is a cut-off rule $\mathscr{T}_R^{fb} = \inf\{t; \theta_t \leq \theta_R^{fb}(t)\}$.*

The proof is in Appendix A. The conjecture could be made from this lemma that in the second-best as well, agents with a history of low productivity shocks retire earlier than agents with a history of high productivity. Hence the composition effect would push for a hump-shaped in age labor wedge in the non-separable utility case as well.

As for retirement consumption, it is constant after retirement as in the separable utility case. However, because the Inverse Euler does not hold, little is known about consumption before retirement and about whether such consumption drops at retirement in the second-best. In the first-best though, the smooth pasting condition implies that marginal utility of consumption is continuous at retirement and consumption drops at retirement $c_{\mathscr{T}_R^+} = c_{\mathscr{T}_R^-} + \frac{\theta_R^{fb}(t)^{1+\varepsilon}}{1+1/\varepsilon}$ to counter the discrete fall in labor.

6.2 Stochastic Lifetime

There is empirical evidence that life expectancy is positively correlated with income.[34] Chetty *et al.* (2016) find that in the United States, between 2001-2014, the gap in life expectancy between the richest 1% and poorest 1% of individuals is 14.6 years.

To model this positive correlation, I assume that there exist an exogenous productivity threshold θ_D such that $T = \mathscr{T}_D = \inf\{t \in \mathbb{R}, \theta_t \leq \theta_D\}$. Then the discounting function after retirement with productivity $\theta \geq \theta_D$ is $g(\theta) = \frac{1}{\rho}\left(1 - \left(\frac{\theta}{\theta_D}\right)^{\gamma-}\right)$ (increasing in current productivity θ) in which $\gamma-$ is the negative solution of $\rho = \mu\gamma + \frac{\sigma^2}{2}\gamma(\gamma-1)$. This modeling choice has the convenience that time is not a state variable of the planner's problem anymore while each agent have a finite expected lifetime.[35] Since the problem is time homogenous, I focus on retirement consumption

[34] Not necessarily causal in one direction or the other.

[35] This allows me in work in progress to have an in-depth look at optimal policies for human capital acquisition in a setting in which life expectancy is positively correlated with income and human capital.

rather than the life-cycle pattern of the wedges. The HJB equation becomes

$$0 = \max_{c_t, y_t, \mathscr{T}_R, \sigma_{\Delta,t}} \Big\{ -K + g(\theta) u_{l=0}^{-1}(\frac{v}{g(\theta)}) \quad , \quad -\rho K + (c_t - y_t) + \mathcal{L}(v, \Delta, \theta, t) \circ K \Big\}$$

where the derivatives operator over state variables $\mathcal{L}(v, \Delta, \theta, t)$ is defined in Appendix A. For a given promised utility v, retirement consumption $u_{l=0}^{-1}(\frac{v}{g(\theta)})$ is decreasing in current productivity. In addition, the net present value of retirement benefits are $g(\theta)u_{l=0}^{-1}(\frac{v}{g(\theta)})$ and for a given promised utility v they are lower for high-productivity agents compared to low-productivity agents.[36] Other things equal, with stochastic lifetime correlated with income, the planner can take advantage of the fact that high-productivity agents have longer life expectancy than the general population in order to give them lower retirement consumption and lower net present value of consumption compared to a model in which the end of the horizon is the average life expectancy $T = \mathrm{E}[\mathscr{T}_D]$.

6.3 Productivity-Dependent Fixed Costs

In this section, I consider the case when the fixed depends on current productivity and age $\phi_t(\theta_t)$. Proof of results on wedges in Appendix A have been done so far under this general case, so that results on wedges are unaffected by this assumption. Only the retirement decisions are left to be determined. The retirement decision depends on ϕ_t', i.e. how fast the fixed cost increases in productivity. I consider two subcases.

6.3.1 Slow-Increasing Fixed Costs

Proposition 5. *(First-best retirement decision) Suppose that for some $\psi > 0$, $\forall(\theta, t)$, $\phi_t'(\theta) \leq \psi\theta^{\varepsilon}$. There exists a time-dependent deterministic productivity threshold $\theta_R^{fb}(t)$ such that, in the first-best, retirement occurs if and only if productivity falls below it: $\mathscr{T}_R^{fb} = \inf\{t; \theta_t \leq \theta_R^{fb}(t)\}$.*

The proof is in Appendix A. Proposition 1 generalizes to productivity-dependent fixed costs as long as the fixed cost of staying in the labor market for high-productivity workers is not too high compared to that of low-productivity workers

Risk Neutrality and Pareto Optimal Retirement To understand how the retirement decision is affected by the dependence of the fixed utility cost in productivity, and compare the first-best

[36] For a concave utility function u, the function $g \mapsto gu^{-1}(v/g)$ is decreasing.

retirement decision to the second-best one, I consider the case where agents are risk neutral that is more tractable than the risk averse agents case.

Consider the case of agents who are risk neutral in consumption and productivity is a GBM. Risk neutrality in consumption implies that consumption need not be distorted. Because of the strict concavity of $u(c)$ in the case of risk-averse agents with a utilitarian planner, the equivalent generalized social marginal welfare weights (as in Saez and Stantcheva (2016)) reflect decreasing marginal utility of consumption. Low-productivity agents have lower consumption and higher marginal utility and therefore higher social welfare weights. To ensure comparability between the risk-averse utilitarian and the risk neutral cases, I assume that the planner puts Pareto welfare weights $\alpha(\theta_0)$ on each agent with initial type θ_0. Since with concave utility, marginal utility of consumption is non-increasing, I assume the function $\alpha : \Theta_0 \mapsto (0; +\infty)$ is non-increasing. I normalize the sum of Pareto weights to one $\int_0^\infty \alpha(\theta_0) dF(\theta_0) = 1$ and call the summand of weights $\Lambda(\theta) = \int_0^\theta \alpha(\theta_0) dF(\theta_0)$.

The following lemma formulates the retirement decision problem by subtituting optimal allocations in the planner's problem.

Lemma 4. *(Allocations and wedges) The labor wedges are time invariant and depend only on initial heterogeneity and the welfare weights*

$$\frac{\tau_t^L}{1-\tau_t^L} = \frac{\tau(\theta_0)}{1-\tau(\theta_0)} = (1+\frac{1}{\varepsilon})\frac{1}{\theta_0}\frac{\Lambda(\theta_0) - F(\theta_0)}{f(\theta_0)} \tag{29}$$

In addition, the planner's problem is to choose the retirement rule so as to solve:

$$\max_{\mathscr{T}_R} \int_0^\infty \mathrm{E}\Big\{ \int_0^{\mathscr{T}_R} e^{-\rho t} \Big[(1-\tau(\theta_0))^\varepsilon [y_t^{fb} - \kappa \frac{(\frac{y_t^{fb}}{\theta_t})^{1+\frac{1}{\varepsilon}}}{1+\frac{1}{\varepsilon}}] - [\phi_t - \frac{\tau(\theta_0)}{1-\tau(\theta_0)} \frac{\varepsilon}{1+\varepsilon} \theta_t \phi_t'(\theta_t)] dt \Big\} dF(\theta_0) \tag{30}$$

The proof of the lemma is in Appendix A. The normalization of Pareto weights and the assumption of non-increasing weights implies that $\Lambda(\theta_0) - F(\theta_0)$ is always non-negative. The labor wedges are therefore non-negative. In the risk neutral case, with GBM productivity, the labor wedges only depend on the inverse intensive Frisch elasticity of labor supply, initial heterogeneity, and the welfare weights of the planner. Because there is no income effect, consumption can be allocated freely over time without distorting the labor margin.

In the context of private information, labor distortions are such that the flow utility of con-

sumption and disutility of labor is lower than it is in the first-best. This is captured by the factor $(1-\tau(\theta_0))^{\varepsilon} < 1$ in front of $[y_t^{fb} - \kappa \frac{(y_t^{fb}/\theta_t)^{1+1/\varepsilon}}{1+1/\varepsilon}]$ in the planner's objective. These labor distortions create incentives for the agents to retire early. However, the virtual fixed cost either increases or decreases depending on the sign of $\phi_t^{'}(\theta_t)$.

If $\phi_t^{'}$ is negative, the virtual fixed cost increases compared to the first-best. Its effect goes in the same direction as the decrease in output y and agents retire earlier than in the first-best. Therefore, if $\phi_t^{'}$ is negative, all agents retire earlier in the second-best compared to the first-best. In addition, retirement is a cut-off rule. If $\phi_t^{'}$ is positive, the virtual fixed cost decreases compared to the first-best and depends negatively on the intensive Frisch elasticity of labor and the labor wedge. Its effect goes in the opposite direction as the decrease in y. Therefore, the distortion on the retirement rule is ambiguous. Suppose there exists $\psi > 0$ such that $\phi_t(\theta_t) = \psi\theta_t$. Having solved the retirement decision problem in the first-best case, the derivation of the analogous rule for the second-best scenario is relatively simple. Dividing the planner's objective by $(1-\tau(\theta_0))^{\varepsilon}$, one can observe that the choice of the retirement rule in the second-best is equivalent to the choice of the retirement rule in the first-best when the fixed utility cost is replaced by a virtual cost $\tilde{\phi}$ defined as $\tilde{\phi}(t,\theta_t) = \frac{\phi(t,\theta_t)}{(1-\tau(\theta_0))^{\varepsilon}}(1 - \frac{\tau(\theta_0)}{1-\tau(\theta_0)}\frac{\varepsilon}{1+\varepsilon})$. In contrast to the first-best case, the retirement rule depends on initial productivity. Defining $S(\tau(\theta_0)) \equiv \tilde{\phi}(t,\theta_t)/\phi(t,\theta_t)$, the following proposition summarizes the results on retirement distortions.

Proposition 6. *(Retirement distortions)*

1. *There exists a time-dependent and initial productivity dependent deterministic retirement threshold $\theta_R^{sb}(t,\theta_0)$ such that $\mathscr{T}_R^{sb} = \inf\{t; \theta_t \leq \theta_R^{sb}(t,\theta_0)\}$.*

2. *Suppose $\phi_t(\theta_t) = \psi\theta_t$ with $\psi \in \mathbb{R}$, at the infinite horizon limit, $T = +\infty$ the retirement thresholds are time-invariant $\hat{\theta}_R^{sb} : \Theta_0 \mapsto \mathbb{R}^{+*}$, $\mathscr{T}_R^{sb} = \inf\{t; \theta_t \leq \theta_R^{sb}(\theta_0)\}$ and*

$$\theta_R^{sb}(\theta_0) = \theta_R^{fb} S(\tau(\theta_0))^{\frac{1}{\varepsilon}}.$$

3. *If $\psi \leq 0$, retirement occurs earlier in the second-best compared to the first-best for all agents $\theta_R^{sb}(t,\theta_0) \geq \theta_R^{fb}(t)$. If $\psi > 0$, a criterion for whether retirement happens early or is delayed compared to the first-best is*

$$S(\theta_0) = \frac{1}{(1-\tau(\theta_0))^{\varepsilon}}(1 - \frac{\tau(\theta_0)}{1-\tau(\theta_0)}\frac{\varepsilon}{1+\varepsilon}).$$

For a given $T < +\infty$, retirement occurs earlier in the second-best compared to the first-best: $\theta_R^{sb}(t, \theta_0) \geq \theta_R^{fb}(t)$ for all $t \leq T$ if and only if $S(\theta_0) \geq 1$.

Point 1 of the proposition highlights that retirement thresholds depend on the initial productivity of the agents. Again, the option of continued work compared to retiring is negative at retirement. The second point gives an explicit formula for the optimal retirement threshold at infinite horizon as in the discussion after Corollary 1.[37] Point 2 gives an explicit expression for the retirement thresholds at infinite horizon.

Point 3 of the proposition states that if the fixed utility cost is increasing in productivity, there is a force that pushes for delayed retirement. High types have a high fixed cost and lower information rents than in the case when the fixed cost is independent of productivity. This creates an effect that goes in the opposite direction of the income tax. Depending on the strength of this effect retirement may occur early or be delayed compared to the first-best. The proposition shows that the relative weight of the two forces depends on the criterion S that in turn depends on the intensive Frisch elasticity of labor and the welfare weights of the planner. This criterion allows one to determine what productivity types should be induced to retire before $S(\theta_0) \geq 1$ or after the first-best $S(\theta_0) < 1$.

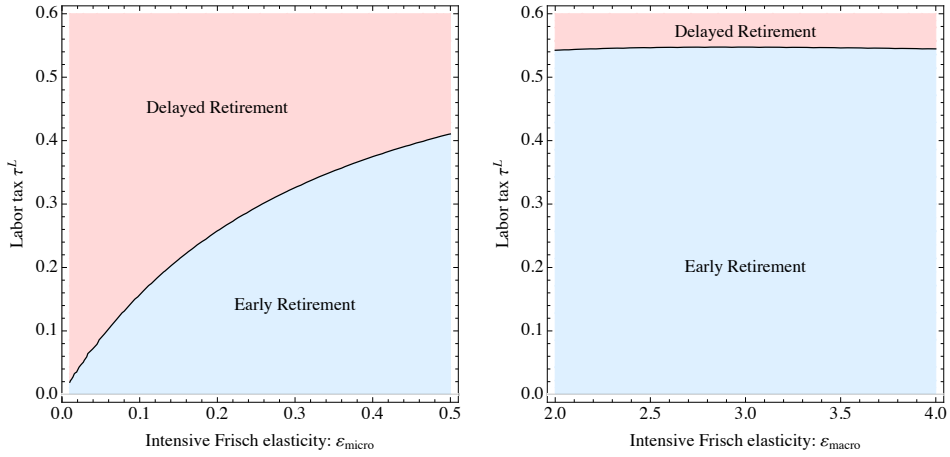

Figure 6: $\{\tau : S(\tau) \geq 1\}$ as a function of ε. On the vertical axis $\tau(\theta_0)$ and on the horizontal axis $\varepsilon_{\text{micro}} \in [0; 0.5]$ on the left and $\varepsilon_{\text{macro}} \in [2; 4]$ on the right. Early retirement in blue (bottom), delayed retirement in red (top).

Figure 6 shows that the size of the intensive Frisch elasticity of labor is important in determining the individual retirement decisions and therefore the optimal hazard rate and labor force

[37] There is no concern for immiseration at infinite horizon here since, with risk neutrality in consumption, consumption is not pinned down by first order conditions.

participation rate of the elderly. The larger the intensive Frisch elasticity, the more agents there are who delay retirement compared to the first-best. Reichling and Whalen (2012) and Peterman (2016) provide a survey of the estimates of the Frisch elasticity of labor supply in the micro and in the macro literature.

Figure 6's left panel illustrates the optimal deviations of retirement compared to the first-best for typical values of the intensive Frisch elasticity of labor supply from the micro literature, with $\varepsilon_{\text{micro}} \in [0; 0.5]$. When ε is small, agents' labor is inelastic and the incentive effect of labor distortions is small. The effect of distortions through rents induced by the fixed cost dominates once an agent faces labor distortions that lie in the red (upper) region. There is a large disparity in optimal retirement behavior. For instance, for a an intensive Frisch elasticity of labor supply of $\varepsilon = 0.2$, agents facing a marginal labor income tax rate[38] below 26% retire early while agents facing a marginal tax rate above 26% delay retirement optimally.

Figure 6's right panel illustrates the optimal deviations of retirement compared to first-best for typical values of the Frisch elasticity of labor from the macro literature, with $\varepsilon_{\text{macro}} \in [2; 4]$. When ε is large, agents' labor is elastic and the incentive effect of labor distortions is large. Therefore, most agents retire earlier than in the first-best. One need a high optimal tax rate, above 54%, for the distortions through rents induced by the fixed costs increasing in productivity for the agents to delay retirement compared to first-best. The curve $\{S(\tau) = 1\}$ asymptotes to around $\tau^L = 54\%$ for large values of ε up to infinity.

This discussion highlights that an accurate estimate of the intensive Frisch elasticity of labor supply and the variations in the extensive elasticity through $\phi^{'}(\theta)$ are important in determining individual retirement decisions and therefore the optimal hazard rate and labor force participation rate of the elderly.

6.3.2 Fast-Increasing Fixed Costs

I assumed that the fixed utility cost of staying in the labor market grows slowly in productivity i.e there exists $\psi > 0$, such that $\forall(\theta, t),\ \phi_t^{'}(\theta) \leq \psi\theta^{\varepsilon}$. This section relaxes this assumption and shows that if the fixed utility cost of staying in the labor market grows fast in productivity, when agents promised utility becomes high, they become too costly to incentivize to work and they retire.

[38] In this setting, allocations can be implemented by non-linear labor income taxes equal to the wedges.

Lemma 5. *Suppose there exists $\psi > 0$ such that $\phi_t(\theta_t) \geq \psi\theta_t^{1+\varepsilon}$. Then, for each t there exists a promised utility v_t^* such that if $v_t \geq v_t^*$, the planner collects more revenue from retiring the agent than from making him work.*

The proof is in Appendix A. The argument of the proof is mechanical and comes directly from the fast growth in $\phi_t(\theta_t)$. The lemma applies to any allocations, even non-incentive compatible ones.

Note that the lemma does not imply directly that under the conditions specified there is an upper retirement boundary since promised utility is an endogenous state variable of the problem. The existence of such a boundary depends on how big the government exogenous revenue $-G$ is to achieve high promised utility. Indeed, if ψ is high it becomes more and more costly to incentivize high types who need to be retired whenever they have accumulated a high promised utility.[39] Under these conditions, both agents with a history of low productivity shocks and agents with a history of high productivity shocks retire earlier than agents with a history of average productivity. Therefore the composition effect is less strong than when $\phi_t(\theta_t)$ is constant or slowly growing in productivity. If ever[40] high-productivity agents experience fixed costs of staying in the labor market much higher than low-productivity agents, or equivalently when the extra benefit of retirement leisure is much higher for the former than the latter, the composition effect pushes for increasing ever more the labor wedge for old workers.

7 Conclusion

This paper studies the optimal design of taxes and retirement benefits in a dynamic life-cycle model with an endogenous flexible retirement age. Individuals adjust their labor supply both through the number of hours worked, an intensive margin, and the timing of their retirement, an extensive margin. A utilitarian government provides insurance and redistributes resources across agents who experience persistent idiosyncratic productivity shocks. Productivity and its evolution are private information of the workers, and the government's goal is to design an incentive compatible mechanism. I obtain the allocations of this second-best problem by following a First Order Approach that

[39] For instance, following the notation in the proof in Appendix A, for log utility the highest promised fixed consumption before retirement occurs is $\bar{c}(t, v_t^*) = 1/K$. This quantity decreases with ψ; therefore when ψ is high the likelihood of an upper retirement boundary being endogenously hit is higher.

[40] On another planet.

relaxes the incentive constraints on the government. I derive the formulas of the optimal implicit taxes, or wedges and describe the forces that determine their evolution. Finally, I describe the optimal retirement decision as a solution of an optimal stopping problem.

A standard result in the dynamic optimal taxation literature is that, when the variance of individual productivity increases with age, and retirement is exogenous, the optimal labor tax increases with age. I qualify this result accounting for an endogenous flexible retirement age. I show through two effects that the optimal labor tax decreases with age for old workers and that the labor tax curve is hump-shaped in age. First, the elasticity effect implies a pattern of the optimal labor tax that is flatter in age relative to an equivalent model without an extensive margin. Second, workers with a history of low productivity shocks retire earlier than workers with a history of high productivity shocks. Through selection, the labor force becomes increasingly productive in old age. When the forces of this novel composition effect are strong enough, the distribution of productivity in the old age labor force features higher mean and lower variance than the one in general population. Setting a decreasing-in-age labor tax for old workers increases the efficiency of the intensive labor supply of this highly productive subpopulation.

Another standard result in the dynamic optimal taxation literature is that simple linear (in income) age-dependent taxes achieve the bulk of the welfare gains from the history-dependent fully optimal system. My simulations show that simple linear age-dependent taxes that mimic the history-dependent system achieve modest welfare gains under the current SS system. I augment this tax reform with a simple SS reform that increases the delayed retirement credits, and I find that this pairing achieves sizeable welfare gains from the fully optimal system. These calibrations suggest that when the endogeneity of retirement is accounted for, introducing age-dependency into the tax code alone is not enough, and one needs reform the SS system as well in order to capture the bulk of welfare gains from optimal policies. Further work would be needed to investigate how robust this result is to other modeling choices such as a different process for productivity or different values of the coefficient of risk aversion.

The theory proposed in this paper leads to two open empirical questions that are important in quantifying the magnitude of optimal policies. Empirical estimates of the time fixed costs and monetary fixed costs of work would improve the calibration of macro models to match micro evidence on extensive margin elasticities. Furthermore, an empirical estimate of the mean and variance of hourly wages among full-time workers age 60-70, would help quantify the strength of

the novel composition effect highlighted in this paper.

References

Aguila, Emma, Attanasio, Orazio, and Meghir, Costas. 2011. Changes in consumption at retirement: evidence from panel data. *Review of Economics and Statistics*, **93**(3), 1094–1099.

Albanesi, Stefania, and Sleet, Christopher. 2006. Dynamic optimal taxation with private information. *The Review of Economic Studies*, **73**(1), 1–30.

Alpert, Abby, and Powell, David. 2013. Estimating Intensive and Extensive Tax Responsiveness: Do Older Workers Respond to Income Taxes?

Atkinson, Anthony Barnes, and Stiglitz, Joseph E. 1976. The design of tax structure: direct versus indirect taxation. *Journal of public Economics*, **6**(1-2), 55–75.

Banks, James, Blundell, Richard, and Tanner, Sarah. 1998. Is there a retirement-savings puzzle? *American Economic Review*, 769–788.

Bergemann, Dirk, and Strack, Philipp. 2015. Dynamic revenue maximization: A continuous time approach. *Journal of Economic Theory*, **159**, 819–853.

Bismut, Jean-Michel. 1973. Conjugate convex functions in optimal stochastic control. *Journal of Mathematical Analysis and Applications*, **44**(2), 384–404.

Chang, Yongsung, Kim, Sun-Bin, Kwon, Kyooho, Rogerson, Richard, *et al.* . 2014. Individual and aggregate labor supply in a heterogeneous agent economy with intensive and extensive margins. *Unpublished Manuscript*.

Chetty, Raj. 2012. Bounds on elasticities with optimization frictions: A synthesis of micro and macro evidence on labor supply. *Econometrica*, **80**(3), 969–1018.

Chetty, Raj, Guren, Adam, Manoli, Day, and Weber, Andrea. 2012. Does Indivisible Labor Explain the Difference between Micro and Macro Elasticities? A Meta-Analysis of Extensive Margin Elasticities. *NBEr macroeconomics Annual*.

Chetty, Raj, Stepner, Michael, Abraham, Sarah, Lin, Shelby, Scuderi, Benjamin, Turner, Nicholas, Bergeron, Augustin, and Cutler, David. 2016. The association between income and life expectancy in the United States, 2001-2014. *Jama*, **315**(16), 1750–1766.

Choné, Philippe, and Laroque, Guy. 2014. Income tax and retirement schemes.

Conesa, Juan Carlos, Kitao, Sagiri, and Krueger, Dirk. 2009. Taxing capital? Not a bad idea after all! *The American economic review*, **99**(1), 25–48.

Cremer, Helmuth, Lozachmeur, Jean-Marie, and Pestieau, Pierre. 2004. Social security, retirement age and optimal income taxation. *Journal of Public Economics*, **88**(11), 2259–2281.

Deaton, Angus, and Paxson, Christina. 1994. Intertemporal choice and inequality. *Journal of political economy*, **102**(3), 437–467.

Di Nunno, Giulia, Øksendal, Bernt Karsten, and Proske, Frank. 2009. *Malliavin calculus for Lévy processes with applications to finance*. Vol. 2. Springer.

Diamond, Peter Arthur, and Mirrlees, James A. 1978. A model of social insurance with variable retirement. *Journal of Public Economics*, **10**(3), 295–336.

Dixit, Avinash. 1993. Art of Smooth Pasting. Vol. 55. *Fundamentals of Pure and Applied Economics*.

Erosa, Andres, and Gervais, Martin. 2002. Optimal taxation in life-cycle economies. *Journal of Economic Theory*, **105**(2), 338–369.

Farhi, Emmanuel, and Werning, Iván. 2013. Insurance and taxation over the life cycle. *The Review of Economic Studies*, **80**(2), 596–635.

French, Eric. 2005. The effects of health, wealth, and wages on labour supply and retirement behaviour. *The Review of Economic Studies*, **72**(2), 395–427.

Golosov, Mikhail, and Tsyvinski, Aleh. 2015. Policy implications of dynamic public finance. *economics*, **7**(1), 147–171.

Golosov, Mikhail, Tsyvinski, Aleh, Werning, Ivan, Diamond, Peter, and Judd, Kenneth L. 2006. New Dynamic Public Finance: A User's Guide [with Comments and Discussion]. *NBER macroeconomics annual*, **21**, 317–387.

Golosov, Mikhail, Troshkin, Maxim, and Tsyvinski, Aleh. 2016. Redistribution and social insurance. *The American Economic Review*, **106**(2), 359–386.

Gomes, Renato, Lozachmeur, Jean-Marie, and Pavan, Alessandro. 2017. Differential taxation and occupational choice. *The Review of Economic Studies*, rdx022.

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

Grochulski, Borys, and Zhang, Yuzhe. 2016. Optimal Contracts with Reflection.

Gruber, Jonathan, and Wise, David. 1998. Social security and retirement: An international comparison. *The American Economic Review*, **88**(2), 158–163.

Gruber, Jonathan, and Wise, David A. 2002 (December). *Social Security Programs and Retirement Around the World: Micro Estimation.* Working Paper 9407. National Bureau of Economic Research.

Hartman, Philip. 2002. *Ordinary differential equations.*

Heathcote, Jonathan, Storesletten, Kjetil, and Violante, Giovanni L. 2005. Two views of inequality over the life cycle. *Journal of the European Economic Association*, **3**(2-3), 765–775.

Heathcote, Jonathan, Perri, Fabrizio, and Violante, Giovanni L. 2010. Unequal we stand: An empirical analysis of economic inequality in the United States, 1967–2006. *Review of Economic dynamics*, **13**(1), 15–51.

Heathcote, Jonathan, Storesletten, Kjetil, and Violante, Giovanni L. 2014. *Optimal tax progressivity: An analytical framework.* Tech. rept. National Bureau of Economic Research.

Heathcote, Jonathan, Storesletten, Kjetil, Violante, Giovanni L, *et al.* . 2017. *Optimal Progressivity with Age-Dependent Taxation.* Tech. rept. Federal Reserve Bank of Minneapolis.

Jacka, SD, and Lynn, JR. 1992. Finite-horizon optimal stopping, obstacle problems and the shape of the continuation region. *Stochastics Stochastics Rep*, **39**(25-42).

Jacquet, Laurence, Lehmann, Etienne, and Van der Linden, Bruno. 2013. Optimal redistributive taxation with both extensive and intensive responses. *Journal of Economic Theory*, **148**(5), 1770–1805.

Kapička, Marek. 2013. Efficient allocations in dynamic private information economies with persistent shocks: A first-order approach. *The Review of Economic Studies*, rds045.

Karabarbounis, Marios. 2016. A road map for efficiently taxing heterogeneous agents. *American Economic Journal: Macroeconomics*, **8**(2), 182–214.

Kushner, Harold, and Dupuis, Paul G. 2013. *Numerical methods for stochastic control problems in continuous time*. Vol. 24. Springer Science and Business Media.

Lazear, Edward P, and Moore, Robert L. 1988. Pensions and turnover. *Pages 163–190 of: Pensions in the US Economy*. University of Chicago Press.

Leland, Hayne E. 1994. Corporate debt value, bond covenants, and optimal capital structure. *The journal of finance*, **49**(4), 1213–1252.

Makris, Miltiadis, and Pavan, Alessandro. 2017. Taxation under Learning-by-Doing.

Michau, Jean-Baptiste. 2014. Optimal redistribution: A life-cycle perspective. *Journal of Public Economics*, **111**, 1–16.

Mirrlees, James A. 1971. An exploration in the theory of optimum income taxation. *The review of economic studies*, **38**(2), 175–208.

Munnell, Alicia H, and Soto, Mauricio. 2005. What replacement rates do households actually experience in retirement?

Pavan, Alessandro, Segal, Ilya, and Toikka, Juuso. 2014. Dynamic mechanism design: A myersonian approach. *Econometrica*, **82**(2), 601–653.

Peterman, William B. 2016. Reconciling micro and macro estimates of the Frisch labor supply elasticity. *Economic Inquiry*, **54**(1), 100–120.

Prescott, Edward C, Rogerson, Richard, and Wallenius, Johanna. 2009. Lifetime aggregate labor supply with endogenous workweek length. *Review of Economic Dynamics*, **12**(1), 23–36.

Reichling, Felix, and Whalen, Charles. 2012. Review of estimates of the Frisch elasticity of labor supply.

Rogerson, Richard, and Wallenius, Johanna. 2013. Nonconvexities, retirement, and the elasticity of labor supply. *The American Economic Review*, **103**(4), 1445–1462.

Rothschild, Casey, and Scheuer, Florian. 2013. Redistributive taxation in the roy model. *The Quarterly Journal of Economics*, **128**(2), 623–668.

Rust, John P. 1989. A dynamic programming model of retirement behavior. *Pages 359–404 of: The economics of aging*. University of Chicago Press.

Saez, Emmanuel. 2001. Using elasticities to derive optimal income tax rates. *The review of economic studies*, **68**(1), 205–229.

Saez, Emmanuel. 2002. Optimal income transfer programs: intensive versus extensive labor supply responses. *The Quarterly Journal of Economics*, **117**(3), 1039–1073.

Saez, Emmanuel, and Stantcheva, Stefanie. 2016. Generalized social marginal welfare weights for optimal tax theory. *The American Economic Review*, **106**(1), 24–45.

Sannikov, Yuliy. 2008. A continuous-time version of the principal-agent problem. *The Review of Economic Studies*, **75**(3), 957–984.

Sannikov, Yuliy. 2014. Moral hazard and long-run incentives. *Unpublished working paper, Princeton University*.

Shourideh, Ali, and Troshkin, Maxim. 2015. *Incentives and efficiency of pension systems*. Tech. rept. Mimeo.

Stantcheva, Stefanie. 2017. Optimal Taxation and Human Capital Policies over the Life Cycle. *Journal of Political Economy*, **125**(6).

Stock, James H, and Wise, David A. 1988. *Pensions, the option value of work, and retirement*.

Storesletten, Kjetil, Telmer, Christopher I, and Yaron, Amir. 2004. Consumption and risk sharing over the life cycle. *Journal of monetary Economics*, **51**(3), 609–633.

Strack, P, and Kruse, T. 2013. *Optimal stopping with private information*. Tech. rept. Mimeo.

Toossi, Mitra. 2015. Labor force projections to 2024: the labor force is growing, but slowly. *Monthly Lab. Rev.*, **138**, 1.

Weinzierl, Matthew. 2011. The surprising power of age-dependent taxes. *The Review of Economic Studies*, **78**(4), 1490–1518.

Williams, Noah. 2011. Persistent private information. *Econometrica*, **79**(4), 1233–1275.

A - Analytic Appendix

1 Proof of Propositions 1 and 5

Proof. The planner's problem is

$$\max_{\{\lambda,c_t,l_t,\mathscr{T}_R\}} \mathrm{E}\Big\{\int_0^T e^{-\rho s}[u(c_t)-\lambda c_t]dt+\int_0^{\mathscr{T}_R} e^{-\rho s}[\lambda\theta_t l_t-\kappa\frac{(l_t)^{1+\frac{1}{\varepsilon}}}{1+\varepsilon}-\phi_t(\theta_t)]dt\Big\}$$

subject to the law of motion of productivity (1). From the optimal allocations $u'(c)=\lambda$ and $\kappa l_t^{\frac{1}{\varepsilon}}=\lambda\theta_t$, denote $\mathrm{E}\Big\{\int_0^T e^{-\rho s}[u(c_t)-\lambda c_t]dt\Big\}=h(\lambda)$. Then the above objective rewrites as

$$\max_{\{\lambda,\mathscr{T}_R\}} h(\lambda)+\mathrm{E}\Big\{\int_0^{\mathscr{T}_R} e^{-\rho t}[\lambda^{1+\varepsilon}\frac{(\theta_t)^{1+\varepsilon}}{\kappa^{\varepsilon}(1+\varepsilon)}-\phi_t(\theta_t)]dt\Big\}.$$

Denote a maximizer by λ^*. By an envelope condition, the expected change in the payoff if retirement is delayed an infinitesimal short time is $\lambda^{*1+\varepsilon}\frac{(\theta_t)^{1+\varepsilon}}{\kappa^{\varepsilon}(1+\varepsilon)}-\phi_t(\theta_t)$. Taking $\psi<\frac{\lambda^{*1+\varepsilon}}{\kappa^{\varepsilon}}$ in the condition of growth bounded from above of $\phi_t(\theta)$ in Proposition 1 or assuming that G is high enough such that marginal utility of consumption $\lambda^{*1+\varepsilon}$ is high and the inequality holds, then the expected change in payoff is increasing in productivity. The dynamic single crossing condition in Strack and Kruse (2013) holds and Theorem 4.3 of Jacka and Lynn (1992) implies that the shape of the stopping region (retirement rule) is determined by a time-varying threshold.

Note that when ϕ_t is independent of productivity, or nonincreasing in productivity, the "bounded growth from above" condition in the Propositon holds, implying Proposition 1. □

2 Proof of Corollary 1

Proof. Consider the infinite horizon model, $T=+\infty$. To ensure convergence of social welfare, I assume

$$\rho>(1+\varepsilon)(\mu+\frac{1}{2}\sigma^2\varepsilon). \tag{31}$$

Social welfare is now time-independent and replacing the HJB equation in this setting is

$$\max\{0-w(\theta),-\rho w(\theta)+\mu\theta w_\theta+\frac{\sigma^2\theta^2}{2}w_{\theta\theta}+\frac{\theta^{1+\varepsilon}}{\kappa^{\varepsilon}(1+\varepsilon)}-\phi\}. \tag{32}$$

I conjecture that the solution is of the following form: there is a threshold θ_R^{fb} such that an agent is retired if and only if his productivity falls below the threshold $\theta_t \leq \theta_R^{fb}$. This implies that $w(\theta) = 0$ for all $\theta \leq \theta_R^{fb}$ and for $\theta > \theta_R^{fb}$, w is a nonnegative solution to the equation

$$-\rho w(\theta) + \mu\theta w_\theta + \frac{\sigma^2\theta^2}{2} w_{\theta\theta} = -\frac{\theta^{1+\varepsilon}}{\kappa^\varepsilon(1+\varepsilon)} + \phi. \tag{33}$$

Moreover, w must be C^1 on its entire domain. This implies that $w(\theta_R^{fb}) = 0$ a value matching condition and $w_\theta(\theta_R^{fb}) = 0$, a smooth pasting condition. Finally, observe that, for $\theta \leq \theta_R^{fb}$, the second term in the right hand side of (32) implies that $\dfrac{\theta^{1+\varepsilon}}{\kappa^\varepsilon(1+\varepsilon)} \leq \phi$ i.e. at retirement and afterwards, the marginal social value of continued work is negative. In particular $\hat{\theta}_R^{fb} \leq \theta^*$.

Define the quadratic polynomial $P(x) = -\rho + \mu x + \frac{\sigma^2}{2}x(x-1)$. The homogeneous equation

$$-\rho w(\theta) + \mu\theta w_\theta + \frac{\sigma^2\theta^2}{2} w_{\theta\theta} = 0 \tag{34}$$

admits the general solution

$$w(\theta) = C_-\theta^{x_-} + C_+\theta^{x_+} \tag{35}$$

in which x_- and and x_+ are the negative and positive roots of P. I find a particular solution for each non-homogenous term, respectively denoted $A\theta^{1+\varepsilon}$ and B in which $A = -\dfrac{1}{\kappa^\varepsilon(1+\varepsilon)P(1+\varepsilon)}$ and $B = -\dfrac{\phi}{\rho}$. By the assumption in (31), $P(1+\varepsilon) < 0$. The sum of these particular solutions $A\theta^{1+\varepsilon} + B$ is the value of social welfare if agents never retire.

By the superposition principle of linear homogenous ODEs the solution takes the form

$$w(\theta) = A\theta^{1+\varepsilon} + B + C_-\theta^{x_-} + C_+\theta^{x_+} \tag{36}$$

for $\theta > \theta_R^{fb}$ and $w(\theta) = 0$ for $\theta \leq \theta_R^{fb}$. From (31) I ensure that $x_+ > 1+\varepsilon$. Since $l^{fb} - \kappa\frac{(l^{fb})^{1+\frac{1}{\varepsilon}}}{1+\frac{1}{\varepsilon}} = \dfrac{\theta^{1+\varepsilon}}{\kappa^\varepsilon(1+\varepsilon)}$ I can conjecture that $w(\theta) =_{\theta\to+\infty} \mathcal{O}(\theta^{1+\varepsilon})$. Therefore $D_+ = 0$.

By the value matching and smooth pasting conditions:

$$A(\theta_R^{fb})^{1+\varepsilon} + B + C_-(\theta_R^{fb})^{x_-} = 0 \tag{37}$$

$$(1+\varepsilon)A\frac{(\theta_R^{fb})^{1+\varepsilon}}{\theta_R^{fb}} + x_-C_-\frac{(\theta_R^{fb})^{x_-}}{\theta_R^{fb}} = 0. \tag{38}$$

Multiplying (37) by x_- and (38) by $\hat{\theta}_R^{fb}$ and subtracting the two yields

$$(1+\varepsilon-x_-)A(\theta_R^{fb})^{1+\varepsilon}=x_-B. \tag{39}$$

Thus the expression of θ_R^{fb} and w in Corollary 1 follow by replacing the values of A and B.

Now in finite horizon, the problem is time dependent and thresholds are time dependent. When time goes to T, the value of waiting for productivity to improve decreases and thresholds converge to θ^*. Only the dynamic single crossing property of the derivative operator is needed in finite horizon for this to hold. This is again an application of Jacka and Lynn (1992). □

3 The First Order Approach

3.1 First Order Approach under Risk Neutrality

I first introduce the First Order Approach (FOA) in the simpler setting in which agents are risk neutral in consumption and productivity is a GBM. I relax incentive compatibility by considering a family of deviations that Bergemann and Strack (2015) call *consistent deviations*. The effect of these deviations on promised utility can be summarized by what Pavan *et al.* (2014) call the *impulse response function*. This FOA is standard in the dynamic contracting literature with persistent shocks.

The value of the agent's productivity if he reports his productivity truthfully is

$$\theta_t=\theta_0\exp((\mu-\frac{\sigma^2}{2})t+\sigma B_t).$$

I define Φ by $\theta_t\equiv\Phi(t,\theta_0,B_t)$ and set the following definition, which is motivated by Bergemann and Strack (2015).

Definition 2. (Consistent deviations). A deviation is called *consistent* if an agent, with real productivity $\theta_t=\Phi(t,\theta_0,B_t)$ and associated initial shock θ_0, misreports his initial shock by announcing $\tilde{\theta}_0\in\Theta_0$ at $t=0$ and continues to misreport $\tilde{\theta}_t=\Phi(t,\tilde{\theta}_0,B_t)$ instead of his true productivity θ_t at all future dates $t\leq T$.

With this definition, an agent who follows a consistent deviation misreports his true type in all future periods. An agent's reported productivity $\tilde{\theta}_t=\Phi(t,\tilde{\theta}_0,B_t)$ would be equal to the productivity

he would have had if his initial shock had been $\tilde{\theta}_0$ instead of θ_0. From these misreports, the planner can infer the true realized path of Brownian shocks B_t. However, since the allocations depend on the history of productivities instead of the Brownian shocks, the inference on the Brownian shocks is not of immediate use for the principal. Bergemann and Strack (2015) show that incentive compatibility with respect to consistent deviations—which is a one-dimensional class of deviations—is sufficient for full incentive compatibility in the risk-neutral and GBM case. This result allows me to derive the incentive-compatible optimal allocations and retirement distortions.

Consider the ex-ante utility at time 0 of an agent with initial productivity θ_0 who announces $\tilde{\theta}_0$ and follows consistent deviations; denoting it $v(\theta_0, \tilde{\theta}_0)$. Then

$$v(\theta_0, \tilde{\theta}_0) = \mathrm{E}^{\{\tilde{\theta}\}}\Big\{ \int_0^T e^{-\rho t} c_t(\tilde{\theta}_0) dt - \int_0^{\mathscr{T}_R(\tilde{\theta}_0)} e^{-\rho t} [\kappa \frac{\Big(\frac{y_t(\tilde{\theta}_0)}{\Phi(t,\theta_0,B_t)}\Big)^{1+\frac{1}{\varepsilon}}}{1+\frac{1}{\varepsilon}} + \phi_t\Big(\Phi(t,\theta_0,B_t)\Big)] dt \Big| \tilde{\theta}_0 \Big\}. \quad (40)$$

Restricting attention to consistent deviations alone, the incentive problem turns into a static one. Truthful reports at time zero are necessary for incentive compatibility, i.e. $v(\theta_0) = \max_{\tilde{\theta}_0} v(\theta_0, \tilde{\theta}_0)$ and an envelope condition allows me to obtain the derivative of ex-ante utility. The sensitivity of ex-ante utility with respect to initial reports satisfies:

$$v_\theta(\theta_0) = \mathrm{E}\Big\{ \int_0^{\mathscr{T}_R} e^{-\rho t} [(1+\frac{1}{\varepsilon})(\frac{\Phi_\theta(t,\theta_0,B_t)}{\theta_t})\kappa \frac{(\frac{y_t}{\theta_t})^{1+\frac{1}{\varepsilon}}}{1+\frac{1}{\varepsilon}} - \Phi_\theta(t,\theta_0,B_t)\phi_t'(\theta_t)] dt \Big| \theta_0 \Big\}. \quad (41)$$

$\Phi_\theta(t,\theta_0,B_t)$ is what Pavan *et al.* (2014) call the *impulse response function* and Bergemann and Strack (2015) call the *stochastic flow* in continuous-time. Here with GBM productivity the stochastic flow is the ratio of current productivity to initial productivity, that is,

$$\Phi_\theta(t,\theta_0,B_t) = \exp((\mu - \frac{\sigma^2}{2})t + \sigma B_t) = \theta_t/\theta_0.$$

Then the incentive compatibility constraint simplifies to

$$v_\theta(\theta_0) = \frac{1}{\theta_0}\mathrm{E}\Big\{ \int_0^{\mathscr{T}_R} e^{-\rho t} [(1+\frac{1}{\varepsilon})\kappa \frac{(\frac{y_t}{\theta_t})^{1+\frac{1}{\varepsilon}}}{1+\frac{1}{\varepsilon}} - \theta_t \phi_t'(\theta_t)] dt \Big| \theta_0 \Big\}. \quad (42)$$

3.2 First Order Approach under Risk Aversion

Here, I relax incentive compatibility by considering specific types of deviations as in the risk neutral case. Suppose the agent has reported his type truthfully until time t, $\{\tilde{\theta}^t\} = \{\theta^t\}$ and then decides

to misreport his type. Since the planner observes continuous reports from the agent, she can construct a process $B_t^{\tilde{\theta}}$ from the reports that evolves according to $dB_t^{\tilde{\theta}} = \frac{d\tilde{\theta}_t - \mu_t \tilde{\theta}_t dt}{\sigma_t \tilde{\theta}_t}$. Under truth-telling, $B_t^{\tilde{\theta}} = B_t$. Therefore, the agent is restricted to reports that make $B_t^{\tilde{\theta}}$ a Brownian motion. The Girsanov Theorem implies that there exist misreports $-\eta_t$ such that $dB_t = dB_t^{\tilde{\theta}} + \eta_t dt$ under the measure $\mathcal{Q}$ of the Brownian motion $B_t^{\tilde{\theta}}$ and gives the formula for the change of measure from $\mathcal{P}$ to $\mathcal{Q}$. An incentive compatible mechanism must be immune to these deviations.

Lemma 6. *(Sensitivity of promised utility) $IC \subseteq FOA$. Moreover, If an allocation $\{c, y, \nu\} \in FOA$ then there exists a process $\{\sigma_{\Delta,t}\}$ such that the sensitivity process $\{\Delta_t\}$ has the integral form:*

$$\Delta_t = \mathrm{E}\Big\{ \int_t^{\mathscr{T}_{\mathscr{R}}} e^{-\rho s}[\mu_s \Delta_s + u_\theta(c_s, \frac{y_s}{\theta_s}) - \phi_t^{'}(\theta_t) + \sigma_{\Delta,s}\sigma_s]ds \Big| \mathcal{F}_t \Big\} \tag{43}$$

Proof. Denote $\{\tilde{\theta}\}$ the process reported by the agent. Let $\theta_t = \theta$ at time t. By Girsanov's theorem, there exists a process $\{\eta\}$ is adapted to $\mathcal{F}_t$ such that

$$d\tilde{\theta}_t = d\theta_t + \eta_t dt = (\theta_t \mu_t + \eta_t)dt + \theta_t \sigma_t dB_t. \tag{44}$$

The agent's problem is to choose controls η_t to maximize promised utility for given allocations $\{c, y\}$ and retirement rule $\mathscr{T}_R$. Denote $\{\theta^\eta\} \equiv \{\tilde{\theta}\}$ the misreports generated by $\{\eta\}$. Global incentive compatibility is equivalent to the fact that the optimal report is truth-telling i.e $\eta_t^\star = 0 \ \forall t$. Now with the FOA, assume that all the controls $\eta_s, \forall s \in [0, t)$ have been equal to 0 so far. Promised utility at time t given the control η is

$$w_t(\theta, \theta^\eta) = \sup_{\{\eta\}} E\Big\{ \int_t^{\mathscr{T}_R(\eta)} e^{-\rho(s-t)} \Big[u\Big(c_s(\eta), \frac{y_s(\eta)}{\theta_s}\Big) - \phi_s(\theta_s) \Big] ds + \int_{\mathscr{T}_R(\eta)}^T e^{-\rho(s-t)}[u(c_s(\eta), 0)]ds \Big| \mathcal{F}_t^\eta \Big\}. \tag{45}$$

The expectation above is taken with respect to the realization of the process $\{\tilde{\theta}\}$, since it is reports that determines the allocation and the retirement rule. If the agent follows a process η then

$$dB_t^\eta = \frac{d\theta_t^\eta - ((\theta_t^\eta - \int_0^t \eta_s ds)\mu_t + \eta_t)dt}{(\theta_t^\eta - \int_0^t \eta_s ds)\sigma_t} \tag{46}$$

forms a standard Brownian motion. Therefore, there is exists nonnegative process γ^η and some sensitivity process $Y^{'\eta}$ such that

$$dw_t(\theta_t, \theta_t^\eta) = (\rho w_t(\theta_t, \theta_t^\eta) - u(c_t, \frac{y_t}{\theta_t}) + \phi_t(\theta_t))dt - \gamma_t^\eta dt + \sigma_t Y_t^{'\eta} dB_t^\eta.$$

Then replacing the standard Brownian from (46) in this equation we have

$$dw_t(\theta_t,\theta_t^\eta)=(\rho w_t(\theta_t,\theta_t^\eta)-u+\phi)dt-\gamma_t^\eta dt+\sigma_t Y_t^\eta[d\theta_t^\eta-((\theta_t^\eta-\int_0^t\eta_s ds)\mu_t+\eta_t)dt]. \tag{47}$$

Since the dependence on past controls $\eta=0$ is completely captured by the current value of θ^η, $v_t=w_t(\theta_t,\theta^{\eta=0})$. Ito's formula implies that

$$dw_t(\theta_t,\theta_t^\eta)=\partial_t w_t(\theta_t,\theta^\eta)dt+\partial_{\theta^\eta}w_t(\theta_t,\theta_t^\eta)(\theta_t\mu_t+\eta_t)dt+\partial_{\theta^\eta}w_t(\theta_t,\theta_t^\eta)\theta_t\sigma_t dB_t+\frac{1}{2}\partial^2_{(\theta^\eta)^2}w_t(\theta_t,\theta_t^\eta)\theta_t^2\sigma_t^2 dt. \tag{48}$$

The equation (47) becomes with the FOA $\eta_s=0,\forall s\in[0,t)$:

$$dw_t(\theta_t,\theta_t^\eta)=(\rho w_t(\theta_t,\theta_t^\eta)-u(c_t,\frac{y_t}{\theta_t})+\phi_t(\theta_t))dt-\gamma_t^\eta dt+\theta_t^\eta\sigma_t Y_t^\eta dB_t.$$

Comparing equations (48) and (47) and equalizing their drifts yield:

$$\partial_t w_t(\theta_t,\theta_t^\eta)+\partial_{\theta^\eta}w_t(\theta_t,\theta_t^\eta)(\theta_t\mu_t+\eta_t)+\frac{1}{2}\partial^2_{(\theta^\eta)^2}w_t(\theta_t,\theta_t^\eta)\theta_t^2\sigma_t^2=(\rho w_t(\theta_t,\theta_t^\eta)-u(c_t,\frac{y_t}{\theta_t})+\phi_t(\theta_t))dt-\gamma_t^\eta dt.$$

Now I obtain the Hamilton-Jacobi-Bellman equation for w_t

$$\rho w_t(\theta_t,\theta_t^\eta)=\sup_{\eta_t}\left\{\partial_t w_t(\theta_t,\theta^\eta)+\partial_{\theta^\eta}w_t(\theta_t,\theta_t^\eta)(\theta_t\mu_t+\eta_t)+\frac{1}{2}\partial^2_{(\theta^\eta)^2}w_t(\theta_t,\theta_t^\eta)\theta_t^2\sigma_t^2+u(c_t,\frac{y_t}{\theta_t})-\phi_t(\theta_t)\right\}.$$

Therefore following Theorem 3.1, p. 95 in Hartman (2002), The envelope theorem implies[41]

$$\rho\partial_\theta w_t(\theta_t,\theta_t^\eta)=\partial_{t,\theta}w_t(\theta_t,\theta^\eta)+\partial^2_{\theta^\eta,\theta}w_t(\theta_t,\theta_t^\eta)(\theta_t\mu_t+\eta_t)+\partial_{\theta^\eta}w_t(\theta_t,\theta_t^\eta)\mu_t+\frac{1}{2}\partial^3_{(\theta^\eta)^2,\theta}w_t(\theta_t,\theta_t^\eta)\theta_t^2\sigma_t^2$$

$$+\partial^2_{(\theta^\eta)^2}w_t(\theta_t,\theta_t^\eta)\theta_t\sigma_t^2+u_\theta(c_t,\frac{y_t}{\theta_t})-\phi_t'(\theta_t).$$

This expression can be evaluated at $\eta_t=0$, writing $\frac{\partial w_t(x,\theta)}{\partial\theta}=\Delta_t(x,\theta)$ and considering the fact that when $\eta_t=0$ we have $\partial w_{\theta^\eta}(\theta,\theta^\eta)=\Delta_t$, so that

$$\rho\Delta_t=\partial_t\Delta_t+\partial_\theta\Delta_t(\theta_t\mu_t+0)+\Delta_t\mu_t+\frac{1}{2}\partial^2_{(\theta)^2}(\Delta_t)\theta_t^2\sigma_t^2+\partial_\theta\Delta_t\theta_t\sigma_t^2+u_\theta(c_t,\frac{y_t}{\theta_t})-\phi_t'(\theta_t).$$

[41] For a fully rigorous argument, one needs to make regularity assumptions on $\mathscr{T}_R$ and use Mallliavin calculus to differentiate with respect to stochastic processes. See Di Nunno *et al.* (2009).

The Feynman-Kac formula applies to this differential equation and we deduce that

$$\Delta_t = \mathrm{E}\Big\{ \int_0^{\mathscr{T}_R} e^{-\rho s}[\Delta_s \mu_s - u_\theta(c_t, \frac{y_t}{\theta_t}) + \phi_t'(\theta_t) + \partial_\theta \Delta_s \theta_s \sigma_s^2]ds + \Delta_{\mathscr{T}_R} \Big| \mathcal{F}_\mathrm{t} \Big\}.$$

After retirement, an optimal allocation must give constant consumption. Therefore the sensitivity is zero at retirement. This with $\partial_\theta \Delta_s \theta_s = \sigma_{\Delta,s}$, implies the result:

$$\Delta_t = \mathrm{E}\Big\{ \int_0^{\mathscr{T}_R} e^{-\rho s}[\Delta_s \mu_s - u_\theta(c_t, \frac{y_t}{\theta_t}) + \phi_t'(\theta_t) + \sigma_{\Delta,s} \sigma_s^2]ds \Big| \mathcal{F}_\mathrm{t} \Big\}.$$

□

Hamilton-Jacobi-Bellman Equation First, for the sake of legibility we drop the state 4-tuple (v, Δ, θ, t) from the notation. The associated Hamilton-Jacobi-Bellman equation to this problem is then:

$$0 = \max_{c_t, y_t, \sigma_{\Delta,t}} \Big\{ -K + g(t) u_{l=0}^{-1}\Big(\frac{v}{g(t)}\Big) \quad , \quad -\rho K + (c_t - y_t) + \mathcal{L}(v, \Delta, \theta, t) \circ K \Big\} \tag{49}$$

in which $\mathcal{L}(v, \Delta, \theta, t)$ is the derivative operator with respect to state variables:

$$\begin{aligned} \mathcal{L}(v, \Delta, \theta, t) \circ K = {} & K_v[\rho v_t - u + \phi_t] + K_\Delta[(\rho - \mu)\Delta_t - u_\theta + \phi_t^{'} - \sigma_{\Delta,t}\sigma] + K_t + K_\theta \theta_t \mu \\ & + \frac{1}{2} K_{vv} \theta_t^2 \Delta_t^2 \sigma^2 + \frac{1}{2} K_{\Delta\Delta} \sigma_{\Delta,t}^2 \sigma^2 + \frac{1}{2} K_{\theta\theta} \theta_t^2 \sigma^2 \\ & + K_{v\Delta} \theta_t \Delta_t \sigma_{\Delta,t} \sigma^2 + K_{v\theta} \theta_t^2 \Delta_t \sigma^2 + K_{\Delta\theta} \theta_t \sigma_{\Delta,t} \sigma^2. \end{aligned} \tag{50}$$

The first component of the right-hand side of this dynamic equation captures that once an agent is retired with promised utility v, the cost of providing such utility is the discounted value of the flow consumption $u_{l=0}^{-1}(\frac{v}{g(t)})$. The second component captures the fact that before retirement, the flow cost over an infinitesimal time dt is the discounted cost $-\rho K dt$, flow consumption minus output, and the derivatives of the cost function with respect to state variables. By optimality, these should sum up to zero in the working region.

4 Proof of Lemma 2

Proof. For given consumption, output, $\{c, y\}$ and retirement rule $\mathscr{T}_R$, the expected utility of an agent is at time t is:

$$v_t = \mathrm{E}\Big\{ \int_t^{\mathscr{T}_R} e^{-\rho(s-t)} u(c_s, \frac{y_s}{\theta_s}) ds + \int_{\mathscr{T}_R}^{T} e^{-\rho(s-t)} u(c_s, 0) ds \Big| \mathcal{F}_t \Big\}$$

Then

$$e^{-\rho t} v_t + \int_0^t e^{-\rho s} u(c_s, \frac{y_s}{\theta_s}) ds = \mathrm{E}\Big\{ \underbrace{\int_0^{\mathscr{T}_R} e^{-\rho s} u(c_s, \frac{y_s}{\theta_s}) ds + \int_{\mathscr{T}_R}^{T} e^{-\rho s} u(c_s, 0) ds}_{W} \Big| \mathcal{F}_t \Big\} \equiv W_t.$$

By iterated expectation, W_t is a martingale. By the Martingale Representation Theorem, there exists a square integrable process such that $W_t = \mathrm{E}[W] + \int_0^t \sigma_s'^v dBs$. This implies that $e^{-\rho t} v_t = \mathrm{E}[Y] - \int_0^t e^{-\rho s} u(c_s, \frac{y_s}{\theta_s}) ds + \int_0^t \sigma_s'^v dBs$. Therefore $e^{-\rho t} v_t$ is an Ito process. Applying Ito's lemma,

$$dv_t = (\rho v_t - u + h) dt + \sigma_t^v dB_t$$

in which $\sigma_t^v = e^{rt} \sigma_t'^v$. By Feynman-Kac, $\sigma_t^v = \theta_t \Delta_t \sigma_t$ and

$$dv_t = (\rho v_t - u + h) dt + \theta_t \Delta_t \sigma_t dB_t$$

with the initial value condition

$$v_0 = v.$$

The law of motion of the sensitivity process is a direct application of this idea to Lemma (6). □

5 Proof of Proposition 2

Proof. Applying Ito's lemma to $\lambda_t = K_v(v_t, \Delta_t, \theta_t, t)$ yields

$$d\lambda_t = \mathcal{L}(v_t, \Delta_t, \theta_t, t) \circ K_v dt + (K_{vv} \theta_t \Delta_t + K_{v\Delta} \sigma_{\Delta,t} + K_{v\theta} \theta_t) \sigma_t dB_t.$$

Using the envelope theorem, differentiate HJB with respect to v to get $-\rho K_v - \mathcal{L}(v_t, \Delta_t, \theta_t, t) \circ K_v + \rho K_v = 0$, i.e $\mathcal{L}(v_t, \Delta_t, \theta_t, t) \circ K_v = 0$. Therefore, the drift of $d\lambda_t$ is zero and λ_t is a martingale.

The drift process is determined by $\sigma_{c,t} = K_{vv}\theta_t\Delta_t + K_{v\Delta}\sigma_{\Delta,t} + K_{v\theta}\theta_t$. □

6 Proof of Proposition 3

Proof. Applying Ito's lemma to $y_t = K_\Delta(v_t, \Delta_t, \theta_t, t)$ yields

$$d\gamma_t = \mathcal{L}(v_t, \Delta_t, \theta_t, t) \circ K_\Delta dt + (K_{\Delta v}\theta_t\Delta_t + K_{\Delta\Delta}\sigma_{\Delta,t} + K_{\Delta\theta}\theta_t)\sigma_t dB_t.$$

Using the envelope theorem, differentiate HJB with respect to Δ to get

$$-\rho K_\Delta - \mathcal{L}(v_t, \Delta_t, \theta_t, t) \circ K_v + (\rho - \mu_t)K_\Delta + K_{vv}\theta_t^2\Delta_t\sigma_t^2 + K_{v\Delta}\theta_t\sigma_{\Delta,t}\sigma_t^2 = 0$$

using this equation, the first order condition for $\sigma_{\Delta,t}$ and the expression for $\sigma_{c,t}$, the drift of γ_t is $(-\theta_t\lambda_t\sigma_{c,t}\sigma_t^2 dt + \mu_t\gamma_t)dt$ and the drift is $\gamma_t\sigma_t dB_t$. Hence the result. □

7 Proof of Lemma 3

Proof. Denote λ the Lagrangian on the government's resource constraint. The first order condition on c_t when an agent works is $\left(c_t - \frac{l_t^{1+\frac{1}{\varepsilon}}}{1+\frac{1}{\varepsilon}}\right)^{-\nu} = \lambda$ and $c_t^{-\nu} = \lambda$ when an agent is retired. The first order condition for the labor of workers is $l_t^{\frac{1}{\varepsilon}}\lambda = \lambda\theta_t$ so that $l_t = \theta_t^\varepsilon$. After rearranging and simplifying, the terms inλ cancel out and the planner's retirement problem is rewritten as:

$$\max_{\{\lambda, \mathscr{T}_R\}} \mathrm{E}\Big\{\int_0^{\mathscr{T}_R} e^{-\rho t}[\lambda\frac{(\theta_t)^{1+\varepsilon}}{(1+\varepsilon)} - \phi_t(\theta_t)]dt\Big\}.$$

The proof ends as in the proof of Propositon 1 applying Theorem 4.3 in Jacka and Lynn (1992). □

8 Proof of Lemma 4

Proof. The problem of the planner is to choose allocations $\{c, y\}$ and a retirement rule $\mathscr{T}_R$ to maximize social welfare subject to the definition of ex-ante utility, the resource constraint (4), the relaxed incentive compatibility constraint (42) and the law of motion of productivity (1). I rewrite

the problem below for reading convenience.

$$
\begin{aligned}
&\max_{\{c,y,v,\mathscr{T}_R\}} \int_0^{\infty} \alpha(\theta_0) v(\theta_0) dF(\theta_0) \\
&\text{s.to } \frac{d\theta_t}{\theta_t} = \mu dt + \sigma dB_t \\
&\quad v(\theta_0) = \mathrm{E}_0\Big\{ \int_0^T e^{-\rho t} c_t dt - \int_0^{\mathscr{T}_R} e^{-\rho t} [\kappa \frac{(\frac{y_t}{\theta_t})^{1+\frac{1}{\varepsilon}}}{1+\frac{1}{\varepsilon}} + \phi_t] dt \Big| \theta_0 \Big\} \\
&\quad 0 \leq \mathrm{E}\Big\{ \int_0^{\mathscr{T}_R} e^{-\rho t} y_t dt \Big\} - \mathrm{E}\Big\{ \int_0^T e^{-\rho t} c_t dt \Big\} \\
&\quad v_\theta(\theta_0) = \frac{1}{\theta_0} \mathrm{E}_0\Big\{ \int_0^{\mathscr{T}_R} e^{-\rho t} [(1+\frac{1}{\varepsilon}) \kappa \frac{(\frac{y_t}{\theta_t})^{1+\frac{1}{\varepsilon}}}{1+\frac{1}{\varepsilon}} - \theta_t \phi_t'(\theta_t)] dt \Big| \theta_0 \Big\} \qquad \text{(FOA)}
\end{aligned}
$$

Eliminate consumption from the problem by plugging the definition of ex-ante utility at time zero into the feasibility constraint (4). The feasibility constraint then becomes:

$$
\int_0^{\infty} \Big(v(\theta_0) + \mathrm{E}_0\Big\{ \int_0^{\mathscr{T}_R} e^{-\rho t} \Big[\kappa \frac{(\frac{y_t}{\theta_t})^{1+\frac{1}{\varepsilon}}}{1+\frac{1}{\varepsilon}} + \phi_t \Big] dt \Big| \theta_0 \Big\} \Big) dF(\theta_0) \leq \int_0^{\infty} \mathrm{E}_0\Big\{ \int_0^{\mathscr{T}_R} e^{-\rho t} y_t dt \Big| \theta_0 \Big\} dF(\theta_0). \tag{51}
$$

Denote by λ the multiplier on the new feasibility constraint (51). If $v(\theta_0)$ is interior, the first order conditions on v: $\alpha(\theta_0) f(\theta_0) - \lambda f(\theta_0) = 0$ integrated over Θ_0 yields $\lambda = 1$. The problem si then to maximize the Lagrangian

$$
\begin{aligned}
\int_0^{\infty} \alpha(\theta_0) v(\theta_0) dF(\theta_0) - \Big[\int_0^{\infty} \Big(v(\theta_0) + \mathrm{E}_0\Big\{ \int_0^{\mathscr{T}_R} e^{-\rho t} [\kappa \frac{(\frac{y_t}{\theta_t})^{1+\frac{1}{\varepsilon}}}{1+\frac{1}{\varepsilon}} + \phi_t] dt \Big| \theta_0 \Big\} \Big) dF(\theta_0) \\
- \int_0^{\infty} \mathrm{E}_0\Big\{ \int_0^{\nu} e^{-\rho t} y_t dt \Big| \theta_0 \Big\} dF(\theta_0) \Big]
\end{aligned}
$$

subject to the incentive constraints from the FOA (42)and the law of motion of productivity (1). By partial integration

$$
\begin{aligned}
\int_0^{\infty} v(\theta_0) dF(\theta_0) &= \int_0^{\infty} \frac{1 - F(\theta_0)}{f(\theta_0)} v_\theta(\theta_0) dF(\theta_0) + \lim_{\theta \to 0} v(\theta) \\
\int_0^{\infty} \alpha(\theta_0) v(\theta_0) dF(\theta_0) &= \int_0^{\infty} \frac{1 - \Lambda(\theta_0)}{f(\theta_0)} v_\theta(\theta_0) dF(\theta_0) + \lim_{\theta \to 0} v(\theta).
\end{aligned}
$$

Eliminating v from the Lagrangian using partial integration and the expression of v_θ from in the incentive compatibility constraint, the planner's problem becomes

$$\int_0^\infty \mathrm{E}_0\Big\{\int_0^{\mathscr{T}_R} e^{-\rho t}\Big[y_t-\kappa\frac{(\frac{y_t}{\theta_t})^{1+\frac{1}{\varepsilon}}}{1+\frac{1}{\varepsilon}}\big[1+(1+\frac{1}{\varepsilon})\frac{\Lambda(\theta_0)-F(\theta_0)}{f(\theta_0)}\frac{1}{\theta_0}\big]-[\phi_t-\frac{\Lambda(\theta_0)-F(\theta_0)}{f(\theta_0)}\frac{\theta_t}{\theta_0}\phi_t'(\theta_t)]\Big]dt\Big|\theta_0\Big\}\Big\}dF(\theta_0). \tag{52}$$

The first order condition for y_t implies that the labor wedge is time invariant and depends only on initial heterogeneity and the welfare weights.

$$\frac{\tau_t^L}{1-\tau_t^L}=\frac{\tau(\theta_0)}{1-\tau(\theta_0)}=(1+\frac{1}{\varepsilon})\frac{1}{\theta_0}\frac{\Lambda(\theta_0)-F(\theta_0)}{f(\theta_0)}.$$

Since $y_t^{fb}-\kappa\frac{(\frac{y_t^{fb}}{\theta_t})^{1+\frac{1}{\varepsilon}}}{1+\frac{1}{\varepsilon}}=\frac{\theta_t^{1+\varepsilon}}{\kappa^\varepsilon(1+\varepsilon)}$ and $y_t^{sb}-\kappa\frac{(\frac{y_t^{sb}}{\theta_t})^{1+\frac{1}{\varepsilon}}}{1+\frac{1}{\varepsilon}}\big[1+(1+\frac{1}{\varepsilon})\frac{\Lambda(\theta_0)-F(\theta_0)}{f(\theta_0)}\frac{1}{\theta_0}\big]=(1-\tau(\theta_0))^\varepsilon\frac{\theta_t^{1+\varepsilon}}{\kappa^\varepsilon(1+\varepsilon)}$ then I can replace y^{sb} in the planner's objective (52) to obtain

$$\max_\nu\int_{\underline{\theta}}^\infty \mathrm{E}\Big\{\int_0^{\mathscr{T}_R} e^{-\rho t}\Big[(1-\tau(\theta_0))^\varepsilon[y_t^{fb}-\kappa\frac{(\frac{y_t^{fb}}{\theta_t})^{1+\frac{1}{\varepsilon}}}{1+\frac{1}{\varepsilon}}]-[\phi_t-\frac{\tau(\theta_0)}{1-\tau(\theta_0)}\frac{\varepsilon}{1+\varepsilon}\theta_t\phi_t'(\theta_t)]dt\Big\}dF(\theta_0). \tag{53}$$

□

9 Proof of Lemma 5

Proof. For a fixed θ, the function $y\mapsto\frac{h(\frac{y}{\theta})+\phi_t(\theta_t)}{y}$ is minimized at a y that satisfies $\frac{1}{\theta}h'(\frac{y}{\theta})=\frac{h(\frac{y}{\theta})+\phi_t(\theta_t)}{y}$ (marginal utility cost equals average utility cost). This yields $\frac{y_{min}}{\theta}=\Big(\frac{\phi_t(\theta)(1+\varepsilon)}{\kappa}\Big)^{\frac{\varepsilon}{1+\varepsilon}}$ and the minimum value of average cost is $\frac{1}{\theta}h'(\frac{y_{min}}{\theta})=\kappa^{\frac{\varepsilon}{1+\varepsilon}}\frac{((1+\varepsilon)\phi_t(\theta_t))^{\frac{1}{1+\varepsilon}}}{\theta_t}$. With the assumption on ϕ_t I have uniformly on θ and t, $h(\frac{y_t}{\theta_t})+\phi_t(\theta_t)\geq Ky_t$ in which $K=\kappa^{\frac{\varepsilon}{1+\varepsilon}}((1+\varepsilon)\psi)^{\frac{1}{1+\varepsilon}}$.

For any v_t and t define $\bar{c}$ the constant consumption level which, given continually to the agent after t, gives him an expected utility of v_t: $g(t)u(\bar{c}(t,v_t))=v_t$. Also define v_t^* by $u'(\bar{c}(t,v_t^*))=K$. Such a level exists provided that $u'(0)>K$, a condition without which the agent would never work even in the full information solution (and which is true by definition for log utility). Then for $v_t\geq v_t^*$ the agent does not work and the optimal contract is $c_{t'}=\bar{c}(t,v_t)$ for all $t'\geq t$. To see this, let $v_t\geq v_t^*$, then $u'(\bar{c}(t,v_t))\leq K$. From concavity of u and inequality on h,

$$v_t=\mathrm{E}\Big(\int_t^T e^{-r(s-t)}(u(c_s)-1_{s\leq\mathscr{T}_R}[h(\frac{y_s}{\theta_s})+\phi_s(\theta_s)])ds\Big)\leq\mathrm{E}\Big(\int_t^T e^{-r(s-t)}(u(\bar{c}(t,v_t))$$

$$+(c_s - \bar{c}(t, v_t))u'(\bar{c}(t, v_t)) - 1_{s \leq \mathscr{T}_R} K y_s) ds \Big)$$

$$\leq g(t)u(\bar{c}(t, v_t)) - u'(\bar{c}(t, v_t))\mathrm{E}\Big(\int_t^T e^{-r(s-t)}(1_{s \leq \mathscr{T}_R} y_s - c_s) ds + g(t)\bar{c}(t, v_t) \Big).$$

Since $v_t = g(t)u(\bar{c}(t, v_t))$ and $u' \geq 0$, the revenue from any allocation (c, y) is less than $-g(t)\bar{c}(t, v_t)$ which is the revenue from retiring the agent with constant consumption $\bar{c}(t, v_t)$. It follows that for $v_t \geq v_t^*$ the agent does not work. □

B - Computational Appendix

1 Dynamic Mirrlees Model Numerical Algorithm

1.1 Planning Problem

I do a numerical simulation of a discrete time version of the model. I present the discrete time model and the algorithm of the numerical simulation below. An agent working until time t, reports a productivity history θ^t and the planner recommends $\{c(\theta^t), y(\theta^t), v(\theta^t), \Delta(\theta^t), s(\theta^t)\}$. A retirement decision s equal to zero means the agent works work in period $t+1$ and equal to one means the agents retires forever independently of θ_{t+1}.

Define $u(c, y; \theta) = u(c, \frac{y}{\theta})$ and $f^t(\theta_t|\theta_{t-1})$ the conditional density of θ_t. With the savings rate denoted q^{-1}, the planner's problem is to minimize the cost K such that, for a working agent $s = 0$:

$$K(v, \Delta, \theta_-, t, 0) = \min \Big[\int \{c(\theta) - y(\theta) + qK(v(\theta), \Delta(\theta), \theta, t+1, \tau(\theta)) f^t(\theta_t|\theta_-) d\theta \Big]$$

subject to for all $\theta \in \Theta$

$$w(\theta) = u(c(\theta), y(\theta); \theta) - \phi_t(\theta) + \beta v(\theta)$$

$$\dot{w}(\theta) = u_\theta(c(\theta), y(\theta); \theta) - \phi_\theta(\theta) + \beta \Delta(\theta)$$

And

$$v = \int w(\theta) f^t(\theta|\theta_-) d\theta$$

$$\Delta = \int w(\theta^t) \partial_{\theta_-} f^t(\theta|\theta_-) d\theta.$$

Define

$$\beta^t_{fact} = \frac{1 - \beta^{T+1-t}}{1 - \beta}.$$

For a retired agent $s = 1$ and $\Delta = 0$:

$$K(v, 0, \theta, t+1, 1) = \beta_{fact}^{t+1} u^{-1}\Big(\frac{v}{\beta_{fact}^{t+1}}\Big).$$

The relaxed planning problem can be recovered by setting $t = 1$ and treating Δ a as control variable:

$$K(v) = \min_{\Delta} K(v, \Delta, \theta_0, 1, 0).$$

1.2 Normalization

The process for productivity is a geometric random walk: $\theta_t = \theta_{t-1}\varepsilon_t$ in which ε_t is log-normal $\log \varepsilon_t \sim N(-\frac{\sigma^2}{2}, \sigma^2)$. Preferences are separable in consumption and labor and $u(c_t) = \log(c_t)$ and I denote $h(y_t/\theta_t)$ the disutility of labor. The fixed cost of staying in the labor market is a funtion of age $\phi(t)$. To reduce the number of state variables I re-normalize $\tilde{y}_t \equiv y_t/\theta_{t-1}$, $\tilde{c}_t \equiv c_t/\theta_{t-1}$, $h(y_t/\theta_t) = h(\tilde{y}_t/\varepsilon_t)$.

Denote g the density of ε_t. The densities of θ_t and ε_t are linked by $f(\theta_t|\theta_{t-1})d\theta_t = g(\varepsilon_t)d\varepsilon_t$ and $\partial_{\theta_{t-1}} f(\theta_t|\theta_{t-1})d\theta_t = \frac{1}{\theta_{t-1}}(g(\varepsilon_t) + \varepsilon_t g'(\varepsilon_t))d\varepsilon_t$ (See the derivation in Stantcheva (2017)). Denote $\tilde{g}(\varepsilon_t) = g(\varepsilon_t) + \varepsilon_t g'(\varepsilon_t)$.

Normalized continuation variables are defined as:

$$\begin{aligned}\tilde{v}_t \equiv & \mathrm{E}\Big(\sum_{s=t+1}^{\mathscr{T}_R(\theta^t)} \beta^{s-t-1}(\log(c_s/\theta_t) - h(y_s/\theta_s) - \phi(s)) + \sum_{s=\tau(\theta^t)+1}^{T} \beta^{s-t-1}\log(c_s/\theta_t)\Big) \\ = & v_t - \beta_{t+1}^{fact}\log(\theta_t),\end{aligned}$$

$$\begin{aligned}\tilde{w}_t(\theta^t) \equiv & u(\tilde{c}_t) - h(\tilde{y}_t/\varepsilon_t) - \phi(t) + \beta\Big(\sum_{s=t+1}^{\tau(\theta^t)} \beta^{s-t-1}(\log(c_s/\theta_{t-1}) - h((y_s/\theta_{t-1})/(\theta_s/\theta_{t-1})) - \phi(s)) \\ & + \sum_{s=\tau(\theta^t)+1}^{T} \beta^{s-t-1}\log(c_s/\theta_{t-1})\Big) \\ = & u(\tilde{c}_t) - h(\tilde{y}_t/\varepsilon_t) - \phi(t) + \beta\tilde{v}_t + \beta_t^{fact}\log(\varepsilon_t) \\ = & w_t - \beta_t^{fact}\log(\theta_{t-1}),\end{aligned}$$

$$\tilde{\Delta}_{t-1} \equiv \Delta_{t-1}/\theta_{t-1}.$$

Renormalized constraints The promise-keeping constraint

$$v_{t-1} = \int w_t(\theta_t) f^t(\theta_t|\theta_{t-1}) d\theta_t$$

implies

$$\tilde{v}_{t-1} + \beta_t^{fact}\log(\theta_{t-1}) = \int [\tilde{w}_t(\theta_t) + \beta_t^{fact}\log(\theta_{t-1})] f^t(\theta_t|\theta_{t-1}) d\theta_t.$$

Therefore

$$\tilde{v}_{t-1} = \int \tilde{w}_t(\varepsilon_t) g_\epsilon(\varepsilon_t) d\varepsilon_t.$$

Sensitivity of promised utility

$$\Delta_{t-1} = \int w_t(\theta_t) \partial_{\theta_{t-1}} f^t(\theta_t|\theta_{t-1}) d\theta_t$$

becomes

$$\Delta_{t-1} = \int [\tilde{w}_t(\varepsilon_t) + \beta_t^{fact}\log(\theta_{t-1})] g^t(\theta_t|\theta_{t-1}) d\theta_t.$$

The integral in log is zero because it's the derivative of the expectation of a constant. Therefore

$$\Delta_{t-1} = \int \tilde{w}_t(\varepsilon_t) \frac{\tilde{g}(\varepsilon_t)}{\theta_{t-1}} d\varepsilon_t$$

and

$$\tilde{\Delta}_{t-1} = \int \tilde{w}_t(\varepsilon_t)\tilde{g}(\varepsilon_t)d\varepsilon_t.$$

In addition

$$\frac{\partial \tilde{w}(\varepsilon_t)}{\partial \varepsilon_t} = \frac{\tilde{y}_t}{\varepsilon_t^2}h'(\frac{\tilde{y}_t}{\varepsilon_t}) + \beta\frac{\tilde{\Delta}_t}{\varepsilon_t}.$$

1.3 Normalized Planning Problem

Let $\tilde{K} = K/\theta_{t-1}$. The planner's problem is then

$$\tilde{K}(\tilde{v}, \tilde{\Delta}, t, 0) = \min\Big[\int \{\tilde{c}(\varepsilon) - \tilde{y}(\varepsilon) + q\varepsilon\tilde{K}(\tilde{v}(\varepsilon), \tilde{\Delta}(\varepsilon), t+1, \tilde{s}(\varepsilon))g(\varepsilon_t)d\varepsilon_t\Big]$$

Subject to

$$\tilde{w}_t(\varepsilon_t) = u(\tilde{c}_t) - h(\tilde{y}_t/\varepsilon_t) - \phi(t) + \beta\tilde{v}_t + \beta_t^{fact}\log(\varepsilon_t)$$

$$\frac{\partial \tilde{w}(\varepsilon_t)}{\partial \varepsilon_t} = \frac{\tilde{y}_t}{\varepsilon_t^2}h'(\frac{\tilde{y}_t}{\varepsilon_t}) + \beta\frac{\tilde{\Delta}_t}{\varepsilon_t}$$

$$\tilde{v}_{t-1} = \int \tilde{w}_t(\varepsilon_t)g(\varepsilon_t)d\varepsilon_t$$

$$\tilde{\Delta}_{t-1} = \int \tilde{w}_t(\varepsilon_t)\tilde{g}(\varepsilon_t)d\varepsilon_t$$

and for retired agents:

$$\tilde{K}(\tilde{v}, 0, t, 1) = \min\Big[\int \{\tilde{c}(\varepsilon) + q\varepsilon\tilde{K}(\tilde{v}(\varepsilon), 0, t+1, 1)g(\varepsilon_t)d\varepsilon_t\Big]$$

Subject to

$$\tilde{w}_t(\varepsilon_t) = u(\tilde{c}_t) + \beta\tilde{v}_t + \beta_t^{fact}\log(\varepsilon_t)$$

$$\tilde{v}_{t-1} = \int \tilde{w}_t(\varepsilon_t)g(\varepsilon_t)d\varepsilon_{t.}$$

1.4 Hamiltonian and First Order Conditions

Dropping the tildes, the Hamiltonian of the normalized problem is, while working:

$$\begin{aligned}
&[C^t(y(\varepsilon), w(\varepsilon) - \beta v(\varepsilon), \varepsilon) - y(\varepsilon)]g(\varepsilon) \\
&+q[K(v(\varepsilon), \Delta(\varepsilon), \varepsilon, t+1, s(\varepsilon))]g(\varepsilon) \\
&+\lambda[v - w(\varepsilon)g(\varepsilon)] + \gamma[\Delta - w(\varepsilon)\tilde{g}(\varepsilon)] \\
&+p(\varepsilon)[u_\theta^t(C^t(y(\varepsilon), w(\varepsilon) - \beta v(\varepsilon), \varepsilon), y(\varepsilon), \varepsilon) + \beta\Delta(\varepsilon)]
\end{aligned}$$

And the limits of the co-state $p(\varepsilon)$ are zero at zero and infinity. The co-state satisfies:

$$\frac{dp(\varepsilon)}{d\varepsilon} = -\Big[\frac{1}{u'(c(\varepsilon))} - \lambda - \gamma\frac{\tilde{g}(\varepsilon_t)}{g(\varepsilon_t)}\Big]g(\varepsilon_t) \tag{54}$$

The FOCs for $\Delta(\varepsilon), v(\varepsilon)$ and $y(\varepsilon)$ are:

$$\begin{aligned}
\frac{p(\varepsilon)}{\varepsilon^2 g(\varepsilon_t)} &= -\frac{q}{\beta}\gamma(\varepsilon) \\
\frac{1}{u'(c(\varepsilon))} &= \frac{q}{\beta}\varepsilon\lambda(\varepsilon) \qquad (55) \\
1 - \frac{1}{\varepsilon}\frac{h'(\frac{\tilde{y}(\varepsilon)}{\varepsilon})}{u'(c(\varepsilon))} &= \frac{p(\varepsilon)}{\varepsilon^2 g(\varepsilon_t)}h'(\frac{\tilde{y}(\varepsilon)}{\varepsilon})[1 + \frac{\tilde{y}(\varepsilon)}{\varepsilon}\frac{h''(\frac{\tilde{y}(\varepsilon)}{\varepsilon})}{h'(\frac{\tilde{y}(\varepsilon)}{\varepsilon})}]. \qquad (56)
\end{aligned}$$

In these equations, I denote the extensions of λ and γ to retired states with the same notation.

1.5 Algorithm

Since the model is in finite horizon, the algorithm solves policy functions backwards from $t = T$, $v_T(\varepsilon) = 0, \Delta_T(\varepsilon) = 0, s_T(\varepsilon) = 1$.

The algorithm takes as state space the dual $(\lambda_-\gamma_-, \varepsilon, s_-)$. I truncate ε between the first percentile and the 99% percentile. The algorithm goes in the following steps:

- If in working state at time t: $s_- = 0$
    1. Start with a guess for the promised utility of the lowest type in a given period: $w_t(\varepsilon_{\text{low}})$

(a) Solve for $y_t(\lambda_t, s_t, \varepsilon_t, p_t, w_t(\varepsilon_{\text{low}}))$ using (56) and (55).

(b) Solve for $\lambda_t(s_t, \varepsilon_t, p_t, w_t(\varepsilon_{\text{low}}))$ from (55), replacing c as a function of w and v using the solution for $y_t(\lambda_t, s_t, \varepsilon_t, p_t, w_t(\varepsilon_{\text{low}}))$ computed in 1(a).

(c) Solve for $\gamma_t(s_t, \varepsilon_t, p_t, w_t(\varepsilon_{\text{low}}))$.

(d) Replace $1/u'(c)$ using (55) in the ODE (54)satisfied by the co-state p and solve the ODE.

i. While solving the ODE compare $K_{t+1}(\lambda_t(s_t = 0), \gamma_t(s_t = 0), \varepsilon, 0)$ to $K_{t+1}(\lambda_t(s_t = 1), \gamma_t(s_t = 1), \varepsilon, 1)$ and set s_t equal to the work status with lowest cost.

2. Check the boundary condition $p(\varepsilon_{\text{high}})$.

(a) If the boundary condition is not met within the tolerance level change $w_t(\varepsilon_{\text{low}})$ and go to 1.

3. Once the boundary condition is met, follow 1. in reverse order to compute policy functions.

(a) Compute $\tilde{w}_t, \tilde{v}_-, \tilde{\Delta}_-$ using their integral definitions.

- If in retired state at time t: $s_- = 0$

  – Set $\lambda_t = \lambda_-/\varepsilon$, $\gamma_t = 0$, $s_t = 1$, $\tilde{c}_t = \lambda_-, \tilde{y}_t = 0$.

2 Baseline Economy Numerical Algorithm

I present the income fluctuation of the model in the baseline US economy. In this economy, agents who face idiosyncratic productivity shocks, consume and save in a risk-free asset, choose their working hours and the age at which they retire. I define retirement as an irreversible exit of the labor force. I assume that the retirement age and the SS benefits claiming age are the same. Denote s the last working period of an agent, i.e $s = t$ if the agent works at time t and $s < t$ if the agent retired before t. The productivity θ_t represents current productivity if $s = t$ and last working productivity if $s < t$, $\theta_t = \theta_s$. With log utility, agents never hit their borrowing constraints because they consume at each period a constant fraction of their net worth. Denote $T(y_t)$ the Heathcote *et al.* (2014) income tax function and $b(\{y_{t'}\}_{t' \in [0,s]}, s)$ the SS benefits as a function of the history of earning and the retirement age. I make a Tauchen approximation of the productivity process $\theta_t = \theta_{t-1}^{\rho}\varepsilon_t$ where $\rho = 0.999$ and denote the transition matrix π.

For a given asset level a_t and productivity θ_t, a working agent's continuation utility is

$$v_t\left(a_t,\theta_t,t\right)=\max_{c_t,y_t,a_{t+1},s_{t+1}}\left\{\ln\left(c_t\right)-\frac{\kappa}{1+\frac{1}{\varepsilon}}(\frac{y_t}{\theta_t})^{1+\frac{1}{\varepsilon}}-\phi(t)+\beta\sum_{\theta_{t+1}|\theta_t}V_{t+1}\left(a_{t+1},\theta_{t+1},s_{t+1}\right)\pi\left(\theta_{t+1}|\theta_t\right)\right\}$$
$$s.t.\quad c_t+\frac{q}{1-\tau^K}a_{t+1}=a_t+y_t-T(y_t).$$

For $s<t$, a retired agent's continuation utility is:

$$v\left(a_t,\theta_t,s\right)=\max_{c_t,y_t,a_{t+1}}\left\{\ln\left(c_t\right)+\beta V_{t+1}\left(a_{t+1},\theta_{t+1},s\right)\right\}$$
$$s.t.\quad c_t+\frac{q}{1-\tau^K}a_{t+1}=a_t+b(\{y_{t'}\}_{t'\in[0,s]},s).$$

Then the intertemporal Euler equation holds, $\frac{1}{c_t}=\frac{\beta q}{1-\tau^K}\mathrm{E}[\frac{1}{c_{t+1}}]$ and for workers, the intratemporal equation holds $\kappa\frac{y_t^{1/\varepsilon}}{\theta_t^{1+1/\varepsilon}}=\frac{1}{c_t}(1-T'(y_t))$.

The algorithm follows these steps.

- Set $a_{T+1}=0, s_{T+1}=T$.
- For each t, if $s=t$:
  1. For given a_{t+1} and $s_{t+1}\in\{t,t+1\}$ solve for c_t using the Euler equation
  2. Solve for y_t using the intratemporal equation
  3. Set s_{t+1} to the work status that yields higher v_t
  4. Solve for a_t using the budget constraint of the workers, $c_t(a_{t+1},s_{t+1})$ and $y_t(a_{t+1},s_{t+1})$
  5. Interpolate the policy functions for the missing values a_t
- For each t, if $s<t$:
  1. For given a_{t+1} and $s_{t+1}=s$ solve for c_t and c_s using the Euler equation
  2. Solve for y_s using the intratemporal equation at time $s$and comput $b(\{y_{t'}\}_{t'\in[0,s]},s)$ taking $\{y_{t'}\}_{t'\in[0,s]}=\{y_s\}$
  3. Solve for a_t using the budget constraint of the retired $c_t(a_{t+1},s)$ and $y_t(a_{t+1},s)$
  4. Interpolate the policy functions for the missing values a_t

At the end of the algorithm I check that $|y(a_t, \theta_t, t) - y(a'(a_t, \theta_t, t), \theta_{t+1}, t+1)| < \eta$ for some tolerance level $\eta$to make sure that agents do not overwork just before retirement to validate the assumption in Step 2, when $s < t$.

3 Alternative Calibration

Constant Fixed Cost With a calibration of a constant fixed cost $\phi(t) = \phi$, over the life-cycle, I match the labor force participation rate for ages 65 to 69 with a fixed cost that is the utility equivalent of 6.8 hours per day. The average retirement age (ARA) in the baseline economy is 63.73 years old.

The left panel of Figure 7 plots the labor force participation rate as a function of age. In the Flexible Retirement model, the optimal ARA is large and equal to 76 years old and is much larger than in the baseline economy. With a constant fixed cost, the only force for an extensive Frisch elasticity of labor supply that increases with age is the decreasing option value of staying in the labor market. With the medium instantaneous variance of productivity of $\sigma_M^2 = 0.0095$, this option value is low. The retirement region doe not evolve much over time and 27.12% of agents work at age 79.

The right panel of Figure 7 plots the average labor wedge for the general population and subpopulations of workers with a history of low productivity shocks and high productivity shocks respectively. The average labor wedge is slightly hump-shaped, increasing from 2.11% at age 25 to 47.82% at age 75 then decreasing to 46.22% at age 79. The small size of the hump is consistent with the low retirement rate in the population and a small composition effect.

As a result from the large gap between the optimal ARA and the ARA in the baseline economy, there are large welfare gains from the second-best optimum with the welfare equivalent of +7.36% consumption at each history at each time compared to baseline economy. However, because this same gap, a reform of the tax system alone only captures 6% of those welfare gains.

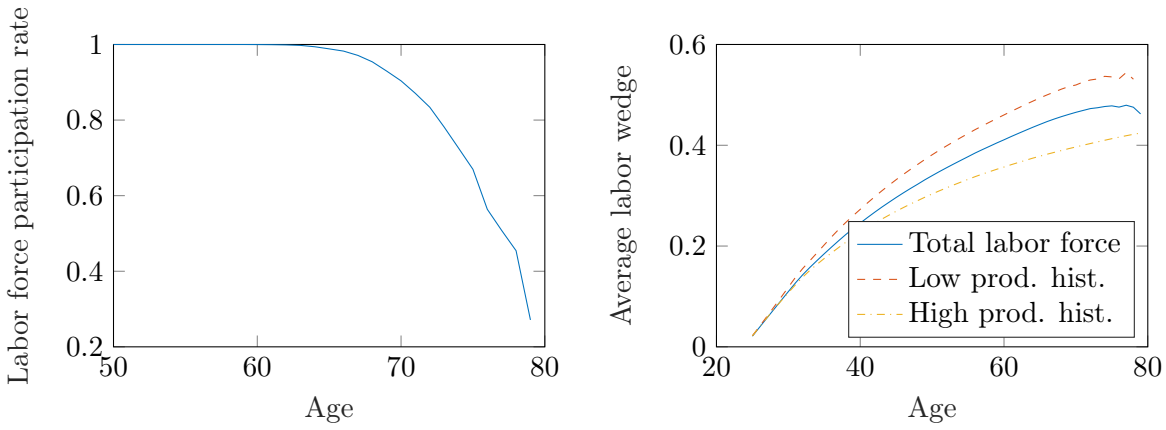

Figure 7: Left: Labor force participation rate as a function of age. Right: Average labor wedge as a function of age.