Seeking the Rent Seekers: Optimal Life Cycle Taxation with Differentiated Human Capital

Ross M. Batzer

May 2026

Abstract

This paper studies optimal dynamic income taxation in a life cycle model with differentiated skills. Imperfect substitution in worker types gives rise to spillover effects in general equilibrium wages from higher aggregate output. A novel Monte Carlo method is developed using multilayered neural networks to compute optimal history-dependent income taxes within a very general class of tax functions. The welfare gains from history-dependent taxation are found to be large, equivalent to a 2 percent increase in lifetime consumption compared to the optimal tax on current income. The welfare gains are found to be close to zero without these features.

Keywords: optimal taxation · deep learning · tax progressivity · labor supply · skill investment · income inequality · welfare

JEL Classification: C45 · C68 · E62 · H20 · J42

For helpful discussions and suggestions on previous versions of this paper, I want to thank Aniket Baksy, Serdar Birinci, Daniele Caratelli, Salim Syed Chanto, V.V. Chari, Kadidja Doucouré, Eugenia Gonzalez-Aguado, Jonathan Heathcote, Amy Handlan, William Jungerman, Patrick Kehoe, Kaja Kierulf, Frederic Martenet, Abdoulaye Ndiaye, Elena Pastorino, Chris Phelan, Sandra Segovia, Martin Souchier, and Kjetil Storesletten, as well as seminar participants from the Bank of Canada, FHFA, Institute for Defense Analyses, and University of Minnesota. This paper is heavily revised and condensed from its previous version titled "Optimal History-Dependent Income Taxation in General Equilibrium". Please address correspondence to rmbatzer@gmail.com.

1 Introduction

A central concern in public finance is how to design a tax and transfer system that provides insurance while minimally distorting labor supply and investment. To this end, a large literature has sought to characterize the the maximally efficient feasible allocation of resources in both static and dynamic models of labor supply. Within dynamic settings, the implementation of constrained efficient allocations can be very complex, depending nonlinearly on an individual's entire history of earnings. However, a standard finding is that indexing tax rates by age alone captures the large majority of potential welfare gains from fully optimal history-dependent policies (Stantcheva, 2020).

This paper revisits optimal life cycle income taxation in an environment with differentiated human capital. Imperfectly substitutable labor has a long been used to model wage inequality, being with Katz and Murphy (1992).[1] Unlike in economies where different types of workers are perfect substitutes in production, higher output produces spillover effects in general equilibrium wages when workers are imperfectly substitutable. Since workers with rarer skill types have higher marginal productivity, the government wants to incentivize labor supply from workers with rare skills 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.

In this context, policymakers want to identify which workers have high marginal productivity so they can extract more output from them. Income history is a powerful tool to screen which workers have the rarest skills. Unlike with history-independent tax schemes, high productivity workers cannot hide their current productivity from the planner by working fewer hours if the planner has already observed the worker earning large amounts

---

[1] see also Acemoglu (2002), Acemoglu and Autor (2011), Heathcote et al. (2017, 2020)

in previous periods. While this provides little benefit in economies with simplified wage dynamics, this paper demonstrates that quantitatively realistic wage dynamics –with differentiated human capital investment and transitory shocks– give rise to potentially large welfare effects from history-dependent policy. However, computing the optimal taxation problem to quantitatively measure these gains is intractable with standard approaches.

Existing studies of optimal dynamic income taxation largely use one of two approaches. The first, the dynamic Mirrlees approach, characterizes the optimal taxation problem as a constrained planning problem subject to the planner being unable to directly observe workers' heterogeneous labor productivities. This approach puts no restrictions on the tax instruments available to the government, allowing for the computation of fully non-linear and history-dependent optimal taxes. The main drawback of this approach is that it is unable to accommodate many quantitatively realistic features of life cycle wages, including multiple wage components and general equilibrium effects. The second, referred to as the parametric and quantitative Ramsey approach by Stantcheva (2020), restricts tax instruments to simple parametric forms, e.g., log-linear. These parametric restrictions accommodate the computation of optimal taxes in quantitatively complex environments featuring credit constraints, incomplete markets, or human capital investments. However, this approach becomes exponentially time consuming to compute with additional complexity in the chosen parametric forms for taxes.

Computing optimal history-dependent income taxes in an economy with differentiated skills presents a challenge because it is difficult to allow for history-dependent taxes with simple parametric tax instruments. To address this computational hurdle, this paper develops a substantially more flexible variant of the parametric approach to optimal taxation by using multilayered neural networks to parameterize optimal income taxes. This allows optimal taxes to be any continuous function of observable information available to the government. Neural networks are able to approximate high-dimensional, nonlinear functions with much fewer parameters than standard methods like perturbation or polynomial projection. Also, because of recent advancements in computing and numerical libraries, neural networks can be estimated very quickly. Using this approach, agents' choices and optimal taxes can be computed simultaneously as the solution to a constrained planning problem where choices and taxes are both computed as separate neural networks. The planner's problem is solved by Monte Carlo simulation of workers' entire life cycle profiles. This approach makes it feasible to compute nonlinear optimal taxes over the life cycle in a quantitatively rich model of life cycle labor supply with human capital investment, incomplete markets, transitory wage shocks, and general equilibrium effects in wages.

This approach is applied to compute optimal history-dependent income taxes in a quantitatively complex model of life cycle labor supply based on Heathcote, Storesletten and Violante (2020). Unlike in more stylized models, the benefits of history-dependent taxation are found to be potentially large, equivalent to a 2.1 percent increase in lifetime consumption. These benefits are shown to come from three features in the model: transitory wage shocks, endogenous human capital investment, and general equilibrium effects in wages arising from differentiated human capital. Transitory wage shocks make it more difficult for the planner to screen workers for their permanent labor productivity using current income alone. Endogenous human capital investment makes it more valuable to screen for worker's types because the government can use the information to extract more revenue. Further, imperfectly substitutable human capital in production makes the value of learning worker's types even more valuable to the planner since higher output increases general equilibrium wages in addition to effects on revenue. Each of these features is found to contribute about one third of the total potential welfare gain from history-dependent taxation. In a version of the model with these features turned off, the potential welfare benefit of history-dependent taxation is found to be close to zero, as found in previous studies (Stantcheva, 2020).

Related Literature Since Mirrlees (1971), a large literature has studied how the government should use observable information to redistribute and pay for public goods without overly distorting output. More recently, a literature starting with Golosov, Kocherlakota and Tsyvinski (2003) has studied optimal taxation in dynamic economies where individuals' wages change over their life cycles, including Weinzeirl (2011), Farhi and Werning (2013), Golosov, Troshkin and Tsyvinski (2016), Stantcheva (2017), and Ndiaye and Yu (2025). This line of work has found that dynamic optimal income taxes are in general very complex and depend on a taxpayer's detailed income history. However, a common finding in this literature is that the potential welfare benefits from achieving the fully optimal constrained efficient allocations are quantitatively small compared to simple history-independent tax instruments, with linear age-dependent taxes reaping the large majority of potential gains.

While the dynamic Mirrlees approach to optimal taxation is able to characterize fully optimal constrained efficient allocations, it is very restrictive on the types of economies it can consider. Because of this, another line of literature places parametric restrictions on tax instruments and quantitatively assesses optimal policies. Imposing parametric restrictions allows this literature to consider more complex and realistic economies, including those with incomplete markets (Conesa and Krueger (2006), Conesa, Kitao and Krueger (2009), Karabarbounis (2016), Peterman and Sager (2022), and many others), human capital investment (Krueger and Ludwig (2013, 2016), Peterman (2016), and others), and general equilibrium effects in the distribution of wages (Heathcote, Storesletten and Violante (2017, 2020)).[2]

This paper partially bridges the gap between these two literatures by using neural networks to parameterize the tax system. While this approach is not able to compute the true constrained efficient allocations, it allows for substantially more flexibility than standard

---

[2] In related work, Kapička (2020) uses the parametric approach to study history-dependent taxation in an environment without savings or human capital investment, where optimal taxes can be characterized analytically. However, Kapička (2020) studies an environment with infinitely lived agents, so the results are not easily comparable to this paper or other studies of optimal life cycle taxation.

parametric approaches.[3] This paper adds to a growing body of work on applying methods involving neural networks to compute structural models that are infeasible to compute with other functional approximation methods, including Maliar, Maliar and Winant (2021), Azinovic, Gaegauf and Scheidegger (2022), Kahou, Fernández-Villaverde, Perla and Sood (2022), Fernández-Villaverde, Hurtado and Nuño (2023), Jungerman (2023), Adenbaum, Babalievsky and Jungerman (2024), Duarte, Duarte and Silva (2024), Kase, Melosi and Rottner (2025), Payne, Rebei and Yang (2025), and Han, Yang and E (2026), among others.

The rest of this paper proceeds as follows: Section 2 demonstrates how differentiated skill types can potentially lead to welfare gains from history-dependent taxation in a simple two period model. Section 3 describes the quantitative model used for the main analysis. Section 4 develops a solution method using multilayered neural networks to compute optimal history-dependent taxes in the model. Section 5 explains how the model is parameterized to fit data on life cycle income, labor supply, and earnings, and then summarizes the main results from computing optimal taxes in the calibrated model. Finally, Section 6 concludes.

2 Two Period Example

This section presents a simple two period model to demonstrate how history dependence in income taxation can generate higher aggregate output, even with balanced growth preferences where changes in history-independent taxes have no effect on hours worked. Assume there two types of workers, $i = 1, 2$, who each live for two periods, $a = 1, 2$. Given wages $\{p_i\}_{i=1,2}$, each type i chooses their efficiency units $\{n_{ia}\}_{a=1,2}$ to maximize their lifetime utility,

$$\max_{n_{i1},n_{i2}} u(c_{i1}, n_{i1}) + u(c_{i2}, n_{i2})$$

---

[3] Appendix B demonstrates that the neural network approach produces virtually identical allocations to the Mirrlees planning problem in a simple static economy.

subject to their budget constraints,

$$
\begin{aligned}
c_{i1} &= (1-\tau_1)p_i n_{i1} \\
c_{i2} &= (1-\tau_2)p_i n_{i2} + \lambda p_i n_{i1},
\end{aligned}
$$

where τ_1 and τ_2 are linear tax rates on income received in each period. Taxes are made history-dependent by giving each worker a subsidy in their second period of life proportional to their output in the first period equal to $\lambda p_i n_{i1}$. If quasi-linear utility is assumed, $u(c,n) = \log c - n$, then the optimal level of efficiency units supplied by workers are given by:

$$n_{i1} = 1 + \frac{\lambda}{1-\tau_1} \text{ and } n_{i2} = 1 - \frac{\lambda}{1-\tau_2}\left(1 + \frac{\lambda}{1-\tau_1}\right)$$

Without history-dependent taxes ($\lambda = 0$), the level of efficiency units supplied by each type of worker is invariant to the choice of tax rates. However, with history dependence ($\lambda \neq 0$), total output is larger than without history dependence as long as

$$\frac{\lambda}{1-\tau_2}\left(1 + \frac{\lambda}{1-\tau_1}\right) < \frac{\lambda}{1-\tau_1}$$

The right-hand side of this inequality represents the incentive effects of history dependence on labor supply in period 1: λ increases labor supply through a substitution effect. The left-hand side of this inequality represents the disincentive effects of history dependence on labor supply in period 1: λ increases consumption and reduces labor supply through an income effect. Total output increases when the substitution effect of history dependence in the initial period is larger than the income effect in the second period. Therefore, if a planner wants to use a history-dependent tax to incentivize high output, they will set λ and τ_1 to be relatively large and set τ_2 to be relatively small.

While higher output produces higher revenue, the additional negative distortions on labor supply may outweigh any benefits from history-dependent tax schemes. However, there are good reasons to believe the benefits from may be large enough to generate positive net gains in welfare. To illustrate this, consider two specifications for aggregate production that lead to different wage structures. Denote the total efficiency units of labor supplied by each type of worker as $N_i = n_{i1} + n_{i2}$. First, suppose the efficiency units of each type of worker are perfectly substitutable in production,

$$Y = N_1 + N_2.$$

In this case, wages are constant ($p_j = 1$), so history-dependent tax schemes will have no effect on prices. Alternatively, suppose that skill types are imperfectly substitutable so that total output is given by

$$Y = \left(N_1^{\frac{\theta-1}{\theta}} + N_2^{\frac{\theta-1}{\theta}}\right)^{\frac{\theta}{\theta-1}},$$

where $\theta > 1$ is the elasticity of substitution between skill types. This specification for aggregate production has long been very commonly used in the literature on wage inequality. Importantly, under this specification, wages for each type of worker are increasing in total output: $p_j = \partial Y/\partial N_j = (Y/N_j)^{1/\theta}$. Therefore, with imperfect substitution of worker types in production, higher output increases wages for all workers symmetrically. In this economy, a benevolent social planner can potentially achieve welfare gains from increasing output via history-dependent tax schemes. The rest of this paper explores the quantitative significance of these potential welfare gains in a full life cycle model calibrated to generate realistic life cycle profiles for wages, consumption and labor supply.

3 Model

A similar economy as the one studied in Heathcote et al. (2020) is used to model life cycle labor supply and wage inequality. However, unlike Heathcote et al. (2020), taxes are allowed to be any continuous function and individuals are allowed to save in a risk-free bond to provide private insurance in addition to insurance provided through the tax system. In Section **??**, this modified model is calibrated to match the same life cycle moments for consumption, wages, and hours worked as Heathcote et al. (2020).

3.1 Economic Environment

Demographics Individuals enter the economy at age $a = 0$ and live for A periods. The total population of workers is of mass one, and therefore each cohort is of mass $1/A$. There are no inter-generational links. Individuals are indexed by $i \in [0,1]$.

Life Cycle Upon entering the economy at age $a = 0$, workers have a chance to invest in skills, s_i. Once the worker has chosen s_i, they enter the labor market. During the labor market, the worker provides $h_i \geq 0$ hours of labor supply and consumes a private consumption good c_i. Each period, they face stochastic fluctuations in idiosyncratic labor productivity ω_i. Labor productivity drops to zero at age $A^* < A - 1$, at which point workers consume their savings.[4]

Preferences Expected lifetime utility from consumption c, hours worked h, and skill investment s for an 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

---

[4] For simplicity, it is assumed there are no public pensions, but workers can save their after-tax earnings and face no additional taxes on the accrued returns. Essentially, in place of social security, workers are given Roth individual retirement accounts with no limit on their annual contributions or penalty on early withdrawals. This simplification can be made because in this environment —without hyperbolic discounting or health shocks in retirement —there is no benefit from public pensions that cannot be achieved through optimaly chosen life cycle income taxes.

$s_i \geq 0$ takes the form

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

where the parameter $\psi \geq 0$ controls the elasticity of skill investment with respect to the marginal return to skills, and $\kappa_i \geq 0$ is an individual-specific parameter that determines the utility cost of acquiring skills. This cost is distributed according to a standard exponential distribution, $\kappa_i \sim Exp(1)$. Skill investment decisions are irreversible, so skills are fixed throughout the life cycle.

As in Heathcote et al. (2020), workers have separable balanced growth preferences,

$$u_i(c_{ia}, h_{ia}) = \log c_{ia} - \exp \phi_{ia} \frac{h_{ia}^{1+1/\sigma}}{1+1/\sigma}$$

The term ϕ_{ia} scales the disutility of supplying labor. This term is the sum of a normally distributed individual fixed effect and an age varying effect shared by all workers: $\phi_{ia} = \bar{\phi}_i + \phi(a)$, $\bar{\phi}_i \sim N(0, v_\phi)$. The fixed effect is used to match the variance of hours worked in the data. This is necessary because variation in wages alone account for less than half of the variation in observed hours worked.

Technology Output is a constant elasticity of substitution aggregate of effective hours supplied by the continuum of skill types $s \geq 0$:

$$Y = Z \left(\int_0^\infty [N(s)m(s)]^{\frac{\theta-1}{\theta}} ds \right)^{\frac{\theta}{\theta-1}},$$

where Z is the (constant) aggregate productivity, $\theta > 1$ is the elasticity of substitution across skill types, $m(s)$ is the density over skill types, and $N(s)$ is the aggregate demand for efficiency units of skill type s.[5] In equilibrium this is equal to the efficiency units supplied by workers. Note that all skill levels enter symmetrically in the production technology, and thus

---

[5] The value of Z is arbitrary and is set to normalize the average value of s to be equal to one in the calibration.

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

$$p(s) = Z^{(\theta-1)/\theta} \left[\frac{Y}{N(s)m(s)} \right]^{1/\theta} \tag{1}$$

Here, each workers' wage $p(s)$ is decreasing in the relative measure of workers with the same skill type $m(s)$ and increasing in total output Y. The fact that prices for all skill types are linked to total output turns out to be crucial in 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 productivity ω_{ia} is the sum of three orthogonal components

$$\omega_{ia} = x(a) + \alpha_{ia} + \varepsilon_{ia}$$

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

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

with innovations that are distributed i.i.d. normal $\eta_{ia} \sim N(0, v_\eta)$ and have an initial value of $\alpha_{i0} = \eta_{i0}$. 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:

$$y_{ia} = \underbrace{p(s_i)}_{\text{skill price}} \times \underbrace{\exp\{\omega_{ia}\}}_{\text{labor productivity}} \times \underbrace{h_{ia}}_{\text{hours worked}}$$

The first component $p(s_i)$ is the equilibrium price for the type of labor supplied by a worker with skills s_i. The second component is the labor productivity. Finally, the third component h_{ia} is the 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 life cycle, (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.

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 savings, b. The stock of assets earns a market return $r = 1/\beta - 1$. Households are assumed to 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

$$G = gY \leq \sum_{a=0}^{A-1} \int T\left(a, y_{ia}; \{y_{it}\}_{t=0}^{a-1}\right) di \tag{2}$$

Here, $T\left(a, y; \{y_t\}_{t=0}^{a-1}\right)$ is the net tax owed by an individual worker of age a with current income y and a previous history of incomes $\{y_t\}_{t=0}^{a-1}$.[6]

Individual Worker Problem The vector of state variables for an individual household at age a is $(b_{ia}, \alpha_{ia}, \varepsilon_{ia}, s_i, \phi_i, a, \{y_t\}_{t=0}^{a-1})$. 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,

$$\max_{s_i\{c_{ia}, h_{ia}, b_{ia+1}\}_{a=0}^{A-1}} -v_i(s_i) + E_0\left[\left(\frac{1-\beta}{1-\beta^A}\right)\sum_{a=0}^{A-1}\beta^a u_i(c_{ia}, h_{ia})\right] \tag{3}$$

subject to their budget constraint,

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

and non-negativity constraints on their choices,

$$s_i, c_{ia}, b_{ia+1} \geq 0,\ 0 \leq h_{ia} \leq 1.$$

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

1. Households solve their problem (3)
2. The firm maximizes profits so that the skill price schedule satisfies equation (1)
3. The governments' budget constraint (2) is satisfied

---

[6] While the computational method developed in Section 4 is able to accommodate taxes that depend on the entire history of income, the main results use only an equally weighted lifetime average of log-income. This is because more sophisticated tax functions were found to produce no observable benefits compared to a simple average.

4. Markets clear for goods,

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

and labor,

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

3.2 Optimal Taxation

Social Welfare Function I follow standard practice in the parametric optimal life cycle taxation literature and define social welfare as the ex-ante (before ability is realized) expected (with respect to uninsurable productivity shocks) lifetime utility of a newborn in a stationary equilibrium.[7] The government's social welfare function is therefore given by

$$SWF=E_{-1}[U]=\int\left\{-v_i(s_i)+E_0\left[\left(\frac{1-\beta}{1-\beta^A}\right)\sum_{a=0}^{A-1}\beta^a u_i(c_{ia},h_{ia})\right]\right\}di. \tag{4}$$

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 policymaker has to trade off the benefits of redistribution and funding public goods against potential distortions on output.

Optimal Tax Problem The government chooses its tax function T to maximize the social welfare function (4) subject to allocations and prices being a stationary competitive equilibrium given the tax function. Therefore, the governments' budget constraint

---

[7] e.g., Conesa and Krueger (2006), Conesa, Kitao and Krueger (2009), Karabarbounis (2016), and Peterman and Sager (2022)

(2), individual workers' choices (3), and profit maximization by the firm (1) constitute implementability constraints on the government's planning problem.

4 Solution Method

This section describes how neural networks are used to approximate the model described in Section 3 and compute optimal income taxes. This involves an iterative process where an income tax function represented as a neural network is chosen to maximize social welfare, which is derived from household choices that are also represented as separate neural networks. The weights of the tax function are changed until a perturbation of any weight cannot produce an increase in social welfare.[8]

4.1 Approximating Optimal Taxes

The process for computing optimal tax functions is conceptually similar to the variational method described by Saez (2001) and used more recently by Findeisen and Sachs (2017), Saez and Stantcheva (2018), Sachs, Tsyvinski and Werquin (2020), and Chang and Park (2021). The optimal tax function is computed by finding the function for which an arbitrary perturbation cannot improve welfare. A neural network is used to perform this process automatically without needing to take derivatives by hand. The tax function is estimated so that changing any of the weights does not alter equilibrium allocations or prices in a way that increases social welfare.

The neural networks are estimated through Monte Carlo simulation, where full life-cycle profiles of idiosyncratic shocks are drawn for a large number of households. These households are divided into smaller groups called *batches* and the social planner's objective function is computed for each batch. The gradient of the social welfare function with respect to network weights is taken iteratively and used to update the weights of the networks using a gradient

---

[8] Additional details on the estimation of neural networks and how they compare to standard polynomial approximation are available in Appendix A.

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 computation slower and make it more difficult to observe groups households for which the planner can use the tax system to improve welfare. It is not necessary that the social welfare function and its gradient are close to their true values on a given iteration, they just must be correct in expectation. Networks are estimated using millions of iterations and an individual iteration will have very little effect on the final values of weights.

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

1. Setup neural networks for allocations, prices, and taxes:
    (a) Network with weights w_s for the skill investment choice: $\mathcal{S}(\kappa; w_s)$
    (b) Network with weights w_c for hours and savings: $(\mathcal{H},\mathcal{B})(s,b,z,\varepsilon,\phi,a,y^{a-1};w_c)$
    (c) Network with weights w_p for skill prices: $\mathcal{P}(s;w_p)$
    (d) Network with weights w_T for taxes: $\mathcal{T}(a,y,y^{a-1};w_T)$
2. Given weights, use the networks to compute allocations for each individual

$$\begin{aligned}
s_i(\mathbf{w}) &= \mathcal{S}(\kappa_i; w_s) \\
b_{ia+1}(\mathbf{w}) &= \mathcal{B}_{ia}(s_i(\mathbf{w}), b_{ia}(\mathbf{w}), \alpha_{ia}, \varepsilon_{ia}, \phi_i, a, y_i^{a-1}(\mathbf{w}); w_c) \\
h_{ia}(\mathbf{w}) &= \mathcal{H}(s_i(\mathbf{w}), b_{ia}(\mathbf{w}), \alpha_{ia}, \varepsilon_{ia}, \phi_i, a, y_i^{a-1}(\mathbf{w}); w_c) \\
y_{ia}(\mathbf{w}) &= \mathcal{P}\left(s_i(\mathbf{w}); w_p\right) \exp(x(a) + \alpha_{ia} + \varepsilon_{ia}) h_{ia}(\mathbf{w}) \\
c_{ia}(\mathbf{w}) &= (1+r) b_{ia}(\mathbf{w}) + y_{ia}(\mathbf{w}) - \mathcal{T}(a, y_{ia}(\mathbf{w}), y_i^{a-1}(\mathbf{w}); w_T) - b_{ia+1}(\mathbf{w})
\end{aligned}$$

    where $\mathbf{w} \equiv (w_s, w_c, w_p, w_T)$. Note that, even though allocations and prices are computed from separate networks, they depend on each other and the tax system through the evolution of endogenous state variables.

3. For n individuals, draw a disutility parameter for skill investment: $\{\kappa_i\}_{i=1}^n$

4. For each individual, simulate two life cycle profiles of wage shocks $\left\{\phi_i^j, \{\eta_{ia}^j, \varepsilon_{ia}^j\}_{a=0}^{A-1}\right\}_{i=1}^n$

5. Use the allocations to compute aggregate allocations, including

   (a) Government revenue for each draw of life cycle wage profiles:

   $$g^j(\mathbf{w}) = \left(\frac{1-\beta}{1-\beta^A}\right)\sum_{i=1}^n \beta^a \left(\frac{y_{ia}^j(\mathbf{w}) - c_{ia}^j(\mathbf{w})}{Y(\mathbf{w})}\right)$$

   (b) Output for each skill type:

   $$N^j(s_i(\mathbf{w})) = \sum_{a=0}^{A-1} \exp(x(a) + \alpha_{ia} + \varepsilon_{ia}) h_{ia}(\mathbf{w})$$
   $$\log p^j(s_i(\mathbf{w})) = \frac{1}{\theta}\log\left(\frac{Y(\mathbf{w})}{N^j(s_i(\mathbf{w}))m(s_i(\mathbf{w}))}\right)$$

   where $m(\mathcal{S}(\kappa_i)) = \exp(-\kappa_i)/\mathcal{S}'(\kappa_i)$

   (c) Lifetime utility for each individual and wage profile:

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

   (d) Social welfare within the sample:

   $$SWF(\mathbf{w}) = \sum_{i=1}^n \log\left(\exp U_i^1(\mathbf{w}) \times \exp U_i^2(\mathbf{w})\right)$$

6. Construct implementability constraints for the planner's objective function using first

order conditions for individual choices, prices, and the government budget constraint:[9]

$$\begin{aligned}
IC_s(\mathbf{w}) &= \sum_{i=1}^{n} \frac{\partial U_i^1(\mathbf{w})}{\partial s_i} \times \frac{\partial U_i^2(\mathbf{w})}{\partial s_i} \\
IC_c(\mathbf{w}) &= \frac{\partial E_i U_i^1(\mathbf{w})}{\partial w_c} \times \frac{\partial E_i U_i^2(\mathbf{w})}{\partial w_c} \\
IC_p(\mathbf{w}) &= \Big(\log p^1(s_i(\mathbf{w})) - \log \mathcal{P}\left(s_i(\mathbf{w}); w_p\right)\Big) \times \Big(\log p^2(s_i(\mathbf{w})) - \log \mathcal{P}\left(s_i(\mathbf{w}); w_p\right)\Big) \\
GBC(\mathbf{w}) &= (1 - g^1(\mathbf{w})/g) \times (1 - g^2(\mathbf{w})/g)
\end{aligned}$$

7. Calculate the social planner's loss function,

$$L(\mathbf{w}) = IC_s(\mathbf{w}) + IC_c(\mathbf{w}) + IC_p(\mathbf{w}) + GBC(\mathbf{w}) - SWF(\mathbf{w})$$

8. Compute gradient of the loss function with respect to every weight of each network,

$$\nabla L(\mathbf{w}) = \left(\frac{\partial L(x+\delta)}{\partial \delta}\right)_{x \in \mathbf{w}}$$

9. If social welfare doesn't improve after many iterations, i.e., $\nabla L(\mathbf{w}) \approx 0$, then stop. Otherwise go back to step 2.

5 Quantitative Analysis

This section describes how the model in Section 3 is calibrated, and then summarizes optimal life cycle taxation in the calibrated model. Further, it is explored how the optimal tax system changes as key features of the model are removed.

---

[9] The specification of the implementability constraints as the product of two residuals of random realizations of first order conditions comes from applying the all-in-one expectations operator introduced by Maliar et al. (2021).

Table 1: Summary of Fixed Parameters

(a) Uncalibrated Parameters

	Description	Value	Source
β	Discount Factor	0.97	Heathcote et al. (2020)
A^*	Years of working life	36	–
σ	Inverse Frisch elasticity	0.5	–
ψ	Elasticity of skill investment	0.65	–
g	Government spending (% of output)	0.189	–
x_1	Linear life cycle wage profile	0.059	–
x_2	Quadratic life cycle wage profile	-1.25×10^{-3}	–
x_3	Cubic life cycle wage profile	5.64×10^{-6}	–
θ	Elasticity of substitution across skills	2.0	Saez (2001)

(b) Calibrated Parameters

	Description	Value	Target
ϕ_0	Constant labor disutilty	0.275	$\text{Avg}(h) = 1/3$
ϕ_1	Linear labor disutilty	-5.46×10^{-3}	$\text{Avg}(h_{15})/\text{Avg}(h_0) = 1.08$
ϕ_2	Quadratic labor disutilty	1.47×10^{-4}	$\text{Avg}(h_{25})/\text{Avg}(h_0) = 1.03$
ϕ_3	Cubic labor disutilty	1.33×10^{-5}	$\text{Avg}(h_{35})/\text{Avg}(h_0) = 0.83$
v_ϕ	Variance of labor disutilty	0.041	$\text{Corr}(\log h_0, \log w_0) = 0.349$
$v_\varepsilon^{1/2}$	Variance of transitory shocks	0.293	$\text{Std}(\log w_0) = 0.30$
Δv_ε	Growth in variance of wages	0.0066	$\text{Std}(\log w_{25}) = 0.59$
$v_\eta^{1/2}$	Variance of permanent shocks	0.06	$\text{Std}(\log c_{25})/\text{Std}(\log w_{25}) = 0.65$
ι	Lump-sum transfer	0.155	$\text{Std}(\log c_0)/\text{Std}(\log w_0) = 0.70$

5.1 Parameterization

The model described in Section 3 is parameterized to match a selection of life cycle moments taken from Heathcote et al. (2020), which are based on a sample from the Panel Study of Income Dynamics from 2000 to 2006. Values for the discount factor, life-cycle wage profile, and the elasticities of labor supply and skill investment are borrowed directly from Heathcote et al. (2020). The parameter governing the Pareto tail of the wage distribution, θ, is set to 2 to be consistent with administrative tax records for the US, as shown by Saez (2001).

Remaining parameters are calibrated inside the model. To do this, following Boar and Midrigan (2022), the tax function in the calibration economy is set so that after-tax income is the sum of a lump-sum transfer and a log-linear function of before-tax income: $\widehat{T}(y) = y - \lambda y^{1-\tau} - \iota$. Then, the progressivity parameter is chosen to match Congressional Budget Office data on the shares of income before and after taxes and transfers, which gives $\tau = 0.086$. The lump-sum transfer is calibrated inside the model to match the initial variance of consumption while the tax level is set residually to satisfy the government's budget constraint, giving $\iota = 0.155$ and $\lambda = 0.654$. The variance of wage shocks in the model are set to match the variance of consumption and wages in the data, and their relative growth over the life cycle. The variance of labor disutility is set to match the correlation of hours worked and wages in the data, and its life cycle profile is calibrated to match the dynamics of average hours worked. All parameter values and their sources are summarized in Table 1.

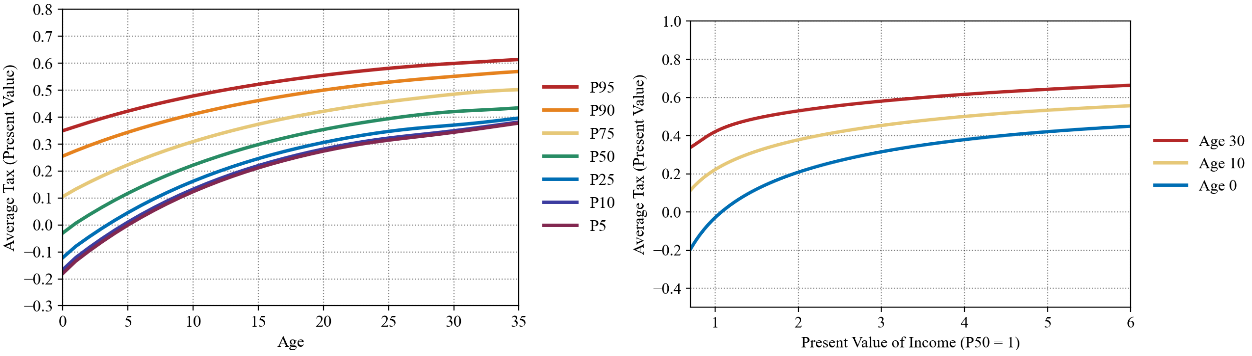
5.2 Optimal Taxation

This section describes the optimal history-dependent tax system in the calibrated model using the method from Figure 4. The results in this section use a simple average of previous log-incomes to tag individual's for their income history. The computational method can be easily extended to allow for dependence on the detailed history of earnings. However, experimentation with more sophisticated representations of income history were shown to have no observable difference in outcomes compared to a simple average. This suggests that a simple average is sufficient for the social planner to screen for a worker's skill type.

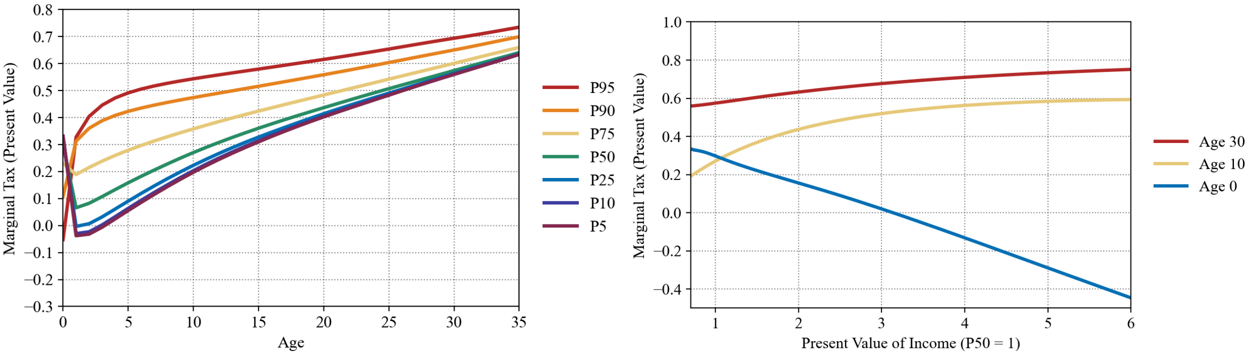
Figure 1 plots the optimal history-dependent tax system computed in the baseline model. The top panels show how the average present value of taxes paid as a fraction of the present value of income increases with both age and income. The bottom panels show how the marginal tax rate evolves over the life cycle. Notably, the marginal tax rate on income earned in the initial period is strongly decreasing in income, before switching to being increasing for the remainder of the life cycle. This suggests the social planner has strong incentives to learn workers' types by subsidizing labor supply in the initial period. Once the planner screens the workers' types, it is able to extract more revenue from the rarest skill types for the remainder of the life cycle.

For comparison, Figure 3 plots the optimal tax system when the planner can only

Figure 1: Optimal History-Dependent Tax Function, Baseline Model



(i) Average Tax by Income Percentile[b] (ii) Average Tax by Age



(iii) Marginal Tax by Income Percentile[c] (iv) Marginal Tax by Age

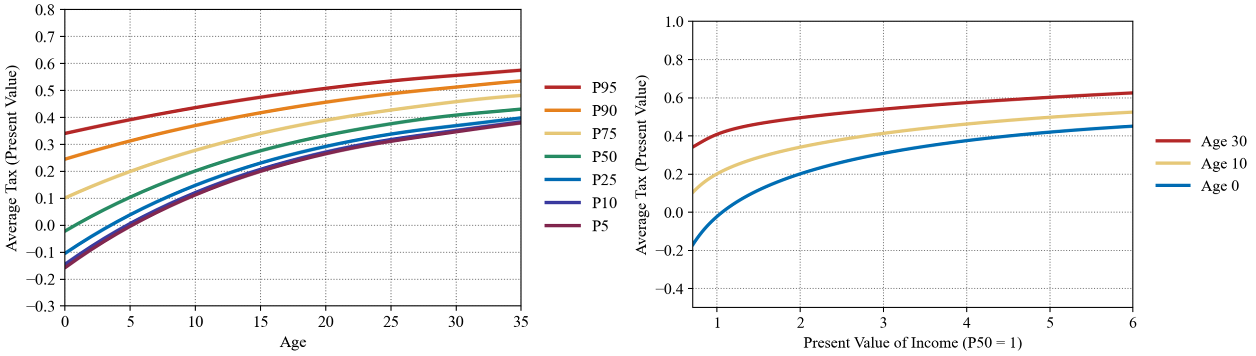
[a] The tax schedules in this figure are computed by simulating the economy with all wage and preference shocks set to zero so that the only heterogeneity comes from skill type.

[b] Average tax rates are computed as the present value of taxes to be paid by each household as a fraction of the present value of future earnings: $\sum_{t=a}^{A-1} \beta^t (y_t - c_t) / \sum_{t=a}^{A-1} \beta^t y_t$.

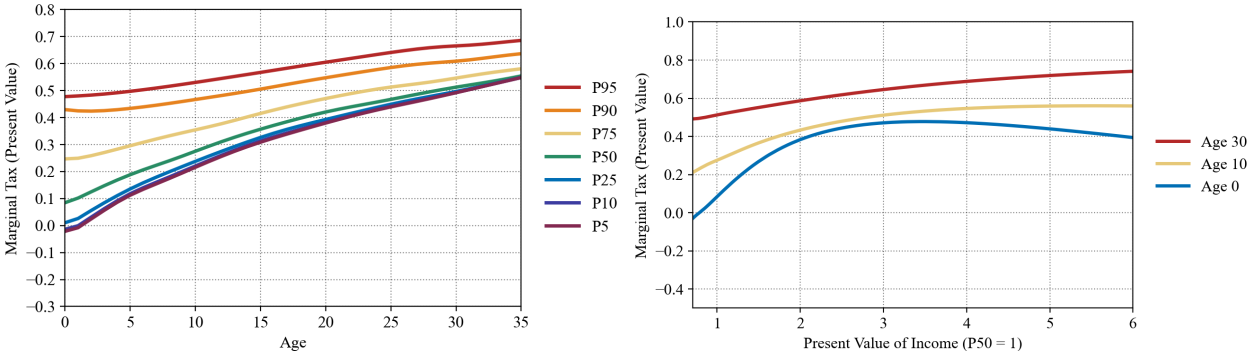
[c] Marginal tax rates are computed as the marginal change in the present value of lifetime earnings from income in the current period: $\partial \left(\sum_{t=a}^{A-1} \beta^t y_t \right) / \partial y_a$

condition taxes on current income and age. As with history-dependent taxation, average taxes increase with age and income. However, without being able to use information from previous years, the planner no longer chooses a decreasing marginal tax rate on earnings in the initial period. The tax system overall is slightly less redistributive than the history-dependent system, which is due to the planner not being able to use income history to simultaneously extract higher revenue while maintaining incentives to work. History-dependence allows the planner to achieve both higher labor supply and revenue from high income workers. This can be seen in Table 2, which summarizes the changes

Figure 3: Optimal Age-Dependent Tax Function, Baseline Model



(i) Average Tax by Income Percentile[b]   (ii) Average Tax by Age



(iii) Marginal Tax by Income Percentile[c]   (iv) Marginal Tax by Age

[a] The tax schedules in this figure are computed by simulating the economy with all wage and preference shocks set to zero so that the only heterogeneity comes from skill type.

[b] Average tax rates are computed as the present value of taxes to be paid by each household as a fraction of the present value of future earnings: $\sum_{t=a}^{A-1} \beta^t (y_t - c_t) / \sum_{t=a}^{A-1} \beta^t y_t$.

[c] Marginal tax rates are computed as the marginal change in the present value of lifetime earnings from income in the current period: $\partial \left(\sum_{t=a}^{A-1} \beta^t y_t \right) / \partial y_a$

in consumption-equivalent welfare, income and consumption when moving from the optimal age-dependent tax system to the optimal history-dependent tax. Overall, history-dependence increases welfare by 2.1 percent, but the gains are not evenly distributed. The largest gains are for households with the lowest present value of lifetime earnings, and consistently decrease with higher levels of income. The top five percent of lifetime incomes experience a two percent loss in welfare on average. The history-dependent tax system is able to slightly increase these workers total output while increasing their tax burden. This increases wages while also providing additional revenue to redistribute to lower income

Table 2: Percent Change in Allocations by Percentile of Income (Present Value),[a]
Optimal History-Dependent Tax Relative to Optimal Age-Dependent Tax

Endogenous and Differentiated Skills

	Overall	<P5	P5–P25	P25–P75	P75–P95	>P95
Welfare[b]	**2.1**	2.9	2.8	2.5	1.0	-2.0
$PV(y)$[a]	-1.7	-4.5	-3.0	-1.5	-0.8	0.2
$PV(c)$	0.7	0.3	0.9	1.3	0.1	-4.1

Endogenous Skills

	Overall	<P5	P5–P25	P25–P75	P75–P95	>P95
Welfare	**1.3**	0.8	1.1	1.5	1.4	-0.5
$PV(y)$	0.5	-3.1	-1.7	0.6	3.1	2.1
$PV(c)$	1.4	-1.0	-0.1	1.5	3.4	0.8

Exogenous Skills

	Overall	<P5	P5–P25	P25–P75	P75–P95	>P95
Welfare	**0.6**	-4.9	-1.9	1.0	3.0	1.3
$PV(y)$	3.8	4.9	4.5	4.1	2.9	1.8
$PV(c)$	2.8	-2.2	0.4	3.5	5.0	2.2

Exogenous Skills, No Transitory Wage Shocks

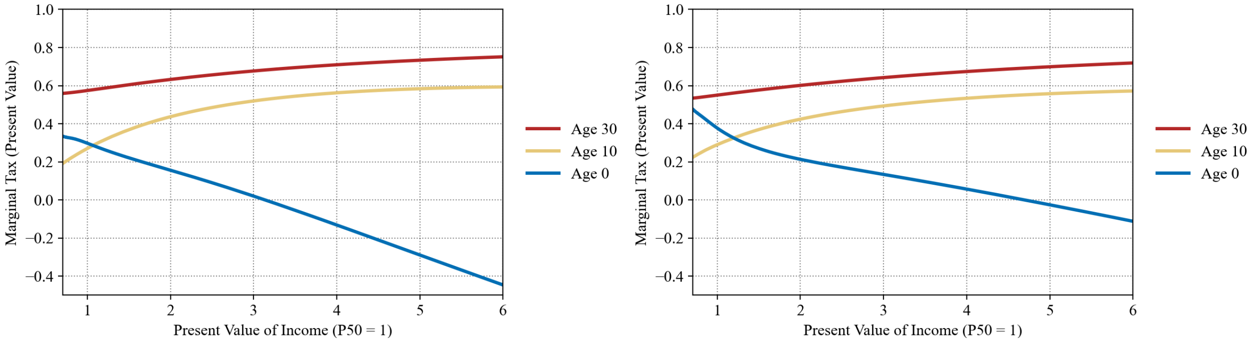
	Overall	<P5	P5–P25	P25–P75	P75–P95	>P95
Welfare	**0.1**	-0.5	-0.3	0.1	0.5	-0.3
$PV(y)$	0.6	0.6	0.8	0.7	0.1	-0.2
$PV(c)$	0.3	-0.3	0.0	0.4	0.6	-0.6

[a] The present value of income is computed as the sum of discounted lifetime earnings and is denoted as $PV(y) \equiv \sum_{a=0}^{A-1}(1+r)^{-a}y_{ia}$. Similarly, the present value of consumption is denoted as $PV(c) \equiv \sum_{a=0}^{A-1}(1+r)^{-a}c_{ia}$.

[b] Welfare is expressed in consumption-equivalent units. In this case, the welfare gain is the percent change in lifetime consumption needed to make an individual indifferent between living under the optimal history-dependent tax and the optimal age-dependent tax system.

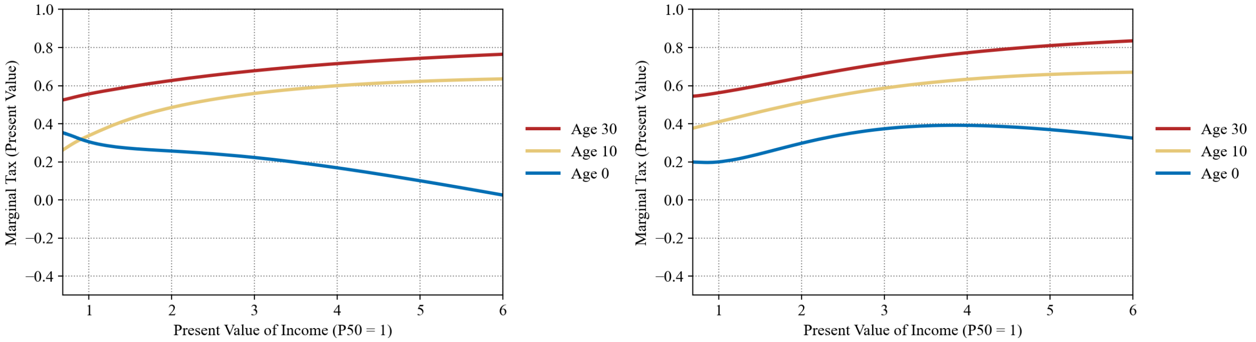
workers. While lower income households receive slightly higher consumption under the optimal tax system, the bulk of welfare benefits come from lower labor supply. Essentially, the planner uses history-dependence to concentrate work effort in the workers with the highest marginal productivities. The rest of Table 2 shows how the benefits of doing this decrease as rent-seeking behaviors are removed from the model.

Figure 5: Optimal History-Dependent Marginal Tax Rates by Skill Investment Technology



(i) Endogenous and Differentiated Skills

(ii) Endogenous Skills



(iii) Exogenous Skills

(iv) Exogenous Skills, No Transitory Shocks

[a] The tax schedules in this figure are computed by simulating the economy with all wage and preference shocks set to zero so that the only heterogeneity comes from skill type. Marginal tax rates are computed as the marginal change in the present value of remaining lifetime earnings from income in the current period: $\partial\left(\sum_{t=a}^{A-1}\beta^t y_t\right)/\partial y_a$.

The second case summarized in Table 2 is where skill investment is still endogenous, but general equilibrium effects are turned off. This is done by setting the price schedule to be an increasing function of skills alone: $p(s) = \exp(s/\theta)$. This choice of price schedule maintains the same level of wage inequality as the baseline model, but removes the effects from aggregate output that are in equation (1). With this modification to the model, the planner still chooses to use history dependence to concentrate labor supply in higher productivity workers, but it does much less redistribution in consumption. The result is a 0.8 percent lower gain in welfare from history dependence compared to the baseline model, with gains more evenly distributed across income levels and the largest gains in the middle of the income distribution.

The third case shown in Table 2 is where skill investment is exogenous, which is done by setting the skill prices to simply be an exponential distribution with tail parameter θ. In this case, the planner can still extract moderate welfare benefits from using history-dependence. However, without endogenous or differentiated human capital, there are no spillover effects to lower income workers, so the benefits of screening go exclusively to workers in the middle and top of the income distribution. The last case shown in Table 2, turns off transitory wage shocks. This makes it much easier for the planner to learn workers' types from current income alone, which reduces the welfare benefits from history-dependence close to zero.

Figure 5 plots the optimal marginal tax rates in each of the cases summarized in Table 2. It can be seen that the marginal tax rates on income in the initial period decrease less strongly as the welfare benefits of history-dependence decrease. Meanwhile, marginal tax rates for earnings later in the life cycle remain largely similar as rent-seeking behaviors are removed from the model. This demonstrates how the optimal tax system imposes larger distortions on workers as the benefits of extracting their type becomes larger.

6 Conclusion

This paper studied optimal life cycle income taxation in an environment with differentiated human capital. It demonstrated that policymakers can achieve potentially large welfare gains from using history-dependent tax schemes to extract output from workers with high marginal productivity and take advantage of potential spillover effects in revenue and general equilibrium wages. To do this, the paper developed a novel method to compute optimal taxes using multilayered neural networks. This approach provided a tractable way to compute arbitrarily complex tax functions in an quantitatively realistic model economy with general equilibrium effects, endogenous human capital investment, and incomplete markets.

While this paper studied highly stylized age-dependent policies that are not likely to be implementable if taken literally, the government has many tools available to subsidize or tax activities based indirectly on age or income history. For example, targeted subsidies in higher education, housing, or childcare would be ways to differentiate effective tax rates on younger workers without literally varying income taxes by age and income history. Furthermore, the method described in this paper is highly flexible and 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 policies to study.

References

Acemoglu, Daron (2002) “Changes in relative wages, 1963-1987: supply and demand factors,” *Journal of Economic Literature*, 40 (1), 7â72, 10.1257/0022051026976.

Acemoglu, Daron and David Autor (2011) “Skills, tasks and technologies: implications for employment and earnings,” *Handbook of Labor Economics*, 4B (12), 1043–1171, 10.1016/S0169-7218(11)02410-5.

Adenbaum, Jacob, Fil Babalievsky, and William Jungerman (2024) “Deep reinforcement learning for economics,” *Working paper*.

Azinovic, Marlon, Luca Gaegauf, and Simon Scheidegger (2022) “Deep equilibrium nets,” *International Economic Review*, 63 (4), 1471–1525, 10.1111/iere.12575.

Boar, Corina and Virgiliu Midrigan (2022) “Efficient redistribution,” *Journal of Monetary Economics*, 131, 78–91, 10.1016/j.jmoneco.2022.07.003.

Brumm, Johannes and Simon Scheidegger (2017) “Using adaptive sparse grids to solve high dimensional dynamic models,” *Econometrica*, 85 (5), 1575–1612, 10.3982/ECTA12216.

Chang, Yongsung and Yena Park (2021) “Optimal taxation with private insurance,” *The Review of Economic Studies*, 88 (6), 2766–2798, 10.1093/restud/rdab003.

Conesa, Juan Carlos, Sagiri Kitao, and Dirk Krueger (2009) “Taxing capital? Not a bad idea after all!,” *American Economic Review*, 99 (1), 25–48, 10.1257/aer.99.1.25.

Conesa, Juan Carlos and Dirk Krueger (2006) “On the optimal progressivity of the income tax code,” *Journal of Monetary Economics*, 53 (7), 1425–1450, 10.1016/j.jmoneco.2005.03.016.

Diamond, Peter A. and James A. Mirrlees (1971) “Optimal taxation and public production II: tax rules,” *American Economic Review*, 61 (3), 261–278.

Duarte, Victor, Diogo Duarte, and Dejanir H Silva (2024) “Machine learning for continuous-time finance,” *The Review of Financial Studies*, 37 (11), 3217â3271, 10.1093/rfs/hhae043.

Farhi, Emmanuel and Iván Werning (2013) “Insurance and taxation over the life cycle,” *The Review of Economic Studies*, 80 (2), 596â635, 10.1093/restud/rds048.

Fernández-Villaverde, Jesús, Samuel Hurtado, and Galo Nuño (2023) “Financial frictions and the wealth distribution,” *Econometrica*, 91 (3), 869–901, 10.3982/ECTA18180.

Findeisen, Sebastian and Dominik Sachs (2017) “Redistribution and insurance with simple tax instruments,” *Journal of Public Economics*, 146, 58–78, 10.1016/j.jpubeco.2016.12.006.

Golosov, Mikhail, Narayana Kocherlakota, and Aleh Tsyvinski (2003) “Optimal indirect and capital taxation,” *The Review of Economic Studies*, 70 (3), 569–587, 10.1111/1467-937X.00256.

Golosov, Mikhail, Maxim Troshkin, and Aleh Tsyvinski (2016) "Redistribution and social insurance," *American Economic Review*, 106 (2), 359–386, 10.1257/aer.20111550.

Han, Jiequn, Yucheng Yang, and Weinan E (2026) "DeepHAM: a global solution method for heterogeneous agent models with aggregate shocks," *Quantitative Economics*, 17 (2), 297–341.

Heathcote, Jonathan, Kjetil Storesletten, and Giovanni L. Violante (2017) "Optimal tax progressivity: an analytical framework," *Quarterly Journal of Economics*, 132 (4), 1693–1754, 10.1093/qje/qjx018.

——— (2020) "Optimal progressivity with age-dependent taxation," *Journal of Public Economics*, 189 (104074), 10.1016/j.jpubeco.2019.104074.

Heathcote, Jonathan and Hitoshi Tsujiyama (2021) "Optimal income taxation: mirrlees meets ramsey," *Journal of Political Economy*, 129 (11), 10.1086/715851.

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, 10.1016/j.jedc.2014.03.003.

Jungerman, William (2023) "Dynamic monopsony and human capital," *Working paper*.

Kahou, Mahdi Ebrahimi, Jesús Fernández-Villaverde, Jesse Perla, and Arnav Sood (2022) "Exploiting symmetry in high-dimensional dynamic programming," *Working paper*.

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 (2), 182–214, 10.1257/mac.20140274.

Kase, Hanno, Leonardo Melosi, and Matthias Rottner (2025) "Estimating nonlinear heterogeneous agent models with neural networks," *Working paper*.

Katz, Lawrence F. and Kevin M. Murphy (1992) "Changes in relative wages, 1963-1987: supply and demand factors," *Quarterly Journal of Economics*, 107 (1), 35–78, 10.2307/2118323.

Krueger, Dirk and Felix Kubler (2004) "Computing equilibrium in OLG models with stochastic production," *Journal of Economic Dynamics and Control*, 28 (7), 1411–1436, 10.1016/S0165-1889(03)00111-8.

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, 10.1257/aer.103.3.496.

——— (2016) “On the optimal provision of social insurance: Progressive taxation versus education subsidies in general equilibrium,” *Journal of Monetary Economics*, 77, 72–98, 10.1016/j.jmoneco.2015.11.002.

Maliar, Lilia, Serguei Maliar, and Pablo Winant (2021) “Deep learning for solving dynamic economic models,” *Journal of Monetary Economics*, 122, 76–101, 10.1016/j.jmoneco.2021.07.004.

Mirrlees, James A. (1971) “An exploration in the theory of optimum income taxation,” *The Review of Economic Studies*, 38 (2), 175–208, 10.2307/2296779.

Ndiaye, Abdoulaye and Zhixiu Yu (2025) “Flexible retirement and optimal taxation,” *Working paper*.

Payne, Jonathan, Adam Rebei, and Yucheng Yang (2025) “Deep learning for search and matching models,” *Working paper*.

Peterman, William B. (2016) “The effect of endogenous human capital accumulation on optimal taxation,” *Review of Economic Dynamics*, 21, 46–71, 10.1016/j.red.2016.04.001.

Peterman, William B. and Erick Sager (2022) “Optimal public debt with life cycle motives,” *American Economic Journal: Macroeconomics*, 14 (4), 404–437, 10.1257/mac.20180247.

Sachs, Dominik, Aleh Tsyvinski, and Nicolas Werquin (2020) “Nonlinear tax incidence and optimal taxation in general equilibrium,” *Econometrica*, 88 (2), 469–493, 10.3982/ECTA14681.

Saez, Emmanuel (2001) “Using elasticities to derive optimal income tax rates,” *The Review of Economic Studies*, 68 (1), 205–229, 10.1111/1467-937X.00166.

Saez, Emmanuel and Stefanie Stantcheva (2018) “A simpler theory of optimal capital taxation,” *Journal of Public Economics*, 162, 120–142, 10.1016/j.jpubeco.2017.10.004.

Scheuer, Florian and Iván Werning (2016) “Mirrlees meets Diamond-Mirrlees,” *NBER Working Paper* (22076).

Stantcheva, Stefanie (2017) “Optimal taxation and human capital policies over the life cycle,” *Journal of Political Economy*, 125 (6), 10.1086/694291.

——— (2020) “Dynamic taxation,” *Annual Review of Economics*, 12, 801–831, 10.1146/annurev-economics-100119-013035.

Stiglitz, Joseph E. (1982) “Self-selection and pareto efficient taxation,” *Journal of Public Economics*, 17 (2), 213–240, 10.1016/0047-2727(82)90020-2.

Weinzeirl, Matthew (2011) “The suprising power of age-dependent taxes,” *The Review of Economic Studies*, 78 (4), 1490–1518, 10.1093/restud/rdr001.

Appendix

A Estimating Neural Networks

A neural network with $\ell \geq 3$ hidden layers is constructed by repeatedly performing nonlinear transformations of outputs of $\ell - 1$ intermediate networks,

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

This network will have $n \times p + (\ell - 1)p^2 + p \times m$ parameters, or 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 why neural networks are able to approximate high dimensional functions quickly.

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

$$\mathbf{w}_{new} = \mathbf{w} + \lambda \frac{\partial L(x;\mathbf{w})}{\partial \mathbf{w}}$$

Here, λ 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 $\partial L(x;\mathbf{w})/\partial \mathbf{w} = 0$, which means that there is no deviation in any weight that creates a lower value of the loss function. The values of weights $\mathbf{w}$ that satisfy $\partial L(x;\mathbf{w})/\partial \mathbf{w} = 0$ for a given network are not unique. Similarly, the number of nodes and layers capable of approximating a certain function are also not unique.

A.1 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

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

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. Meanwhile, 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.

B Comparison to Mechanism Design in a Static Economy

This appendix shows how the the computational approach used in the paper compares to the true optimal tax system in a simple static economy where optimal taxes can be computed as the solution to a constrained planning problem. The two methods are shown to coincide almost exactly, as demonstrated in Figure 7.

B.1 Optimal Taxation in a Static Economy

Consider a static economy with a measure one of households. Worker types, α, are drawn from a distribution $F(\alpha)$. Households have preferences $u(c, y/\alpha)$ over their choices of consumption c and income y. Assume that α is not observable, so the government can only impose taxes based on income.

Mechanism Design As described by Mirrlees (1971), optimal taxation in this environment is the solution to a planning problem with the planner constrained by workers truthfully reporting their types. The government's problem is to choose allocations for each type to maximize utilitarian welfare,

$$\max_{c,y} \int u\left(c(\alpha), y(\alpha)/\alpha\right) dF(\alpha),$$

subject to incentive compatibility,

$$u\left(c(\alpha), y(\alpha)/\alpha\right) \geq u\left(c(\alpha'), y(\alpha')/\alpha\right), \text{ for every, } \alpha, \alpha'$$

and the government's budget constraint,

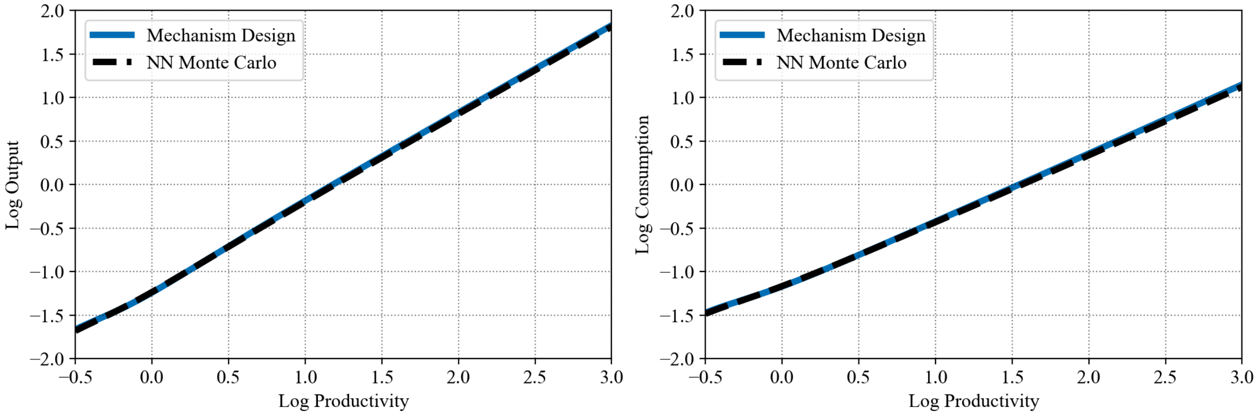
$$G = \int \left(y(\alpha) - c(\alpha)\right) dF(\alpha), \tag{5}$$

where G is the level of government expenditures.

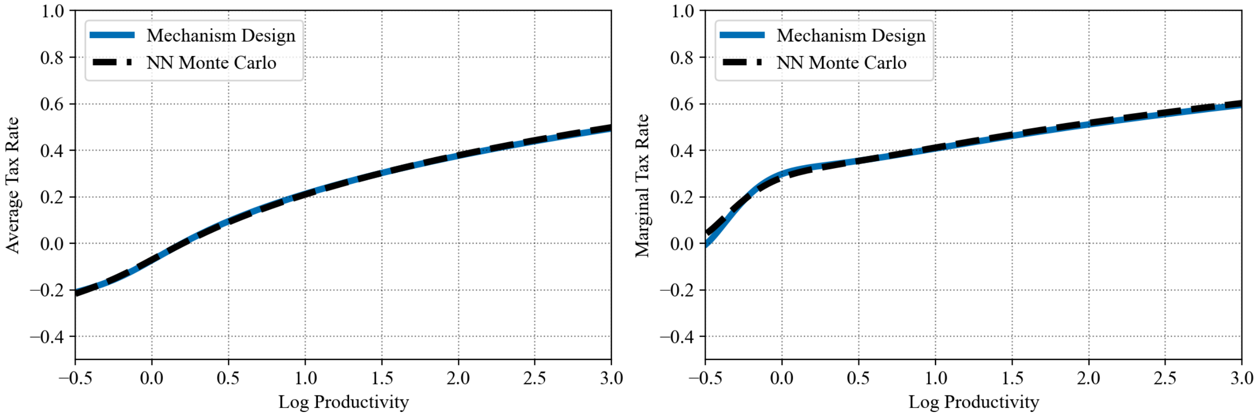
Optimal Tax Function As described by Saez (2001), an equivalent way to cast the optimal taxation problem is to solve for a tax function that maximizes welfare given the individual optimization of workers. In particular, given a tax function T, a worker of type α chooses their allocations to solve

$$\max_{c,y} u(c, y/\alpha)$$

Figure 7: Comparison of Solutions to Static Optimal Tax Problem



(i) Income (ii) Consumption



(iii) Average Tax Rate (iv) Marginal Tax Rate

subject to their individual budget constraint,

$$c = y - T(y).$$

Denote the solution to this problem as $c(\alpha;T)$ and $y(\alpha;T)$. Then, the government's problem is to choose a tax function

$$\max_{T} \int u\left(c(\alpha;T), h(\alpha;T)\right) dF(\alpha)$$

subject to the government's budget constraint,

$$G = \int T\left(y(\alpha;T)\right) dF(\alpha).$$

Here, the individual choice functions place implementability constraints on the government's problem so that the government cannot achieve allocations inconsistent with individual optimization. The method described in Section 4 can be easily applied to solve this optimal taxation problem, where allocations and the tax function are represented as separate neural networks. Utilitarian welfare is maximized with penalty functions added to enforce individual first order conditions. The results of this are shown in Figure 7. It can be seen that the allocations and implied tax rates coincide with the solution to the Mirrlees planning problem. Similarly they produce no noticeable differences in social welfare. To produce this figure, α is assumed to follow an exponentially modified gaussian distribution, with parameter values borrowed from Heathcote and Tsujiyama (2021).[10]

C Production in a Mirrlees Optimal Tax Problem

As noted by Scheuer and Werning (2016), the specific form of the production function does not change the optimal income tax formulas from Diamond and Mirrlees (1971). This appendix will demonstrate using a 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 optimal tax formula from Diamond and Mirrlees (1971) holds. The reason for this is that the Diamond and Mirrlees (1971) 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, as formulated by Stiglitz (1982), so that they enter the production function in different ways, income is no longer just a transformation of efficiency units. 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.

Static Economy Consider a static economy with a measure one of households. Worker types, α, are drawn from a distribution $F(\alpha)$. A worker of type α who works h hours produces $I(\alpha,h)$ units of income. All households have preferences $u(c,h)$ over consumption

---

[10] Specifically, $u(c,h) = \log c - h^{1+\sigma}/(1+\sigma)$ with $\sigma = 2$, and $\alpha \sim EMG(1/\zeta, +v_\alpha/2, v_\alpha, \zeta)$ with $v_\alpha = 0.142$ and $\zeta = 2.2$.

c and hours worked h. Assume that neither α or h are observable, so the government can only impose taxes on income. A household of type α solves

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

subject to

$$c = I(\alpha,h) - T\left(I(\alpha,h)\right) \tag{6}$$

The first order condition of this problem is

$$-\frac{u_h}{u_c} = \left[1 - T'\left(I(\alpha,h)\right)\right] y_h(\alpha,h) \tag{7}$$

Optimal Tax Problem Suppose the government wants to maximize utilitarian welfare across households

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

subject to the government's budget constraint,

$$G = \int T\left(I(\alpha,h)\right) dF(\alpha), \tag{8}$$

where G is the level of government expenditures. Optimal taxes can be 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 ($\partial SWF / \partial \epsilon = 0$). For simplicity, assume that household utility is GHH, $u(c,h) = \frac{1}{1-\gamma}\left[c - v(h)\right]^{1-\gamma}$. First, compute $\partial SWF/\partial \epsilon$ by differentiating welfare with respect to a perturbation and set it to zero to get the optimality condition of the government,

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

Use the GHH utility to simplify this to

$$0 = \int u_c(\alpha) \left[\frac{\partial c(\alpha)}{\partial \epsilon} - v'(h(\alpha)) \frac{\partial h(\alpha)}{\partial \epsilon} \right] dF(\alpha) \tag{9}$$

Now, two specifications for production will be considered that will have different implications for how general equilibrium effects show up in the formulas for optimal taxes.

C.1 Diamond and Mirrlees (1971)

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

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

where θ is the elasticity of substitution between different levels of efficiency units. With this production function, household income is given by $I(h,\alpha) = p(\alpha h)$, where $p(y)$ is the marginal product of households with efficiency units y:

$$p(y) = \frac{\partial Y}{\partial [y f_y(y)]} = [y f_y(y)]^{\frac{\theta-1}{\theta}-1} \left[\int [y f_y(y)]^{\frac{\theta-1}{\theta}} dy \right]^{\frac{\theta}{\theta-1}-1} = \left[\frac{Y}{y f_y(y)} \right]^{\frac{1}{\theta}}. \tag{10}$$

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

$$c(\alpha) = p(\alpha h(\alpha)) - T(p(\alpha h(\alpha))) - \epsilon H(p(\alpha h(\alpha))) \tag{11}$$

$$v'(h(\alpha)) = \left[1 - T'(p(\alpha h(\alpha))) - \epsilon H'(p(\alpha h(\alpha))) \right] p'(\alpha h(\alpha)) \alpha \tag{12}$$

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

$$c(\alpha) = p(\alpha h(\alpha)) - T(p(\alpha h(\alpha))) - \epsilon H(p(\alpha h(\alpha)))$$

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

Differentiating these equations with respect to ϵ gives

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

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

These two equations, together with (10), (11) and (12) are five equations in five unknowns (given the tax function): $c(\alpha), h(\alpha), p\left(\alpha h(\alpha)\right), \frac{\partial c(\alpha)}{\partial \epsilon}, \frac{\partial h(\alpha)}{\partial \epsilon}$. Here, wages p are just a (possibly nonlinear) transformation of efficiency units $\alpha h(\alpha)$. 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.

C.2 Stiglitz (1982)

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

$$Y = \left(\int \left[h(\alpha) f(\alpha)\right]^{\frac{\theta-1}{\theta}} d\alpha\right)^{\frac{\theta}{\theta-1}} \tag{13}$$

where θ 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 $I(h, \alpha) = p(\alpha)h(\alpha)$, where $p(\alpha)$ is the marginal product of households of type α:

$$p(\alpha) = \frac{\partial Y}{\partial\left[h(\alpha) f(\alpha)\right]} = \left[h(\alpha) f(\alpha)\right]^{\frac{\theta-1}{\theta}-1}\left[\int \left[h(\alpha) f(\alpha)\right]^{\frac{\theta-1}{\theta}} d\alpha\right]^{\frac{\theta}{\theta-1}-1} = \left[\frac{Y}{h(\alpha) f(\alpha)}\right]^{\frac{1}{\theta}}. \tag{14}$$

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

$$c(\alpha) = p(\alpha)h(\alpha) - T\left(p(\alpha)h(\alpha)\right) - \epsilon H\left(p(\alpha)h(\alpha)\right) \tag{15}$$

$$v'\left(h(\alpha)\right) = \left[1 - T'\left(p(\alpha)h(\alpha)\right) - \epsilon H'\left(p(\alpha)h(\alpha)\right)\right] p'(\alpha)\alpha \tag{16}$$

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

$$c(\alpha) = p(\alpha)h(\alpha) - T(p(\alpha)h(\alpha)) - \epsilon H(p(\alpha)h(\alpha))$$

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

Differentiating these equations with respect to ϵ gives

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

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

These two equations, together with (14), (15) and (16) are five equations in *six* unknowns (given the tax function): $c(\alpha), h(\alpha), p(\alpha), \frac{\partial c(\alpha)}{\partial \epsilon}, \frac{\partial h(\alpha)}{\partial \epsilon}, \frac{\partial p(\alpha)}{\partial \epsilon}$. 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, $\frac{\partial p(\alpha)}{\partial \epsilon}$. This effect is given by differentiating the labor market clearing equation (14),

$$\frac{\partial p(\alpha)}{\partial \epsilon} = \frac{1}{\theta} \left[\frac{Y}{h(\alpha)f(\alpha)}\right]^{\frac{1}{\theta}-1} \left[\frac{1}{h(\alpha)f(\alpha)} \int p(\tilde{\alpha}) \frac{\partial h(\tilde{\alpha})}{\partial \epsilon} dF(\tilde{\alpha}) - \frac{Y}{h(\alpha)^2 f(\alpha)} \frac{\partial h(\alpha)}{\partial \epsilon}\right] \tag{17}$$

This sixth equation completes the characterization of $(c(\alpha), h(\alpha), p(\alpha), \frac{\partial c(\alpha)}{\partial \epsilon}, \frac{\partial h(\alpha)}{\partial \epsilon}, \frac{\partial p(\alpha)}{\partial \epsilon})$ 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 $\tilde{\alpha}$, i.e., $\partial h(\tilde{\alpha})/\partial \epsilon > 0$, it will increase the wage of type α. This spillover effect was described by Stiglitz (1982). The benefit to the government of increasing households' labor supply is highest from the rarest skill types, who have the highest wages $p(\tilde{\alpha})$. This effect becomes stronger when the elasticity of substitution between skill types is lower. In fact, it can be seen in equation (17) that the effect of a tax reform on wages is directly proportional to the inverse of the elasticity of substitution, θ. Importantly, the spillover effects from the first term in the brackets mean that $\partial p(\alpha)/\partial \epsilon$ is not just a nonlinear transformation of type α'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.