Letting Your Past Define Your Taxes: Optimal History-Dependent Income Taxation in General Equilibrium

Ross Batzer*

Click here for most recent draft

November 12, 2021

Abstract

I study how labor income taxation should vary with a household's income history in a life-cycle model with differentiated skill types and general equilibrium wages. I consider a very general class of tax functions that can depend on the entire income history, which requires solving a dynamic maximization problem with up to 41 state variables. To maintain the feasibility of the optimal taxation problem with a large state space, I use a novel neural network approach to *simultaneously* solve the model and compute the labor income tax function that maximizes steady state welfare. I find that the optimal history-dependent tax function behaves very differently depending on a household's income history. For most households, average tax rates increase over the life-cycle. However, for households with very high income histories, average tax rates decrease over the life-cycle. The welfare gains from history-dependent taxation are large, equivalent to about a 2 percent increase in lifetime consumption compared to a tax on current income. This suggests that history-dependent taxation may have large welfare benefits in reality, but the exact benefit will depend critically on the production structure of the economy and how complementary different types of labor are. Finally, I show that a parametric function that varies with just the average of previous income levels mimics the full history-dependent tax system by increasing taxes more slowly over the life cycle with higher levels of average past income.

Keywords: optimal taxation, machine learning, neural networks, tax progressivity, life cycle, labor supply

JEL Codes: C45, C68, E62, H20, J42

*Department of Economics, University of Minnesota and Federal Reserve Bank of Minneapolis. 1925 Fourth South St., Minneapolis, MN 55455. Email: batz0025@umn.edu. Website: https://sites.google.com/view/ross-batzer. The views expressed here are those of the author and not of the Federal Reserve Bank system. I would like to express appreciation to my committee - V.V. Chari, Chris Phelan and Kjetil Storesletten - for their valuable advice and support. For their helpful comments and suggestions, I also wish to thank Aniket Baksy, Serdar Birinci, Daniele Caratelli, Kadidiatou Doucouré, Eugenia Gonzalez-Aguado, Jonathan Heathcote, Amy Handlan, Clare Harklau, Larry Jones, Frederic Martenet, Martin Souchier, Salim Syed Chanto as well as seminar participants from the University of Minnesota.

1 Introduction

Allowing labor income taxes to depend on both a household's current income and also previous levels of income has a long history in both economic policy and theory. For example, income averaging, where income taxes are conditioned on the average of previous income levels, was actually implemented as a part of the United States' income tax policy between 1964 and 1986.[1] The intent of this policy is to limit the effect of temporary fluctuations in income on taxes paid in a given year. Income averaging is still available to specific sectors with highly variable income, such as farming. Additionally, social security benefits depend on the entire history of a household's income. Given that history dependent income taxes are a feasible option for the government, we should understand the welfare implications of such a policy.

An existing literature – known as the Dynamic Mirrlees literature – has studied optimal taxation when taxes can depend on the entire previous history of a household's income.[2] This literature has generally found that the welfare benefits of history-dependent taxation are very small compared to taxes on just current income. However, it is an open question whether this result is robust to alternate modeling choices. In particular, for tractability, this literature does not usually consider general equilibrium effects on household wages and assumes that wages are invariant to the government's tax policy.

In this paper, I study history-dependent labor income taxation in a general equilibrium model of life-cycle labor supply. The environment I study is similar to that studied by Heathcote, Storesletten and Violante (2020). The model focuses on the labor supply of middle income households: those who make almost all their income from labor and generally earn between 15,000 and 500,000 dollars a year. In the model, households are differentiated in skills, and the labor supplied by the skill types are complementary in production. This makes the production technology in the economy an important aspect of the government's problem. Households are also heterogeneous in age, wealth, productivity and labor disutility. Additionally, history-dependent taxes make previous incomes state variables in the household problem. This implies that after 35 years of work, the household solves a maximization problem with up to 41 state variables.

I consider a very general class of tax functions where I impose minimal restrictions so that the tax function implies unique equilibrium allocations. In particular, I allow taxes to be any continuously differentiable function of a household's entire income history, but I restrict

[1] Income averaging began with the Revenue Act of 1964, and ended with the 1986 Tax Reform Act, which restructured the income tax code.

[2] Prominent examples of this approach include Weinzierl (2011), Farhi and Werning (2013), Golosov, Troshkin and Tsyvinski (2016) and Stantcheva (2017).

taxes to be zero when income is zero. I use neural networks to approximate the income tax function that maximizes social welfare in a stationary equilibrium. A recent literature has applied neural networks to solve high-dimensional structural economic models. I build on this literature by developing a nested method where I use neural networks to both solve the model *and* to optimize the labor income tax function. This method allows me to quickly approximate the optimal history-dependent tax function using standard algorithms from machine learning.

I find that that the optimal history-dependent tax function behaves very differently depending on a household's income history. For the large majority of households, tax rates – both average and marginal – increase over the life cycle.[3] However, for the households with income histories near the top of the income distribution, average tax rates decrease over the life cycle. The government uses history-dependent tax rates to separate skill types and reward higher labor supply from households with high income history. This is because households with high income history in the model are usually the ones with rare skill types. Since rare skill types have very high marginal productivity and all skill types are complementary in production, the government wants to incentivize labor supply from high income history households to increase general equilibrium wages for all households. This incentive was first described by Stiglitz (1982): when labor is differentiated in production, the government wants to increase output from households with the highest marginal productivity. This allows the government to increase resources for lower income households so that it does not have to redistribute as much under the optimal tax system as it would absent the general equilibrium effects. If the government is allowed to condition taxes on households' skill type, it still gives history-dependent rewards for high output, but the rewards are smaller since it no longer needs to learn which households are capable of producing the most output.

An existing branch of the optimal income taxation literature – known as the Quantitative Ramsey literature – has extensively studied optimal labor income taxation with general equilibrium effects. To do this, these papers use a parametric approach where they optimize within a specific parametric class of tax functions. This approach allows them to study rich models of life-cycle labor supply, but optimal taxes are restricted to be simple parametric functions in only current income.[4] Computational challenges have prevented this literature

[3] I will study age-dependent tax systems throughout this paper, which is common in the optimal taxation literature. Taken literally, this would mean a person's tax rate on their salary would vary with their birth date. However, the government has many tools available to subsidize or tax activities correlated with age. For example, subsidizing higher education or childcare would be ways to reduce effective tax rates on younger workers without literally varying income tax rates by age.

[4] Some papers have also studied optimal taxes that can vary with other characteristics of households, such as age, wealth or marital status. However, these studies still impose restrictions so that the tax system is defined by only several parameters. In this paper, I mainly focus on varying taxes with age and income

from studying more general tax functions or history-dependent taxation since computing optimal tax functions with more than a few parameters has been infeasible. Because of this, it is an open question how labor income taxes should change with a household's previous history of income in the presence of general equilibrium effects on wages.

In order to study optimal history-dependent labor income taxation in a model with general equilibrium wages, I need taxes to be able to depend on a household's entire income history while also maintaining the flexibility of the parametric approach. To do this, I apply a neural network method to approximate the optimal tax system. Neural networks are able to approximate high-dimensional, nonlinear functions with much fewer parameters than standard methods, e.g. perturbation, polynomial projection. Also, because of recent advancements in computing and new numerical libraries, e.g. Google's Tensorflow, neural networks can be estimated very quickly.

In the baseline model, the welfare gains from history-dependent taxation are large, equivalent to about a 2 percent increase in lifetime consumption compared to a tax on current income. The welfare benefits of history-dependent taxation are driven entirely by general equilibrium effects. In fact, if the elasticity of substitution between skill types is set to infinity so that skill types are made to be perfect substitutes, the optimal tax schedule is virtually invariant to income history and there are no welfare benefits from allowing for history-dependent taxation. This welfare gain from history-dependent taxation is substantially larger than what has been found in the models studied in the Dynamic Mirrlees literature. This suggests that history-dependent taxation may have large welfare benefits in reality, but the exact benefit will depend critically on the production structure of the economy and how complimentary different types of labor are. In this paper, I assume constant elasticity of substitution between types of labor. If instead low wage types of labor are not complementary with high wage labor, then the welfare benefits of history-dependence will be minimal.

Finally, I show that the full history-dependent policy can be used to see which simpler policies are able to achieve similar levels of welfare. In particular, I consider a parametric function that varies with just the average of previous income levels. This policy turns out to achieve similar welfare as the fully nonlinear history-dependent tax system. It mimics the full history-dependent tax system by increasing taxes slower over the life-cycle with higher levels of average past income. For most income histories, this makes tax rates increase substantially over the life-cycle to smooth after-tax income across the life-cycle. However, for households with high income history, tax rates stay close to constant across the life-cycle.

history. For a study of how incorporating various forms of heterogeneity affects the optimal labor income tax system, see Karabarbounis (2016).

The rest of this paper is organized as follows: In section 2, I review the related literature and discuss how this paper contributes to the existing literature. In section 3, I describe the economic environment that I study and the optimal taxation problem faced by the government. In section 4, I explain how the parameters of the model are selected. In section 5, I show how I solve the optimal taxation problem. In section 6, I describe the results of the optimal tax approximation. Finally, I conclude and discuss how the results should be interpreted.

2 Related Literature

This paper is related to multiple segments of the optimal taxation literature. Especially related is the quantitative public finance literature that has studied optimal progressive income taxation within a specific class of parametric tax functions. This is a very large literature that has studied many different models with different sources of heterogeneity, preferences and technologies. An incomplete list of examples from this literature includes Kindermann and Krueger (2021), Heathcote, Storesletten and Violante (2017, 2020), Stantcheva (2020), Krueger and Ludwig (2016), Karabarbounis (2016), Peterman (2016), Gervais (2012), Huggett and Parra (2010), Conesa, Kitao and Krueger (2009), Conesa and Krueger (2006), and Erosa and Gervais (2002). My main contribution to this literature is to study a very general tax function that allows for history-dependence.

However, this is not the first study of history-dependent or nonparametric taxation. Notably, the Dynamic Mirrlees literature has studied optimal taxation when the government has no restrictions on what it can tax. An incomplete list of examples from this literature include Ndiaye (2020), Stantcheva (2020), Stantcheva (2017), Golosov, Troshkin and Tsyvinski (2016), Golosov and Tsyvinski (2015), Farhi and Werning (2013), Fukushima (2011), Weinzierl (2011), Albanesi and Sleet (2006) and Golosov, Kocherlakota and Tsyvinski (2003). In this literature, instead of optimizing over an optimal tax function, the government chooses optimal allocations subject to constraints on what information is available. The actual optimal tax system under this method is not unique: there are many different tax systems that can implement the optimal allocation. Also, this method requires simplifying the economic environment enough to apply mechanism design methods to analytically solve the government's social planning problem. Instead, I impose minimal restrictions on an income tax function to ensure that equilibrium allocations are unique under a given tax function. This allows for straightforward interpretation of the optimal income tax system and allows me to study the rich models of lifecycle labor supply studied by the quantitative public finance literature. Therefore, even though I study nonlinear and history-dependent

taxation, my paper is much more closely related to the parametric literature.

Another literature that takes a different approach to either the quantitative or Mirrlees literatures is the sufficient statistic approach. This approach has been applied in both dynamic (Chang and Park (2020), Saez and Stantcheva (2018), Findeisen and Sachs (2017)) and static (Heathcote and Tsujiyama (2020), Sachs, Tsyvinski and Werquin (2020), Saez (2001)) environments. This literature uses a variational approach to compute nonlinear optimal taxes on current income that can usually be expressed in terms of easily computed elasticities. This gives simple and easily interpretable analytic expressions for the optimal income tax function, but it becomes intractable without simplifying the economic environment. The method used in this paper is very similar to the variational approach used in the sufficient statistic literature, except that the optimal tax function is computed automatically by a computer. Optimal taxes computed with the neural network method are less easily interpretable, but the method can be applied to a much wider range of models.

My model is most closely related to the overlapping generations models studied by Heathcote, Storesletten and Violante (2017, 2020). The main difference between my model and these papers is that I allow for savings and nonseparable utility. Allowing for savings is important when studying history-dependent taxation since history-dependent taxation makes labor supply a forwarding looking decision. When labor supply solves an inter-temporal maximization problem, it cannot be assumed that savings behavior will be relatively unaffected by changes in the tax system. Nonseparable utility turns out to not matter much for the key results, but assuming leisure is not complementary with consumption causes the government to provide relatively little consumption insurance, especially for older households.[5] Also closely related is the study of history-dependent taxation by Kapička (2020). Kapička (2020) studies optimal history-dependent taxation in a partial equilibrium environment similar to Heathcote, Storesletten and Violante (2017, 2020) where they restrict income taxes to be a parametric function of a weighted sum of past income. This allows them to derive analytical expressions for optimal tax function and weights on past income. They find that the welfare benefits of history-dependent taxation are large. The source of welfare gains are different than what is found in this paper. They find that it is optimal to have a history dependent tax system that is more progressive with respect to the current income than a history independent tax system, but more regressive with respect to past incomes. The effects described by Kapička (2020) do not seem to be important in the model I study. It is unclear why the results are so different, but it might have to do with either the stochastic processes for productivity that I consider or households being able to transfer resources across periods. Instead, I find that virtually all benefits from history-dependent taxation come from

[5] I describe the results under separable utility in Appendix A.7

spillovers in general equilibrium prices. Although, in this paper, the government chooses to use history-dependent rewards to manipulate general equilibrium wages, the intuition for why they do this is similar to the static optimal tax problems described by Rothschild and Scheuer (2013), Sachs, Tsyvinski and Werquin (2020), Stiglitz (1982, 1987), and others: the government wants to increase before-tax income so that it needs to redistribute less, and subsequently create less distortions, under the optimal tax system.

Finally, this paper is related to recent papers that have shown that neural networks can be used to solve high dimensional economic models. These papers include Azinovic, Gaegauf and Scheidegger (2019), Chen, Joseph, Kumhof and Pan (2021), Duarte (2018), Fernández-Villaverde, Nuño, Sorg-Langhans and Vogler (2020), and Maliar, Maliar and Winant (2019). This paper builds on these methods by solving a nested problem: neural networks are used not only to solve the model itself but also to compute the general tax function that maximizes social welfare.

3 Model

I will use a similar economy as Heathcote, Storesletten and Violante (2020). However, I will allow individuals to save in a risk-free bond to provide private insurance in addition to insurance provided through the tax system.

3.1 Economic Environment

Demographics Individuals enter the economy at age $a = 0$ and live for A periods. The total population is of mass one, and therefore each cohort is of mass $1/A$. There are no inter-generational links. I index individuals by $i \in [0, 1]$. Since I only study stationary economies, I will only index age and not time.

Life Cycle Upon entering the economy at age $a = 0$, individuals have a chance to invest in skills, s_i. Once the individual has chosen s_i, they enter the labor market. During the labor market, the individual provides $h_i \geq 0$ hours of labor supply, consumes a private consumption good c_i and a receives utility from a publicly provided good G. Each period, they face stochastic fluctuations in labor productivity θ_i.

Preferences Expected lifetime utility over consumption, hours worked, publicly provided goods and skill investment for individual i is given by

$$U_i = -v_i(s_i) + E_0\left[\left(\frac{1-\beta}{1-\beta^A}\right)\sum_{a=0}^{A-1}\beta^a u_i(c_{ia}, h_{ia})\right]$$

where $\beta \leq 1$ is the discount factor, common to all individuals, and the expectation is taken over future idiosyncratic productivity shocks. The disutility from initial skill investment $s_i \geq 0$ takes the form

$$v_i(s_i) = \kappa_i^{-\frac{1}{\psi}} \frac{s_i^{1+\frac{1}{\psi}}}{1+\frac{1}{\psi}}$$

where the parameter $\psi \geq 0$ controls the elasticity of skill investment with respect to the marginal return to skill, and $\kappa_i \geq 0$ is an individual-specific parameter that determines the utility cost of acquiring skills. This cost is distributed according to an exponential distribution $\kappa_i \sim Exp(\lambda)$. Skill investment decisions are irreversible, so skills are fixed through the life cycle.

The period utility function u_i is given by

$$u_i(c_{ia}, h_{ia}) = \frac{\left[c_{ia}^{\phi_i}(1-h_{ia})^{1-\phi_i}\right]^{1-\gamma}}{1-\gamma} \tag{1}$$

This utility function has complementarity in consumption and leisure: higher levels of consumption raise the marginal utility of leisure. The term ϕ_i is a fixed individual effect that scales the utility of leisure and γ determines risk aversion. The fixed effect ϕ_i will be used to match the variance of hours worked in the data. This is necessary because variation in wages alone can only account for less than half of the variation in hours worked observed in the data.

Technology Output is a constant elasticity of substitution aggregate of effective hours supplied by the continuum of skill types $s \in [0, \infty)$

$$Y = \left(\int_0^\infty [N(s)f_s(s)]^{\frac{\omega-1}{\omega}} ds\right)^{\frac{\omega}{\omega-1}}$$

where $\omega > 1$ is the elasticity of substitution across skill types, $f(s)$ is the density over skill types and $N(s)$ is the average effective labor supply of individuals of skill type s,

$$N(s) = \int \exp\{\theta_i\} h_i di.$$

Note that all skill levels enter symmetrically in the production technology, and thus any equilibrium differences in skill prices will reflect relative scarcity. This will be reflected in equilibrium skill prices, which are the marginal products of labor supplied by each type s. The skill price, or skill premium, of a skill type s is given by

$$p(s) = \left[\frac{Y}{N(s)f_s(s)}\right]^{\frac{1}{\omega}}$$

Here, the skill price $p(s)$ is higher when 1) the measure $f_s(s)$ of households with the skill type is lower and 2) total output is higher. The fact that prices for all skill types are linked to total output will be crucial to the government's optimal taxation problem. Since all households benefit from higher total output, the government will have a large incentive to increase output.

Labor Productivity and Income Log individual labor efficiency θ_{ia} is the sum of three orthogonal components

$$\theta_{ia} = x(a) + z_{ia} + \varepsilon_{ia}$$

The first component $x(a)$ captures the deterministic age profile of labor productivity common to all individuals. The second component z_{ia} captures permanent idiosyncratic shocks that follow a unit root process

$$z_{ia} = z_{ia-1} + \eta_{ia}$$

with innovations that are distributed i.i.d. normal $\eta_{ia} \sim N(0, v_\eta)$ and have an initial value of $z_{i0} \sim N(0, v_z)$. The third component ε_{ia} captures transitory idiosyncratic shocks that are also distributed i.i.d. normal $\varepsilon_{ia} \sim N(0, v_\varepsilon)$. Since individuals exist in a continuum, a standard law of large numbers implies that individual shocks induce no aggregate uncertainty in the economy.

Individual labor income y_{ia} is the product of four components:

$$y_{ia} = \underbrace{p(s_i)}_{\text{skill price}} \times \underbrace{\exp\{x(a)\}}_{\text{age-productivity profile}} \times \underbrace{\exp\{z_{ia} + \varepsilon_{ia}\}}_{\text{labor market shocks}} \times \underbrace{h_{ia}}_{\text{hours worked}}$$

The first component $p(s_i)$ is the equilibrium price for the type of labor supplied by an individual with skills s_i. The second component $x(a)$ is the life cycle profile of average labor productivity. The third component $z_{ia} + \varepsilon_{ia}$ is individual stochastic labor productivity. Finally, the fourth component h_{ia} is the number of hours worked by the individual. Therefore,

total individual labor income is determined by (i) skill investment made before entering the labor market, which reflects innate disutility of acquiring skills κ_i, (ii) productivity that evolves exogenously over the lifecycle, (iii) labor market outcomes determined by the realization of stochastic idiosyncratic shocks to productivity, (iv) time spent working, which reflects disutility of labor because of individual preferences ϕ_i.

In this economy, taxation affects income by distorting both skill investment and hours worked. Heathcote, Storesletten and Violante (2017, 2020) are able to study both these margins because they restrict their taxes to be a log-linear function of current income. Krueger and Ludwig (2013, 2016) and Peterman (2016) also use parametric restrictions to jointly study the effects of progressive taxation on skill investment and labor supply. The neural network approach described in the next section allows me to compute fully nonlinear optimal taxes while maintaining both margins. This allows me to directly compare the parametric tax instruments commonly used in these papers to the unrestricted optimal tax system.

Asset Markets Asset markets are incomplete and agents cannot fully insure against the idiosyncratic shocks by trading state-contingent assets. However, they can partially self-insure against these risks by accumulating precautionary asset holdings, b. The stock of assets earns a market return R. I assume that households enter the economy with zero assets and are not allowed to borrow against future income, so that $b_{i0} = 0$ and $b_{ia} \geq 0$ for all i and a.

Government The government runs a tax and transfer scheme and uses the revenue from the tax system to fund its expenditures G. Let g denote government expenditures as a fraction of aggregate output so that $G = gY$. The government must run a balanced budget, so the government's budget constraint is therefore

$$G = gY \leq \sum_{a=0}^{A-1} \int T\left(y_{ia}; \{y_{it}\}_{t=0}^{a-1}, a\right) di$$

Here, $T_a\left(y; \{y_t\}_{t=0}^{a-1}\right)$ is the net tax owed by a household of age a with current income y and a previous history of incomes $\{y_t\}_{t=0}^{a-1}$. The only restriction I will put on the function T is that average tax rates τ must be a differentiable function of age a and history $\{y_t\}_{t=0}^{a}$,

$$T_a\left(y; \{y_t\}_{t=0}^{a-1}\right) = \tau_a\left(y; \{y_t\}_{t=0}^{a-1}\right) y \tag{2}$$

As I show in Appendix A.1, this form for the tax function ensures that the tax function implies unique equilibrium allocations. Therefore, the allocations under the optimal tax system will be unique. This form of tax function allows me to directly compare the optimal history-dependent tax function to the parametric tax function used by Heathcote, Storesletten and Violante (2020),

$$T_a^{hsv}(y) = y - (1-\tau_a)y^{\rho_a} \tag{3}$$

The age-invariant version of this parametric tax function has a long history in public finance dating back to Feldstein (1969) and can do a very good job at matching the actual US tax schedule for labor income (as noted by Blundell, Pistaferri and Saporta-Eksten (2016) and Heathcote, Storesletten and Violante (2017)). More recent studies that have used this function include Benabou (2000, 2002), Karabarbounis (2016) and Heathcote, Storesletten and Violante (2017, 2020). With this tax function, a household's after-tax labor income $\tilde{y}$ is given by

$$\tilde{y}(y) = (1-\tau_a)y^{\rho_a}$$

The parameter τ_a determines the average tax rate faced by households of age a. The parameter ρ_a determines the progressivity of the tax system. If $\rho_a < 1$, then the tax system is progressive: average tax rates are increasing in income y and marginal tax rates for the lowest income households are negative. If $\rho_a = 1$, then the tax system is linear and everyone has the same tax rate τ_a. If $\rho_a > 1$, then average tax rates are decreases in income, so the tax system is regressive.

I will focus on comparing five possible tax systems: 1) non-parametric history dependent taxes: $T_a\left(y_{ia}; \{y_{it}\}_{t=0}^{a-1}\right)$, 2) non-parametric age dependent taxes: $T_a\left(y_{ia}\right)$, 3) the parametric age-dependent taxes specified in equation (3) where I assume the tax level and progressivity parameters are quadratic functions of age,

$$\tau(a) = \tau_0 + \tau_1 a + \tau_2 a^2 \text{ and } \rho(a) = \rho_0 + \rho_1 a + \rho_2 a^2$$

4) non-parametric taxes on only current income: $T(y)$, and 5) parametric taxes on only current income also using the tax function specified in equation (3) where the tax level and progressivity are constants.

Individual Problem The vector of state variables for an individual household at age a is $(b_{ia}, z_{ia}, \varepsilon_{ia}, s_i, \phi_i, a, \{y_t\}_{t=0}^{a-1})$. Individuals choose skill investment s_i and sequences of consumption $\{c_{ia}\}_{a=0}^{A-1}$, hours worked $\{h_{ia}\}_{a=0}^{A-1}$ and savings $\{b_{ia+1}\}_{a=0}^{A-1}$ to maximize their

expected lifetime utility,

$$\max_{s_i\{c_{ia},h_{ia},b_{ia+1}\}_{a=0}^{A-1}} -v_i(s_i) + E_0\left[\left(\frac{1-\beta}{1-\beta^A}\right)\sum_{a=0}^{A-1}\beta^a u_i(c_{ia},h_{ia})\right] \tag{4}$$

subject to their budget constraint,

$$c_{ia} + b_{ia+1} = Rb_{ia} + y_{ia} - T_a(y_{ia};\{y_{it}\}_{t=0}^{a-1})$$

and non-negativity constraints on their choices,

$$s_i, c_{ia}, b_{ia+1} \geq 0,\ h_{ia} \in [0,1]$$

3.2 Equilibrium

A stationary competitive equilibrium is allocation functions $\left(s,\{c_a,h_a,b_{a+1}\}_{a=0}^{A-1}\right)$ and prices $p(s)$ such that

1. Households solve their problem
2. Skill price $p(s)$ is the marginal product of type s

$$p(s) = \left[\frac{Y}{N(s)f_s(s)}\right]^{\frac{1}{\omega}}$$

3. Densities for skills f_s and savings f_b are consistent with individual choices
4. Government budget is satisfied

$$G \leq \sum_{a=0}^{A-1}\int T\left(y_{ia};\{y_{it}\}_{t=0}^{a-1},a\right)di$$

5. Markets clear

$$\sum_{a=0}^{A-1}\int c_{ia}di + G = r\sum_{a=0}^{A-1}\int b_{ia}di + Y$$

and

$$N(s) = \sum_{a=0}^{A-1}\int \exp\{x(a)+z_{ia}+\varepsilon_{ia}\}h_{ia}(s)di$$

3.3 Optimal Taxation Problem

Social Welfare Function I follow Conesa, Kitao and Krueger (2009) and Karabarbounis (2016) and choose taxes to maximize the ex-ante (before ability is realized) expected (with respect to uninsurable productivity shocks) lifetime utility of a newborn in a stationary equilibrium. The government's social welfare function is therefore given by

$$W = \int U_i di = \int \left\{ -v_i(s_i) + E_0 \left[\left(\frac{1-\beta}{1-\beta^A} \right) \sum_{a=0}^{A-1} \beta^a u_i(c_{ia}, h_{ia}) \right] \right\} di \tag{5}$$

This welfare criterion embeds a concern by the policy maker for insurance against idiosyncratic shocks and redistribution between agents of different productivity levels. Since lifetime utility is strictly concave in ability to generate income, taking a dollar from higher income households and giving it to lower income households, all else equal, increases social welfare. However, higher taxes on labor income create disincentive effects on households' skill investment and labor supply decisions. The policy maker has to trade off the benefits of redistribution and funding public goods against the distortionary effects of taxes on labor supply and skill investment.

The Government's Problem The government chooses its tax function T to maximize the social welfare function (5) subject to its budget constraint

$$gY \leq \sum_{a=0}^{A-1} \int T\left(y_{ia}; \{y_{it}\}_{t=0}^{a-1}, a\right) di$$

and that households solve their respective problems (4) given the tax function.

4 Parameter Selection

4.1 Labor Market Shocks

I use data from the Panel Study of Income Dynamics (PSID) for survey years 2000, 2002, 2004, and 2006. I estimate the processes for the two labor market shocks, permanent shocks z and transitory shocks ε, by matching moments of log wages in the PSID data. I compute household wages by dividing total before-tax household labor income by total hours worked by all household members. To obtain the stochastic component of log wages, $z + \varepsilon$, I run a regression of log wages on a quadratic in age and dummy variables for education, family

size, number of children and state of residence,

$$\log w_{ia} = x_0 + x_1 a + x_2 a^2 + D_i + \epsilon_{ia}$$

I use the estimated coefficients on the age variables as my lifecycle profile of log wages, $x(a) = x_1 a + x_2 a^2$. I assume that the residual of this regression consists of a permanent component z and a transitory component ε,

$$\epsilon_{ia} = z_{ia} + \varepsilon_{ia}$$

The process for permanent shocks is assumed to be

$$z_{ia+1} = z_{ia} + \eta_{ia+1},\ \eta_{ia} \sim N(0, v_\eta)$$

where the initial value of z is distributed as

$$z_{i0} \sim N(0, v_z)$$

and the process for transitory shocks is assumed to be

$$\varepsilon_{ia} \sim N(0, v_\varepsilon)$$

Therefore, I need to estimate three parameters: $(v_z, v_\eta, v_\varepsilon)$, so I need three moments. I choose (v_η, v_ε) to match the variance of the growth of log wages between consecutive survey years, $var(\epsilon_{ia+2} - \epsilon_{ia})$, and between every other survey year, $var(\epsilon_{ia+4} - \epsilon_{ia})$. Using the assumed distributions of η and ε, this gives two equations in two unknowns,

$$var(\epsilon_{ia+2} - \epsilon_{ia}) = 2v_\eta + 2v_\varepsilon$$
$$var(\epsilon_{ia+4} - \epsilon_{ia}) = 4v_\eta + 2v_\varepsilon$$

This two equation system can be solved for v_η and v_ε. Then, I use that the total variance of log wages, $var(\epsilon_{ia})$, is equal to

$$var(\epsilon_{ia}) = v_z + a v_\eta + v_\varepsilon$$

which can be solved for v_z given values for v_η and v_ε. I summarize the values of the parameters governing the wage process below,

Parameter	Description	Value
x_1	Linear component of life cycle profile	0.03
x_2	Quadratic component of life cycle profile	-0.0005
v_z	Variance of initial condition z_0	0.12
v_η	Variance of permanent shocks z	0.003
v_ε	Variance of transitory shocks ε	0.135

Table 1: Summary of Parameters for Wage Process

4.2 Summary of Model Parameters

Once I estimate the process for labor market shocks, $(v_z, v_\eta, v_\varepsilon)$, I can solve the model given values for $\beta, R, \omega, \psi, \gamma, \nu, \phi$. I set the discount factor to be a standard value, $\beta = R^{-1} = 0.98$ (e.g. Golosov, Troshkin and Tsyvinski (2016)). I take parameters for skill investment and substitutability of skills in production from Heathcote, Storesletten and Violante (2017). In particular, the substitutability parameter for skills is set to be $\omega = 3.124$ and the elasticity parameter for skill investment is set to $\psi = 0.65$. These parameters were chosen to match the distribution of income and trends in the skill premium observed in the data. I assume a standard value for risk aversion $\gamma = 2$ and assume that the parameter governing utility of leisure, ϕ, is distributed as

$$\phi = \frac{1}{1 + \exp \tilde{\phi}}, \ \tilde{\phi} \sim N(m_\phi^{ns}, v_\phi^{ns})$$

This transformation forces the final value of ϕ to be between zero and one. I calibrate the mean and variances of this process to match average hours worked, $\int h_i di$, to be 0.33 and the variance of log hours, $var(\log h_i)$, to be its value in the PSID.

I summarize the values of all fixed parameters and how they were selected in the following table,

5 Solution Method

In this section, I will describe how neural networks can be used to approximate the model described in section 3 and compute optimal nonparametric income taxes. First, I will describe the general structure of neural networks. Then, I will describe how to estimate the parameters of neural networks to approximate a general function. Next, I will explain how neural networks can be applied to solve my model and compute optimal policies. Finally, I will detail the exact algorithm to compute optimal income taxes in my model.

Parameter	Description	Value	Source/Target
β	Discount Factor	0.98	Golosov et al. (2016)
R	Return on savings	1/0.98	
A	Years of working life	36	Heathcote et al. (2020)
ν	Inverse Frisch elasticity	0.5	
ψ	Elasticity of skill investment to return	0.65	
ω	Elasticity of substitution across skills	3.124	
g	Government spending (% of output)	0.19	
m_ϕ^{ns}	Mean of leisure disutilty	0.275	$H = 0.33$
v_ϕ^{ns}	Variance of labor disutilty	0.041	$var(\log h_i) = 0.124$

Table 2: Summary of Fixed Parameters

5.1 Neural Networks

Neural networks are constructed by computing weighted sums of inputs, $x = (x_1, \ldots, x_m)$, and then transforming the weighted sums using some function, e.g. $f(x) = \tanh(x)$, $f(x) = \max\{0, x\}$. The simplest neural network is simply a weighted sum of inputs,

$$y_i^0(x; w) = \sum_{j=1}^{m} w_{i,j}^0 x_j, \text{ for } i = 1, \ldots, n$$

where m is the number of input variables and n is the number of output variables. Here, there is a weight for each input and output variable, so there are $m \times n$ weights that can be chosen

$$w^0 = \begin{bmatrix} w_{1,1} & \ldots & w_{1,n} \\ \vdots & \ddots & \vdots \\ w_{m,1} & \ldots & w_{m,n} \end{bmatrix}$$

A simple weighted sum will generally not approximate a nonlinear function well. In order to get a better approximation for the function we want, the weighted sum of inputs can be transformed using some nonlinear function f and then these transformations are be used to compute a new weighted sum,

$$\begin{aligned} y_i^1(x; w) &= \sum_{j=1}^{p} w_{i,j}^1 f\left(y_j^0(x; w)\right) \\ &= \sum_{j=1}^{p} w_{i,j}^1 f\left(\sum_{k=1}^{m} w_{j,k}^0 x_k\right), \text{ for } i = 1, \ldots, n \end{aligned}$$

Here, p is the number of *nodes* in the hidden layer, which determines the number of intermediate transformations of inputs that can be used to compute the outputs. This is referred to as a neural network with one *hidden layer* since there is a single intermediate transformation of inputs before they are used to compute outputs. There is a weight for each node and input as well as each node and output, so there are $n \times p + p \times m$ weights that can be chosen

$$w^1 = \begin{bmatrix} w_{1,1} & \dots & w_{1,p} \\ \vdots & \ddots & \vdots \\ w_{n,1} & \dots & w_{n,p} \end{bmatrix} \text{ and } w^0 = \begin{bmatrix} w_{1,1} & \dots & w_{1,m} \\ \vdots & \ddots & \vdots \\ w_{p,1} & \dots & w_{p,m} \end{bmatrix}$$

A neural network with one hidden layer can in theory approximate any continuous function given that is has a sufficiently large number of intermediate nodes p. This result is known as the *universal approximation theorem*. However, in practice, better approximations can often be achieved more easily by adding additional hidden layers instead of more nodes. A neural network with two hidden layers can be constructed by transforming the outputs of a network with one hidden layer,

$$\begin{aligned} y_i^2(x;w) &= \sum_{j=1}^{p} w_{i,j}^2 f\left(y_j^1(x;w)\right) \\ &= \sum_{j_2=1}^{p} w_{i,j_2}^2 f\left(\sum_{j_1=1}^{p} w_{j_2,j_1}^1 f\left(\sum_{k=1}^{m} w_{j_1,k}^0 x_k\right)\right), \text{ for } i = 1, \dots, n \end{aligned}$$

Using more than one hidden layer is commonly referred to as *deep learning*. Adding additional layers allows the network to better approximate nonlinear features of the target function, but it requires $p \times p$ additional weights to be estimated. In the above network with two hidden layers, there are $n \times p + p^2 + p \times m$ weights that can be chosen

$$w^2 = \begin{bmatrix} w_{1,1} & \dots & w_{1,p} \\ \vdots & \ddots & \vdots \\ w_{n,1} & \dots & w_{n,p} \end{bmatrix} \text{ and } w^1 = \begin{bmatrix} w_{1,1} & \dots & w_{1,p} \\ \vdots & \ddots & \vdots \\ w_{p,1} & \dots & w_{p,p} \end{bmatrix} \text{ and } w^0 = \begin{bmatrix} w_{1,1} & \dots & w_{1,m} \\ \vdots & \ddots & \vdots \\ w_{p,1} & \dots & w_{p,m} \end{bmatrix}$$

A neural network with $L \geq 3$ hidden layers is constructed by repeatedly transforming outputs of $L-1$ intermediate networks,

$$\begin{aligned} y_i^L(x;w) &= \sum_{j=1}^{p} w_{i,j}^L f\left(y_j^{L-1}(x;w)\right) \\ &= \sum_{j_L=1}^{p} w_{i,j_L}^L \left(\cdots \sum_{j_2=1}^{p} w_{j_3,j_2}^2 f\left(\sum_{j_1=1}^{p} w_{j_2,j_1}^1 f\left(\sum_{k=1}^{m} w_{j_1,k}^0 x_k \right) \right) \cdots \right), \text{ for } i = 1,\ldots,n \end{aligned}$$

This network will have $n \times p + (L-1)p^2 + p \times m$ weights that must be estimated. Notice that even with many layers, the number of weights that need to be estimated still grows linearly in the number of inputs m. This is the main reason why neural networks are able to approximate high dimensional functions quickly. While alternate methods of functional approximation like splines or Chebychev polynomials can be updated faster, the number of parameters increases either polynomially or exponentially. This makes approximating high dimensional functions infeasible. The weights of neural networks cannot be updated analytically, but thanks to new numerical libraries for machine learning like Google's Tensorflow, these weights can be updated much quicker than they could be before.

5.1.1 Estimating Neural Networks

Neural networks are estimated by specifying a *loss function*, $\ell(x;w)$, which takes the input vector $x \in \mathbb{R}^m$ and returns a scalar for a given sets of network weights w. Weights are updated through gradient descent: the loss function is differentiated with respect to every weight in the network $\frac{\partial \ell(x;w)}{\partial w}$ and the weights are updated in the direction of the gradient,

$$w_{new} = w + \alpha \frac{\partial \ell(x;w)}{\partial w}$$

Here, α is the *learning rate,* which determines how aggressively the weights are updated at each iteration. A high learning rate will reach the area around a minimum faster, but it might overshoot the true global minimum because it moves the weights too much. A low learning rate will be less aggressive in updating network weights, but the small movement will make it more likely that the weights get caught at a local minimum or flat region of the loss function. The goal is to find a set of weights such that $\frac{\partial \ell(x;w)}{\partial w} = 0$, which means that there is no deviation in any weight that creates a lower value of the loss function. The values of weights w that satisfy $\frac{\partial \ell(x;w)}{\partial w} = 0$ for a given network are not unique. The number of nodes and layers capable of approximating a certain function are also not unique. The learning rate and network structure that most quickly and accurately approximate a given

function must be found through experimentation.

5.1.2 Comparison with Polynomial Approximation

A more common method for functional approximation is to use polynomial regression, where a function is approximated as a weighted sum of polynomial functions of inputs. Similar to neural networks, a polynomial function can in theory approximate any continuous function arbitrarily well with a sufficiently high degree of approximation. A polynomial regression is structured as

$$y(x;w) = \sum_{j_m=0}^{p} \cdots \sum_{j_1=0}^{p} w_{j_1,\ldots,j_m} \times f_{j_1,\ldots,j_m}(x)$$

where m in the dimension of the inputs $x = (x_1, \ldots, x_m)$ and p is the degree of polynomial being used in the approximation. Here, there are p^m weights to be estimated. This means that the number of weights to be estimated grows exponentially in the number of inputs. These weights can generally be updated faster than weights in a neural network, and with certain types of orthogonal polynomials, like Chebyshev polynomials, they can be updated analytically. However, since the number of weights grows exponentially in the number of input variables, this method is only feasible for functions with a small number of inputs.

With neural networks, the network can self-allocate parameters to approximate the most nonlinear parts of a function without needing many additional weights. As Fernández-Villaverde, Nuño, Sorg-Langhans and Vogler (2020) describe, neural networks are *compositional* while polynomial approximation is *additive*. With standard polynomial approximation, the approximation must at least double the number of parameters to capture the new dimension and the new parameters do not improve the approximation on the existing dimensions. When more dimensions are added to a neural network approximation, the network reallocates parameters to better approximate the added dimension without loosing much accuracy in the existing dimensions. In practice, neural networks perform dimension reduction by finding weighted sums of inputs that can best represent the relevant features of the data. For example, if a simple average of inputs is sufficient to predict outputs, a neural network will identify this relationship easily, but standard functional approximation methods would still require computing many combinations of input values.

The exponential growth in parameters associated with standard functional approximation can be avoided with sparse grid methods like those described by Krueger and Kubler (2004), Judd, Maliar, Maliar and Valero (2014) and Brumm and Scheidegger (2017). However, these methods will still require a very large number of parameters to be estimated with high-dimensional problems. This is because polynomial functions do not extrapolate well to values outside of the input values the approximation has observed (i.e. grid points). Since

neural networks instead use bounded and monotonic functions, they are much better at accurately predicting values outside of the data that they have already been estimated with. These two features of neural networks: linear growth in parameters with respect to inputs and accurate extrapolation make them able to accurately approximate high dimensional functions with a relatively small amount of data.

5.2 Approximating Optimal Taxes

In this subsection, I will describe how neural networks can be used to both solve my model and compute optimal policies. This will involve an iterative process where an income tax function represented as a neural network is chosen to maximize utilitarian welfare which is derived from household choices that are also represented as different neural networks. The weights of the tax function are changed until a perturbation of any weight cannot produce an increase in welfare.

5.2.1 Algorithm for the Optimal Taxation Problem

Nonparametric Taxes The process for computing optimal tax functions is similar to the variational method described by Saez (2001) and used more recently by Chang and Park (2020), Findeisen and Sachs (2017), Saez and Stantcheva (2018) and Sachs, Tsyvinski and Werquin (2020). The optimal tax function is found by finding the function for which an arbitrary perturbation cannot improve welfare. However, I use a neural network to perform this process automatically without needing to take any derivatives by hand. The neural network is trained so that it finds the tax function where changing any of the weights does not alter allocations or prices in a way that increases social welfare.

The neural networks are trained through simulation, where I draw full lifecycle profiles of idiosyncratic shocks for a large number of households, I. I divide these households into smaller groups called *batches* and compute the loss functions for each batch. I take the gradient of the loss function with respect to network weights and update the weights of the networks using a gradient descent method. The gradients taken at each iteration are a noisy measure of the true gradient since it is only taken with the data for a single batch. The accuracy of the gradient can be increased by increasing the size of each batch, but this will also make it harder for the computer to observe groups households for which it can use the tax system to improve welfare. I found that a batch size of 25 households did a good job at giving reasonable aggregate values while also allowing the computer to find places where nonlinear taxes can improve welfare. With neural networks, it is not necessary that the loss function and its gradient are close to their true values on a given iteration, they should just

be correct on average. An individual iteration will have very little effect on the final values of weights after training, the network uses information from tens of thousands of iterations to obtain its final weights. Using noisy measurements of the loss function on small samples are how neural networks are capable of quickly getting accurate approximations while using large amounts of data.

The steps to solving the optimal taxation problem are described below,

1. Draw I values for fixed effects $\{\kappa_i, \phi_i, z_{i0}\}_{i=1}^{I}$ and profiles of wage shocks $\{\eta_{ia}, \varepsilon_{ia}\}_{a=0}^{A-1}$ for each individual

2. Setup neural networks for individual choices

   (a) Setup neural network with weights w_c for the consumption and hours worked

   $$c(b, z, \varepsilon, s, \phi, a, y^a; w_c, w_T), h(b, z, \varepsilon, s, \phi, a, y^a; w_c, w_T)$$

   (b) Setup neural network with weights w_s for the skill investment choice

   $$s(\kappa; w_s, w_c, w_T)$$

3. Create loss function: weights for individual choices and skill investment are chosen to maximize

   $$U(w_s, w_c; w_T) = \frac{1}{I} \sum_{i=1}^{I} \Bigg[- v_i \left(s_i(w_s; w_T) \right) + \left(\frac{1-\beta}{1-\beta^A} \right) \sum_{a=0}^{A-1} \beta^a u_i \left(c_{ia}(w_c; w_T), h_{ia}(w_c; w_T) \right) \Bigg]$$

   so that

   $$\frac{\partial U}{\partial w_c} = 0 \text{ and } \frac{\partial U}{\partial w_s} = 0$$

   taking the tax system, w_T, public spending G and prices $p(s)$ as given.

4. Setup neural network with weights w_T for the tax function

   $$T(y^a, a; w_T)$$

5. Consider a perturbation to one of the weights of the tax function: $w_T + \Delta = (w_1, \ldots, w_i + \delta, \ldots, w_n)$

6. Compute skill prices for each κ_i given current guess for $h(b, z, \varepsilon, s(\kappa), \phi, a, y^a; w_c)$

$$p\left(s_i(\kappa_i); w_c\right) = \left[\frac{Y}{N(s(\kappa_i); w_c) f_s(s(\kappa_i))}\right]^{\frac{1}{\omega}}$$

where

$$N(s(\kappa_i); w_c) = \frac{1}{A}\sum_{a=0}^{A-1} \exp \theta_{ia} h_{ia}(w_c)$$

and $f_s(s(\kappa_i))s'(\kappa_i) = f_\kappa(\kappa_i) = \exp(-\kappa)$

7. Update individual choice functions under the perturbed tax function (and implied prices for skills) by gradient descent

$$\tilde{w}_c = w_c + \frac{\partial U(w_s, w_c; w_T + \Delta)}{\partial w_c} \text{ and } \tilde{w}_s = w_s + \frac{\partial U(w_s, \tilde{w}_c; w_T + \Delta)}{\partial w_s}$$

8. Update skill prices under new function for hours worked

$$p\left(s_i(\kappa_i); \tilde{w}_c\right) = \left[\frac{Y}{N(s(\kappa_i); \tilde{w}_c) f_s(s(\kappa_i))}\right]^{\frac{1}{\omega}}$$

9. Update individual choices again under new prices

10. Compute welfare under the perturbed tax function

$$\begin{aligned} W(w_s, w_c, w_T + \Delta) = \frac{1}{I}\sum_{i=1}^{I} \Bigg[& - v_i\left(s_i(\tilde{w}_s; w_T + \Delta)\right) \\ & + \left(\frac{1-\beta}{1-\beta^A}\right) \sum_{a=0}^{A-1} \beta^a u_i\left(c_{ia}(\tilde{w}_c; w_T + \Delta), h_{ia}(\tilde{w}_c; w_T + \Delta)\right) \Bigg] \end{aligned}$$

11. The optimal tax system is the tax function w_T such that a perturbation produces no welfare gain

$$\frac{\partial W(w_s, w_c, w_T + \Delta)}{\partial \Delta} = 0$$

Parametric Taxes I will compare the nonparametric tax function to the parametric tax function (3), which is common in the quantitative public finance literature. I use a neural network to represent the parameters of the function,

$$T(y, t; w_T) = y - \left[1 - \tau(t; w_T)\right] y^{\rho(t; w_T)} \tag{6}$$

where w_T are the weights of the neural network. This tax function makes after-tax income a log-linear function of before-tax income. This allows the tax function to vary by age so that the tax function at each age is a parametric function, but there are no restrictions on how the parameters vary over the lifecycle.

In both this case and the full history-dependent tax function (2), I do not allow for unconditional (lump-sum) transfers that are distributed evenly to all households independent of their income. This is consistent with many real-life transfer policies such as the Earned Income Tax Credit of he Child Tax Credit. However, unconditional transfers are commonly included in studies of optimal taxation. Introducing lump-sum transfers in this problem leads to a "practical indeterminacy", in that multiple solutions produce almost identical levels of welfare even though there is no indeterminacy in theory. Including lump-sum transfers give the government two policy tools for redistribution: more progressive marginal tax rates or higher lump-sum transfers. The government can achieve very similar tax systems with either tool. Since my solution method maximizes welfare directly using noisy simulations, the computer is not able to differentiate between the two policies. This probably would not be true if my model included many extremely poor households with near zero wages, but this model focuses on PSID households outside the tails of the income distribution.

5.2.2 Testing the Approximation

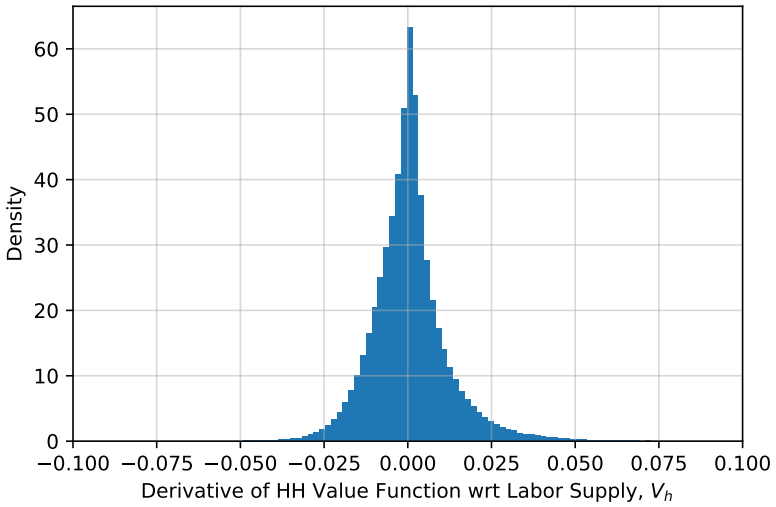


Figure 1: Histogram of Derivatives of HH Value Function ($\partial V/\partial h$)

To check if the neural network approximation is actually finding a maximum of the households' value function, I can manually take derivatives of the value function around the

final allocations. The algorithm should find allocations for which the first derivatives are on average equal to zero and the second derivatives are negative.[6] In Table 3, I show the distribution of derivatives and second derivatives with respect to the labor supply decision of households. Notably, first derivatives are close to zero and second derivatives are negative and far away from zero. While this does not ensure that the neural network is finding a global maximum of the household problem, it does seem to find an optimum and the value function is concave at the optimal allocation.

	Percentile of Distribution					
	10%	25%	50%	75%	90%	Average
First ($\partial V/\partial h$)	-1.4 $\times 10^{-2}$	-6.5 $\times 10^{-3}$	-1.6 $\times 10^{-4}$	6.4 $\times 10^{-3}$	1.5 $\times 10^{-2}$	7.9 $\times 10^{-4}$
Second ($\partial^2 V/\partial h^2$)	-0.228	-0.193	-0.134	-0.110	-0.086	-0.151

Table 3: Derivatives of Household Value Function at Approximation

In Figure 1, I plot the density of the first derivatives of the household problem with respect to their labor supply. This plot shows that not only are the derivatives close to zero on average, but they also have relatively small variance and skewness. This suggests that the neural network is in fact solving the correct maximization problem. Although there is no way to ensure that the approximation of optimal taxes given by the neural network approach is close to the true optimum, this seems to suggest that the algorithm described in the previous subsection is maximizing the correct object and that household welfare seems to be a concave function in allocations. Since the social welfare function is an aggregate of the concave individual value functions, the neural network should also be capable of finding a tax function close to the true optimal system.

6 Results

In this section, I describe the results of the neural network approximation of optimal taxes. First, I show how optimal taxes change under different restrictions on the tax function. Then, I show how welfare and allocations vary under each tax function. I show that the government does not choose to condition taxes on income history when types are perfect substitutes in the production function. Finally, I show that a simpler parametric tax function that varies with just the average of lifetime income can mimic the full history-dependent tax schedule and achieves almost the same level of welfare.

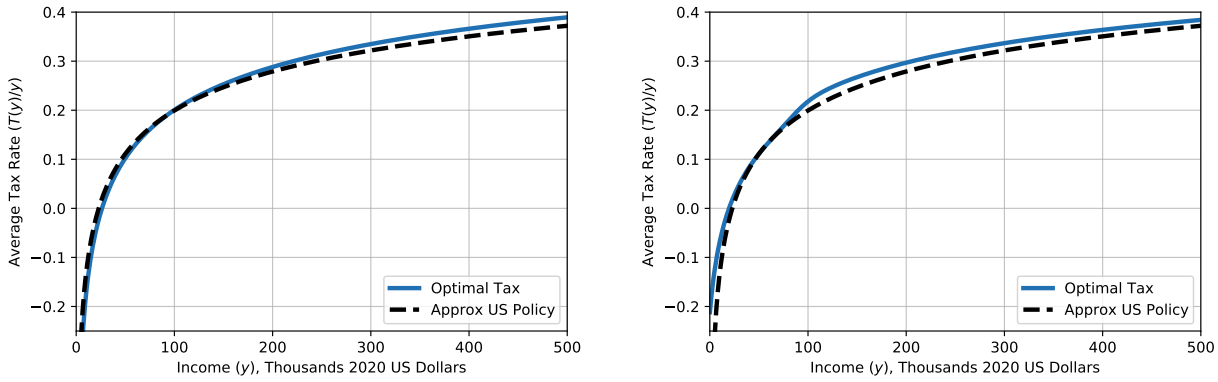
[6] With history-dependent taxes, labor supply solves a forward-looking optimization problem, so first derivatives should not be expected to always be exactly equal to zero.

6.1 Optimal Tax Functions

In this section, I describe the approximations of the optimal taxation exercise described in the section 3.3. I approximate optimal tax schedules under various restrictions of the tax function available to the government. First, I only allow the government to put taxes on current labor income. I then allow the government to condition its tax function on age. Finally, I let the government condition taxes on a household's entire previous history of income.

6.1.1 Taxes on Current Income Only

In Figure 2, I plot the average tax rates for the optimal tax functions when I restrict the tax function to depend only on current income.[7] In the lefthand panel I plot the optimal tax function under the parametric function (6) where the parameters of the tax function cannot vary with age. I also plot an approximation of the labor income tax function in the Unites States using the same parametric function (6), where I set $\tau = 0.151$, which was used by **?** and Golosov et al. (2016) to approximate the US tax income policy. The optimal tax under this parametric function is slightly more progressive than the approximation of the US policy. In the righthand panel, I plot the optimal tax function when I remove the parametric restriction on the tax function. Instead, I allow the average tax function to be any differentiable function of current income. This results in a roughly similar tax function as the parametric case. The main difference is that tax rates dip slightly for middle income households. This U-shape in tax rates is consistent with what previous studies of nonlinear income taxation, e.g. Saez (2001), have found in the optimal system.



(a) Optimal Parametric Tax, $T(y) = y - (1-\tau)y^{\rho}$ (b) Optimal Nonparametric Tax, $T(y)$

Figure 2: Optimal Tax on Current Income

[7] I summarize marginal tax rates for taxes on current income in Appendix A.5.1

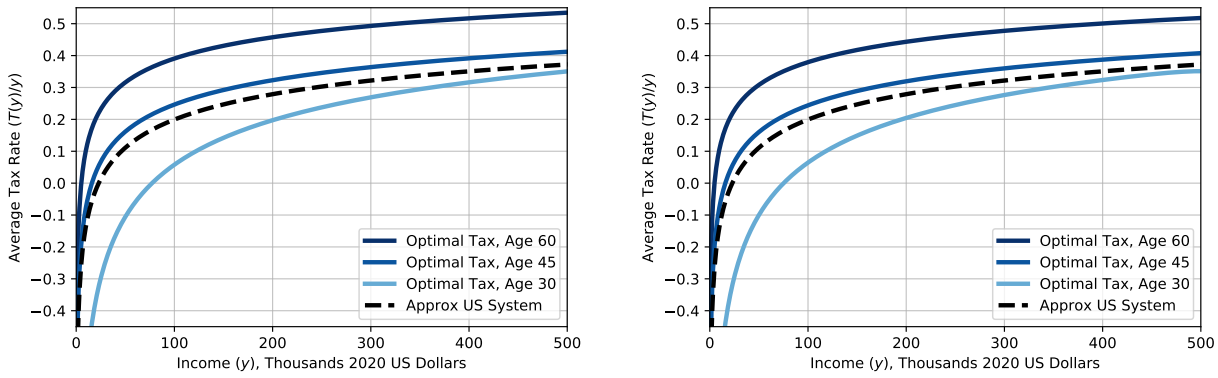
6.1.2 Age-Dependent Taxes on Current Income Only

In Figure 3, I plot the average tax rates for the optimal tax functions when I allow the tax function to vary with age.[8] The lefthand panel shows the optimal tax function under the parametric function (6), where parameters of the tax function are assumed to be a simple polynomial in age,

$$T_a(y) = y - (1 - \tau(a))y^{\rho(a)}$$

$$\tau(a) = \tau_0 + \tau_1 a + \tau_2 a^2 \text{ and } \rho(a) = \rho_0 + \rho_1 a + \rho_2 a^2$$

The optimal tax function under this tax function increases with age and is more progressive for the youngest households. The shape of optimal taxes over the life cycle in this model closely follows the shape of the deterministic age profile of productivity. The government uses taxes to smooth after-tax wages over the lifecycle. The righthand panel of Figure 3 plots the average tax rates when the average tax rate is allowed to be any differentiable function of current income and age. This less restrictive function turns out to be very similar to the parametrically restricted optimal tax. The main difference is that the tax function becomes flatter for higher incomes, which is barely visible when looking at the tax functions themselves. In Figure 9, I plot the present value of taxes paid under both tax functions. In present value terms, it is much clearer that the nonparametric tax function is flatter for households with incomes above two hundred thousand dollars. However, it turns out that removing parametric restrictions on the optimal age-dependent tax results in minimal gains to social welfare.



(a) Optimal Parametric Tax, $T(y,a) = y - (1 - \tau(a))y^{\rho(a)}$

(b) Optimal Nonparametric Tax, $T(y,a)$

Figure 3: Optimal Age-Dependent Tax on Current Income

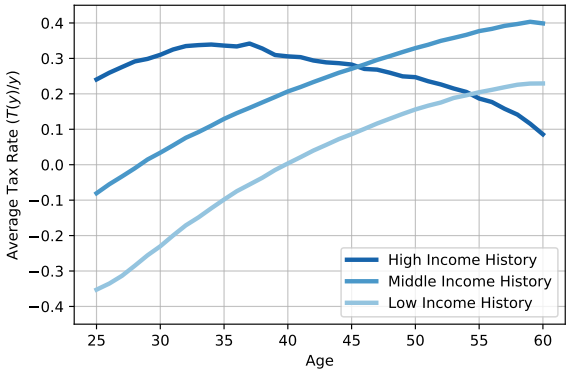
[8] I summarize marginal tax rates for age-dependent taxes in Appendix A.5.2

6.1.3 History-Dependent Taxes

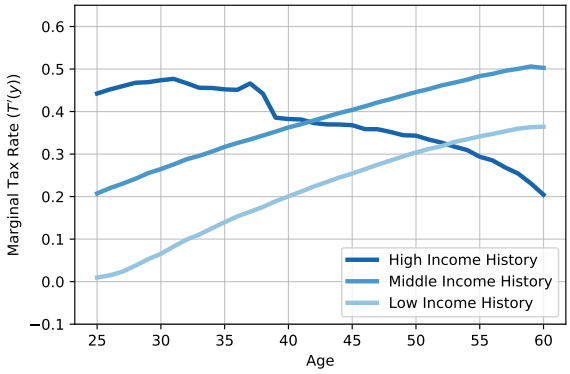
Now I allow taxes to depend on both age and the entire history of past income. To clarify how income history affects the amount of taxes paid, I follow three different income histories. I set income to be a constant value every single year and then compute the tax function at different ages. In particular I look at three levels of income: a "low income history" that receives the 25th percentile of income (within their age group) every period, a "middle income history" that receives the 75th percentile of income every period and a "high income history" that receives the 99th percentile of income every period. The low income history makes around 30,000 dollars on average during their working life, the middle income history earns on average around 100,000 dollars and the high income history earns on average around 500,000 thousand dollars.

Tax Rates In Figure 4, I plot the tax rates faced by three households who receive the 25th, 75th and 99th percentiles of the income distribution (within their age group) each period.[9] In the left panel, I plot the average tax rates. While the average tax rate is increasing for the histories close to average incomes, average tax rates for high income history begin to decrease in age after ten years. In fact, by age sixty the high income history is paying less than ten percent in taxes. In the right panel, shows the marginal tax rates for the same three income histories. Similarly to average tax rates the government increases marginal tax rates with age for the more average income histories, but decreases tax rates for the high income history. This suggests that the government has a motive to give tax breaks to high income households, but only after many years of high levels of output.

[9]Tax rates for the 99th percentile are much less smooth than in the middle of the income distribution. This comes from the fact that the optimal tax system is approximated through simulation. Since less data is provided in the estimation, the final optimal tax function is less smooth for extreme levels of income.

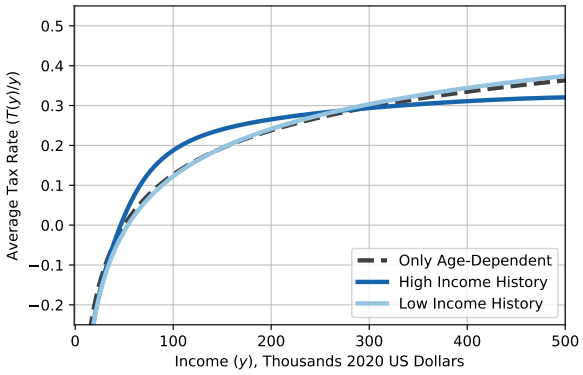


(a) Average Tax Rates

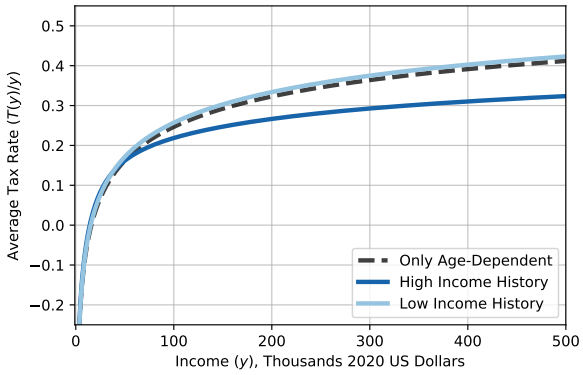


(b) Marginal Tax Rates

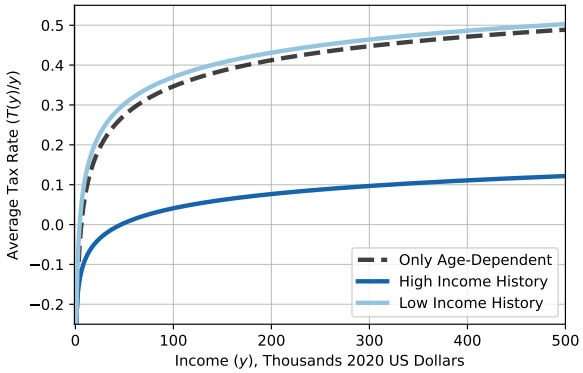
Figure 4: Optimal History-Dependent Tax by Age



(a) Optimal Tax, Age 30



(b) Optimal Tax, Age 45



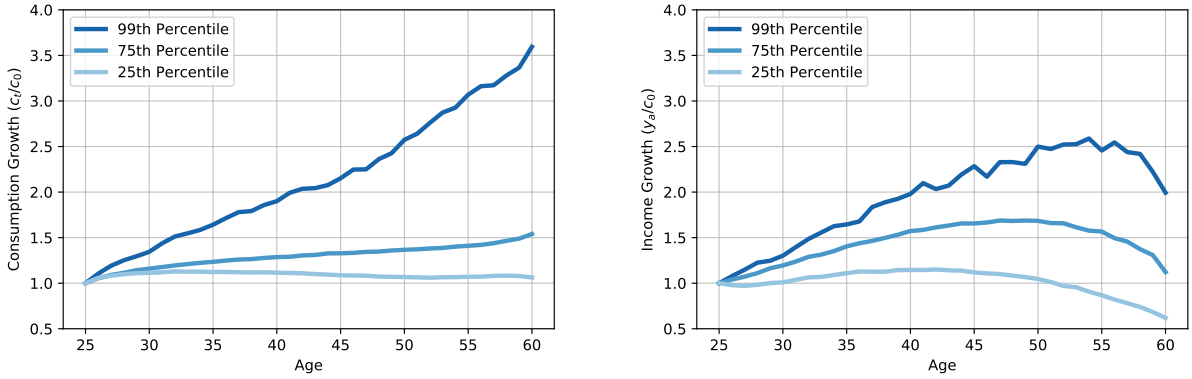
(c) Optimal Tax, Age 60

Figure 5: Optimal Non-parametric History-Dependent Tax

Tax Schedules In Figure 5, I plot the entire tax schedules faced by the low income history and the high income history. I plot the tax schedules faced by each history after five, twenty and thirty-five years in the labor market. I also plot the optimal tax function when taxes are not allowed to depend on income history and can only vary depend on age and current income. Notice that the low income history experiences similar tax schedules as when taxes only vary with age: taxes generally increase over time and are more progressive when young. However, the schedules for the high income history change very differently from the age-dependent case. Instead, taxes generally decrease with time spent in the labor market. From a pure redistribution point of view, this is counterintuitive since it might be expected that the government would tax households with a history of high income more. As I will show in the following sections, this intuition is no longer true in a general equilibrium model with differentiated labor. When labor is differentiated in skill types, as it is in this model and in Heathcote, Storesletten and Violante (2017, 2020), then the government can get large benefits from increasing output from the rarest skill types. This is because rare skill types have very high marginal productivities, so incentivizing them to increase their output has potentially large spillover effects through general equilibrium wages. How large these effects are depends entirely on how complementary labor is. I show in section 6.1.5 that the government does not choose to condition taxes on past income if skill types are perfect substitutes. This highlights what might be the key takeaway from these results: the elasticity of substitution between skill types is critical in how much benefit comes from nonlinear or history-dependent policy. This paper assumes skill types are permanent after entering the labor market and that the elasticity of substitution between skills is constant, which implies there may be large benefits from history-dependence. Taken literally, this implies that households with a history of high income should be taxed at lower rates. However, this assumes that high labor income in reality is mainly a result of acquiring rare skills and that rare skill types are highly complementary with more common skill types. A much richer model of how skills are formed and interact in production is necessary to determine the quantitative significance of history-dependence in optimal taxation. It might also be possible that the government is better off focusing on policies that result in more equitable initial skill formation than manipulating prices after skill choices have been made.

Discussion Why do history-dependent rewards increase labor supply? Changing tax rates on current income will have very small effects on labor supply with balanced growth preferences, which are used in this model. This is because the income effect – the disincentive effects of being able to work less for the same amount of income – and the substitution effect – the incentive effect of making more income per hour of work – of higher wages cancel so

that increasing after-tax wages has no effect on labor supply. Unlike a tax on current income, a history-dependent tax can increase labor supply with a pure substitution effect. The government is able to achieve this by promising higher consumption in the future in exchange for higher labor supply today. This increases the effective after-tax wage without changing consumption, which produces a substitution effect without any countervailing income effect. The government exploits this to increase labor supply by younger households. At the same time, history-dependent taxation also produces an income effect later in life when households receive the higher consumption for their past output. However, since household wages increase deterministically over the life-cycle, the income effects from the history-dependent tax for older workers are much smaller than the substitution effects for younger workers. The result is that a history-dependent tax can substantially increase total output over a tax on current income.[10]



(a) Percentiles of Consumption by Age

(b) Percentiles of Income by Age

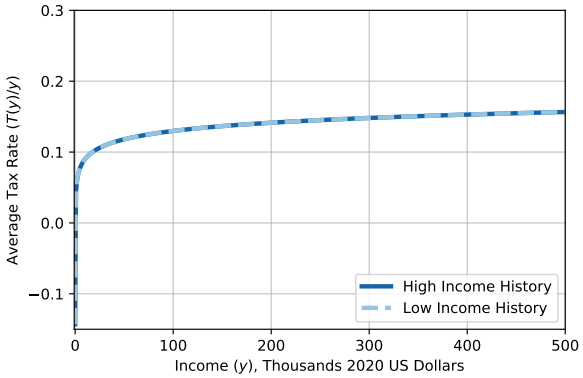
Figure 6: Life-cycle Percentiles of Consumption and Income

Figure 6 shows the 25th, 75th and 99th percentiles of consumption and income growth by age under the optimal history-dependent tax system. These are shown as allocations relative to their levels at age 25. Notice that consumption is relatively smooth over the life-cycle for lower income levels, but consumption for the highest income levels is increasing strongly with age. This reflects how the history-dependent tax function smoothes consumption for low and middle income households but creates an increasing consumption profile for high income households through history-dependent rewards. The government wants redistribute consumption by high wage households to later periods in order to maximize average labor supply over their lifetime. This is highly distortionary for the high wage households, but

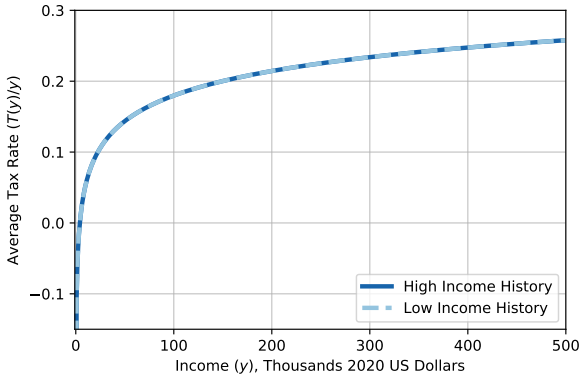
[10]In Appendix A.3, I demonstrate in a simple two period example how history-dependent taxation can increase total output.

results in increased wages for all other households.

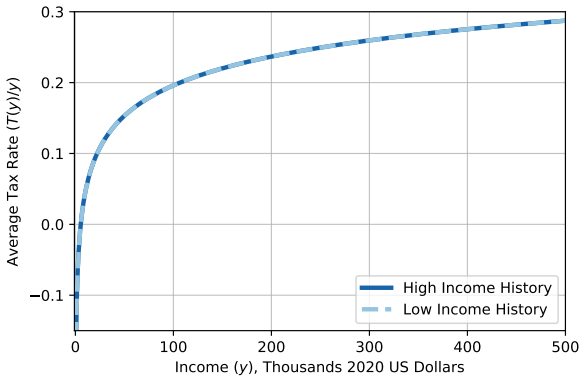
6.1.4 Skills Are Perfect Substitutes in Production



(a) Optimal Tax After 10 Years



(b) Optimal Tax After 20 Years



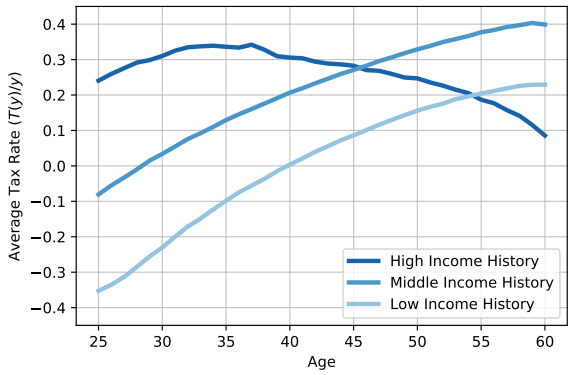
(c) Optimal Tax After 30 Years

Figure 7: Optimal History-Dependent Tax, Perfect Substitutes ($\omega = \infty$)

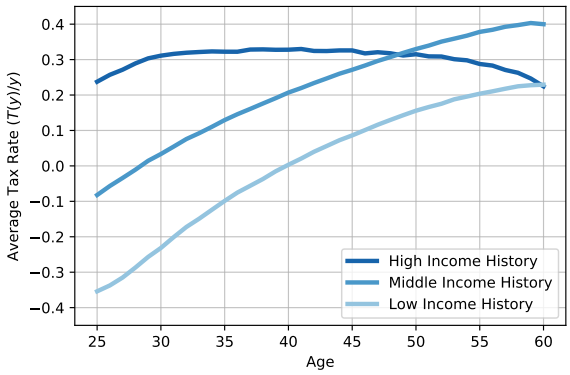
For the results so far, I have used the elasticity of substitution between skill types estimated by Heathcote et al. (2017). This elasticity implies a relatively high level of complementarity between skill types. In this subsection, I will show that this low elasticity drives all benefits from history-dependent taxation. In fact, if I consider the extreme case where all skill types are perfect substitutes in production, the government essentially chooses to not condition taxes on past income at all. In Figure 7, I again plot the tax schedules for two households with different income histories in the same way I plotted the history-dependent schedules for the baseline model. The first household, which I call the "low income history" household makes fifty thousand dollars every single year and the second household, which I

call the "high income history" household makes four hundred thousand dollars every single year. I plot the tax schedules faced by each household after ten, twenty and thirty years in the labor market. Notably, when I make skill types perfect substitutes, the schedules faced by each household are basically identical. In fact, when I simulate the economy under history-dependent taxes and just age-dependent taxes, welfare under the two tax systems are virtually identical. Unlike in the baseline model, progressivity increases with age when skills are perfect substitutes. This is because, in this case, skill prices are all equal to one and wages are simply equal to household productivity. Since the mean and variance of productivity increases with age, the government improves welfare by increasing both average taxes and progressivity over the life cycle.

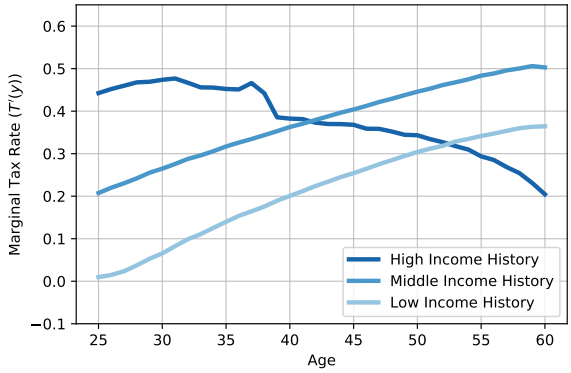
6.1.5 Skills Types are Observable



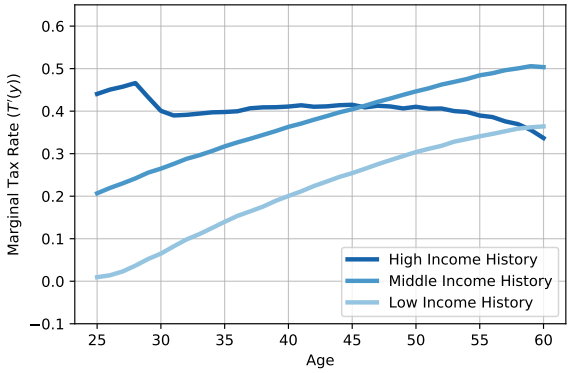
(a) Average Tax Rates, Unobservable Skills



(b) Average Tax Rates, Observable Skills



(c) Marginal Tax Rates, Unobservable Skills



(d) Marginal Tax Rates, Observable Skills

Figure 8: Optimal History-Dependent Tax by Age

Another motive the government has for using history-dependent taxation is to learn which skill type each household is. Now, consider the case where households' skill type s is observable to the government and the government can make taxes depend on the skill type. This will demonstrate how relatively important learning a household's type is versus rewarding high levels of output. Figure 8 shows the average and marginal tax rates under both the baseline model where skill types are unobservable and when skills are observable to the government. With observable types, I compute the taxes paid by each percentile of income at the skill type associated with each percentile. Notably, both the average and marginal tax rates are flatter for the high income history when skill types are observable. Meanwhile, average and marginal tax rates for the low and middle income history are virtually identical in each case. Average and marginal tax rates for the high income history begin at similar levels in both cases, but by the final period of work, both the average and marginal tax rates are almost fifteen points higher in the observable case. This demonstrates how the government still chooses to use history-dependence to incentivize output when types are observable, but the government can use smaller rewards if they do not need to separate types.

6.2 Welfare Comparison

Now I describe welfare and allocations under each of the optimal tax systems. I first describe how welfare is converted to consumption equivalent units for direct comparison across tax systems. Then, I show how much welfare is gained by progressively removing restrictions on the tax function. I next show how allocations change under history-dependent versus just age-dependent taxes. Finally, I show how the distribution of the present value of taxes paid changes under the different tax functions.

6.2.1 Consumption-Equivalent Welfare

To compare the welfare effects of each policy, I compute consumption equivalent welfare by calculating the percent change in consumption that delivers equal levels of utilitarian welfare. Consumption-equivalent welfare gain of moving to a new tax function T^* from T is the g such that

$$W\left(s^*, \{(c_a^*, h_a^*\}_{a=0}^{A-1}; T^*\right) = W\left(s, \{(1+g)c_a, h_a\}_{a=0}^{A-1}; T\right)$$

where W is the social welfare function,

$$W\left(s, \{c_a, h_a\}_{a=0}^{A-1}; T\right) = \int \left\{ -v_i\left(s_i(T)\right) + E_0\left[\left(\frac{1-\beta}{1-\beta^A}\right)\sum_{a=0}^{A-1}\beta^a u_i\left(c_{ia}(T), h_{ia}(T)\right)\right]\right\} di$$

That is, g is the percent increase in lifetime consumption, under the old tax function T, necessary to deliver same level of utilitarian welfare under the new tax function, T^*. Converting welfare to consumption equivalent units provides a more easily interpretable way to compare welfare across policies. For example, a one percent increase in lifetime consumption would imply, on average, between five hundred to one thousand dollars in additional consumption per year.

6.2.2 Welfare Under Optimal Taxes

Here, I report welfare in the stationary equilibrium under each tax optimal tax system. In Table 4, I summarize the different tax functions that I study. The first column shows the history-dependent tax function that is allowed to be any continuous function of the household's entire history of income. For now, I only allow history-dependent taxes to be nonparametric; in section 6.3 I study a simple parametric tax function that can depend on the average of previous incomes. The second column shows the age-dependent tax functions I study, the top row allows the average tax function to be any differentiable function of age and current income, while the bottom row restricts the function so that after-tax income is a log-linear transformation of before-tax income. The third column shows the tax functions that can only depend on current income.

In Table 4, I report the welfare gains from each tax function compared to the simplest policy: a parametric tax on only current income. The welfare gains from removing parametric restrictions are 0.8 percent of lifetime consumption for taxes on current income, but only about 0.15 percent for taxes that can depend on age. The welfare gains from allowing taxes to depend on age are about four percent of lifetime consumption, while the gains from additionally allowing for history-dependent taxation are about two percent of lifetime consumption. Welfare gains from age-dependent taxation being much larger than history-dependence is consistent with previous studies of history-dependent taxation, but the size of the welfare gain from history-dependence is much larger. Also, as I will show in the following subsections, the reason for welfare gains is much different than in previous studies. While previous papers have focused on the redistributive gains of taxation holding wages constant, the main welfare gains in this model seem to come from the government using the tax system to manipulate general equilibrium wages. In fact, as I show in section 6.1.5, when I make skill types in the model perfect substitutes in production, there is virtually no benefit from history-dependent taxes.

	History of Income	Age and Current Income	Only Current Income
Nonparametric	$T(y;\{y_t\}_{t=0}^{a-1},a)$	$T(y;a)$	$T(y)$
Parametric		$y-(1-\tau(a))y^{\rho(a)}$	$y-(1-\tau)y^{\rho}$

Table 4: Summary of Tax Functions

	Income History	Age and Current Income	Only Current Income
Nonparametric	6.31%	4.41%	0.82%
Parametric		4.25%	0.0%

Table 5: Welfare Gain of Moving from Parametric Tax on Current Income

6.2.3 Optimal Allocations

In this section, I will compare steady state allocations under the optimal history-dependent tax system to allocations under the optimal parametric age-dependent tax. I will compare allocations across different values of the present value of income, which is computed for each household i as

$$PV_i(y) = \left(\frac{1-R^{-1}}{1-R^{-A}}\right)\sum_{a=0}^{A-1} R^{-a}y_{ia}$$

Table 6 shows the percent increase in allocations under the optimal history-dependent tax system compared to the optimal age-dependent tax system.[11]Most notably the percent increase in income is huge for households in the top quartile of the present value of income distribution while incomes for all other quartiles fall slightly. Consumption for the top quartile also increases substantially under the history-dependent system while consumption for all other quartiles changes very little. Also, leisure for the top quartile decreases substantially while it increases a large amount for the lower quartiles. When reading these numbers, it is important to note that the utility weight on consumption, ϕ, is on average equal to 0.275, which means households value leisure about three and a half times consumption. Essentially, the history-dependent tax system is able to increase leisure for lower income households without a large change in consumption. To understand how this is possible, notice that skill prices rise on average even though skill investment decreases substantially for the lowest skill types. This highlights a key feature of the production technology described in section 3: skill types are complementary to each other with constant elasticity of substi-

[11]These allocations can be viewed in levels for each tax system in section A.6.

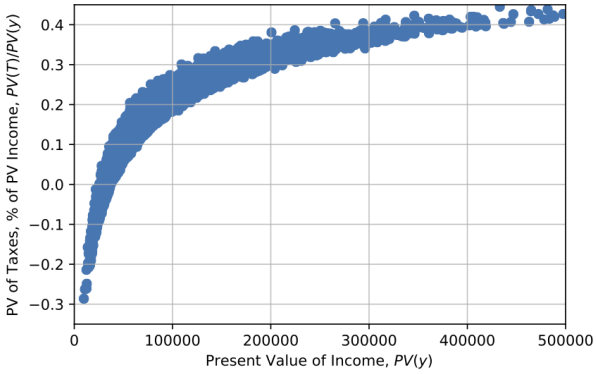
tution. If an especially rare skill type raises their labor supply, it causes a relatively large increase in output. Because of complementarity in production, an increase in output from rare skill types substantially raises the marginal product of all other households. The government uses history-dependence to take advantage of this complementarity in skill types. It uses history-dependent rewards to incentivize output from rare skill types to increase skill prices. This allows it to increase leisure and lower skill investment without substantially decreasing consumption. This is similar to the effects described by Stiglitz (1982, 1987): when labor is differentiated in production, the government wants to use taxes to increase labor supply from types with higher marginal productivity so that it can increase general equilibrium wages. Then, higher wages allow it to redistribute less under the optimal tax system.

	Quartile, Present Value of Income				
	0-25%	25-50%	50-75%	75-100%	Total
Income ($PV(y)$)	-0.82%	-0.56%	-0.17%	5.20%	2.12%
Consumption ($PV(c)$)	0.35%	-0.10%	-0.23%	7.08%	2.85%
Leisure ($PV(1-h)$)	0.38%	0.40%	0.38%	-1.15%	0.01%
Skills (s)	-2.26%	-0.66%	0.65%	1.64%	0.51%
Skill Price ($p(s)$)	-0.39%	0.19%	0.75%	1.72%	0.70%

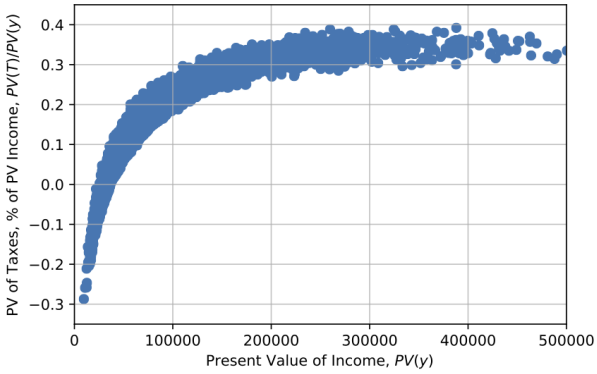
Table 6: Percent Gain in Average Allocations by Quartile of PV of Income, AD to HD Tax

6.2.4 Present Value of Taxes Paid

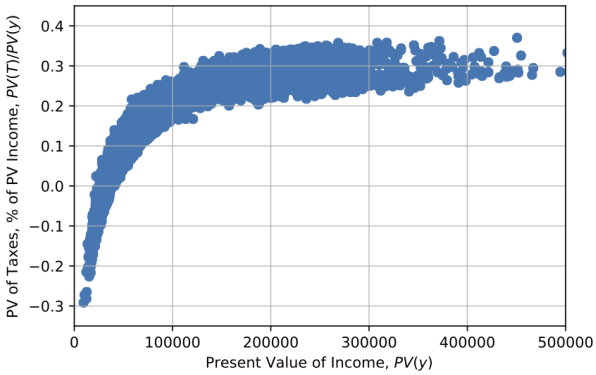
In Figure 9, I plot the distribution of the present value of taxes paid by each household as a fraction of their present value of income. I simulate twenty-five thousand households under each tax system and create a scatter plot of their present value of taxes paid. In the first panel, I plot the present value of taxes paid under parametric age-dependent taxes. In this case, households' present value of taxes are roughly a log-linear function of their present value of income. In the second panel, I plot the distribution of present values of taxes paid under the nonparametric age-dependent tax system. This tax function is very similar to the parametric tax system for incomes below two hundred thousand dollars. However, the present value of taxes paid do not increase as much for households with present values of income above two hundred thousand, instead staying between thirty and forty percent of their present value of income. The third panel shows the present values of taxes paid under the history-dependent tax system. Besides being relatively flat for higher levels of income, the variance of taxes paid within households with similar present values of income is larger



(a) Parametric Age-Dependent Taxes



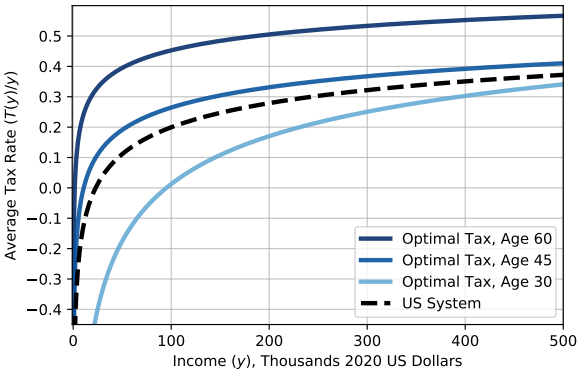
(b) Nonparametric Age-Dependent Taxes



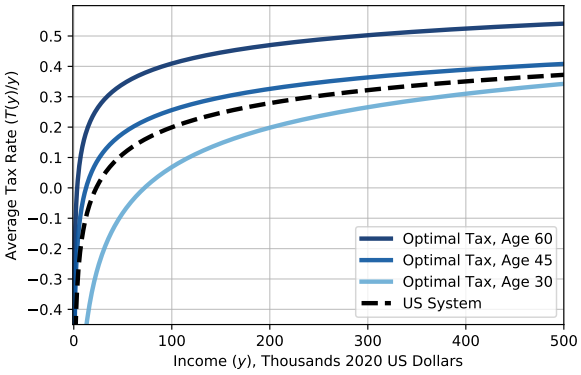
(c) History-Dependent Taxes

Figure 9: Present Value of Taxes Paid by Present Value of Income

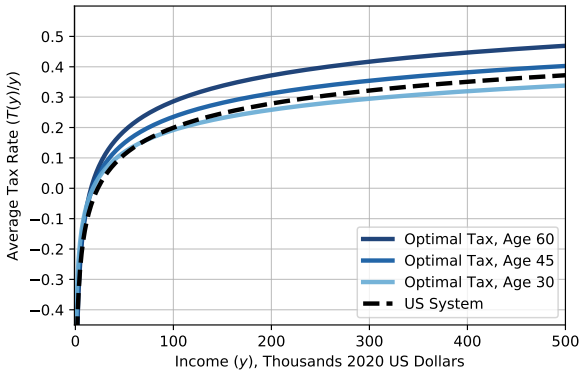
than in the age-dependent case. This is especially true for the households with relatively high present values of income, with households under one hundred thousand seeming to face similar taxes. This is consistent with the allocations described above: the government mostly uses history-dependent taxes to incentivize work for households with high present values of income, who are usually the households with the rarest skill types. History dependent taxation creates both intra-temporal and inter-temporal distortions on households, so the government will generally avoid using them unless the benefit of incentivizing work is very high. Also, the government can redistribute well with just age-dependent taxes. Because of this, they generally will not use history-dependence with low and middle income households.



(a) Optimal Tax by Age, $\bar{y} = \$25,000$



(b) Optimal Tax by Age, $\bar{y} = \$100,000$



(c) Optimal Tax by Age, $\bar{y} = \$500,000$

Figure 10: Optimal Parametric History-Dependent Tax

6.3 Simpler Implementation of History-Dependent Taxes

The history-dependent tax function studied in the previous sections is difficult to interpret or use for insights into how tax policy should be structured. Parametric tax rules are useful because they give much more easily interpretable results. In this section, I consider a parametric function in just the average of previous income

$$T(y;a,\bar{y}) = y - (1-\tau(a,\bar{y}))y^{\rho(a,\bar{y})}$$

where

$$\tau(a,\bar{y}) = \tau_0 + \tau_1 a + \tau_2 a^2 + \tau_3 a\bar{y} + \tau_4 a^2\bar{y} \text{ and } \rho(a,\bar{y}) = \rho_0 + \rho_1 a + \rho_2 a^2 + \rho_3 a\bar{y} + \rho_4 a^2\bar{y}$$

and $\bar{y} = \frac{1}{a-1}\sum_{t=0}^{a-1} y_t$. Notice that the level and progressivity of the tax function now vary with average lifetime income as well as age. This parametric function allows me to more directly compute how the progressivity and levels of taxes should change with income history. Also, it turns out that this much simpler parametric tax function achieves almost the same level of welfare as the full history-dependent tax function. The full history-dependent tax function only provides the equivalent of a 0.2% gain in lifetime consumption, so the parametric function captures about 90 percent of the potential welfare gains from history-dependence.

In Figure 10, I plot the optimal taxes under this simpler parametric tax function. The most noticeable feature of this optimal tax function is that taxes increase with age across all levels of previous income, but they start higher and increase slower with higher levels of previous income. This mimics the key features of the full history-dependent taxes, where households with high income history are rewarded for high output with less taxes in the future. For households with low income history, the government chooses a highly redistributive tax function that increases substantially with age. Households with high income history instead start their lives with higher average tax rates, but their tax rates increase very slowly with age.

7 Conclusion

In this paper, I used neural networks to compute optimal history-dependent taxes in an overlapping generations economy with general equilibrium wages. I found that the welfare gains from history-dependence are potentially large and that the elasticity of substitution between skill types is critical to optimal policy. In fact, virtually all potential gains from history-dependent taxation disappear if skill types are perfect substitutes. In this paper,

I followed existing literature and assumed that skill types are permanent after entering the labor market and that the elasticity of substitution between skills is constant, which implies there may be large benefits from history-dependence. Taken literally, this implies that households with a history of high income should be taxed at lower rates. However, this assumes that high labor income in reality is mainly a result of acquiring rare skills and that rare skill types are highly complementary with more common skill types. A much richer model of how skills are formed and interact in production is necessary to determine the quantitative significance of history-dependence in optimal taxation. It might also be possible that the government is better off focusing on policies that result in more equitable initial skill formation than manipulating prices after skill choices have been made. Therefore, incorporating skill formation and how different skills enter production into the study of optimal taxation seems like an important avenue for future research.

The method I describe in this paper can be applied to many different models to solve for optimal policies. Being able to compute more general optimal policies in a given model is useful for considering which more easily implementable (or interpretable) policies to study. For example, in this paper I showed that a simple tax function in age and average income gets very close to fully nonlinear history-dependent optimal tax system.

References

Albanesi, Stefania and Christopher Sleet (2006) "Dynamic Optimal Taxation with Private Information," *Review of Economic Studies*, 73 (1), 1–30.

Azinovic, Marlon, Luca Gaegauf, and Simon Scheidegger (2019) "Deep Equilibrium Nets," *Working Paper*.

Benabou, Roland (2000) "Unequal Societies: Income Distribution and the Social Contract," *American Economic Review*, 90 (1), 96–129.

——— (2002) "Tax and Education Policy in a Heterogeneous-Agent Economy: What Levels of Redistribution Maximize Growth and Efficiency?" *Econometrica*, 70 (2), 481–517.

Blundell, R., L. Pistaferri, and I. Saporta-Eksten (2016) "Consumption Inequality and Family Labor Supply," *American Economic Review*, 106, 387–435.

Brumm, Johannes and Simon Scheidegger (2017) "Using Adaptive Sparse Grids to Solve High Dimensional Dynamic Models," *Econometrica*, 85, 1575–1612.

Chang, Yongsung and Yena Park (2020) "Optimal Taxation with Private Insurance," *Working Paper*.

Chen, Mingli, Andreas Joseph, Michael Kumhof, and Xinlei Pan (2021) "Deep Reinforcement Learning in a Monetary Model," *Working paper*.

Conesa, Juan Carlos, Sagiri Kitao, and Dirk Krueger (2009) "Taxing Capital? Not a Bad Idea after All!," *American Economic Review*, 99 (1), 25–48.

Conesa, Juan Carlos and Dirk Krueger (2006) "On the Optimal Progressivity of the Income Tax Code," *Journal of Monetary Economics*, 53 (7), 1425–1450.

Duarte, Victor (2018) "Machine Learning for Continuous Time Finance," *Working Paper*.

Erosa, Andres and Martin Gervais (2002) "Insurance and Taxation over the Life Cycle," *Review of Economic Studies*, 105 (2), 338–369.

Farhi, Emmanuel and Ivan Werning (2013) "Insurance and Taxation over the Life Cycle," *Review of Economic Studies*, 810 (2), 596–635.

Feldstein, Martin (1969) "The Effects of Taxation on Risk Taking," *Journal of Political Economy*, 77 (5), 755–764.

Fernández-Villaverde, Jesús, Galo Nuño, George Sorg-Langhans, and Maximilian Vogler (2020) "Solving High-Dimensional Dynamic Programming Problems using Deep Learning," *Working Paper*.

Findeisen, Sebastian and Dominik Sachs (2017) "Redistribution and Insurance with Simple Tax Instruments," *Journal of Public Economics*, 146, 58–78.

Fukushima, Kenichi (2011) "Quantifying the Welfare Gains from Flexible Dynamic Income Tax Systems," *Global COE Hi-Stat Discussion Paper Series* (176).

Gervais, Martiin (2012) "On the optimality of age-dependent taxes and the progressive US tax system," *Journal of Economic Dynamics and Control*, 36 (4), 682–691.

Golosov, Mikhail, Narayana Kocherlakota, and Aleh Tsyvinski (2003) "Optimal Indirect and Capital Taxation," *Review of Economic Studies*, 70, 569–587.

Golosov, Mikhail, Maxim Troshkin, and Aleh Tsyvinski (2016) "Redistribution and Social Insurance," *American Economic Review*, 106 (2), 359–386.

Golosov, Mikhail and Aleh Tsyvinski (2015) "Policy Implications of Dynamic Public Finance," *Annual Review of Economics*, 7, 147–171.

Heathcote, Jonathan, Kjetil Storesletten, and Giovanni L. Violante (2017) "Optimal Tax Progressivity: An Analytical Framework," *Quarterly Journal of Economics*, 132 (4), 1693–1754.

——— (2020) "Optimal Progressivity with Age-Dependent Taxation," *Journal of Public Economics*.

Heathcote, Jonathan and Hitoshi Tsujiyama (2020) "Optimal Income Taxation: Mirrlees Meets Ramsey," *Working Paper*.

Huggett, Mark and Juan Carlos Parra (2010) "How Well Does the U.S. Social Insurance System Provide Social Insurance?" *Journal of Political Economy*, 118 (1).

Judd, Kenneth L., Lilia Maliar, Serguei Maliar, and Rafael Valero (2014) "Smolyak method for solving dynamic economic models: Lagrange interpolation, anisotropic grid and adaptive domain," *Journal of Economic Dynamics and Control*, 44, 92–123.

Kapička, Marik (2020) "Quantifying the Welfare Gains from History Dependent Income Taxation," *Working paper*.

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

Kindermann, Fabian and Dirk Krueger (2021) "High Marginal Tax Rates on the Top 1*American Economic Journal: Macroeconomics*, forthcoming.

Krueger, Dirk and Felix Kubler (2004) "Computing equilibrium in OLG models with stochastic production," *Journal of Economic Dynamics and Control*, 28 (7), 1411–1436.

Krueger, Dirk and Alexander Ludwig (2013) "Optimal Progressive Labor Income Taxation and Education Subsidies When Education Decisions and Intergenerational Transfers Are Endogenous," *American Economic Review*, 103 (3), 496–501.

——— (2016) "On the optimal provision of social insurance: Progressive taxation versus education subsidies in general equilibrium," *Journal of Monetary Economics*, 77, 72–98.

Maliar, Lilia, Serguei Maliar, and Pablo Winant (2019) "Will Artificial Intelligence Replace Computational Economists Any Time Soon?" *CEPR Working Paper*.

Ndiaye, Abdoulaye (2020) "Flexible Retirement and Optimal Taxation," *Working Paper*.

Peterman, William (2016) "The effect of endogenous human capital accumulation on optimal taxation," *Review of Economic Dynamics*, 21, 46–71.

Rothschild, Casey and Florian Scheuer (2013) "Redistributive Taxation in the Roy Model," *Quarterly Journal of Economics*, 128 (2), 623–668.

Sachs, Dominik, Aleh Tsyvinski, and Nicolas Werquin (2020) "Nonlinear Tax Incidence and Optimal Taxation in General Equilibrium," *Econometrica*, 88 (2), 469–493.

Saez, Emmanuel (2001) "Using Elasticities to Derive Optimal Income Tax Rates," *Review of Economic Studies*, 68, 205–229.

Saez, Emmanuel and Stefanie Stantcheva (2018) "A Simpler Theory of Optimal Capital Taxation," *Journal of Public Economics*, 162, 120–142.

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

——— (2020) "Dynamic Taxation," *Annual Review of Economics*, 12, 801–831.

Stiglitz, Joseph (1982) "Self-Selection and Pareto Efficient Taxation," *Journal of Public Economics*, 17, 213–240.

——— (1987) "Pareto Efficient and Optimal Taxation and the New New Welfare Economics," *Handbook of Public Economics*, 2, 991–1042.

Weinzierl, Matthew (2011) "The Surprising Power of Age-Dependent Taxes," *Review of Economic Studies*, 78 (4), 1490–1518.

A Appendix

A.1 Proof that History-Dependent Tax Functions are Unique

Consider an economy where households live for two periods, $t = 0, 1$. They choose consumption for each period, $\{c_0, c_1\}$, savings b and income for each period $\{y_0, y_1\}$ to maximize their lifetime utility,

$$\max_{c_0, c_1, b, y_0, y_1} u(c_0, y_0) + \beta u(c_1, y_1)$$

subject to

$$\begin{aligned} c_0 + b &= y_0 - T_0(y_0) \\ c_1 &= Rb + y_1 - T_1(y_0, y_1) \end{aligned}$$

where tax functions each period are given by

$$T_0(y_0) = \tau_0(y_0) y_0$$

$$T_1(y_0, y_1) = \tau_1(y_0, y_1) y_1$$

and $\{\tau_0, \tau_1\}$ are differentiable functions.

The solution to the household's problem is characterized by the following five equations,

$$\begin{aligned} c_0 &: c_0 + b = y_0 - T_0(y_0) \\ c_1 &: c_1 = Rb + y_1 - T_1(y_0, y_1) \\ b &: u_{c0} = \beta R u_{c1} \\ y_0 &: 0 = u_{y0} + \left[1 - T_0'(y_0)\right] u_{c0} - \beta T_{1,y_0}(y_0, y_1) u_{c1} \\ y_1 &: 0 = u_{y1} + \left[1 - T_{1,y_1}(y_0, y_1)\right] u_{c1} \end{aligned}$$

where

$$\begin{aligned} T_0'(y_0) &= \tau_0'(y_0) y_0 + \tau_0(y_0) \\ T_{1,y_0}(y_0, y_1) &= \tau_{1,y_0}(y_0, y_1) y_0 + \tau_1(y_0, y_1) \\ T_{1,y_1}(y_0, y_1) &= \tau_{1,y_1}(y_0, y_1) y_1 + \tau_1(y_0, y_1) \end{aligned}$$

Simplify the equilibrium conditions by combining the two budget constraints to get the

present value budget constraint,

$$c_0 + \frac{1}{R}c_1 = y_0 - T_0(y_0) + \frac{1}{R}\left[y_1 - T_1(y_0, y_1)\right]$$

Next, assume $\beta = 1/R$, so that the household's Euler condition, $u_{c0} = \beta R u_{c1}$, implies that consumption is constant in the two periods, $c_0 = c_1 = c$. The equilibrium conditions can then be reduced to three equations in (c, y_0, y_1),

$$c: \ \left(1 + \frac{1}{R}\right)c = y_0 - T_0(y_0) + \frac{1}{R}\left[y_1 - T_1(y_0, y_1)\right] \tag{7}$$

$$y_0: \ 0 = u_{y0} + \left[1 - T_0'(y_0) - \beta T_{1,y_0}(y_0, y_1)\right] u_c \tag{8}$$

$$y_1: \ 0 = u_{y1} + \left[1 - T_{1,y_1}(y_0, y_1)\right] u_c \tag{9}$$

where $u_c \equiv u_{c0} = u_{c1}$.

Theorem: Under tax functions T_0, T_1 of the form

$$T_0(y_0) = \tau_0(y_0) y_0$$

$$T_1(y_0, y_1) = \tau_1(y_0, y_1) y_1$$

where τ_0 and τ_1 are differentiable functions, the tax functions associated with a given equilibrium allocation (c, y_0, y_1) are unique.

Proof (by contradiction): Suppose that there exists an alternate set of tax functions $\tilde{T}_0, \tilde{T}_1$ of the form

$$\tilde{T}_0(y_0) = \tilde{\tau}_0(y_0) y_0$$

$$\tilde{T}_1(y_0, y_1) = \tilde{\tau}_1(y_0, y_1) y_1$$

such that the tax functions imply an identical allocation (c, y_0, y_1) as under the original tax functions T_0, T_1 . This is true if and only if all three of the equilibrium equations hold under the alternate tax functions,

$$c: \ \left(1 + \frac{1}{R}\right)c = y_0 - \tilde{T}_0(y_0) + \frac{1}{R}\left[y_1 - \tilde{T}_1(y_0, y_1)\right] \tag{10}$$

$$y_0: \ 0 = u_{y0} + \left[1 - \tilde{T}_0'(y_0) - \beta \tilde{T}_{1,y_0}(y_0, y_1)\right] u_c \tag{11}$$

$$y_1: \ 0 = u_{y1} + \left[1 - \tilde{T}_{1,y_1}(y_0, y_1)\right] u_c \tag{12}$$

Perturbation in y_1 dimension First, consider an arbitrary perturbation of the tax functions in the y_1 dimension, $\delta(y_1)$, so that

$$\tilde{\tau}_0(y_0) = \tau_0(y_0)$$
$$\tilde{\tau}_1(y_0, y_1) = \tau_1(y_0, y_1) + \delta(y_1)$$

Immediately, it is true that $\tilde{T}_0(y_0) = \tilde{\tau}_0(y_0)y_0 = \tau_0(y_0)y_0 = T(y_0)$. Subtracting the present value budget constraint under the original tax function (7) from the budget constraint under the perturbed tax function (10) gives

$$\begin{aligned}\left(1+\frac{1}{R}\right)c - \left(1+\frac{1}{R}\right)c = y_0 - \tilde{T}_0(y_0) + \frac{1}{R}\left[y_1 - \tilde{T}_1(y_0, y_1)\right] \\ - \left[y_0 - T_0(y_0) + \frac{1}{R}\left[y_1 - T_1(y_0, y_1)\right]\right]\end{aligned}$$

Using $\tilde{T}_0(y_0) = T(y_0)$, this implies that

$$\tilde{T}_1(y_0, y_1) = T_1(y_0, y_1)$$

So the only perturbation in the y_1 dimension that delivers the same allocations as the original tax functions is zero ($\delta(y_1) = 0$).

Perturbation in y_0 dimension Now, consider arbitrary perturbations $\epsilon(y_0), \delta(y_0)$ to the tax functions in the y_0 dimension so that

$$\tilde{\tau}_0(y_0) = \tau_0(y_0) - \epsilon(y_0)$$
$$\tilde{\tau}_1(y_0, y_1) = \tau_1(y_0, y_1) + \delta(y_0)$$

Then, the marginal tax functions are given by

$$\begin{aligned}\tilde{T}_0'(y_0) &= \left[\tau_0'(y_0) - \epsilon'(y_0)\right] y_0 + \tau_0(y_0) - \epsilon(y_0) \\ \tilde{T}_{1,y_0}(y_0, y_1) &= \left[\tau_{1,y_0}(y_0, y_1) + \delta'(y_0)\right] y_0 + \tau_1(y_0, y_1) + \delta(y_0) \\ \tilde{T}_{1,y_1}(y_0, y_1) &= \tau_{1,y_1}(y_0, y_1)y_1 + \tau_1(y_0, y_1) + \delta(y_0)\end{aligned}$$

First, substitute the expression for the perturbed marginal tax function, $\tilde{T}_{1,y_1}(y_0, y_1)$, into the household's first order condition (12) for y_1

$$0 = u_{y1} + (1 - [\tau_{1,y_1}(y_0, y_1)y_1 + \tau_1(y_0, y_1) + \delta(y_0)])\, u_c$$

Subtract the same condition under the original tax functions (9) to get

$$u_{y1} + (1 - [\tau_{1,y_1}(y_0, y_1)y_1 + \tau_1(y_0, y_1)])\, u_c = u_{y1} + (1 - [\tau_{1,y_1}(y_0, y_1)y_1 + \tau_1(y_0, y_1) + \delta(y_0)])\, u_c$$

which implies that

$$\delta(y_0) = 0$$

This implies that $\tilde{T}_1(y_0, y_1) = T_1(y_0, y_1)$. Now, subtract the present value budget constraint under the original tax function (7) from the budget constraint under the perturbed tax function (10),

$$\begin{aligned}\left(1 + \frac{1}{R}\right) c - \left(1 + \frac{1}{R}\right) c = y_0 - \tilde{T}_0(y_0) + \frac{1}{R}\left[y_1 - \tilde{T}_1(y_0, y_1)\right] \\ - \left[y_0 - T_0(y_0) + \frac{1}{R}\left[y_1 - T_1(y_0, y_1)\right]\right]\end{aligned}$$

Using $\tilde{T}_1(y_0, y_1) = T_1(y_0, y_1)$, this reduces to

$$\tilde{T}_0(y_0) = T_0(y_0)$$

which in turn implies that $\epsilon(y_0) = 0$. That is, the only perturbation in the y_0 dimension to the tax function that delivers the same allocations is no perturbation ($\delta(y_0) = \epsilon(y_0) = 0$).

Since there is no perturbation to the original tax functions that can deliver the same allocation (c, y_0, y_1), the tax functions associated with the allocation are unique. □

A.2 Production in a Standard Static Optimal Tax Problem

In the standard Mirrlees optimal tax problem, the production function does not directly enter into the optimal tax formulas. Here, I'll show using a standard static economy with a continuum of types why this does not apply when skills enter the production function as separate inputs. However, if the government is allowed to condition taxes on type, then the standard Mirrlees optimal tax equation holds. The reason for this is that the standard Mirrlees formulation is only nonlinear in efficiency units of labor and not types, so workers are only differentiated if they supply different amounts of labor. Under this assumption, before-tax wages are just a transformation of efficiency units of labor.

When workers are differentiated by type and enter the production function in different ways, income is no longer just a transformation of labor supply. Instead, different types receive different income for the same level of labor supply depending on the relative scarcity of their skills. This means changes in taxes can directly effect multiple skills types with

different levels of labor supply. This adds additional terms to the optimal tax formula to account for how taxes will affect each skill type and the associated general equilibrium spillover effects across skill types. These terms will not disappear if the government is allowed to condition its taxes on skill type. The effects come from skill types being differentiated, not the information available to the government.

A.2.1 Environment

Consider a static economy with a measure one of households. Household types, $\theta \in \mathbb{R}_+$, are drawn from a distribution $F(\theta)$. Household of type θ who works h hours produces $I(\theta, h)$ units of income. All households have preferences $u(c, h)$ over consumption c and hours worked h. Assume that neither θ or h are observable, so the government can only impose taxes on y. Let the tax on labor income be $T(y)$

Household Problem

A household of type θ solves

$$\max_{c,h} u(c, h)$$

subject to

$$c = I(\theta, h) - T\left(I(\theta, h)\right) \tag{13}$$

The first order condition of this problem is

$$-\frac{u_h}{u_c} = \left[1 - T'\left(I(\theta, h)\right)\right] I_h(\theta, h) \tag{14}$$

A.2.2 Optimal Tax Problem

Suppose the government wants to maximize utilitarian welfare across households

$$W = \max_T \int u\left(c(\theta; T), h(\theta; T)\right) dF(\theta)$$

Note: although the planner cannot observe θ for an individual household, it knows the distribution of types across households, $F(\theta)$. The government must satisfy it's budget constraint

$$G = \int T\left(I(\theta, h)\right) dF(\theta) \tag{15}$$

where E is the level of government expenditures. Optimal taxes are computed using a variational approach where the government starts with an arbitrary tax function, T, and consider a small perturbation of the tax function: $T(y) + \epsilon H(y)$, where ϵ is a scalar and

H is a function in y. $T(0)$ is adjusted to satisfy the government's budget constraint. The specification of the perturbation function H is arbitrary, but a useful normalization is

$$H(y) = \begin{cases} 0 & y < y^* \\ \dfrac{1}{1 - F_y(y^*)} & y \geq y^* \end{cases}$$

where F_y is the cumulative distribution of income and y^* is an arbitrary level of income at which the perturbation is performed. Note that this function has the property that

$$\int g(y) H'(y) f_y(y) dy = \frac{g(y^*) f_y(y^*)}{1 - F_y(y^*)}$$

where g is some function of y. The optimal tax reform is the T such that welfare cannot be improved by any perturbation ($\frac{\partial W}{\partial \epsilon} = 0$). For simplicity, assume that household utility is GHH, $u(c,h) = \frac{1}{1-\gamma}\left[c - v(h)\right]^{1-\gamma}$. First, compute $\frac{\partial W}{\partial \epsilon}$ by differentiating welfare with respect to a perturbation and set it to zero to get the optimality condition of the government,

$$\frac{\partial W}{\partial \epsilon} = 0 = \int \left[u_c(\theta) \frac{\partial c(\theta)}{\partial \epsilon} + u_h(\theta) \frac{\partial h(\theta)}{\partial \epsilon} \right] dF(\theta)$$

Use the GHH utility to simplify this to

$$0 = \int u_c(\theta) \left[\frac{\partial c(\theta)}{\partial \epsilon} - v'\left(h(\theta)\right) \frac{\partial h(\theta)}{\partial \epsilon} \right] dF(\theta) \tag{16}$$

Now I will consider two specifications for production that will have different implications for how general equilibrium effects show up in the optimal tax equation.

A.2.3 Standard Mirrlees

First, assume output is a CES function in efficiency units of labor $y(\theta) = \theta h(\theta)$,

$$Y(F) = \left(\int \left[y f_y(y) \right]^{\frac{\omega - 1}{\omega}} dy \right)^{\frac{\omega}{\omega - 1}}$$

where ω is the elasticity of substitution between different levels of efficiency units.

With this production function, household income is given by

$$I(h, \theta) = p\left(\theta h\right)$$

where $p(y)$ is the marginal product of households with efficiency units y

$$p(y) = \frac{\partial Y}{\partial [y f_y(y)]} = [y f_y(y)]^{\frac{\omega-1}{\omega}-1} \left[\int [y f_y(y)]^{\frac{\omega-1}{\omega}} dy\right]^{\frac{\omega}{\omega-1}-1} = \left[\frac{Y}{y f_y(y)}\right]^{\frac{1}{\omega}} \tag{17}$$

Given a tax function T, allocations (c, h) in this model are characterized by the following two equations,

$$c(\theta) = p\left(\theta h(\theta)\right) - T\left(p\left(\theta h(\theta)\right)\right) - \epsilon H\left(p\left(\theta h(\theta)\right)\right) \tag{18}$$

$$v'\left(h(\theta)\right) = \left[1 - T'\left(p\left(\theta h(\theta)\right)\right) - \epsilon H'\left(p\left(\theta h(\theta)\right)\right)\right] p'(\theta h(\theta))\theta \tag{19}$$

Consider a perturbation of the tax function: $T\left(p\left(\theta h(\theta)\right)\right) + \epsilon H\left(p\left(\theta h(\theta)\right)\right)$ so that the household's optimality conditions become

$$c(\theta) = p\left(\theta h(\theta)\right) - T\left(p\left(\theta h(\theta)\right)\right) - \epsilon H\left(p\left(\theta h(\theta)\right)\right)$$

$$v'\left(h(\theta)\right) = \left[1 - T'\left(p\left(\theta h(\theta)\right)\right) - \epsilon H'\left(p\left(\theta h(\theta)\right)\right)\right] p'(\theta h(\theta))\theta$$

Differentiating these equations with respect to ϵ gives

$$\frac{\partial c(\theta)}{\partial \epsilon} = \left[1 - T'\left(p\left(\theta h(\theta)\right)\right)\right] p'(\theta h(\theta))\theta \frac{\partial h(\theta)}{\partial \epsilon} - \frac{\partial T(0)}{\partial \epsilon} - H\left(p\left(\theta h(\theta)\right)\right)$$

$$\begin{aligned} v''\left(h(\theta)\right)\frac{\partial h(\theta)}{\partial \epsilon} = {} & T''\left(p\left(\theta h(\theta)\right)\right)\left[p'(\theta h(\theta))\theta\right]^2 \frac{\partial h(\theta)}{\partial \epsilon} + \left[1 - T'\left(p\left(\theta h(\theta)\right)\right)\right] p''(\theta h(\theta))\theta^2 \frac{\partial h(\theta)}{\partial \epsilon} \\ & - H'\left(p\left(\theta h(\theta)\right)\right) p'(\theta h(\theta))\theta \end{aligned}$$

These two equations, together with (17), (18) and (19) are five equations in five unknowns (given the tax function): $c(\theta), h(\theta), p\left(\theta h(\theta)\right), \frac{\partial c(\theta)}{\partial \epsilon}, \frac{\partial h(\theta)}{\partial \epsilon}$. Here, wages p are just a (possibly nonlinear) transformation of efficiency units $\theta h(\theta)$. This is the only way that production enters the government's problem. The production function will matter in the final optimal tax system, but it won't change the structure of the optimal tax formula: a different Y will just imply a different transformation of efficiency units into income.

A.2.4 Stiglitz (1982)

Now, assume output is a CES function in labor supply of each type, $h(\theta)$,

$$Y = \left(\int [h(\theta) f(\theta)]^{\frac{\omega-1}{\omega}} d\theta\right)^{\frac{\omega}{\omega-1}} \tag{20}$$

where ω is the elasticity of substitution between different types. The assumption here is that the government cannot observe types, but the firm can.

With this production function, household income is given by

$$I(h,\theta) = p(\theta)h(\theta)$$

where $p(\theta)$ is the marginal product of households of type θ.

$$p(\theta) = \frac{\partial Y}{\partial\left[h(\theta)f(\theta)\right]} = \left[h(\theta)f(\theta)\right]^{\frac{\omega-1}{\omega}-1}\left[\int\left[h(\theta)f(\theta)\right]^{\frac{\omega-1}{\omega}}d\theta\right]^{\frac{\omega}{\omega-1}-1} = \left[\frac{Y}{h(\theta)f(\theta)}\right]^{\frac{1}{\omega}} \tag{21}$$

Here, wages are high if the skill type is rare and $f(\theta)$ is small. Given a tax function T, allocations (c,h) in this model are characterized by the following two equations,

$$c(\theta) = p(\theta)h(\theta) - T\left(p(\theta)h(\theta)\right) - \epsilon H\left(p(\theta)h(\theta)\right) \tag{22}$$

$$v'\left(h(\theta)\right) = \left[1 - T'\left(p(\theta)h(\theta)\right) - \epsilon H'\left(p(\theta)h(\theta)\right)\right]p'(\theta)\theta \tag{23}$$

Consider a perturbation of the tax function: $T\left(p(\theta)h(\theta)\right) + \epsilon H\left(p(\theta)h(\theta)\right)$ so that the household's optimality conditions become

$$c(\theta) = p(\theta)h(\theta) - T\left(p(\theta)h(\theta)\right) - \epsilon H\left(p(\theta)h(\theta)\right)$$

$$v'\left(h(\theta)\right) = \left[1 - T'\left(p(\theta)h(\theta)\right) - \epsilon H'\left(p(\theta)h(\theta)\right)\right]p(\theta)$$

Differentiating these equations with respect to ϵ gives

$$\frac{\partial c(\theta)}{\partial\epsilon} = \left[1 - T'\left(p(\theta)h(\theta)\right)\right]p(\theta)\frac{\partial h(\theta)}{\partial\epsilon} + \frac{\partial p(\theta)}{\partial\epsilon}h(\theta) - \frac{\partial T(0)}{\partial\epsilon} - H(p(\theta)h(\theta))$$

$$\begin{aligned} v''\left(h(\theta)\right)\frac{\partial h(\theta)}{\partial\epsilon} = {} & T''\left(p(\theta)h(\theta)\right)p(\theta)^2\frac{\partial h(\theta)}{\partial\epsilon} + \left[1 - T'\left(p(\theta)h(\theta)\right)\right]\left[p(\theta)\theta\right]^2\frac{\partial h(\theta)}{\partial\epsilon} \\ & - H'\left(p(\theta)h(\theta)\right)p(\theta) + \left[1 - T'\left(p(\theta)h(\theta)\right)\right]\frac{\partial p(\theta)}{\partial\epsilon} \end{aligned}$$

These two equations, together with (21), (22) and (23) are five equations in *six* unknowns (given the tax function): $c(\theta), h(\theta), p(\theta), \frac{\partial c(\theta)}{\partial\epsilon}, \frac{\partial h(\theta)}{\partial\epsilon}, \frac{\partial p(\theta)}{\partial\epsilon}$. Here, wages p are not just a transformation of efficiency units, they are a completely different object that the government needs to compute the effect of taxes on, $\frac{\partial p(\theta)}{\partial\epsilon}$. This effect is given by differentiating the labor

market clearing equation (21),

$$\frac{\partial p(\theta)}{\partial \epsilon} = \frac{1}{\omega}\left[\frac{Y}{h(\theta)f(\theta)}\right]^{\frac{1}{\omega}-1}\left[\frac{1}{h(\theta)f(\theta)}\int p(\tilde{\theta})\frac{\partial h(\tilde{\theta})}{\partial \epsilon}dF(\tilde{\theta}) - \frac{Y}{h(\theta)^2 f(\theta)}\frac{\partial h(\theta)}{\partial \epsilon}\right] \tag{24}$$

This sixth equation completes the characterization of $(c(\theta), h(\theta), p(\theta), \frac{\partial c(\theta)}{\partial \epsilon}, \frac{\partial h(\theta)}{\partial \epsilon}, \frac{\partial p(\theta)}{\partial \epsilon})$ under a given tax function T. Here, the production function will explicitly create a separate term in the optimal tax formula representing the general equilibrium effects of a change in taxes. The key term here is the first term in the brackets, which represents the effect of the tax perturbation on all other skill types. This term implies that if the tax perturbation induces higher labor supply by other skill types $\tilde{\theta}$ (i.e. $\frac{\partial h(\tilde{\theta})}{\partial \epsilon} > 0$), it will increase the wage of type θ. This spillover effect was described by Stiglitz (1982, 1987). Notice that the benefit to the government of increasing households' labor supply is highest from the rarest skill types, who have the highest wages $p(\tilde{\theta})$. The government will take advantage of this by giving relatively low marginal tax rates to the households with the rarest skill types. This effect becomes stronger when the elasticity of substitution between skill types is low. In fact, it can be seen in equation (24) that the effect of a tax reform on wages is directly proportional to the inverse of the elasticity of substitution, ω. Importantly, the spillover effects from the first term in the brackets mean that $\frac{\partial p(\theta)}{\partial \epsilon}$ is not just a nonlinear transformation of type θ's efficiency units. In this setup, since households enter the production function by type instead of by efficiency units, there is no direct mapping between efficiency units and income.

If the government could observe skill types θ and give a different tax function to each type, these general equilibrium effects are still present. Even though a change in one type's tax function no longer affects the taxes paid by other types, it would still alter the wages of the other types. Equation (24) is the same in this case: a change in labor supply from one type changes prices for all other types. Therefore, these general equilibrium effects on the optimal tax system are not eliminated when the government can observe skill types as the firm can. Under the production function (20), general equilibrium effects enter the optimal tax system due to skill types being differentiated, not because the government cannot observe the types.

A.3 History-Dependence and Labor Supply: Two Period Example

In this section, I will demonstrate using a two period example why history-dependent taxation can increase total output even when household's have balanced growth preferences and labor supply is invariant to changing taxes on current income.

A worker lives for two periods, $t = 0, 1$, and solves

$$\max_{c_0, c_1, y_0, y_1} U = u(c_0, y_0) + u(c_1, y_1)$$

subject to

$$c_0 = (1 - \tau_0) y_0$$

$$c_1 = (1 - \tau_1) y_1 + \tau_{01} y_0$$

Here, τ_0 is the tax rate on income earned in the initial period, τ_1 is income earned in period 1 and τ_{01} is a tax paid in the first period that depends on income earned in the initial period. For simplicity, use quasi-linear utility,

$$u(c, y) = \log c - y$$

First order conditions of this problem imply

$$1 = (1 - \tau_0 + \tau_{01}) \frac{1}{c_0}$$

$$1 = (1 - \tau_1) \frac{1}{c_1}$$

Substituting for consumption c_0 and c_1 using the budget constraints gives expressions for output in each period,

$$y_0 = 1 + \frac{\tau_{01}}{1 - \tau_0} \text{ and } y_1 = 1 - \frac{\tau_{01}}{1 - \tau_1} \left(1 + \frac{\tau_{01}}{1 - \tau_0} \right)$$

Total output is then given by

$$Y = y_0 + y_1 = 2 + \frac{\tau_{01}}{1 - \tau_0} - \frac{\tau_{01}}{1 - \tau_1} \left(1 + \frac{\tau_{01}}{1 - \tau_0} \right) \tag{25}$$

First, consider a tax on only current income so that $\tau_{01} = 0$. In this case, output in each period is given by

$$y_0 = 1 + \frac{0}{1 - \tau_0} = 1 \text{ and } y_1 = 1 - \frac{0}{1 - \tau_1} \left(1 + \frac{0}{1 - \tau_0} \right) = 1$$

In this case, output is invariant to the tax system: changing τ_0 or τ_1 will have no effect on output. Alternatively, a history-dependent tax can increase output. From the expression for

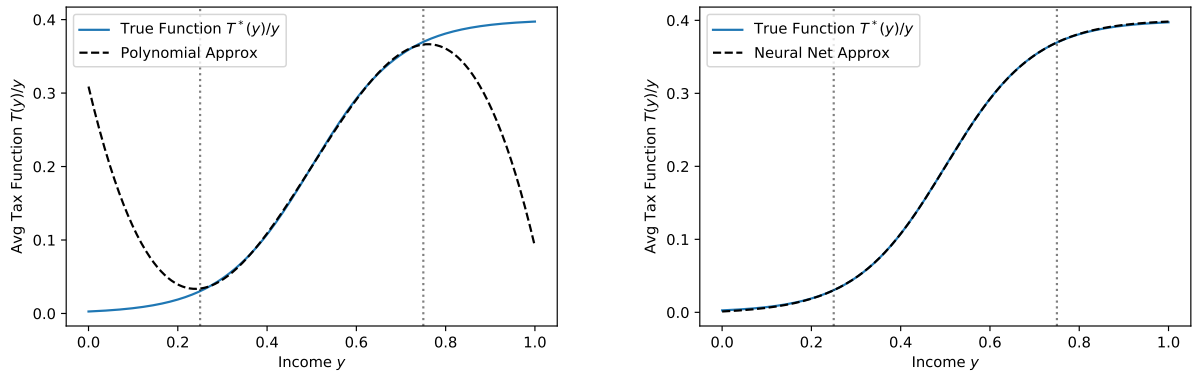
total output (25), this is true whenever

$$\frac{\tau_{01}}{1-\tau_1}\left(1+\frac{\tau_{01}}{1-\tau_0}\right) < \frac{\tau_{01}}{1-\tau_0}$$

The right-hand side of this inequality represents the incentive effects of τ_{01} on labor supply in period 0: τ_{01} increases labor supply through a substitution effect. The left-hand side of this inequality represents the disincentive effects of τ_{01} on labor supply in period 1: τ_{01} increases consumption and reduces labor supply through an income effect. Total output increases when the substitution effect of τ_{01} in the initial period is larger than the income effect in the second period. Notice that the income effect is larger when τ_1 is bigger, in which case τ_{01} becomes a larger fraction of period 1 consumption. Therefore, if the government wants to use a history-dependent tax to incentivize high output, they will set τ_{01} and τ_0 to be relatively large and set τ_1 to be relatively small.

A.4 Simple Example of Neural Network Approximation

In this section, I will demonstrate a simple example to show how neural networks outperform polynomial approximation with smaller amounts of data. Assume that I know that the optimal tax function is given by $T(y) = \frac{0.4y}{1+\exp\{-10(y+0.5)\}}$, which implies that average tax rates increase from zero at $y = 0$ to forty percent at $y = 1$. I will approximate this function first using a polynomial regression and then using a neural network on the same data between $y = 0.25$ and $y = 0.75$.



(a) Polynomial Approximation

(b) Neural Network Approximation

Figure 11: Simple Comparison of Polynomial and Neural Network Approximation

In Figure 11, I plot the results of a polynomial regression and neural network approximation of the example tax function. While the polynomial and neural network approximation

perform similarly in the region between the gray dotted lines where the estimation data was taken from, the neural network does a much better job of matching the function in the regions where it has not been shown any data. The polynomial approximation diverges from the true solution very quickly once it leaves the region where it has been given data. This happens because polynomials are unbounded functions, while neural networks are constructed with bounded functions like tanh or $\max\{0, x\}$. This means neural networks can extrapolate accurately to combinations of inputs that it has not observed, which allows neural networks to obtain accurate approximations with relatively small amounts of observations.

A.5 Tax Rates

Here I summarize the average and marginal tax rates under each of the different optimal tax functions

A.5.1 Taxes on Current Income Only

	Current Income, Thousands of Dollars					
	10	25	50	100	200	500
US System	5%	17%	25%	32%	39%	47%
Parametric	4%	22%	25%	34%	41%	49%
Nonparametric	3%	15%	24%	40%	40%	47%

Table 7: Marginal Tax Rates, Taxes on Current Income

A.5.2 Age-Dependent Taxes on Current Income

	Current Income, Thousands of Dollars					
Age	10	25	50	100	200	500
30	-21%	1%	15%	28%	38%	50%
45	10%	22%	29%	36%	43%	50%
60	26%	36%	43%	49%	55%	61%

Table 8: Marginal Tax Rates, Parametric Age-Dependent Tax

	Current Income, Thousands of Dollars					
Age	10	25	50	100	200	500
30	-20%	0%	16%	28%	39%	19%
45	10%	21%	29%	36%	42%	50%
60	26%	35%	42%	48%	53%	59%

Table 9: Marginal Tax Rates, Nonparametric Age-Dependent Tax

A.5.3 History-Dependent Taxes

	Age			
	30	40	50	60
Low Income History	-17%	11%	26%	33%
Middle Income History	-1%	20%	33%	39%
High Income History	30%	32%	19%	2%

Table 10: Average Tax Rates, History-Dependent Tax

	Age			
	30	40	50	60
Low Income History ($50K)	15%	25%	35%	46%
Middle Income History ($100K)	28%	34%	42%	51%
High Income History ($400K)	30%	37%	22%	7%

Table 11: Marginal Tax Rates, History-Dependent Tax

A.5.4 Parametric History-Dependent Taxes

	Current Income, Thousands of Dollars					
Age	10	25	50	100	200	500
30	-25%	-1%	14%	27%	38%	50%
45	13%	23%	30%	36%	42%	49%
60	33%	41%	47%	52%	57%	62%

Table 12: Marginal Tax Rates, Parametric History-Dependent Tax, $\bar{y} = \$50,000$

	Current Income, Thousands of Dollars					
Age	10	25	50	100	200	500
30	7%	16%	23%	29%	35%	42%
45	9%	20%	28%	35%	42%	49%
60	12%	25%	34%	42%	49%	57%

Table 13: Marginal Tax Rates, Parametric History-Dependent Tax, $\bar{y} = \$500,000$

A.6 Allocations Under Optimal Tax Functions

Here I show the levels of the allocations used to compute the percent changes in section 6.2.3.

	Present Value of Income Quartile				
	0-25%	25-50%	50-75%	75-100%	Total
Income ($PV(y)$)	37,991	58,902	82,549	149,060	82,124
Consumption ($PV(c)$)	36,466	51,568	67,279	106,729	65,510
Taxes ($PV(T)$)	1,525	7,333	15,267	42,315	16,610
Avg. Tax Rates ($PV(T/y)$)	-4.4%	3.9%	9.4%	17.6%	6.6%
Leisure ($PV(1-h)$)	0.648	0.643	0.639	0.634	0.641
Skills (s)	0.221	0.333	0.460	0.823	0.459
Skill Price ($p(s)$)	0.179	0.197	0.218	0.279	0.218

Table 14: Average Allocations by Quartile of PV Income Distribution, AD Tax

	Quartile, Present Value of Income				
	0-25%	25-50%	50-75%	75-100%	Total
Income ($PV(y)$)	37,679	58,573	82,407	156,819	83,868
Consumption ($PV(c)$)	36,588	51,518	67,124	114,281	67,377
Taxes ($PV(T)$)	1,164	7,117	15,288	42,588	16,539
Avg. Tax Rates ($PV(T/y)$)	-5.6%	3.1%	9.0%	17.4%	6.0%
Leisure ($PV(l)$)	0.650	0.646	0.642	0.626	0.641
Skills (s)	0.216	0.331	0.463	0.837	0.462
Skill Price ($p(s)$)	0.178	0.198	0.219	0.284	0.220

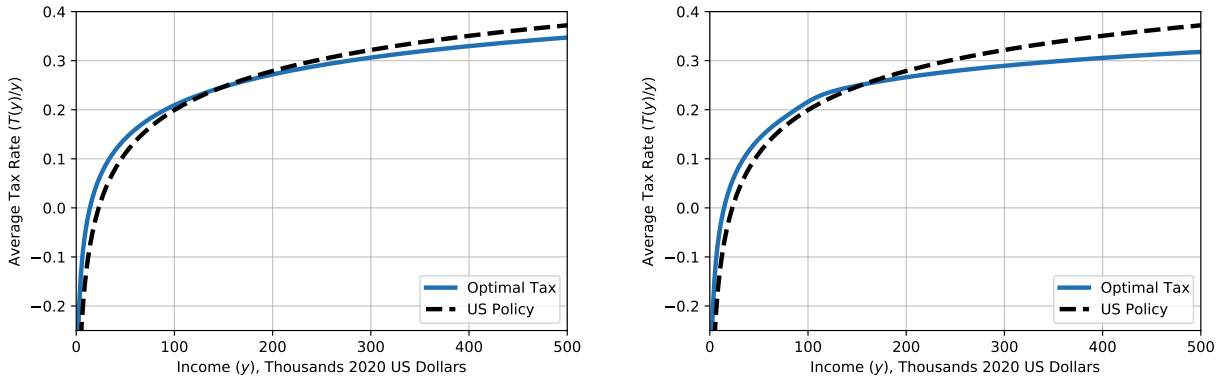
Table 15: Average Allocations by Quartile of PV Income Distribution, HD Tax

A.7 Results With Separable Utility

In this section, I consider an alternate utility function where utility from consumption and hours worked are separate functions. As with the baseline utility function, I calibrate the process for $\varphi \sim N(m_\varphi, v_\varphi)$ to match the average and variance of log hours worked in the PSID.

$$u_i(c,h) = \log c - \exp\varphi_i \frac{h^{1+\frac{1}{\nu}}}{1+\frac{1}{\nu}}$$

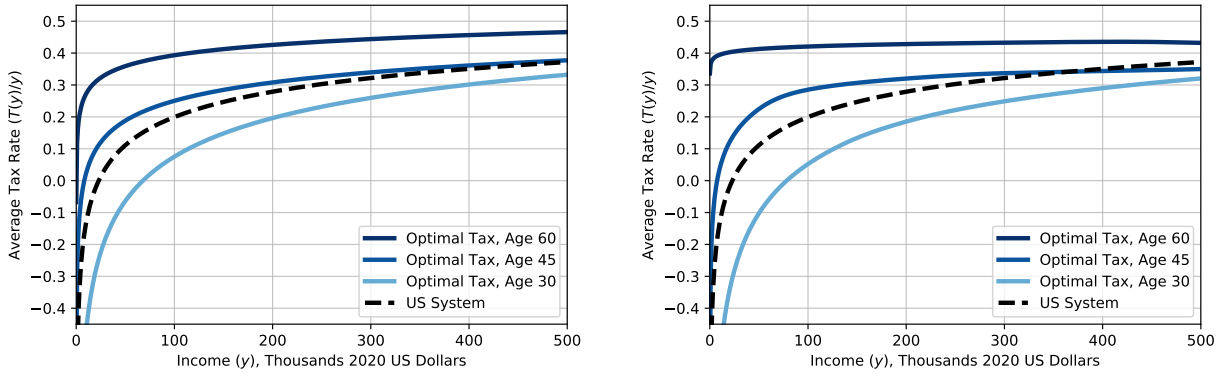
A.7.1 Optimal Taxes on Current Income Only



(a) Optimal Parametric Tax, $T(y) = y - (1-\tau)y^\rho$ (b) Optimal Nonparametric Tax, $T(y)$

Figure 12: Optimal Tax on Current Income, Separable Utility

A.7.2 Age-Dependent Taxes on Current Income Only



(a) Optimal Parametric Tax, $T_a(y) = y - (1 - \tau(a))y^{\rho(a)}$

(b) Optimal Nonparametric Tax, $T_a(y) = \tau_a(y)y$

Figure 13: Optimal Age-Dependent Tax on Current Income, Separable Utility

A.7.3 History-Dependent

Tax Rates Here are tables summarizing optimal history-dependent average and marginal tax rates for low, middle and high income history when utility is separable.

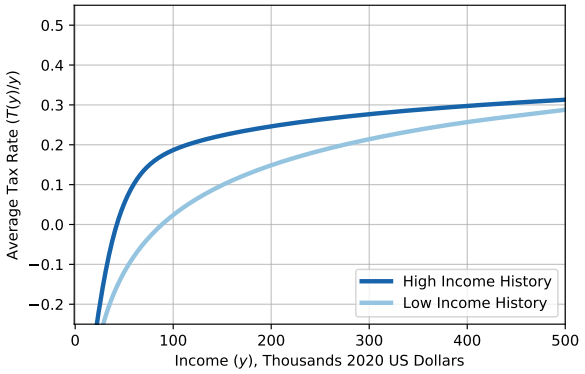
	Age			
	30	40	50	60
Low Income History ($50K)	-25%	30%	35%	49%
Middle Income History ($100K)	-5%	21%	38%	49%
High Income History ($400K)	29%	33%	20%	9%

Table 16: Average Tax Rates, History-Dependent Tax, Separable Utility

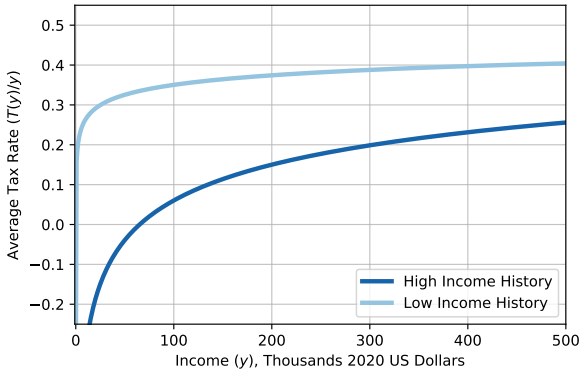
	Age			
	30	40	50	60
Low Income History ($50K)	6%	36%	39%	49%
Middle Income History ($100K)	20%	32%	41%	49%
High Income History ($400K)	37%	40%	32%	26%

Table 17: Marginal Tax Rates, History-Dependent Tax, Separable Utility

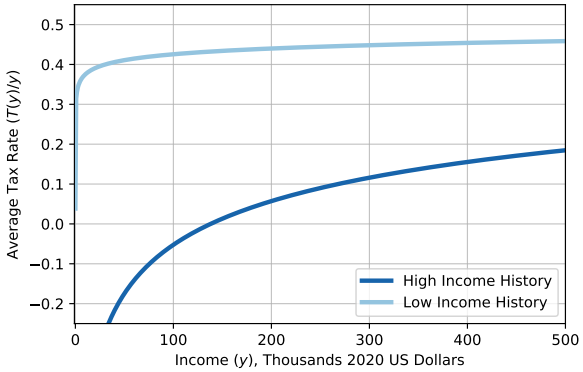
Tax Schedules These are the optimal history-dependent tax schedules for low and high income history when utility is separable.



(a) Optimal Tax After 10 Years



(b) Optimal Tax After 20 Years



(c) Optimal Tax After 30 Years

Figure 14: Optimal Non-parametric History-Dependent Tax, Separable Utility

A.7.4 Welfare Comparison

Here are the consumption equivalent welfare gains of the different tax function when utility is separable.

	History of Income	Age and Current Income	Only Current Income
Nonparametric	$T(y;\{y_t\}_{t=0}^{a-1},a)$	$T(y;a)$	$T(y)$
Parametric		$y-(1-\tau(a))y^{\rho(a)}$	$y-(1-\tau)y^{\rho}$

Table 18: Summary of Tax Functions

	Income History	Age and Current Income	Only Current Income
Nonparametric	0.0%	1.55%	5.20%
Parametric		1.65%	5.33%

Table 19: Welfare Gain of Moving to Nonparametric, History Dependent Taxes