Tulane Economics Working Paper Series

A General Theory of Inverse Welfare Functions

Katy Bergstrom
Tulane University
kbergstrom@tulane.edu

William Dodds
Tulane University
wdodds@tulane.edu

Working Paper 2408
June 2024

Abstract

This paper develops a general theory to recover the inverse welfare function that rationalizes a given tax schedule as optimal. Our theory allows for complex environments including the presence of multidimensional tax schedules, bunching/jumping behavior, optimization frictions, general equilibrium effects, and externalities. We show how inverse welfare functions can be used to assess the desirability of tax reforms and to test for Pareto efficiency. We numerically construct several inverse welfare functions for piecewise-linear income taxation, taxation with optimization frictions, joint income and property taxation, income taxation with labor demand and endogenous wages, and taxation with inequality aversion.

Keywords: inverse welfare function, inverse welfare weights, frictions, multidimensional taxation, general equilibrium, Pareto efficiency
JEL codes: H21, H31, D60

A General Theory of Inverse Welfare Functions[*]

Katy Bergstrom[†] William Dodds[‡]

June 11, 2024

Abstract

This paper develops a general theory to recover the inverse welfare function that rationalizes a given tax schedule as optimal. Our theory allows for complex environments including the presence of multidimensional tax schedules, bunching/jumping behavior, optimization frictions, general equilibrium effects, and externalities. We show how inverse welfare functions can be used to assess the desirability of tax reforms and to test for Pareto efficiency. We numerically construct several inverse welfare functions for piecewise-linear income taxation, taxation with optimization frictions, joint income and property taxation, income taxation with labor demand and endogenous wages, and taxation with inequality aversion.

Keywords: inverse welfare function, inverse welfare weights, frictions, multidimensional taxation, general equilibrium, Pareto efficiency

JEL: H21, H31, D60

[*] The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. We would like to thank Nathan Hendren, Nicolas Werquin, Juan Rios, and conference participants at SAET, SEA, and Tulane University for their very helpful comments and suggestions.

[†] Tulane University. Email: kbergstrom@tulane.edu

[‡] Tulane University. Email: wdodds@tulane.edu

1 Introduction

Since the seminal work of Mirrlees (1971), the vast majority of theoretical work on taxation entails social welfare maximization wherein a planner is endowed with a social welfare function that he/she seeks to maximize subject to constraints. However, in recent years there has been increased interest in solving the so-called "inverse taxation problem" wherein the economist is given a (proposed or actual) tax schedule and attempts to infer the social welfare function that rationalizes this tax schedule as optimal. The "inverse welfare function" recovers the implicit interpersonal welfare comparisons that justify a given tax schedule. From a policy perspective, if the inverse welfare function for a given tax schedule diverges sharply from society's true preferences, this places an onus on the government to change tax policy.

In the past 15 years, there have been numerous papers that recover inverse welfare functions for observed or proposed income tax schedules. For example, Blundell et al. (2009) explore the inverse welfare function that rationalizes the observed tax treatment of single mothers in the U.K. and Germany; Bourguignon and Spadaro (2010) recover the inverse welfare function for the French redistributive system; Bargain et al. (2013) compare inverse welfare functions for income tax systems across 17 European countries and the U.S.; Jacobs, Jongen and Zoutman (2017) explore the inverse welfare functions that justify proposed tax policies for different political parties in the Netherlands; and Hendren (2020) uses inverse welfare functions to explore the impact of income inequality over time and across countries. Importantly, all of these applications have been solely for income tax schedules and typically make a number of assumptions: individuals do not face optimization frictions and respond smoothly to tax schedules, individuals differ only in terms of a unidimensional heterogeneity parameter, there are no general equilibrium effects of taxation, and there are no externalities.

The goal of the present paper is to develop a general theory of inverse welfare functions that can be applied to recover implicit social preferences in much more general taxation settings. This paper has three contributions. First and foremost, we prove two theorems (Theorems 1 and 2) establishing the existence of inverse welfare functions; importantly, these theorems are constructive insofar as they show explicitly how to recover an inverse welfare function. Our methodology to construct inverse welfare functions can be applied to settings that feature, for example: multidimensional tax schedules, bunching/jumping behavior, optimization frictions, general equilibrium effects, and externalities. Second, we illustrate the policy relevance of inverse welfare functions by proving that inverse welfare functions can be used to determine the desirability of (marginal) non-linear tax reforms and can also be used to test for Pareto efficiency of a tax schedule. Third, we demonstrate numerically that various modelling assumptions (such as the presence of optimization frictions, multidimensional tax schedules, general equilibrium wage effects, or externalities) can have large and meaningful impacts on the inverse welfare function. We now provide a brief outline of the paper.

Section 2 first presents a highly simplified model without all of the mathematical machinery present in the rest of the paper to build intuition for a key insight: even in complex tax environments, we can compute inverse welfare functions as long as government revenue is sufficiently smooth as a function of the tax schedule. Section 2 then discusses notation and presents our first main theorem on existence and construction of inverse welfare functions: Theorem 1. Theorem 1 proves that in a partial equilibrium setting, if the government's budget constraint is satisfied with equality and revenue is Gateaux differentiable in the tax schedule (the Gateaux derivative is a generalization of the gradient), then we can compute an inverse welfare function. The proof to Theorem 1 is constructive: we show explicitly how to construct the inverse welfare function for arbitrary multidimensional tax schedules. Intuitively, we compute the inverse welfare weight for individuals making particular choices by equating the revenue effect of an "instantaneous" tax change at that choice level with the mechanical welfare effect of such an "instantaneous" tax change.

Next, Section 3 provides a number of analytical constructions of Gateaux derivatives of government revenue along with associated inverse welfare functions, highlighting that the Gateaux derivative of revenue can be expressed in terms of behavioral responses to tax reforms. The next result of the paper, Proposition 1, provides a set of general sufficient conditions for Gateaux differentiability of government revenue, establishing that Gateaux differentiability of government revenue is a relatively mild restriction: revenue can be Gateaux differentiable even if the tax schedule is non-differentiable (generating bunching) or individuals have multiple optima (e.g., respond on the extensive margin). Hence, Proposition 1 combined with Theorem 1 establishes that inverse welfare functions often exist even if the tax schedule is "pathological" in various ways. To conclude Section 3, we show that Theorem 1 also applies when individuals face "sparsity-based" optimization frictions (which are a general form of frictions discussed in Anagol et al. (2022)) and prove that Gateaux differentiability of government revenue is a mild restriction even when individuals face optimization frictions.

Next, Section 4 illustrates how inverse welfare functions can be used directly to inform tax policy. Specifically, we argue that comparing the inverse welfare weights for a given tax system with society's "true" welfare weights allows one to assess the desirability of arbitrary small tax reforms. Loosely, if society's true welfare weights are larger (smaller) than the inverse welfare weights at a particular income level, then decreasing (increasing) taxes at that income level (and closing the budget via changing the lump-sum transfer) is welfare improving. Section 4 also presents general conditions for Pareto optimality of multidimensional tax schedules: tax schedules with a positive inverse welfare function (e.g., welfare weights that are all positive) are Pareto efficient; conversely, tax schedules that can only be rationalized with inverse welfare functions which are non-positive (so that society implicitly values increasing utility for some types a negative amount) are Pareto inefficient.

Section 5 then conducts a number of numerical simulations to illustrate how various modeling assumptions impact the inverse welfare function. We begin with a baseline exercise: we recover the inverse welfare weights for a smooth approximation to the U.S. income tax schedule under the assumption that individuals face no optimization frictions. We then compute inverse welfare weights for the actual (piecewise linear) U.S. income tax schedule again under the assumption that individuals can perfectly optimize their labor supply and therefore bunch at kink points; we show that rationalizing kinks in the tax schedule requires a pathological welfare function. Next we relax the assumption that individuals respond smoothly to tax reforms on the intensive margin: we compute the inverse welfare function for the actual piecewise linear U.S. income tax schedule assuming agents can either work full-time, part-time, or not at all. Despite evidence from the observed lack of bunching at kink points suggesting that individuals face labor supply frictions (Saez, 2010), to the best of our knowledge, previous inverse welfare function calculations have not incorporated such frictions. We find that the inverse welfare function is no longer pathological and that the presence of frictions increases (decreases) the inverse welfare weights on high income (low income) households. Finally, Section 5 shows how to compute inverse welfare weights for a joint tax system of income and property taxes. The main finding is that society must have substantially different welfare weights for those who spend different amounts on housing *conditional on having the same income* in order to rationalize property taxes.

Next, Section 6 discusses how our theory of inverse welfare functions can be extended to settings with general equilibrium effects. When there are general equilibrium effects of taxation, we prove Theorem 2, which shows that an inverse welfare function can be constructed from the Gateaux derivative of government revenue *and* the Gateaux derivatives of equilibrium objects (e.g., endogenous wages or prices) with respect to the tax schedule. The inverse welfare function in this case is computed as the fixed point of an integral equation. We then illustrate Theorem 2 via a numerical example for a model with both labor supply and labor demand similar to Sachs, Tsyvinski and Werquin (2020). We find that general equilibrium wage effects can have substantial impacts on the inverse welfare function. Ignoring implicit redistribution via general equilibrium wage effects may lead us to vastly overestimate "redistributive tastes": tax schedules that are supported by highly redistributive welfare functionals without general equilibrium wage effects are supported by substantially less redistributive welfare functionals once these effects are taken into account. Finally, Section 6 shows that our definition of "general equilibrium effects" is broad enough so that Theorem 2 can also allow for externalities. We demonstrate numerically how the presence of inequality aversion (wherein other individuals' incomes generate an externality by contributing to inequality) impacts inverse welfare weights.

Section 7 discusses one situation in which inverse welfare functions typically *do not* exist: when the type space is smaller than the choice space. While in reality, the choice space is likely smaller than the type space, Section 7 discusses how this non-existence result has at least one important *theoretical* application: strengthening the Atkinson-Stiglitz Theorem. Specifically, Section 7 illustrates that in the Atkinson-Stiglitz environment, it is often impossible to rationalize indirect taxes (e.g., savings or commodity taxes) even if the government wants to make some individuals as miserable as possible (via negative welfare weights). Section 8 concludes.

2 Construction of Inverse Welfare Functionals

2.1 Simple Example to Build Intuition

This paper will be primarily concerned with the analysis of tax schedules defined over a continuum of choices. As such, we will use several tools from functional analysis which may not be familiar to all economists. We therefore believe it is useful to build intuition for our first main result, Theorem 1, in a simplified setting with a finite choice set which strips away much of the mathematical machinery. Consider a population of individuals who differ in terms of a parameter n, which represents individual productivity. Individuals choose between two income levels, z_1 and z_2, to maximize a quasi-linear utility function $u(z;n) = z - T(z) - v(z/n)$ where $z - T(z)$ is consumption and z/n is labor supply of individual n required to earn income z. The government chooses the tax levied on individuals who earn z_1 or z_2: $\{T_1, T_2\}$. Let p_1 denote the fraction of the population that chooses z_1 (recognizing that p_1 is a function of T_1 and T_2). The government has a budget constraint that revenue, $R(T_1, T_2) \equiv T_1 p_1 + T_2(1 - p_1)$, is zero. Letting $U(n; T_1, T_2)$ denote indirect utility for type n, our goal is to find an inverse welfare function $\int_N \phi(n) U(n; T_1, T_2) dF(n)$ that rationalizes a given tax schedule as optimal, where $F(n)$ is the distribution of n. We form a Lagrangian for the government with Lagrange multiplier λ:

$$\int_N \phi(n) U(n; T_1, T_2) dF(n) + \lambda R(T_1, T_2)$$

Assuming all objects are differentiable, a necessary condition for $\phi(n)$ to rationalize the tax schedule is that $\{T_1, T_2\}$ is a stationary point of the above Lagrangian. Differentiating the Lagrangian with respect to T_1 and T_2, $\phi(n)$ must satisfy the following vector equation in order to be an inverse welfare function (where we have applied the envelope theorem to calculate the impact of a tax change on indirect utility):

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -\int_N \phi(n) dF(n|z_1) p_1 \\ -\int_N \phi(n) dF(n|z_2)\,(1-p_1) \end{bmatrix} + \begin{bmatrix} \lambda \frac{\partial R(T_1,T_2)}{\partial T_1} \\ \lambda \frac{\partial R(T_1,T_2)}{\partial T_2} \end{bmatrix} \equiv \begin{bmatrix} -\bar{\phi}(z_1) p_1 \\ -\bar{\phi}(z_2)\,(1-p_1) \end{bmatrix} + \lambda \nabla R(T_1, T_2) \quad (1)$$

Equation 1 pins down the average welfare weights, $\bar{\phi}(z_1)$ and $\bar{\phi}(z_2)$, for individuals choosing z_1 and z_2 as a function of the gradient of government revenue with respect to the tax rates T_1 and T_2 (note that we can normalize $\lambda = 1$ as this simply scales $\phi(n)$ multiplicatively). As long as we pick *any* $\phi(n)$ that satisfies Equation 1, then the tax schedule will be locally extremal. The high level insight of Theorem 1 is that this intuition holds much more generally: even in settings with choices made over a continuum, multidimensional tax instruments, multidimensional agent heterogeneity, and complicated behavioral responses to tax changes, we can recover inverse welfare functions from the "gradient" (i.e., the Gateaux derivative) of government revenue with respect to the tax schedule.

2.2 Notation

We consider a population of individuals indexed by a type vector $\mathbf{n} = (n_1, n_2, ..., n_K) \in \mathbf{N}$ on compact $\mathbf{N}$ distributed according to some distribution $F(\mathbf{n})$ with density $f(\mathbf{n})$. Individuals choose $\mathbf{z} = (z_1, z_2, ..., z_J) \in \mathbb{R}^J$ to maximize a smooth utility function subject to a budget constraint given a tax schedule, $T(\mathbf{z})$, which is a function of individual choice variables $\mathbf{z}$:

$$\begin{aligned} &\max_{\mathbf{z}} \;\; u(c, \mathbf{z}; \mathbf{n}) \\ &\text{s.t. } \; c = y(\mathbf{z}) - T(\mathbf{z}) \end{aligned} \tag{2}$$

where c is numeraire consumption and is a function of choices $y(\mathbf{z})$ as well as the tax schedule $T(\mathbf{z})$. For example, z_i might represent income from a particular source (e.g., labor or savings) or consumption of a particular good or the z_i's could represent incomes in various time periods. We assume that there is a societal budget constraint that total tax revenue is greater than or equal to some exogenous revenue requirement, E:

$$\int_{\mathbf{N}} T(\mathbf{z}(\mathbf{n}))dF(\mathbf{n}) \geq E \tag{3}$$

where $\mathbf{z}(\mathbf{n})$ denotes optimal choices for type $\mathbf{n}$ under the given tax schedule. While we omit additional arguments to make expressions more readable, it is very important to note that $\mathbf{z}$ is a function of not only $\mathbf{n}$ but also the tax schedule $T(\cdot)$. Next, let us denote $U(\mathbf{n}; T)$ as the indirect utility profile that arises when agents optimize under tax schedule $T(\mathbf{z})$ according to Equation 2. And let us denote $\mathcal{U}$ as the set of all utility profiles that are generated by maximization under some tax schedule that also satisfy the government's budget constraint, Equation 3. Note that continuity of the utility function ensures that $\mathcal{U} \subset C(\mathbf{N})$ (by Berge's Maximum Theorem), where $C(\mathbf{N})$ is the set of continuous functions on $\mathbf{N}$.

A welfare functional, $W(U(\mathbf{n}; T))$, is defined as a *continuous linear functional* which takes the utility profile $U(\mathbf{n}; T)$ as its argument and returns a scalar value which we refer to as welfare.[1]

Definition 1. *$W : C(\mathbf{N}) \mapsto \mathbb{R}$ is a continuous linear functional if $W(a_1 f_1 + a_2 f_2) = a_1 W(f_1) + a_2 W(f_2)$ $\forall a_1, a_2 \in \mathbb{R}, f_1, f_2 \in C(\mathbf{N})$ and for any $f_1, f_2 \in C(\mathbf{N})$, $\forall \epsilon \; \exists \delta$ s.t. $||f_2 - f_1||_\infty < \delta \implies |W(f_2) - W(f_1)| < \epsilon$ where $|| \cdot ||_\infty$ is the supnorm.*

Remark 1. *By the Riesz-Markov-Kakutani representation theorem (Theorem 6.19 of Rudin (1974)), any continuous linear functional W as defined in Definition 1 can be expressed as follows for some distribution $\Phi(\mathbf{n})$:*

$$W(U(\mathbf{n}; T)) = \int_{\mathbf{N}} U(\mathbf{n}; T) d\Phi(\mathbf{n}) \tag{4}$$

By Remark 1, our restriction that welfare functionals be continuous and linear mandates that $W(\cdot)$ is a weighted sum of utilities. Our goal will be to recover the *inverse welfare functional* $W(U(\mathbf{n}; T))$ that rationalizes a given tax schedule $T(\mathbf{z})$ as optimal. Also, note while every continuous linear functional $W(\cdot)$ can be represented as an integral against some distribution $\Phi(\mathbf{n})$ as in Equation 4, it may not always be convenient or necessary to do so. If $\Phi(\mathbf{n})$ is differentiable, then Equation 4 can be expressed via "welfare weights" so that $W(U(\mathbf{n}; T)) = \int_{\mathbf{N}} \phi(\mathbf{n}) U(\mathbf{n}; T) d\mathbf{n}$; however, $W(U(\mathbf{n}; T))$ can also contain mass points as in a Rawlsian welfare

[1] Note that in the introduction we abused language for expositional simplicity by referencing "inverse welfare functions". This paper will be concerned with inverse welfare *functionals*, recalling that a functional is a function whose argument is a function.

function where $\Phi(\mathbf{n})$ is a distribution that puts all weight on the lowest type $\mathbf{n}$ in society.[2]

Finally, much of our analysis will require us to take *Gateaux derivatives* and *Gateaux variations* of various objects, which we define as in the Encyclopedia of Mathematics:

Definition 2. *Let $R : \mathcal{T} \mapsto \mathbb{R}$ be a functional. We say that R is Gateaux differentiable at a $T \in \mathcal{T}$ if $\exists$ a continuous linear functional, DR_T, which we call the Gateaux derivative, such that for any $\tau \in \mathcal{T}$:*[3]

$$\lim_{\epsilon \to 0} \frac{R(T + \epsilon\tau) - R(T)}{\epsilon} = DR_T(\tau)$$

Furthermore, we refer to the object $\lim_{\epsilon \to 0} \frac{R(T+\epsilon\tau)-R(T)}{\epsilon}$ as the Gateaux variation of R at T in the direction of τ, noting that the Gateaux variation need not be a continuous linear functional.

Remark 2. *In finite dimensional cases (i.e., agents choose among a discrete set of options as in Section 2.1), the Gateaux variation (i.e., the directional derivative) is always equal to the Gateaux derivative (i.e., the gradient) multiplied by the direction vector.*

Remark 3. *In Definition 2, the Gateaux derivative must be a continuous linear functional. In following sections, we will occasionally use the standard fact that continuity is equivalent to boundedness for linear functionals.*

2.3 A Constructive Existence Theorem

Our goal is to develop a theory of inverse optimal welfare functionals by tackling two questions: (1) "For a given tax schedule $T(\mathbf{z})$, is there an inverse welfare functional $W(\cdot)$ that rationalizes $T(\mathbf{z})$ as the optimal tax schedule?" and if so (2) "How can we construct such an inverse welfare functional?"

First, recognize that if $W(\cdot)$ rationalizes $T(\mathbf{z})$, this means that the utility profile $U(\mathbf{n}; T)$, generated when agents optimize according to the tax schedule $T(\mathbf{z})$, maximizes $W(\cdot)$ within the set $\mathcal{U}$ of all possible utility profiles generated by optimization under any tax schedule. However, because $\mathcal{U}$ is not in general a convex set, ensuring the existence of an inverse welfare functional $W(\cdot)$ for an arbitrary tax schedule $T(\mathbf{z})$ is difficult.[4] In Appendix B.1 we prove Proposition 6 establishing existence of an inverse welfare functional for a given tax schedule under a concavity condition; however, this result is non-constructive and the concavity condition is not always easy to verify in practice.

Towards rectifying these two issues, we will primarily consider a weaker notion of inverse welfare functionals: *local* inverse welfare functionals. Let us consider the following Lagrangian (with Lagrange multiplier λ on the societal budget constraint) under a welfare functional W:

$$L(T; W) \equiv W(U(\mathbf{n}; T)) + \lambda \left[\int_{\mathbf{N}} T(\mathbf{z}(\mathbf{n})) dF(\mathbf{n}) - E \right] \tag{5}$$

[2] Note that $\Phi(\mathbf{n})$ incorporates the type distribution; for instance, a utilitarian welfare function sets $\Phi(\mathbf{n})$ equal to $F(\mathbf{n})$ so that each type is weighted according to their density: $W(U(\mathbf{n}; T)) = \int_{\mathbf{N}} U(\mathbf{n}; T) f(\mathbf{n}) d\mathbf{n}$.

[3] The definition of Gateaux differentiability in Definition 2 is weaker than Frechet differentiability because we do not require uniform convergence in all directions τ.

[4] Consider two utility profiles $U(\mathbf{n}; T_1)$ and $U(\mathbf{n}; T_2)$ derived from individual optimization under two tax schedules $T_1(\mathbf{z})$ and $T_2(\mathbf{z})$, then the convex combination $U_3(\mathbf{n}) = \alpha U(\mathbf{n}; T_1) + (1 - \alpha) U(\mathbf{n}; T_2)$ may not be consistent with individual optimization under *any* tax schedule.

Definition 3. *$W(\cdot)$ is a local inverse welfare functional for $T(\mathbf{z})$ if $T(\mathbf{z})$ is a stationary point of the Lagrangian $L(T;W)$ so that the Gateaux derivative of $L(T;W)$ is 0.*

Henceforth, when we refer to an "inverse welfare functional" this should be understood as "local inverse welfare functional"; we drop the "local" for brevity. Next, let $R(T) \equiv \int_{\mathbf{N}} T(\mathbf{z}(\mathbf{n}))dF(\mathbf{n})$ denote government revenue as a function of $T(\mathbf{z})$ and let $\mathbf{Z} = \{\mathbf{z}(\mathbf{n})|\mathbf{n} \in \mathbf{N}\}$ denote the set of $\mathbf{z}$ that are chosen by some type $\mathbf{n}$ (hence, $\mathbf{Z}$ is a function of the tax schedule even though we omit it as an argument). This brings us to our first key result:

Theorem 1. *Consider continuous $T(\mathbf{z})$ such that $R(T) = E$, $\mathbf{Z}$ is compact, and for every $\mathbf{z}$ at least one $\mathbf{n}$ that chooses $\mathbf{z}$ has a unique optimum. A local inverse functional exists if $R(T)$ is Gateaux differentiable.*

Proof. We provide a sketch that avoids measure theory and skips over a number of technical details; see Appendix A.1 for a full proof.

For simplicity, suppose that the Gateaux derivative of the budget constraint can be written as the following sum over some partition $\{\mathbf{Z}_i\}$ with $\mathbf{Z}_1 \cup \mathbf{Z}_2 \cup \cdots \cup \mathbf{Z}_M = \mathbf{Z}$ (as we will see later, the Gateaux derivative of revenue is sometimes expressed via a partition of the domain $\mathbf{Z}$, e.g., into the interior and boundary):

$$\lim_{\epsilon \to 0} \frac{R(T+\epsilon\tau) - R(T)}{\epsilon} = \sum_i \int_{\mathbf{Z}_i} \tau(\mathbf{z})\gamma(\mathbf{z})h(\mathbf{z})d\mathbf{Z}_i \tag{6}$$

where $h(\mathbf{z})$ is the density of income (i.e., this sketch assumes there is no bunching) and $d\mathbf{Z}_i$ is the volume element of $\mathbf{Z}_i$. Loosely, $\gamma(\mathbf{z})$ represents the "instantaneous budgetary effect" of an infinitesimal tax perturbation that changes the tax schedule at a given choice level $\mathbf{z}$ (Figure 1a below illustrates such a perturbation for the unidimensional case in which agents choose an income z and consumption is given by $c = z - T(z)$). Section 3 will explore how to construct the Gateaux derivative of revenue.

We want to show how to construct an inverse welfare functional such that the government's Lagrangian has a stationary point at the given $T(\mathbf{z})$. Denoting $\mathbf{N}(\mathbf{z})$ as the set of types $\mathbf{n}$ that choose a given $\mathbf{z}$, we will show how to construct such an inverse welfare functional of the form $W(U(\mathbf{n};T)) = \sum_i \int_{\mathbf{Z}_i} \int_{\mathbf{N}(\mathbf{z})} \phi(\mathbf{n})U(\mathbf{n};T)dF(\mathbf{n}|\mathbf{z})h(\mathbf{z})d\mathbf{Z}_i$.[5] Hence, our objective is to find the weights $\phi(\mathbf{n})$ such that $T(\mathbf{z})$ is a stationary point of the government's Lagrangian:[6]

$$\sum_i \int_{\mathbf{Z}_i} \int_{\mathbf{N}(\mathbf{z})} \phi(\mathbf{n})U(\mathbf{n};T)dF(\mathbf{n}|\mathbf{z})h(\mathbf{z})d\mathbf{Z}_i + \lambda \left[\int_{\mathbf{N}} T(\mathbf{z}(\mathbf{n}))dF(\mathbf{n}) - E\right]$$

Next, we take the Gateaux derivative of the government's Lagrangian. Recall that $U(\mathbf{n};T) \equiv u(y(\mathbf{z}(\mathbf{n})) - T(\mathbf{z}(\mathbf{n})), \mathbf{z}(\mathbf{n}); \mathbf{n})$ and that $z(\mathbf{n})$ is a function of the tax schedule. The envelope theorem implies that for all $\mathbf{n}$ with a unique optima, behavioral responses to tax changes have

[5] If an individual has multiple optimal $\mathbf{z}$, we arbitrarily assign them to one $\mathbf{z}$.

[6] Note that this assumed welfare functional is linear by linearity of the integral operator and is continuous as long as all $\phi(\mathbf{n})$ are bounded by the equivalence of continuity and boundedness for linear functionals.

only second order impacts on indirect utility so that:

$$\lim_{\epsilon \to 0} \frac{U(\mathbf{n}; T + \epsilon\tau) - U(\mathbf{n}; T)}{\epsilon} = -u_c(y(\mathbf{z}(\mathbf{n})) - T(\mathbf{z}(\mathbf{n})), \mathbf{z}(\mathbf{n}); \mathbf{n})\tau(\mathbf{z}(\mathbf{n})) \equiv -u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n}))$$

Hence, as long as almost all $\mathbf{n}$ locating at each $\mathbf{z}$ have a unique optimum, we can apply the envelope theorem to write the Gateaux derivative of the government's welfare functional as:[7]

$$\frac{\partial W(U(\mathbf{n}; T + \epsilon\tau))}{\partial \epsilon} = -\sum_i \int_{\mathbf{Z}_i} \int_{\mathbf{N}(\mathbf{z})} \phi(\mathbf{n}) u_c(\mathbf{n}) \tau(\mathbf{z}) dF(\mathbf{n}|\mathbf{z}) h(\mathbf{z}) d\mathbf{Z}_i \tag{7}$$

Thus, the Gateaux derivative of the government's Lagrangian is given by:

$$-\sum_i \int_{\mathbf{Z}_i} \int_{\mathbf{N}(\mathbf{z})} \phi(\mathbf{n}) u_c(\mathbf{n}) \tau(\mathbf{z}) dF(\mathbf{n}|\mathbf{z}) h(\mathbf{z}) d\mathbf{Z}_i + \lambda \sum_i \int_{\mathbf{Z}_i} \tau(\mathbf{z}) \gamma(\mathbf{z}) h(\mathbf{z}) d\mathbf{Z}_i \tag{8}$$

To ensure that this Gateaux derivative equals zero, it suffices to ensure that for each $\mathbf{z}$:

$$\int_{\mathbf{N}(\mathbf{z})} \phi(\mathbf{n}) u_c(\mathbf{n}) dF(\mathbf{n}|\mathbf{z}) h(\mathbf{z}) = \lambda \gamma(\mathbf{z}) h(\mathbf{z}) \tag{9}$$

Hence, we can construct our inverse welfare functional pointwise for each $\mathbf{z}$ by normalizing $\lambda = 1$ (which simply rescales the inverse welfare functional multiplicatively) and choosing $\phi(\mathbf{n})$ to satisfy Equation 9.[8] If the mapping $\mathbf{n} \mapsto \mathbf{z}$ is not bijective, then in general the associated inverse welfare functional is not unique: any weights that satisfy Equation 9 for each $\mathbf{z}$ will do. For example, one could suppose that $\phi(\mathbf{n}) = \phi(\mathbf{z}(\mathbf{n}))$ so that all types $\mathbf{n}$ that choose the same $\mathbf{z}$ get the same weight. In this case, if we denote $\overline{u_c}(\mathbf{z}) \equiv \int_{\mathbf{N}(\mathbf{z})} u_c(\mathbf{n}) dF(\mathbf{n}|\mathbf{z})$, then one can determine $\phi(\mathbf{z}) = \frac{\gamma(\mathbf{z})}{\overline{u_c}(\mathbf{z})}$, which is the the budgetary impact of raising taxes infinitesimally at $\mathbf{z}$ divided by the average marginal utility of consumption at $\mathbf{z}$.

□

The intuition for Theorem 1 is as follows. In order for a tax schedule to be (locally) optimal, it must be a stationary point of some Lagrangian. By the envelope theorem, the direct welfare impact of a tax change on individuals choosing a given $\mathbf{z}$ is equal to the average welfare weighted marginal utility multiplied by the size of the tax change at that choice level. On the other hand, if government revenue is Gateaux differentiable, then there is also a budgetary impact of an infinitesimal tax perturbation at a given $\mathbf{z}$ equal to $\gamma(\mathbf{z})\tau(\mathbf{z})$. Equating these two terms pointwise, we can choose welfare weights such that the welfare impact of an infinitesimal tax change at each choice level is exactly equal to the budgetary effect of such a tax change.

There are several technical points to discuss. First, note that if many types locate at a given $\mathbf{z}$ so that $\mathbf{N}(\mathbf{z})$ is not a singleton, then there may be many inverse welfare functionals that support a given tax schedule: the Gateaux derivative of the budget only pins down the *average* inverse welfare weight at each $\mathbf{z}$. If many different types $\mathbf{n}$ pool on a particular $\mathbf{z}$ then any welfare functional that puts the requisite amount of weight on these types collectively

[7] In Appendix A.1 we only need *one* $\mathbf{n}$ at each $\mathbf{z}$ to have a unique optimum, but assuming *almost all* $\mathbf{n}$ at each $\mathbf{z}$ have a unique optimum simplifies the exposition.

[8] Rescaling the inverse welfare functional is WLOG: if a tax schedule is (locally) optimal under $W(\cdot)$, then it is also (locally) optimal under a new welfare function equal to kW for a constant k.

(i.e., satisfies Equation 8) will support the given tax schedule. Second, the assumption that $T(\mathbf{z})$ is continuous is mostly WLOG; Lemma 2 in Appendix B.2 establishes that every utility profile derived from individual optimization under some tax schedule can also be derived from individual optimization under a continuous tax schedule as long as indifference surfaces have (uniformly) bounded gradients. Third, we cannot remove the assumption that for every $\mathbf{z}$ at least one $\mathbf{n}$ that chooses $\mathbf{z}$ has a unique optimum. To see why, consider a unidimensional setting with an individual n_1 who has two optimal incomes, z^- and z^+. No other individual finds z^- or z^+ optimal other than the n_1 individual. Consider any small tax perturbation right around the z^- income level: there will be some welfare weight on type n_1 that ensures the Lagrangian remains unchanged. Similarly, consider any small tax perturbation right around the z^+ income level: there will be some welfare weight on type n_1 that ensures the Lagrangian remains unchanged. However, these two welfare weights need not coincide; hence, no matter what welfare weight we choose we can therefore always improve welfare by changing taxes at either z^- or z^+.[9] Conversely, even if type n_1 has two optimal incomes, z^- and z^+, if there are types n_2 and n_3 with a single optimum at z^- and z^+, respectively, then we can set the welfare weight on type n_1 to zero and choose the weights on n_2 and n_3 to ensure that small tax perturbations at z^- and z^+ leave the government's Lagrangian unchanged. Loosely speaking then, the existence of a local inverse welfare functionals for a given tax schedule becomes more likely as the dimension of the type space gets larger relative to the dimension of the choice space. Last and most importantly, Theorem 1 requires that government revenue is Gateaux differentiable in the tax schedule; when this condition is satisfied, Theorem 1 is a powerful result that provides an explicit construction of a local inverse welfare functional for any tax schedule that satisfies the budget constraint with equality.

3 Constructing Inverse Welfare Functionals: Examples

The next natural question then is whether most tax schedules generate a government revenue function that is Gateaux differentiable and, if so, how do we compute this Gateaux derivative? In other words, how do we actually find the object $\gamma(\mathbf{z})$ referenced in the proof to Theorem 1? Previewing ahead, Propositions 1 and 2 will establish general sufficient conditions for Gateaux differentiability of government revenue: we will show that revenue can be Gateaux differentiable even when the tax schedule is a function of multiple choices, agents have multidimensional heterogeneity, the tax schedule is non-differentiable (generating bunching), individuals have multiple optima, and when individuals face optimization frictions. To build towards Propositions 1 and 2, we first provide a number of examples illustrating how to calculate the Gateaux derivative of government revenue and apply Theorem 1 to calculate inverse welfare functionals. In doing so, we highlight how the Gateaux derivative of revenue can be expressed in terms of behavioral responses to tax perturbations and is thus, in principle, an empirically estimable object. Note, Sections 3.1, 3.2, and 3.3 do not contain any new theoretical results; readers who are familiar with computing Gateaux derivatives can skim these sections without much loss.

[9] We work through a numerical example of this situation in Appendix B.3.

3.1 Smooth Unidimensional Example

First, let us consider an example with a unidimensional type $n \in N = [\underline{n}, \overline{n}]$ with utility function $u(c, z/n)$. Suppose that we want to find an inverse welfare functional for a smooth $T(z)$ under which all individuals have a unique optima and the single crossing property holds so that $n \mapsto z$ is a strictly increasing bijection (Mirrlees, 1971). This setting has been analyzed previously (Bourguignon and Spadaro (2010); Bargain et al. (2013); Jacobs, Jongen and Zoutman (2017); Hendren (2020)), but it is useful to start here to build intuition and then move to more complex taxation settings. Concretely, we are searching for a functional $W(U)$ such that $\forall \tau \in C^1$:

$$\lim_{\epsilon \to 0} \frac{W(U(n; T + \epsilon\tau)) - W(U(n; T))}{\epsilon} + \lambda \lim_{\epsilon \to 0} \frac{R(T + \epsilon\tau) - R(T)}{\epsilon} = 0 \tag{10}$$

Let us first calculate the Gateaux variation of $R(T)$ in the direction of some $\tau(z)$ (recall that $z(n)$ is also a function of the tax schedule even though we omit it as an argument for brevity):

$$\lim_{\epsilon \to 0} \frac{R(T + \epsilon\tau) - R(T)}{\epsilon} = \int_N \frac{\partial}{\partial \epsilon} \left(T(z(n)) + \epsilon\tau(z(n))\right) f(n) dn = \int_N \left(T'(z(n)) \frac{\partial z}{\partial \epsilon}(n) + \tau(z(n)) \right) f(n) dn \tag{11}$$

We have the individual first order condition:

$$u_1(z - T(z) - \epsilon\tau(z), z/n) \left(1 - T'(z) - \epsilon\tau'(z)\right) + \frac{1}{n} u_2(z - T(z) - \epsilon\tau(z), z/n) = 0$$

For all individuals with a unique optimum where the tax schedule is twice continuously differentiable, the second order condition holds strictly (Lemma 3 of Bergstrom and Dodds (2021)), hence we can apply the implicit function theorem to determine the impact of a tax perturbation:

$$\frac{\partial z}{\partial \epsilon}(n) = \frac{u_1 \tau'(z) + \left[u_{11}(1 - T'(z)) + \frac{1}{n} u_{12}\right] \tau(z)}{u_{11}(1 - T'(z))^2 + \frac{2}{n} u_{12}(1 - T'(z)) + \frac{1}{n^2} u_{22} - T''(z) u_1} \equiv \underbrace{\xi(n)}_{\substack{\text{Substitution} \\ \text{Effect}}} \times \tau'(z(n)) + \underbrace{\eta(n)}_{\substack{\text{Income} \\ \text{Effect}}} \times \tau(z(n)) \tag{12}$$

Plugging in Equation 12 into Equation 11 and changing the variable of integration from n to z (with $h(z)$ denoting the income density) we find that:[10]

$$\lim_{\epsilon \to 0} \frac{R(T + \epsilon\tau) - R(T)}{\epsilon} = \int_{\underline{z}}^{\overline{z}} \left(T'(z)\xi(z)\tau'(z) + \left[1 + T'(z)\eta(z)\right] \tau(z) \right) h(z) dz \tag{13}$$

where $\overline{z} \equiv z(\overline{n})$ and $\underline{z} \equiv z(\underline{n})$. However, Equation 13 is not linear in $\tau(z)$ and Theorem 1 requires that tax revenue be Gateaux differentiable, which requires $\lim_{\epsilon \to 0} \frac{R(T+\epsilon\tau) - R(T)}{\epsilon}$ to be a *linear* functional of $\tau(z)$. Using integration by parts to get rid of the $\tau'(z)$ term, Equation 13 is equal to:

$$\int_{\underline{z}}^{\overline{z}} \left(-\frac{\partial}{\partial z} \left[T'(z)\xi(z)(z)h(z)\right] + \left[1 + T'(z)\eta(z)\right] h(z) \right) \tau(z) dz + T'(z)\xi(z)h(z)\tau(z) \Big|_{\underline{z}}^{\overline{z}} \tag{14}$$

Note that all $\tau(z)$ terms enter Equation 14 linearly so that Equation 14 is a linear functional of $\tau(z)$. Assuming that behavioral responses are finite (so that all terms in Equation 14 are bounded), then the Gateaux variation of $R(T)$ is a bounded - hence continuous - linear functional of $\tau(z)$; thus, $R(T)$ is Gateaux differentiable. Therefore an inverse welfare functional supporting

[10] By monotonicity, $H(z(n)) = F(n)$ so that $h(z(n)) = f(n) \left(\frac{dz}{dn}\right)^{-1}$ so that the density $h(z)$ accounts for the Jacobian of the change of variables.

$T(z)$ exists by Theorem 1. To see this, consider the following welfare functional:

$$W(U) = \int_N \phi(n)U(n;T)f(n)dn + \overline{\phi}U(\overline{n};T)h(z(\overline{n})) + \underline{\phi}U(\underline{n};T)h(z(\underline{n})) \tag{15}$$

By the envelope theorem, the derivative of indirect utility $U(n;T+\epsilon\tau) = u(z(n) - T(z(n)) - \epsilon\tau(z(n)), z(n)/n)$ with respect to ϵ evaluated at $\epsilon = 0$ equals $u_c(n)\tau(z(n)) \equiv u_c(z(n) - T(z(n)), z(n)/n)\tau(z(n))$. Hence, the Gateaux derivative of $W(U)$ in Equation 15 equals:

$$\begin{aligned}&\int_N -\phi(n)u_c(n)\tau(z(n))f(n)dn - \overline{\phi}u_c(\overline{n})\tau(z(\overline{n}))h(z(\overline{n})) - \underline{\phi}u_c(\underline{n})\tau(z(\underline{n}))h(z(\underline{n})) \\ &= \int_Z -\phi(n(z))u_c(n(z))\tau(z)h(z)dz - \overline{\phi}u_c(\overline{n})\tau(\overline{z})h(\overline{z}) - \underline{\phi}u_c(\underline{n})\tau(\underline{z})h(\underline{z})\end{aligned} \tag{16}$$

From here, we can solve for $\phi(n(z))$ by plugging in Equations 14 and 16 into Equation 10, noting that we can normalize the Lagrange multiplier to equal 1, which simply scales the inverse welfare functional multiplicatively. This yields an equation pinning down $\phi(n(z))$ for all $z \in \text{Int}(Z)$ (Equation 17) and for $z \notin \text{Int}(Z)$ (Equations 18 and 19):

$$\phi(n(z))u_c(n(z))h(z) = -\frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right] + \left[1 + T'(z)\eta(z)\right]h(z) \tag{17}$$

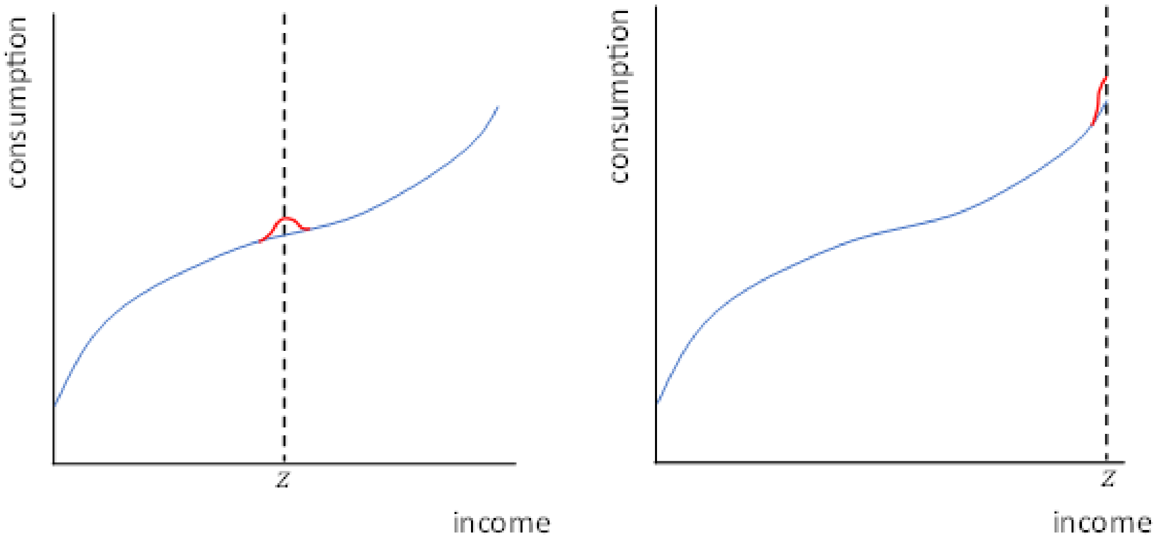
$$\overline{\phi}u_c(\overline{n})h(\overline{z}) = T'(\overline{z})\xi(\overline{z})h(\overline{z}) \tag{18}$$

$$\underline{\phi}u_c(\underline{n})h(\underline{z}) = -T'(\underline{z})\xi(\underline{z})h(\underline{z}) \tag{19}$$

Hence, we have constructed an inverse linear welfare functional such that for all tax perturbations $\tau(z)$, the net impact on the Lagrangian is zero. Intuitively, Equations 17, 18, and 19 ensure that the total impact on the government's Lagrangian of every possible "bump" perturbation is zero.[11] Equation 17 ensures that at each interior z, adding a small bump to the tax schedule as in Figure 1a leaves the Lagrangian unchanged. Conceptually, an interior bump perturbation leads to a mechanical welfare impact which, due to the envelope theorem, equals the left hand side of Equation 17. Moreover, an interior bump perturbation leads to a mechanical budgetary impact along with an income effect, $[1 + T'(z)\eta(z)]\,h(z)$, and also leads to a negative substitution effect to the right of z along with a positive substitution effect to the left of z; in the limit, this difference in substitution effects equals the (negative) derivative of the substitution effect, $-\frac{\partial}{\partial z}[T'(z)\xi(z)h(z)]$.[12] Equation 18 ensures that the impact of a perturbation at the top of the income distribution, as in Figure 1b, has no net effect on the Lagrangian. This perturbation generates a positive substitution effect to the left along with mechanical and income budgetary effects; however, the substitution effect is of higher order than the mechanical and income budgetary effects, hence only this term remains in the limit: $T'(\overline{z})\xi(\overline{z})h(\overline{z})$. If we care a discrete amount about the top income individual, then the mechanical welfare impact of this perturbation also enters the Gateaux derivative of the Lagrangian; this term is given by the left hand side of Equation 18. Identical logic explains the intuition behind Equation 19.

[11] Note, our derivations are all based on arbitrary perturbations $\tau(z)$; thus, we never use any specific bump functions in our derivations. Bump functions are merely a useful device to build intuition.

[12] Note that Equation 17 is just a differentiated version of Equation (19) from Saez (2001).



(a) Interior Bump Function Perturbation

(b) Boundary Bump Function Perturbation

Figure 1: Bump Functions to Smooth Schedule

Note: This figure shows two different "bump function" perturbations to the tax schedule (consistent with the optimal taxation literature, we depict the impact on the consumption schedule, $c = z - T(z)$). Panel 1a shows an interior bump function perturbation and Panel 1b shows a boundary bump function perturbation.

Finally, note that the inverse welfare functional we constructed from Equation 15 may require the government to care a *discrete* amount about the welfare of the individuals at the top and bottom of the income distribution (as in, for example, a Rawlsian welfare functional). These "boundary weights", which are pinned down by Equations 18 and 19, allow us to construct inverse welfare functionals that rationalize non-zero marginal tax rates at the top and bottom.[13] Note that if $h(\overline{z}) = h(\underline{z}) = 0$ then Equations 18 and 19 are vacuously satisfied; in many of our examples we will make this (reasonably sensible) assumption so that these "boundary weights" can be set to zero.

3.2 Non-Smooth Unidimensional Example

Next, let us consider another example with a unidimensional tax schedule $T(z)$ but with two dimensions of heterogeneity so that utility is given by $u(c, z/n; v)$ where the second dimension of heterogeneity is denoted by the parameter v with $(n, v) \in [\underline{n}, \overline{n}] \times [\underline{v}, \overline{v}]$. Suppose that we want to find an inverse welfare functional for a piecewise linear tax schedule with three brackets for which the budget constraint is satisfied with equality; the marginal tax rates in the three brackets are denoted T_1, T_2, T_3 with $T_2 > T_1$ and $T_2 > T_3$ (so that we have one kink point with decreasing marginal rates and one with increasing marginal tax rates; generalizing to an arbitrary number of brackets will therefore be immediate). In other words, we want to find a welfare functional such that this piecewise linear schedule is the optimal *non-linear* tax schedule.

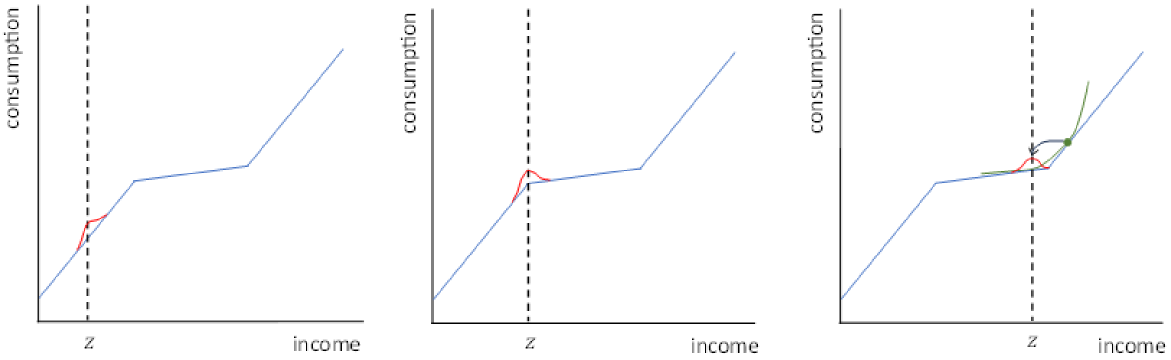
[13] Recall the classic results of Sadka (1976) and Seade (1977) which imply that marginal tax rates must be zero at the top and bottom of the income distribution under any "standard" welfare functional (which does not include mass points) as long as $h(\overline{z}), h(\underline{z}) \neq 0$.

Relative to Section 3.1, the presence of kink points generates additional behavioral responses that must be accounted for when calculating the Gateaux derivative of government revenue. First, there are individuals that bunch at the first kink, denoted K_1, where marginal tax rates increase. Let $M(K_1)$ denote the mass of types bunching at K_1. Second, there are individuals with multiple optima (one optimum in the second tax bracket and one in the third tax bracket, see Figure 2c) around the second kink, K_2; these individuals "jump" in response to tax perturbations. For all individuals with a unique optimum who do not bunch, let us use $\overline{\xi}(z)$ to denote the average substitution effect across types (n, v) locating at a given z, and define $\overline{\eta}(z)$ as the average income effect across types (n, v) locating at a given z. We denote $T_1\overline{\xi}(K_1^-)h(K_1^-)$ as $\lim_{z\to K_1^-} T'(z)\overline{\xi}(z)h(z)$ and $T_2\overline{\xi}(K_1^+)h(K_1^+)$ as $\lim_{z\to K_1^+} T'(z)\overline{\xi}(z)h(z)$. Finally, let $h(z)$ denote the income density for $z \neq K_1$. Under the simplifying assumptions that $\overline{\xi}(z)$ is differentiable except at K_1 and that $z(n, v)$ is monotonic in both arguments, we show in Appendix B.4 that the Gateaux derivative of $R(T)$ with three tax brackets is given by:

$$\begin{aligned}
&\lim_{\epsilon\to 0}\frac{R(T+\epsilon\tau)-R(T)}{\epsilon} = \\
&\underbrace{\int_{Z_1}\left(-\frac{\partial}{\partial z}\left[T_1\overline{\xi}(z)h(z)\right]+\left[1+T_1\overline{\eta}(z)\right]h(z)\right)\tau(z)dz}_{\text{Perturbations in First Bracket}} \\
&+\underbrace{T_1\overline{\xi}(K_1^-)h(K_1^-)\tau(K_1)+M(K_1)\tau(K_1)-T_2\overline{\xi}(K_1^+)h(K_1^+)\tau(K_1)}_{\text{Perturbation at Kink}} \\
&+\underbrace{\int_{Z_2}\left(-\frac{\partial}{\partial z}\left[T_2\overline{\xi}(z)h(z)\right]+\left[1+T_2\overline{\eta}(z)\right]h(z)-J_2(z)\right)\tau(z)dz}_{\text{Perturbations in Second Bracket}} \\
&+\underbrace{\int_{Z_3}\left(-\frac{\partial}{\partial z}\left[T_3\overline{\xi}(z)h(z)\right]+\left[1+T_3\overline{\eta}(z)\right]h(z)+J_3(z)\right)\tau(z)dz}_{\text{Perturbations in Third Bracket}}
\end{aligned} \tag{20}$$

where Z_1, Z_2, Z_3 represent the sets of incomes in the first, second, and third bracket, respectively, and $J_2(z)$ and $J_3(z)$ capture the budgetary impacts of individuals with multiple optima "jumping" in response to tax perurbations in the second and third brackets, respectively (these terms are defined in Appendix B.4). Importantly, Equation 20 is linear in $\tau(z)$ so that $R(T)$ is Gateaux differentiable (assuming that all the terms in Equation 20 are bounded).

Equation 20 collects the impacts of an infinite number of infinitesimal "bump" perturbations on government revenue. Because there is bunching and there are individuals with multiple optima, the impacts of these bump perturbations are more complex than in Section 3.1, but the underlying intuition is unchanged. There are three different "regions" at which we need to consider small bump perturbations, illustrated in Figure 2. First, we need to consider small perturbations to the tax schedule in the first tax bracket as in Figure 2a; because all individuals in the first bracket move smoothly, these perturbations generate a negative derivative of substitution effects (i.e., a positive substitution effect on the left and a negative substitution effect on the right) along with instantaneous mechanical and income effects as discussed in Section 3.1. Second, a bump function perturbation at the kink K_1 (as in Figure 2b) has a mechanical



(a) Perturbation in 1st Bracket (b) Perturbation at Kink (c) Perturbation in 2nd Bracket

Figure 2: Bump Function Perturbations for Piecewise Linear Schedule

Note: This figure shows three different "bump function" perturbations to a piecewise linear tax schedule (consistent with the optimal taxation literature, we depict the impact on the consumption schedule, $c = z - T(z)$). Panel 2a shows a bump function perturbation in the first bracket, Panel 2b shows a perturbation at the kink, and Panel 2c shows a perturbation at an income at which some individual has multiple optima, causing them to "jump" between optima.

effect on the bunching mass along with a positive substitution effect on the left and a negative substitution effect on the right (the elasticities are not continuous across the kink because tax rates change at the kink and because the individual just to the left of the kink is not the same as the individual just to the right). Third, a bump perturbation at an income where someone has multiple optima (as in Figure 2c) generates an additional budgetary effect of some individuals "jumping" between tax brackets. The impacts of these jumping individuals on tax revenue are given by the $J_2(z)$ and $J_3(z)$ terms in Equation 20.[14]

How can we use Equation 20 to find an inverse welfare functional? Suppose welfare is given by $\iint_{N\times V} \phi(n,v)U(n,v;T)dF(n,v) = \int_Z \iint_{N\times V} \phi(n,v)U(n,v;T)dF(n,v|z)dH(z)$. Letting $u_c(n,v) \equiv u_c(c(z(n,v)); z(n,v)/n; v)$, the Gateaux derivative of the government's Lagrangian equals (employing the envelope theorem, assuming that all but a measure zero set of types locating at each z have a unique optima):

$$-\int\limits_Z \iint\limits_{N\times V} \phi(n,v)u_c(n,v)dF(n,v|z)\tau(z)dH(z) + \lambda \lim_{\epsilon\to 0} \frac{R(T+\epsilon\tau) - R(T)}{\epsilon} \tag{21}$$

From here we note that if Equation 21 equals 0 for every $\tau(z)$, then $\phi(n,v)$ are inverse welfare weights. We can find such a set of weights by plugging in the Gateaux derivative of revenue from Equation 20 and then simply matching up all the terms that multiply $\tau(z)$ at each z. For instance, at each z in the first tax bracket, we choose welfare weights $\phi(n,v)$ such that:

$$\iint\limits_{N\times V} \phi(n,v)u_c(n,v)dF(n,v|z)h(z) = -\frac{\partial}{\partial z}\left[T_1\bar{\xi}(z)h(z)\right] + \left[1 + T_1\bar{\eta}(z)\right]h(z) \tag{22}$$

[14] Note, Equation 20 is a differentiated version of the budgetary effects from Equation (43) in Bergstrom and Dodds (2021) that also allows for non-differentiable tax schedules. Also note that Equation 20 does not feature any "boundary terms" as in Equation 14; this is due to the monotonicity assumption of $z(n,v)$ in v which ensure that the income density is zero at the top and bottom incomes.

Or at the kink K_1:

$$\iint\limits_{N\times V} \phi(n,v)u_c(n,v)dF(n,v|K_1)M(K_1) = T_1\overline{\xi}(K_1^-)h(K_1^-) + M(K_1) - T_2\overline{\xi}(K_1^+)h(K_1^+) \tag{23}$$

Or in the second tax bracket (weights for the third bracket are defined analogously):

$$\iint\limits_{N\times V} \phi(n,v)u_c(n,v)dF(n,v|z)h(z) = -\frac{\partial}{\partial z}\left[T_2\overline{\xi}(z)h(z)\right] + \left[1+T_2\overline{\eta}(z)\right]h(z) - J_2(z) \tag{24}$$

Equations 22, 23, and 24 ensure that all such bump perturbations as in Figure 2 generate a zero total impact on the government's Lagrangian. Because there are many different types that locate at each z, there are many choices of weights that satisfy the above equations: any of them will be an inverse welfare functional.[15] Thus, we have shown that even with piecewise linear schedules that generate bunching and individuals with multiple optima that (1) government revenue is Gateaux differentiable and (2) we can recover an inverse welfare functional using the logic of Theorem 1.

3.3 Smooth Multidimensional Example

Next, we consider a higher dimensional setting wherein utility is given by Equation 2. We will construct an inverse welfare functional for a smooth tax schedule $T(\mathbf{z})$. First, individual choices satisfy the following first order conditions under the perturbed schedule $T(\mathbf{z}) + \epsilon\tau(\mathbf{z})$:

$$\begin{aligned}
u_c(y(\mathbf{z}) - T(\mathbf{z}) - \epsilon\tau(\mathbf{z}), \mathbf{z}; \mathbf{n})\left(y_{z_1}(\mathbf{z}) - T_{z_1}(\mathbf{z}) - \epsilon\tau_{z_1}(\mathbf{z})\right) + u_{z_1}(y(\mathbf{z}) - T(\mathbf{z}) - \epsilon\tau(\mathbf{z}), \mathbf{z}; \mathbf{n}) &= 0 \\
&\vdots \\
u_c(y(\mathbf{z}) - T(\mathbf{z}) - \epsilon\tau(\mathbf{z}), \mathbf{z}; \mathbf{n})\left(y_{z_J}(\mathbf{z}) - T_{z_J}(\mathbf{z}) - \epsilon\tau_{z_J}(\mathbf{z})\right) + u_{z_J}(y(\mathbf{z}) - T(\mathbf{z}) - \epsilon\tau(\mathbf{z}), \mathbf{z}; \mathbf{n}) &= 0
\end{aligned} \tag{25}$$

Suppose all individuals have a unique optimum and that second order conditions hold strictly so that $\mathbf{H}(\mathbf{n})$, the Hessian matrix of second partial derivatives of u with respect to $\mathbf{z}$, is invertible. Then we can determine the derivative of $\mathbf{z}(\mathbf{n})$ with respect to ϵ for any given function $\tau(\mathbf{z})$ via the implicit function theorem (again, recall that $\mathbf{z}(\mathbf{n})$ is also a function of the tax schedule):

$$\begin{aligned}
\frac{\partial \mathbf{z}}{\partial \epsilon}(\mathbf{n}) &= \mathbf{H}^{-1}(\mathbf{n})FOC(\mathbf{n})_\epsilon|_{\epsilon=0} = \mathbf{H}^{-1}(\mathbf{n})[\mathbf{a}(\mathbf{n})\tau(\mathbf{z}) + \mathbf{B}(\mathbf{n}) \cdot \nabla_{\mathbf{z}}\tau(\mathbf{z})] \\
&\equiv \vec{\eta}(\mathbf{n})\tau(\mathbf{z}) + \mathbf{X}(\mathbf{n}) \cdot \nabla_{\mathbf{z}}\tau(\mathbf{z})
\end{aligned} \tag{26}$$

where $FOC_\epsilon|_{\epsilon=0}$ is the vector of derivatives of the first order conditions 25 with respect to ϵ. The second equality in Equation 26 follows for some vector $\mathbf{a}(\mathbf{n})$ and a matrix $\mathbf{B}(\mathbf{n})$ given that the derivative of each first order condition with respect to ϵ (evaluated at $\epsilon = 0$) is linear in τ and each component of $\nabla_{\mathbf{z}}\tau(\mathbf{z}) = (\tau_{\mathbf{z_1}}, \tau_{\mathbf{z_2}}, ..., \tau_{\mathbf{z}_J})$. The third equality in Equation 26 simply follows by defining $\vec{\eta}(\mathbf{n}) \equiv \mathbf{H}^{-1}(\mathbf{n})\mathbf{a}(\mathbf{n})$ and $\mathbf{X}(\mathbf{n}) \equiv \mathbf{H}^{-1}(\mathbf{n})\mathbf{B}(\mathbf{n})$. $\vec{\eta}(\mathbf{n})$ represents the vector of income effects (how each component of $\mathbf{z}$ changes with the tax level, τ) and $\mathbf{X}(\mathbf{n})$ represents the matrix of substitution effects (how each component of $\mathbf{z}$ changes with each marginal tax rate).

[15] For example, we could choose weights that are identical for all of those who make the same choices. In this case, the constant weight on types at each choice level in the first bracket, for example, would be equal to $\frac{-\frac{\partial}{\partial z}[T_1\overline{\xi}(z)h(z)]+[1+T_1\overline{\eta}(z)]h(z)}{\iint_{N\times V} u_c(n,v)dF(n,v|z)h(z)}$.

The Gateaux derivative of the government's Lagrangian (given by Equation 5) equals:

$$\begin{aligned}
&W(-u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n}))) + \lambda \int_{\mathbf{N}} \left(\tau(\mathbf{z}) + \nabla_{\mathbf{z}} T(\mathbf{z}(\mathbf{n})) \left[\vec{\eta}(\mathbf{n})\tau(\mathbf{z}) + \mathbf{X}(\mathbf{n}) \cdot \nabla_{\mathbf{z}}\tau(\mathbf{z})\right]\right) dF(\mathbf{n}) \\
&= W(-u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n}))) + \lambda \int_{\mathbf{Z}} \int_{\mathbf{N}(\mathbf{z})} \left(\tau(\mathbf{z}) + \nabla_{\mathbf{z}} T(\mathbf{z}) \left[\vec{\eta}(\mathbf{n})\tau(\mathbf{z}) + \mathbf{X}(\mathbf{n}) \cdot \nabla_{\mathbf{z}}\tau(\mathbf{z})\right]\right) dF(\mathbf{n}|\mathbf{z}) dH(\mathbf{z}) \qquad (27) \\
&= W(-u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n}))) + \lambda \int_{\mathbf{Z}} \left(\tau(\mathbf{z}) + \nabla_{\mathbf{z}} T(\mathbf{z}) \left[\overline{\vec{\eta}}(\mathbf{z})\tau(\mathbf{z}) + \overline{\mathbf{X}}(\mathbf{z}) \cdot \nabla_{\mathbf{z}}\tau(\mathbf{z})\right]\right) dH(\mathbf{z})
\end{aligned}$$

where $u_c(\mathbf{n}) \equiv u_c(c(\mathbf{z}(\mathbf{n})), \mathbf{z}(\mathbf{n}); \mathbf{n})$ and the Gateaux derivative of $W(U(\mathbf{n}; T))$ is calculated via the envelope theorem. The first equality in System 27 first integrates the budget constraint over types $\mathbf{n}$ that choose a given $\mathbf{z}$ and then integrates over $\mathbf{z}$; the second equality evaluates the inner integral, representing the budgetary Gateaux derivative as a function of the average behavioral effects at each $\mathbf{z}$: $\overline{\vec{\eta}}(\mathbf{z})$ and $\overline{\mathbf{X}}(\mathbf{z})$. As before, we need to manipulate Equation 27 to get rid of the derivatives of $\tau(\mathbf{z})$ by appealing to multi-dimensional integration by parts:

Lemma 1 (Multidimensional Integration by Parts)**.** *For a continuously differentiable function $\tau(\mathbf{z})$ and a continuously differentiable vector field $\mathbf{v}(\mathbf{z})$, where $\mathbf{Z} \in \mathbb{R}^J$ is connected, bounded, and open with piecewise smooth boundary $\partial\mathbf{Z}$, we have the following identity:*

$$\int_{\mathbf{Z}} \mathbf{v}(\mathbf{z}) \cdot \nabla_{\mathbf{z}}\tau(\mathbf{z}) d\mathbf{z} = \int_{\partial\mathbf{Z}} \mathbf{v}(\mathbf{z})\tau(\mathbf{z}) \cdot \rho dS - \int_{\mathbf{Z}} [\nabla_{\mathbf{z}} \cdot \mathbf{v}(\mathbf{z})]\tau(\mathbf{z}) d\mathbf{z}$$

where ρ is the outward-pointing unit normal vector to $\partial\mathbf{Z}$ and dS is the boundary element.[16]

Assuming that the average behavioral effects $\overline{\mathbf{X}}(\mathbf{z})$ are smooth and the distribution of incomes $H(\mathbf{z})$ admits a differentiable density function $h(\mathbf{z})$, we can appeal to Lemma 1 (recognizing that $\nabla_{\mathbf{z}} T(\mathbf{z})\overline{\mathbf{X}}(\mathbf{z})h(\mathbf{z})$ is a vector field on $\mathbf{Z}$) to rewrite Equation 27 as:

$$\begin{aligned}
&W(-u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n}))) + \lambda \int_{\mathbf{Z}} \left(\left[1 + \nabla_{\mathbf{z}} T(\mathbf{z})\overline{\vec{\eta}}(\mathbf{z})\right] h(\mathbf{z}) - \nabla_{\mathbf{z}} \cdot \left[\nabla_{\mathbf{z}} T(\mathbf{z})\overline{\mathbf{X}}(\mathbf{z})h(\mathbf{z})\right]\right) \tau(\mathbf{z}) d\mathbf{z} \\
&+ \lambda \int_{\partial\mathbf{Z}} \nabla_{\mathbf{z}} T(\mathbf{z})\overline{\mathbf{X}}(\mathbf{z})\tau(\mathbf{z})h(\mathbf{z}) \cdot \rho dS
\end{aligned} \qquad (28)$$

Importantly, note that Equation 28 is *linear* in $\tau(\mathbf{z})$ so that revenue is Gateaux differentiable in the tax schedule (assuming all terms in Equation 28 are bounded so that the Gateaux derivative is a bounded - hence, continuous - linear functional). To construct an inverse welfare functional, suppose that the inverse welfare functional takes the following form:

$$W(U) = \int_{\mathbf{Z}} \int_{\mathbf{N}(\mathbf{z})} \phi(\mathbf{n}) U(\mathbf{n}; T) dF(\mathbf{n}|\mathbf{z}) h(\mathbf{z}) d\mathbf{z} + \int_{\partial\mathbf{Z}} \int_{\mathbf{N}(\mathbf{z})} \phi(\mathbf{n}) U(\mathbf{n}; T) dF(\mathbf{n}|\mathbf{z}) h(\mathbf{z}) dS$$

Then we can write the Gateaux derivative of $W(U)$ as:

$$-\int_{\mathbf{Z}} \int_{\mathbf{N}(\mathbf{z})} \phi(\mathbf{n}) u_c(\mathbf{n}) \tau(\mathbf{z}) dF(\mathbf{n}|\mathbf{z}) h(\mathbf{z}) d\mathbf{z} - \int_{\partial\mathbf{Z}} \int_{\mathbf{N}(\mathbf{z})} \phi(\mathbf{n}) u_c(\mathbf{n}) dF(\mathbf{n}|\mathbf{z}) \tau(\mathbf{z}) h(\mathbf{z}) dS$$

[16] We often assume $\mathbf{Z}$ is compact; this is not problematic as Lemma 1 can also be applied if $\mathbf{Z}$ is the closure of an open set as the inclusion of the (measure zero) boundary does not impact the integrals $\int_{\mathbf{Z}} \mathbf{v}(\mathbf{z}) \cdot \nabla_{\mathbf{z}}\tau(\mathbf{z}) d\mathbf{z}$ and $\int_{\mathbf{Z}} [\nabla_{\mathbf{z}} \cdot \mathbf{v}(\mathbf{z})]\tau(\mathbf{z}) d\mathbf{z}$.

To satisfy Equation 28, we choose $\phi(\mathbf{n})$ for those locating at each $\mathbf{z} \in \text{Int}(\mathbf{Z})$ such that:[17]

$$\int_{\mathbf{N}(\mathbf{z})} \phi(\mathbf{n}) u_c(\mathbf{n}) dF(\mathbf{n}|\mathbf{z}) h(\mathbf{z}) = \left[1 + \nabla_{\mathbf{z}} T(\mathbf{z}) \overline{\vec{\eta}}(\mathbf{z})\right] h(\mathbf{z}) - \nabla_{\mathbf{z}} \cdot \left[\nabla_{\mathbf{z}} T(\mathbf{z}) \overline{\mathbf{X}}(\mathbf{z}) h(\mathbf{z})\right] \tag{29}$$

and we choose $\phi(\mathbf{n})$ for those locating at each $\mathbf{z} \in \partial\mathbf{Z}$:

$$\int_{\mathbf{N}(\mathbf{z})} \phi(\mathbf{n}) u_c(\mathbf{n}) dF(\mathbf{n}|\mathbf{z}) h(\mathbf{z}) = \nabla_{\mathbf{z}} T(\mathbf{z}) \overline{\mathbf{X}}(\mathbf{z}) h(\mathbf{z}) \cdot \rho \tag{30}$$

If Equations 29 and 30 are satisfied, then Equation 28 is equal to zero for all $\tau(\mathbf{z})$.

An important takeaway from Sections 3.1, 3.2, and 3.3 is that the Gateaux derivative of government revenue is an empirical object that depends on substitution effects, income effects, jumping effects, bunching masses, and the density of choices $\mathbf{z}$. In principle, all of these objects can be estimated given sufficient tax variation. In practice, it is difficult to observe sufficient tax variation to estimate the requisite heterogeneous behavioral responses to tax reforms; hence, practitioners attempting to construct inverse welfare functionals will typically need to make simplifying assumptions such as making structural assumptions on utility and calibrating or assuming elasticities are constant across the choice distribution (which is implicitly a type of structural assumption).

3.4 Sufficient Conditions for $R(T)$ to be Gateaux Differentiable

Section 3.2 shows that $R(T)$ can be Gateaux differentiable even if there is bunching and/or individuals who have multiple optima in a unidimensional setting. Section 3.3 shows that $R(T)$ can be Gateaux differentiable in a multidimensional setting. We now show that we can combine these two scenarios by providing general sufficient conditions for $R(T)$ to be Gateaux differentiable:

Proposition 1. *The following are sufficient conditions for $R(T)$ to be Gateaux differentiable:*

1. *The tax schedule is twice continuously differentiable except across some closed finite set of measure zero surfaces*
2. *Individuals have multiple optima only along a finite set of measure zero surfaces in* $\mathbf{N}$
3. *The set* $\mathbf{Z}$ *of chosen* $\mathbf{z} = (z_1, z_2, ..., z_J)$ *is the closure of an open set in* $\mathbb{R}^J$
4. *The set of individuals whose second order conditions hold weakly is measure zero*
5. *(5 technical regularity conditions discussed in the proof)*

Proof. See Appendix B.5. □

The key takeaway from Proposition 1 is that government revenue can be Gateaux differentiable even if the tax schedule is multidimensional, agent heterogeneity is multidimensional, and the tax schedule features various "non-smooth" properties. For instance, the tax schedule can be non-differentiable causing people to bunch and/or create individuals with multiple optima so that the mapping $\mathbf{n} \mapsto \mathbf{z}$ is not smooth and bijective. This also allows for individuals responding

[17] The first order conditions in Equations 29 and 30 have been derived previously in Golosov, Tsyvinski and Werquin (2014) and Spiritus et al. (2022).

on the extensive margin by entering/exiting the workforce; we illustrate how to compute the Gateaux derivative of revenue with extensive margin responses in Appendix B.6.

While the proof to Proposition 1 is quite long, let us give a sketch of the intuition. The goal is to show that we can express the impact of any tax change as a continuous linear functional of $\tau(\mathbf{z})$. The idea is to split up the set of choices, $\mathbf{Z}$, into regions where people respond smoothly to tax changes, regions where people may "jump" between multiple optima, and regions where the tax schedule is non-differentiable causing bunching. Our assumptions in Proposition 1 ensure that at most $\mathbf{z}$, the tax schedule is smooth and individuals have a unique optimum with their second order condition holding strictly (i.e., their Hessian matrix is negative definite); hence, individuals who choose these incomes respond to tax changes according to the implicit function theorem (i.e., via standard income and substitution effects). For these individuals, we can therefore use multidimensional integration by parts as in Section 3.3 to express the revenue impact of a tax change as a linear functional of $\tau(\mathbf{z})$. There may be surfaces where individuals have multiple optima and thereby react to some tax changes by jumping to a different $\mathbf{z}$; however, under the stated assumptions, we can show that these jumping effects can be expressed as a linear functional of the tax level changes along the surface because the decision to jump only depends on the tax level (not marginal tax rates) at each $\mathbf{z}$. Finally, there are surfaces along which the tax schedule is non-differentiable. For purposes of intuition, consider the case when there are two choice variables. A non-differentiable surface for the tax schedule in this case is a ridge in three dimensional space (visually, imagine a creased piece of paper). Almost all individuals who choose $\mathbf{z}$ on this ridge strictly prefer their chosen $\mathbf{z}$ to any $\mathbf{z}$ that is off the ridge (consider an indifference surface that is tangent to a given $\mathbf{z}$ on the ridge). Hence, in response to any small tax perturbation, these individuals may move *along* the ridge but do not move off the ridge. For small tax perturbations we can then recast the optimization problem for individuals on the ridge as a choice over some parameter t which parameterizes the ridge. Thus, for these individuals, we can reduce their problem to a unidimensional optimization problem wherein we can use integration by parts (integrating over the parameter t) to express the revenue impact of a tax change as a linear functional just as in Section 3.1.[18]

In summary, Proposition 1 proves that Gateaux differentiability of government revenue is a relatively mild restriction. Thus, taken together, Theorem 1 and Proposition 1 establish that one can construct inverse welfare functionals in a wide variety of situations.

3.5 Labor Supply Frictions

Thus far, we have assumed that agents make choices in a frictionless environment, but there is a growing body of evidence suggesting that agents face a variety of labor supply frictions (Chetty, 2012). We now discuss how the theory presented so far can be adapted when agents face "sparsity-based frictions". In particular, suppose that rather than selecting among a continuum of choices, agents select from a sparse set of limited choices. As discussed in Anagol et al.

[18] The same intuition applies in cases where $\mathbf{Z} \subset \mathbb{R}^J$ for $J > 2$: we think of behavioral responses for those who locate along non-differentiable surfaces as smooth responses on a lower dimensional manifold.

(2022), this sparisty based models of frictions is quite general and can accommodate a variety of microfoundations, ranging from choices among full-time/part-time/no work, choices among professions, costly search, rational inattention, or imperfect targeting.[19]

Definition 4. *[Sparsity Based Frictions] Agents face "sparsity based frictions" if they make choices over* $\mathbf{z} = (z_1, z_2, ..., z_J)$ *subject to the restriction that* $\mathbf{z} \in A(\mathbf{n})$ *for some discrete set of choices* A *that may vary across individuals* $\mathbf{n}$*.*

Importantly, the proof of Theorem 1 does not actually make any assumptions about the choice sets of agents so that:

Remark 4. *Theorem 1 holds when agents face sparsity based frictions.*

Given Remark 4, the natural question is then whether Gateaux differentiability of government is a reasonable requirement when agents face heterogeneous discrete choice sets. We have:

Proposition 2. *The following are sufficient conditions for* $R(T)$ *to be Gateaux differentiable if agents face sparsity based frictions:*

1. *Individuals have multiple optima only along a finite set of measure zero surfaces in* $\mathbf{N}$
2. *The set of individuals with more than two optima is measure zero restricted to the set of surfaces of those who have multiple optima (i.e., almost all individuals with multiple optima just have two optima).*

Proof. See Appendix B.7. □

To illustrate Proposition 2, consider agents who make a choice to work full-time, part-time, or not at all. Agents differ in terms of labor productivity n as well as a parameter a that determines their choice set of incomes: $\{0, a/2, a\}$. Conditional on a value of a, as long as utility $u(c, z; n)$ satisfies the single crossing property then choice of income is monotonic in n (Mirrlees, 1971), so that government revenue can be written:

$$\int_A \left\{ \int_{\underline{n}}^{n_1(a)} [T(0) + \epsilon\tau(0)] f(n|a) dn + \int_{n_1(a)}^{n_2(a)} [T(a/2) + \epsilon\tau(a/2)] f(n|a) dn + \int_{n_2(a)}^{\overline{n}} [T(a) + \epsilon\tau(a)] f(n|a) dn \right\} f(a) da$$

where type $n_1(a)$ is indifferent between earning 0 and $a/2$ and type $n_2(a)$ is indifferent between earning $a/2$ and a:

$$u(-T(0), 0; n_1(a)) = u(a/2 - T(a/2), a/2; n_1(a)) \tag{31}$$
$$u(a/2 - T(a/2), a/2; n_2(a)) = u(a - T(a), a; n_2(a)) \tag{32}$$

Taking the Gateaux derivative of revenue, we get the following, recognizing that the indifferent individuals $n_1(a)$ and $n_2(a)$ change with the tax schedule (representing individuals

[19] Our framework can also accommodate other sorts of labor supply frictions as well. For example, suppose agents make income choices over time so that z_i represents income in time period i; if agents face adjustment costs whenever $z_i \neq z_{i-1}$, this simply changes the utility function. In this case, Theorem 1 and Proposition 1 hold without modification. Theorem 1 does not necessarily hold if agents misperceive the tax schedule or their own utility function; we discuss this more in Section 8.

"jumping" between multiple optima):

$$\int_A \Bigg\{ \underbrace{\int_{\underline{n}}^{n_1(a)} \tau(0) f(n|a) dn + \int_{n_1(a)}^{n_2(a)} \tau(a/2) f(n|a) dn + \int_{n_2(a)}^{\overline{n}} \tau(a) f(n|a) dn}_{\text{Mechanical Effect}} \Bigg\} f(a) da$$
$$+ \int_A \Bigg\{ \underbrace{[T(0) - T(a/2)] f(n_1(a)|a) \frac{\partial n_1(a)}{\partial \epsilon}}_{\text{Extensive Effect}} + \underbrace{[T(a/2) - T(a)] f(n_2(a)|a) \frac{\partial n_2(a)}{\partial \epsilon}}_{\text{Intensive Effect}} \Bigg\} f(a) da \tag{33}$$

where $\frac{\partial n_1(a)}{\partial \epsilon}$ and $\frac{\partial n_2(a)}{\partial \epsilon}$ come from plugging in the tax schedule $T(z) + \epsilon\tau(z)$ and differentiating Equations 31 and 32 w.r.t. ϵ. We show in Appendix B.8 that all terms in Equation 33 are linear in $\tau(z)$ and so revenue is, as claimed, Gataux differentiable. From here, we can follow the procedure outlined in Section 3.2 to construct an inverse welfare functional: form the Gateaux derivative of the government's Lagrangian, set the Lagrange multiplier $\lambda = 1$, and collect terms that multiply a given $\tau(z)$ for each z to solve for the aggregate welfare weight that the government must put on types that choose each income z (see Appendix B.8 for more details).

4 Policy Relevance

Next, we discuss how can we use inverse welfare functionals to inform tax policy. First, we show that inverse welfare functionals can be used to determine the desirability of tax reforms:

Proposition 3. *Suppose that the conditions of Theorem 1 hold and suppose that the inverse welfare function can be written as:*[20]

$$\int_{\mathbf{Z}} \int_{\mathbf{N(z)}} \phi(\mathbf{n}) U(\mathbf{n}; T) dF(\mathbf{n}|\mathbf{z}) h(\mathbf{z}) d\mathbf{z}$$

Then if for some compact set $\mathbf{V} \in \mathbf{Z}$*, we have that true societal welfare weights* $\phi^T(\mathbf{n})$ *satisfy:*

$$\int_{\mathbf{N(v)}} \phi^T(\mathbf{n}) u_c(\mathbf{n}) dF(\mathbf{n}|\mathbf{v}) < \int_{\mathbf{N(v)}} \phi(\mathbf{n}) u_c(\mathbf{n}) dF(\mathbf{n}|\mathbf{v}) \ \forall \mathbf{v} \in \mathbf{V} \tag{34}$$

and the true weights and inverse weights have the same normalization:

$$\int_{\mathbf{N}} \phi^T(\mathbf{n}) u_c(\mathbf{n}) dF(\mathbf{n}) = \int_{\mathbf{N}} \phi(\mathbf{n}) u_c(\mathbf{n}) dF(\mathbf{n}) \tag{35}$$

then any small tax perturbation that (weakly) increases taxes at choices in $\mathbf{V}$ *and uses this money to fund a lump sum transfer is welfare improving. If the inequality in Equation 34 is reversed then tax perturbations that decrease taxes at choices in* $\mathbf{V}$ *(and finance these tax decreases via decreasing the lump sum transfer) are welfare improving.*

Proof. See Appendix A.2. □

We also have the following corollary, which holds because the sum of two small welfare improving perturbations is welfare improving by linearity of the Gateaux derivative:

Corollary 3.1. *Suppose that the conditions of Theorem 1 hold and that the inverse welfare*

[20] The proposition holds more generally. Under a continuous, linear welfare function, we can always write welfare as follows by the Disintegration Theorem: $\int_{\mathbf{Z}} \int_{\mathbf{N(z)}} U(\mathbf{n}; T) d\Phi_1(\mathbf{n}|\mathbf{z}) d\Phi_2(\mathbf{z})$. Then the theorem can be shown to hold if we replace Equation 34 with the following "statewise dominance" condition such that for all positive functions $\tau(\mathbf{v})$: $\int_{\mathbf{V}} \tau(\mathbf{v}) \int_{\mathbf{N(v)}} u_c(\mathbf{n}) d\Phi_1^T(\mathbf{n}|\mathbf{v}) d\Phi_2^T(\mathbf{v}) < \int_{\mathbf{V}} \tau(\mathbf{v}) \int_{\mathbf{N(v)}} u_c(\mathbf{n}) d\Phi_1(\mathbf{n}|\mathbf{v}) d\Phi_2(\mathbf{v})$.

function can be written as in Proposition 3. Suppose there are two compact sets $\mathbf{V}_1$ and $\mathbf{V}_2$ s.t.:

$$\int_{\mathbf{N(v)}} \phi^T(\mathbf{n})U(\mathbf{n};T)dF(\mathbf{n}|\mathbf{v}) < \int_{\mathbf{N(v)}} \phi(\mathbf{n})U(\mathbf{n};T)dF(\mathbf{n}|\mathbf{v}) \ \forall \mathbf{v} \in \mathbf{V}_1 \tag{36}$$

$$\int_{\mathbf{N(v)}} \phi^T(\mathbf{n})U(\mathbf{n};T)dF(\mathbf{n}|\mathbf{v}) > \int_{\mathbf{N(v)}} \phi(\mathbf{n})U(\mathbf{n};T)dF(\mathbf{n}|\mathbf{v}) \ \forall \mathbf{v} \in \mathbf{V}_2 \tag{37}$$

Then any tax reform that increases taxes for choices in $\mathbf{V}_1$, decreases taxes for choices in $\mathbf{V}_2$, and balances the budget by changing the lump transfer, is welfare improving.

As an example application of Proposition 3, suppose that society's *true* welfare weights for households earning more than \$500,000 are lower than the *inverse* welfare weights for these households. Then *any* small tax reform that increases taxes on households earning more than \$500,000 and uses this money to fund a universal basic income is welfare improving. If, in addition, society's *true* welfare weights for households earning less than \$30,000 are higher than the corresponding *inverse* welfare weights, then any small budget neutral tax reform that reduces taxes on those earning less than \$30,000 and increases taxes on those earning more than \$500,000 is welfare improving by Corollary 3.1.

Inverse welfare functionals can also be used to test for Pareto efficiency of complex tax schedules. There has been a good amount of recent work characterizing Pareto efficient tax schedules in unidimensional settings (e.g., Werning (2007), Lorenz and Sachs (2016), Scheuer and Werning (2016), Bierbrauer, Boyer and Hansen (2023), Sturm and Sztutman (2023)) as well as Proposition 4 of Spiritus et al. (2022), which proves a necessary condition for Pareto efficiency in a multidimensional setting. Our Proposition 4 below extends these previous results in two ways: (1) our framework places fewer restrictions on the tax schedule and the behavioral responses of individuals (e.g., tax schedules can be non-differentiable generating bunching, individuals can face optimization frictions, individuals can respond on the extensive margin) and (2) we provide a sufficient condition in addition to a necessary condition for Pareto efficiency:

Proposition 4. *The following necessary and sufficient conditions hold for Pareto efficient tax schedules:*

1. *Suppose $R(T)$ is Gateaux differentiable and $T(\mathbf{z})$ satisfies the budget constraint with equality. $T(\mathbf{z})$ is Pareto efficient only if the Gateaux derivative of $R(T)$, $DR_T(\tau)$, is positive: $DR_T(\tau) > 0$ when $\tau(\mathbf{z}) > 0 \ \forall \mathbf{z}$. If, additionally, almost all $\mathbf{n}$ choosing each $\mathbf{z}$ have a unique optima, then $T(\mathbf{z})$ is Pareto efficient only if $\exists$ a positive (local) inverse welfare functional so that $W(U) > 0$ when $U(\mathbf{n}) > 0 \ \forall \mathbf{n}$.*
2. *$T(\mathbf{z})$ is Pareto efficient if there exists a (global) inverse welfare functional for $T(\mathbf{z})$ which can be written as $\sum_i^M \int_{\mathbf{N}_i} \phi_i(\mathbf{n})U(\mathbf{n};T)d\mathbf{N}_i$ with $\phi_i(\mathbf{n}) > 0$ and with mutually disjoint N_i such that $N_1 \cup N_2 \cup \cdots \cup N_M = \mathbf{N}$.*

Proof. See Appendix A.3. □

The intuition for Proposition 4 is as follows: if the Gateaux derivative of $R(T)$ is not positive then there exists a tax perturbation that reduces taxes yet increases government revenue (i.e., portions of the tax schedule are beyond the local Laffer rate). Similarly, if the tax schedule is (globally) optimal under some linear welfare functional with positive welfare weights, then it must be Pareto optimal otherwise the Pareto improvement would be welfare improving as well.[21] There are two takeaways from Proposition 4. First, we can use the inverse welfare functional to determine whether a tax schedule is Pareto efficient. Second, Pareto efficient schedules are associated with positive Gateaux derivatives of government revenue and positive local inverse welfare functionals. Thus, existence of a local inverse functional is *weaker* than Pareto efficiency. There are many tax schedules that are not Pareto efficient yet still have associated local inverse optimal welfare functionals: such schedules simply feature negative welfare weights.

5 Numerical Computation of Inverse Welfare Functionals

As a proof of concept, we will now illustrate numerical computations of inverse welfare functionals for various tax systems. We begin with a baseline case, calculating the inverse welfare functional for a smoothed version of the U.S. income tax schedule as in Hendren (2020). We then show how the inverse welfare functional changes with: (1) a non-smooth tax schedule, (2) sparsity-based frictions, and (3) a multidimensional tax schedule of joint income and property taxation.

5.1 Baseline Scenario

Our examples will take a "structural" approach in that we will use the observed income distribution and estimated elasticities to calibrate a utility function and a distribution of primitives and then calculate the Gateaux derivative of revenue from behavioral responses under this utility function. Alternatively, one could use the observed income distribution and estimated behavioral responses to taxes in a "sufficient statistics" approach to construct inverse welfare functions, recognizing that, in practice, sufficient statistics approaches require making implicit structural assumptions about how elasticities vary across the choice distribution given a lack of heterogeneous elasticity estimates.[22]

Suppose that individuals vary in terms of labor productivity n as well as a fixed cost of working v, which generates extensive margin effects (we show how to calculate the Gateaux derivative of revenue with extensive margin effects in Appendix B.6):

$$\begin{aligned} u(c, z; n, v) &= c - \frac{(z/n)^{1+k}}{1+k} - v\mathbb{1}[z > 0] \\ c &= z - T(z) \end{aligned} \tag{38}$$

[21] While most of our analysis has concerned *local* inverse functionals, under the conditions of Proposition 6 in Appendix B.1, a unique local inverse welfare functional is a global inverse welfare functional. These conditions hold for some common classes of utility functions, see Remark 5.

[22] We employ a structural approach for comparability between exercises given that the sufficient statistics approach is difficult to apply in some situations due to a lack of tax variation. For instance, in the model of joint income and property taxation in Section 5.4, the sufficient statistic approach would require estimates of heterogeneous own-tax and cross-tax elasticities of both income and housing across the joint distribution of income and housing. Given the lack of tax variation to estimate these heterogeneous elasticities, we would have to make assumptions about how they vary across the joint distribution. Such assumptions are arguably less transparent than the structural approach, which makes assumptions on the utility function directly and calibrates to match average reduced-form elasticities.

We choose the parameter k to match the intensive margin taxable income elasticity of 0.3 (Saez, Slemrod and Giertz, 2012). We calibrate the distribution of types $f(n, v)$ to match the U.S. income distribution in 2019 from the ACS, the fraction of the population that is unemployed, and an extensive margin taxable income elasticity of 0.25 (Chetty et al., 2013).[23] We first compute inverse welfare weights for a smooth tax schedule $T(z)$ that approximates the U.S. combined tax schedule of income and payroll taxes for single individuals (shown in Figure 12 in Appendix C).[24] We plot these inverse weights against income in Figure 3.[25] Before we explore how various aspects of the economic environment impact inverse welfare weights it is worthwhile to reiterate how to interpret these inverse weights. The inverse welfare weight at an income level represents the implicit value society places on giving \$1 to households at that income level relative to the value of splitting a dollar equally among the population (i.e., welfare weights are normalized to integrate to 1 so that the welfare increase from splitting a dollar equally among the population equals 1). By Proposition 3, if society's true value of giving money to those with a certain income is higher (lower) than the inverse welfare weight, then a tax reform that locally lowers (raises) taxes at that income level and closes the budget by changing the lump sum transfer is welfare improving. For example, in Figure 3, if society truly values giving a dollar to those earning \$400,000 per year less than 20% as much as splitting a dollar evenly among the population, then this implies that raising taxes for those earning \$400,000 per year is welfare improving.

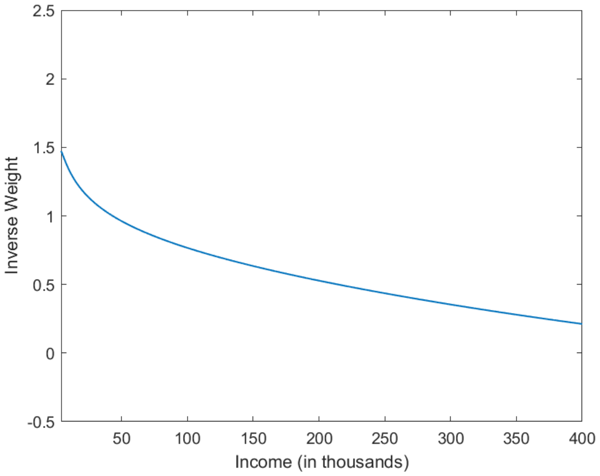


Figure 3: Inverse Welfare Weights for Smooth Approximation to U.S. Income Tax Schedule

Note: This figure shows inverse welfare weights across the income distribution for a 5th degree polynomial $T(z)$ chosen to approximate the U.S. combined tax schedule of income and payroll taxes in Figure 12. Households are assumed to maximize utility function 38 with the calibration described in Section 5.1.

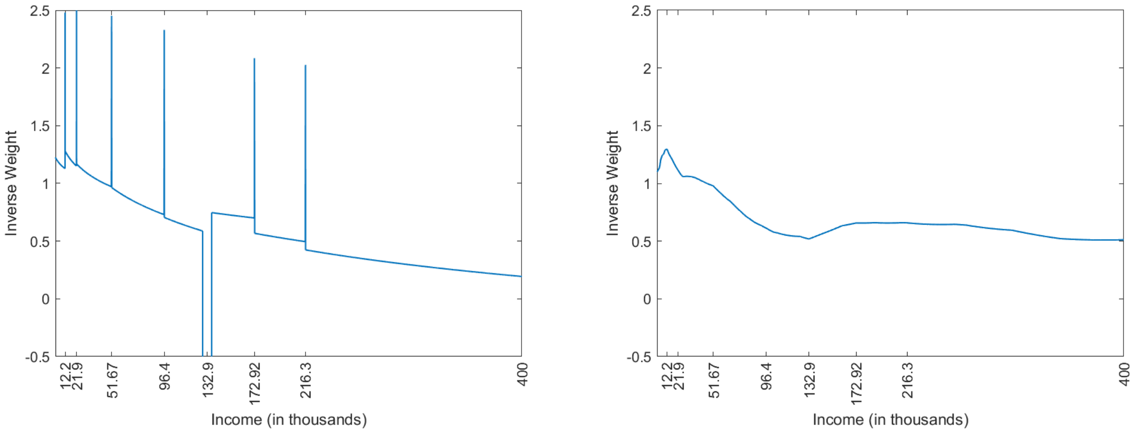
[23] The extensive margin elasticity is defined as the percent change in employment that results from a one percent change in after-tax income.

[24] We abstract from marriage penalty concerns by assuming that couples are just taxed on their average income at the same rate as singles.

[25] Given that multiple (n, v) types pool on each income level z, these inverse welfare weights should be interpreted as *average* welfare weights at each income z.

5.2 Piecewise Linear Income Tax Schedule

Next, we explore how inverse welfare weights change relative to the baseline scenario when we use the actual U.S. piecewise linear income tax schedule instead of a smooth approximation. We continue to assume that households choose income to maximize utility function 38 (the parameter k and the distribution of types $f(n, v)$ are unchanged from the baseline scenario).[26] The presence of kinks in the piecewise linear tax schedule induces bunching (around kinks where marginal tax rates increase) and individuals with multiple optima that "jump" in response to tax changes (at kinks where marginal tax rates decrease). We plot inverse welfare weights for this exercise in Figure 4a.



(a) No Frictions

(b) Sparsity Based Frictions

Figure 4: Inverse Welfare Weights for U.S. Income Tax Schedule

Note: Figure 4a shows inverse welfare weights across the income distribution for the U.S. combined tax schedule of income and payroll taxes shown in Figure 12. Households are assumed to maximize utility function 38 with the calibration described in Section 5.2. Figure 4b shows inverse welfare weights across the income distribution for the U.S. combined tax schedule of income and payroll taxes shown in Figure 12, but agents are assumed to face sparsity-based frictions by solving Equation 39 with the calibration described in Section 5.3.

There are two takeaways from this exercise. First, the inverse welfare weights jump up discretely at kink points where marginal tax rates increase. Intuitively, in order to rationalize kink points that generate bunching (as opposed to smoothing out these kink points and thereby raising taxes on the bunching individuals), we need to have large welfare weights on the bunching individuals. Second, the inverse welfare weights are extremely negative right around the kink point where marginal tax rates decrease (corresponding to the maximum income on which social security taxes are levied): this implies that the presence of this kink point is Pareto inefficient (i.e., the government can raise tax revenue by lowering tax rates around the kink point) as discussed in Proposition 4.[27] Thus, rationalizing the presence of a piecewise linear tax schedule requires a non-smooth, somewhat pathological inverse welfare functional.

[26] We actually assume there is a small amount of heterogeneity in k (elasticities range from 0.29-0.31) for this exercise so that an inverse welfare functional exists; if individuals have multiple optima and there is no k heterogeneity, then inverse welfare functionals typically do not exist as discussed in Section 2.3.

[27] Note, decreasing marginal tax rates lead to individuals with multiple optima, which is always Pareto inefficient for sufficiently limited elasticity heterogeneity by Proposition 3 of Bergstrom and Dodds (2021).

5.3 Sparsity-Based Frictions

A conceptual issue with the analysis in Sections 5.1 and 5.2 (as well as all prior work on inverse welfare weights, to the best of our knowledge) is that inverse weights are derived assuming a frictionless environment in which individuals freely choose their income and can seamlessly adjust this income on the intensive margin in response to small tax changes. However, in reality we have evidence from the lack of bunching at kink points (Saez, 2010) to suggest that individuals face some sorts of labor supply frictions. We now explore how to reconcile this inconsistency with a model of sparsity-based frictions in which households are prevented from bunching as a result of their limited choice set.

Suppose that individuals solve the following maximization problem, choosing whether to work full-time, part-time, or not at all as in Section 3.5, modified to include a fixed cost of working:

$$\begin{aligned} &\max_{z\in\{0,a/2,a\}} c - \frac{(z/n)^{1+k}}{1+k} - v\mathbb{1}[z>0] \\ &c = z - T(z) \end{aligned} \tag{39}$$

Households differ in terms of three parameters: (1) their choice set which is determined by a, (2) their disutility of working which is determined by n, and (3) their fixed cost of working v (we assume the parameter k is homogenous across the population and is equal to the calibrated value from Section 5.1). We calibrate the distribution of a, $f(a)$, to match the distribution of full-time equivalent incomes in the population from the 2019 ACS (i.e., for part-time workers we scale their income by 2 to get a full-time equivalent income). To calibrate $f(n, v|a)$, we first define two elasticities. First, we define the extensive margin elasticity as in Section 5.1, which is the percent change in employment that results from a one percent change in after-tax income. Second, even though workers cannot modify their labor supply on the intensive margin, we define a "modified intensive margin elasticity" as the percent change in average income, conditional on employment, that results from a one percent change in the keep rate (1 minus the marginal tax rate). For instance, if there are 1,000 individuals working full time earning \$100,000 and we change the keep rate by 1% and 6 of these 1,000 individuals change to working part-time (earning \$50,000) in response, the "modified intensive margin elasticity" at \$100,000 equals $0.3 = \frac{6\times\$50{,}000/(1{,}000\times\$100{,}000)}{0.01}$. We calibrate the conditional distribution $f(n, v|a)$ for each a to match four moments: the share of unemployed workers with full-time equivalent income a (where we predict full-time potential income for unemployed individuals using a simple Mincer regression), the share of part-time workers with full-time equivalent income a, an average extensive margin elasticity of 0.25 (Chetty et al., 2013), and a "modified intensive margin elasticity" elasticity of 0.3 (Saez, Slemrod and Giertz, 2012).

Figure 4b shows inverse welfare weights computed under this calibration for the actual piecewise linear U.S. tax schedule. First, note that the inverse weights computed under the sparsity-based frictions model are substantially different to the inverse welfare weights computed under the standard assumption of continuous choices in Figure 4a: with frictions one no longer needs a pathological inverse welfare functional to rationalize a piecewise linear schedule precisely because kinks in the tax schedule do not generate bunching (or missing masses). Second, the high-level qualitative pattern is that inverse welfare weights are *flatter* than in the baseline case (see also Figure 13 in Appendix C which shows inverse welfare weights in the sparsity-based frictions model under the smoothed approximation to the U.S. tax code as in Figure 3). For example, inverse welfare weights for those earning \$400,000 are 2.5 times as high under the model with frictions than under the baseline calibration in Section 5.1. Why is this? Essentially, this results due to non-linearities in the tax schedule and the income density. In both the baseline model and the frictions model, tax changes lead to intensive margin effects and extensive margin effects (remembering that the intensive margin effects in the frictions model arise from individuals moving from part-time to full-time and vice-versa). The extensive margin effects are calibrated to be roughly equivalent across the two models. In the baseline model, the intensive margin effects depend on (1) the intensive margin elasticity and (2) how the income density and marginal tax rates are changing *locally* with income (see Equation 17). In the frictions model the intensive margin effects depend on (1) the "modified intensive margin elasticity" elasticity (which is calibrated to be the same as the intensive margin elasticity in the baseline model) and (2) how the income density and marginal tax rates are changing *between a household's full-time and part-time incomes*. We illustrate in Appendix B.9 that for regions where the income density is convex and tax rates are mostly flat (e.g., near the top of the income distribution), the intensive margin effects of increased taxation are *larger* in the baseline model than in the frictions model.[28] Thus, raising taxes locally for those earning, say, \$400,000 per year is less costly for the government than in the baseline case. In order to rationalize not wanting to raise taxes on these individuals, the government must then have higher inverse welfare weights in a world with frictions than in a world with a continuum of choices. Hence, allowing for frictions is potentially quite important for determining the desirability of tax reforms. For instance, suppose society's true weights on those earning around \$400,000 are inbetween the inverse weights under the baseline model (≈ 0.2) and the inverse weights under the frictions model (≈ 0.5). In this case, the baseline model would imply that increasing taxes on these individuals is welfare *decreasing* whereas the frictions model would imply that increasing taxes on these individuals is welfare *improving* by Proposition 3.

5.4 Nonlinear Income and Property Taxation

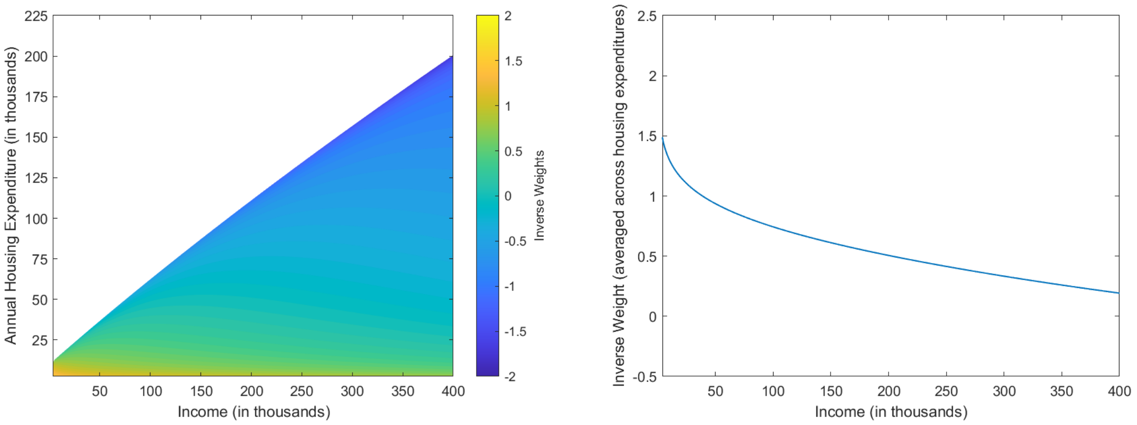
Our next example illustrates how inverse welfare weights change when the tax schedule is multidimensional. We consider a simple model of taxation of both income, z_1, and the amount of money spent on housing, z_2. For households that rent, housing rent is simply equal to the (explicit) amount of money they spend on rent per year; for households that own their home, we assume the implicit housing rent (i.e., rent paid to oneself) is equal to a fraction of the property value. With perfect pass-through of taxes onto renters and a constant rental rate of return,

[28] Loosely, this results because the income density is changing faster, for example, at an income of \$400,000 than it is changing on average between \$400,000 and \$800,000.

a tax on housing rent is therefore equivalent to a property tax. Individuals differ in terms of three dimensions: labor productivity n_1, preferences over housing n_2, and the discrete cost of working, v:

$$\begin{aligned} u(c, z_1, z_2; n_1, n_2) &= \frac{c^{1-g}}{1-g} - \frac{(z_1/n_1)^{1+k_1}}{1+k_1} + n_2 \frac{z_2^{1+k_2}}{1+k_2} - v\mathbb{1}[z_1 > 0] \\ c &= z_1 - z_2 - T(z_1, z_2) \end{aligned} \tag{40}$$

We calibrate k_1 and k_2 to match an average taxable income elasticity of 0.3 (Saez, Slemrod and Giertz, 2012) and an average elasticity of housing rent with respect to the tax rate of -0.83 (Albouy, Ehrlich and Liu, 2016). We assume that $g = 0.5$, which implies that income effects are (on average) approximately -0.1 across the joint distribution of (n_1, n_2), which is in line with estimates of income effects from Gruber and Saez (2002). We calibrate the distribution $f(n_1, n_2)$ to match the empirical joint distribution of labor income and housing rents from the 2019 American Community Survey (ACS) where implicit rent for homeowners is assumed to be 5% of the property value (5% is the median rent-to-price ratio across the 50 largest U.S. cities, SmartAsset (2024)). We calibrate the distribution $f(v)$ to match the fraction of the population that is unemployed in the 2019 ACS along with an extensive margin elasticity of 0.25 (Chetty et al., 2013). We plot inverse weights for a smooth approximation to the U.S. income tax schedule along with a 20% tax on housing over the (income, housing) distribution in Figure 5a; we plot average weights over the income distribution in Figure 5b.[29]



(a) Inverse Weights

(b) Average Inverse Weights

Figure 5: Inverse Welfare Weights for U.S. Income and Property Tax Schedule

Note: This figure shows inverse welfare weights multiplied by marginal utility of consumption $(\phi(n_1, n_2, v)c(n_1, n_2, v)^{-g})$ for a smooth approximation to the U.S. income tax schedule and a 20% tax on housing expenses. Figure 5a plots inverse weights across the distribution of income and housing expenses. Figure 5b shows average inverse weights across the income distribution. The calibration is described in Section 5.4.

There are two key findings from this exercise. First, including a tax on housing expenses leads to substantial variation in implicit welfare weights between individuals with the same income but different housing expenses. For example, the inverse welfare weight for households

[29] The average property tax in the U.S. is approximately 1% (U.S. Census Bureau, 2021). If the rent-to-price ratio of homes is 5% per year, then a 20% tax on implicit rent is equivalent to a 1% property tax.

earning \$100,000 per year and spending \$15,000 per year on housing (equivalent to owning a \$300,000 home with a 5% implicit rental income rate) is 0.6 whereas the inverse welfare weight for households earning \$100,000 per year and spending \$30,000 per year on housing (equivalent to owning a \$600,000 home) is only 0.3.[30] In other words, to rationalize a property tax, one must have substantially different redistributive preferences amongst people with the same income yet different tastes for housing. The second key finding is that while the multidimensional tax system creates variation in inverse weights among those with the same income, the average welfare weights across the income distribution (Figure 5b) are essentially unchanged relative to the baseline scenario in Section 5.1. Thus, taxing housing does not substantially alter society's implicit weights across the income distribution but rather creates variation in implicit weights between individuals with the same income and different tastes for housing.

6 Inverse Welfare Functionals in General Equilibrium

The theory developed so far is quite general in the sense that we made very few assumptions on the utility function, choice variables $\mathbf{z}$, or primitives $\mathbf{n}$. However, one key restriction that we have made, consistent with most of the optimal taxation literature, is that we have only considered a "partial equilibrium" setting in which individual's decisions $\mathbf{z}$ do not impact the economy more broadly. Next, we will show how to compute inverse welfare functionals when there are "general equilibrium" effects. We will state and prove this result in Theorem 2. We will then illustrate the impact of general equilibrium effects on inverse welfare functionals via a simple labor demand/labor supply model with endogenous wages. Note, this example contains all of the key intuition of Theorem 2 and is substantially simpler to understand. Finally, we will show how this expanded framework can be used to explore inverse welfare functionals in the presence of externalities.

6.1 Main Result with General Equilibrium Effects

Suppose that individuals still differ in terms of $\mathbf{n} = (n_1, n_2, ..., n_K) \in \mathbf{N}$ distributed according to some distribution $F(\mathbf{n})$. The government chooses a tax schedule, $T(\mathbf{z})$, which is a function of a set of observable individual choice variables $\mathbf{z} = (z_1, z_2, ..., z_J) \in \mathbf{Z}$. Individual utility also depends on a vector, $\mathbf{w}$, of "general equilibrium" parameters. $\mathbf{w}$ might consist of prices, wages, externalities, or other quantities that are impacted in some way by aggregate individual decisions (and hence depend on taxes). Thus, individuals maximize the following utility function which is assumed smooth in all arguments:

$$\begin{aligned} U(n; T, \mathbf{w}) = \max_{\mathbf{z}} \; & u\left(c, \mathbf{z}; \mathbf{n}, \mathbf{w}\right) \\ \text{s.t.} \; & c = y(\mathbf{z}, \mathbf{w}) - T(\mathbf{z}) \end{aligned} \tag{41}$$

Theorem 2. *Consider $T(\mathbf{z})$ with $R(T) = E$ such that $\mathbf{Z}$ is compact and for every $\mathbf{z}$ $\exists$ an $\mathbf{n}$ with a unique optima. A local inverse functional supporting $T(\mathbf{z})$ exists if:*

[30] Also, note that inverse weights are negative for those with very high housing expenditures, implying that the tax schedule is Pareto inefficient.

1. *$R(T)$ is Gateaux differentiable*
2. *Each $w_i \in \mathbf{w}$ is Gateaux differentiable as a function of T with $\lim_{\epsilon\to 0}\frac{w_i(T+\epsilon\tau)-w_i(T)}{\epsilon} = \int_{\mathbf{Z}}\tau(\mathbf{z})dp_i(\mathbf{z})$*
3. *The direct welfare impacts of changing taxes are larger than the indirect welfare impacts of changing $\mathbf{w}$ that ensue from changing taxes. Technically, the maximum average willingness-to-pay for an increase in each w_i, $\left|\left|\overline{\frac{u_{w_i}}{u_c}}\right|\right|_\infty$ (where the average is taken over $\mathbf{n}$ at each $\mathbf{z}$ and the supnorm is taken over $\mathbf{z}$), multiplied by an upper bound for the impact of a tax change on w_i, denoted $||p_i||_{TV}$ is less than 1:*[31]

$$\sum_i ||p_i||_{TV}\left|\left|\overline{\frac{u_{w_i}}{u_c}}\right|\right|_\infty < 1$$

Proof. See Appendix A.4 □

The high level takeaway from Theorem 2 is that even when taxes have indirect welfare impacts via general equilibrium effects, we can often nonetheless construct inverse welfare functionals under some differentiability restrictions on general equilibrium parameters and government revenue. The proof to Theorem 2 is dense and employs a number of technical tools from functional analysis and measure theory. The key intuition, which will be explored in Section 6.2 with a labor supply/labor demand model, is as follows: if the "direct" impacts of a tax perturbation on utility are larger than the "indirect" impacts of a tax perturbation on utility (via impacts on general equilibrium objects $\mathbf{w}$), then the equation that pins down a local inverse welfare functional is a contraction and hence has a solution. The proof is much more technical because the government's first order condition is an integral equation formulated in a measure space, but all of the intuition for Theorem 2 can be understood from the following example.

6.2 Example: Labor Demand

We now explore an example wherein wages are endogenous because, contrary to much of the optimal taxation literature, the labor demand side of the market is *not* assumed to be infinitely elastic. Much of the analysis of tax perturbations in this section is similar to Sachs, Tsyvinski and Werquin (2020) who explore optimal income taxation with endogenous wages (we illustrate how to compute inverse welfare functionals whereas Sachs, Tsyvinski and Werquin (2020) show how to compute optimal tax schedules). Consider a government that chooses a tax schedule to maximize welfare for a given population of individuals indexed by a unidimensional type n. Individuals choose an income $z = wnl$ where l is labor supply and w is a wage paid on effective effort, nl. Individuals choose z to maximize a quasi-linear iso-elastic utility function:

$$\begin{aligned} U(n;T,w) &= \max_z \; c - \frac{[z/(nw)]^{1+k}}{1+k} \\ \text{s.t. } \; c &= z - T(z) + s(n)\pi(w) \end{aligned} \tag{42}$$

where c is again numeraire consumption, $\pi(w)$ represents firm profits, and $s(n)$ represents the

[31] $||p_i||_{\text{TV}}$ denotes the total variation norm which is the maximum possible Gateaux variation of w_i: $||p_i||_{\text{TV}} \equiv \sup_{||\tau||_\infty \le 1}\int_{\mathbf{Z}}\tau(\mathbf{z})dp_i(\mathbf{z})$.

share of profits owned by a given type n with $\int_N s(n)f(n)dn = 1$. There is also a single firm that produces the consumption good c by hiring labor to maximize profits. Firm output depends on total hired effective effort, L. Thus, firm profits are given by:

$$\pi = Y(L) - wL$$

where $Y(L)$ is the firm's production function. Market clearing requires that:

$$L = \int_N nl(n)dF(n) \tag{43}$$

Suppose that we are interested in calculating an inverse welfare functional in this setting for a smooth income tax schedule. The government's Lagrangian is given by:

$$W(U(n;T,w)) + \lambda\left[\int_N T(z(n))dF(n) - E\right] \tag{44}$$

Now, let us take the Gateaux variation of Equation 44 in the direction of $\tau(z)$, assuming that $n \mapsto z$ is a smooth bijective function, individual second order conditions hold strictly, and $\frac{\partial w}{\partial \epsilon}$ exists (importantly, note that $z(n)$ is also a function of the tax schedule and the wage even though we omit these arguments for clarity):

$$\begin{aligned} &W\left(-\tau(z(n)) + \left(\frac{z(n)}{nw}\right)^{1+k}\frac{1}{w}\frac{\partial w}{\partial \epsilon} + s(n)\pi'(w)\frac{\partial w}{\partial \epsilon}\right) \\ &+ \lambda\int_N\left(\tau(z) + T'(z(n))\left.\frac{\partial z(n)}{\partial \epsilon}\right|_w + T'(z(n))\left.\frac{\partial z(n)}{\partial w}\right|_\epsilon\frac{\partial w}{\partial \epsilon}\right)dF(n) \end{aligned} \tag{45}$$

Next, we need to express $\frac{\partial w}{\partial \epsilon}$ in terms of $\tau(z)$. We show in Appendix B.10 that $\frac{\partial w}{\partial \epsilon}$ is Gateaux differentiable in T and that if $h(z) \to 0$ at the top and bottom of the income distribution (this assumption is not necessary it just simplifies the subsequent expressions), then $\frac{\partial w}{\partial \epsilon}$ can be expressed as follows for a function $p(z)$:

$$\frac{\partial w}{\partial \epsilon} = \int_Z p(z)\tau(z)dz$$

Suppose that the welfare functional takes the form $W(U(n;T,w)) = \int_N \phi(n)U(n;T)dF(n)$. We show in Appendix B.10 that we can express Equation 45 as the following linear functional:

$$\begin{aligned} &-\int_Z\left[\underbrace{\phi(n(z))h(z)}_{\text{Direct Welfare Effect}} \underbrace{-p(z)\left(\int_Z \phi(n(\tilde{z}))\left[\left(\frac{\tilde{z}}{n(\tilde{z})w}\right)^{1+k}\frac{1}{w} + s(n(\tilde{z}))\pi'(w)\right]h(\tilde{z})d\tilde{z}\right)}_{\text{Indirect Welfare Effect}}\right]\tau(z)dz \\ &+\lambda\int_Z\left(\underbrace{h(z) - \frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right]}_{\text{Direct Budgetary Effect}} + \underbrace{p(z)\int_N\left(T'(z(n))\left.\frac{\partial z(n)}{\partial w}\right|_\epsilon\right)dF(n)}_{\text{Indirect Budgetary Effect}}\right)\tau(z)dz \end{aligned} \tag{46}$$

Intuitively, Equation 46 captures two separate impacts: the direct impacts of the tax change and the indirect impacts of the wage change that results from changes in labor supply as a result of the tax change. A local inverse welfare functional is a set of weights $\phi(n)$ such that

Equation 46 equals zero for all possible $\tau(z)$. Normalizing $\lambda = 1$, this will hold as long as:

$$\begin{aligned} &\phi(n(z))h(z) - p(z)\left(\int_Z \phi(n(\tilde{z}))\left[\left(\frac{\tilde{z}}{n(\tilde{z})w}\right)^{1+k}\frac{1}{w} + s(n(\tilde{z}))\pi'(w)\right]h(\tilde{z})d\tilde{z}\right) \\ &= h(z) - \frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right] + p(z)\int_N \left(T'(z(n))\left.\frac{\partial z(n)}{\partial w}\right|_{\epsilon}\right)dF(n) \end{aligned} \tag{47}$$

Equation 47 is more complex to solve for weights than Equation 17 in the analogous partial equilibrium case with a smooth bijective relationship between n and z. For a given z, Equation 17 is linear in $\phi(n(z))$ whereas Equation 47 is an *integral equation* in $\phi(n(z))$. Defining:

$$K(z,\tilde{z}) \equiv \frac{p(z)\left[\left(\frac{\tilde{z}}{n(\tilde{z})w}\right)^{1+k}\frac{1}{w} + s(n(\tilde{z}))\pi'(w)\right]h(\tilde{z})}{h(z)}$$

$$\chi(z) \equiv \frac{h(z) - \frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right] + p(z)\int_N \left(T'(z(n))\frac{\partial z(n)}{\partial w}\Big|_{\epsilon}\right)dF(n)}{h(z)}$$

Equation 47 can be expressed as:

$$\phi(n(z)) = \chi(z) + \int_Z K(z,\tilde{z})\phi(n(\tilde{z}))d\tilde{z} \tag{48}$$

which is a Fredholm integral equation. It is a standard result that this type of integral equation has a solution so long as $\int_Z |K(z,\tilde{z})|d\tilde{z} < 1\ \forall z$; in this case, $\chi(z) + \int_Z K(z,\tilde{z})\phi(n(\tilde{z}))d\tilde{z}$ is a contraction mapping so that existence of a solution to Equation 48 (i.e., a fixed point of the contraction mapping) follows immediately from the contraction mapping theorem.[32] Economically, the condition that $\int_Z |K(z,\tilde{z})|d\tilde{z} < 1\ \forall z$ ensures that the direct welfare effect of a change in taxes is larger than the indirect welfare effect of a change in taxes (defined in Equation 46). Thus, in the context of a labor supply/labor demand model, we have shown that even in the presence of general equilibrium effects, we can solve for the local inverse welfare functional that supports a given tax schedule satisfying the budget constraint under some regularity conditions.

6.3 Numerical Simulation

Let us now explore the extent to which general equilibrium wage effects impact inverse welfare weights numerically. We suppose that individuals have quasi-linear iso-elastic utility as in Equation 42 from Section 6.2. We calibrate the primitive distribution as in the baseline model of Section 5.1 and again choose k to match a taxable income elasticity of 0.3. However, we now suppose that there is a labor demand side with a production function $Y(L) = aL^{\beta}$ so that the labor demand elasticity with respect to the wage is equal to $E^D = 1/(\beta - 1)$.

Suppose we want to find inverse welfare weights that support the smoothed approximation to the actual U.S. tax schedule from Section 5.1. If we assume, as is common in the optimal taxation literature, that labor demand is infinitely elastic (corresponding to a production function with $\beta = 1$) we can recover inverse welfare weights in "partial equilibrium" as in Section 3.1. In contrast, if the labor demand elasticity is finite, we must compute the inverse welfare functional by finding the fixed point of integral equation 47. Figure 6 plots these inverse welfare weights

[32] Numerically, one can solve for this fixed point in a straight-forward way: start with an arbitrary set of weights $\phi(n(z))$ and then iterate on Equation 48 until convergence.

for various values of E^D.

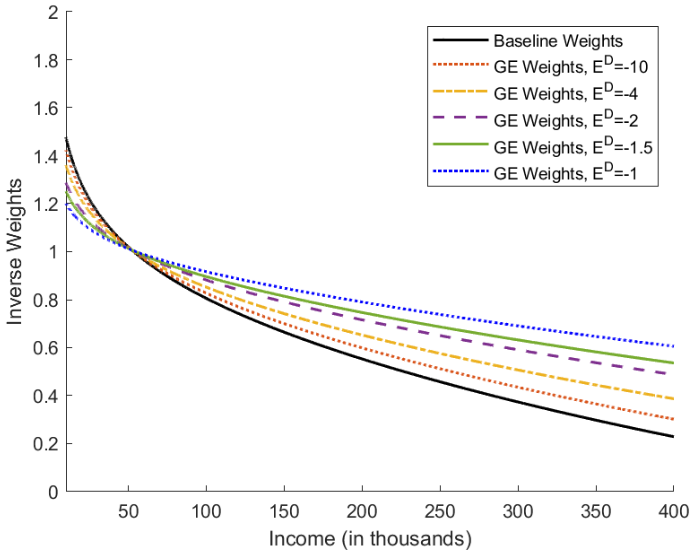


Figure 6: Inverse Welfare Weights with Finite Labor Demand Elasticity

Note: This figure shows the inverse welfare weights for a smooth approximation to the U.S. income tax schedule under various assumptions about the size of the labor demand elasticity. Individuals have utility function 42 and the calibration is the same as in Section 5.1. The labor demand side has a production function aL^β, which implies that the labor demand elasticity with respect to the wage is equal to $E^D = 1/(\beta - 1)$.

The key takeaway from Figure 6 is that more inelastic labor demand (i.e., smaller labor demand elasticities in absolute value) corresponds to inverse welfare weights that are much higher for high productivity types and lower for low productivity types. For instance, assuming that labor demand is infinitely elastic, the baseline calibration yields that the implicit inverse welfare weight on individuals earning $400,000 is about 0.2 whereas if the labor demand elasticity is in fact close to -1, then the inverse weight on individuals earning $400,000 is about 0.7.[33] In other words, the U.S. tax schedule is rationalized with far weaker implicit redistributive preferences if labor demand is relatively inelastic. Intuitively, if labor demand is more inelastic, then wages are more responsive to reductions in labor supply. As a result, raising taxes on high income individuals generates two redistributive effects: the direct effect of increased tax revenue and the indirect effect of increased wages from high income individuals reducing their labor supply. Therefore more inelastic labor demand implies that we need larger inverse welfare weights for high income individuals to justify a given tax schedule. The larger inverse welfare weights implied by inelastic labor demand therefore have important policy ramifications. For example, suppose that society's true welfare weight on individuals earning around $400,000 is 0.4. *Decreasing* taxes for these individuals would be welfare improving if labor demand is infinitely elastic as the inverse welfare weight on these individuals is ≈ 0.2 whereas *increasing* taxes for these individuals would be welfare improving if the labor demand elasticity is -2

[33] Recall that these inverse welfare weights are normalized to integrate to 1 so that the value of splitting a dollar equally among the population equals 1. Thus, for example, the baseline calibration implies society values giving a dollar to those earning $400,000 20% as much as splitting a dollar equally among the population.

because the inverse welfare weights on these individuals is ≈ 0.6.[34]

One concern with the simplistic model of Section 6.2 is that labor of high productivity types is perfectly substitutable with labor of low skilled types (because the production function only depends on aggregate labor supply); thus, one may wonder whether these findings would change substantially if the production function features complementarity between high- and low-skilled labor. We augment the model from Section 6.2 using ideas discussed in the more general Theorem 2 to allow for a CES production function $Y(L_l, L_h) = (a_l L_l^\sigma + a_h L_h^\sigma)^{\frac{v}{\sigma}}$ with low-skilled labor L_l and high-skilled labor L_h along with two general equilibrium wages: one for high-skilled workers (those above median productivity) and one for low-skilled workers (those below median productivity). Details of how to compute the inverse welfare functional for this more complicated model are given in Appendix B.12; we show in Figure 14 in Appendix C how the inverse welfare weights vary with the degree of complementarity between L_l and L_h. Allowing for complementarity between high- and low-skilled labor has minimal impacts on the inverse welfare functional and does not change the takeaways from this exercise: tax schedules that are supported by highly redistributive welfare functionals without GE wage effects are supported by substantially less redistributive welfare functionals when GE wage effects are taken into account. Further investigation using more sophisticated labor demand models is certainly warranted.

6.4 Externalities Example

We conclude this section by discussing another application of Theorem 2: externalities. Externalities occur when certain choices $\mathbf{z}$ have indirect impacts on utility of others; this can be modeled in Theorem 2 by including the choices of others in $\mathbf{w}$, the vector of "general equilibrium" parameters. For example, if good z_i creates pollution and individual utility depends on total societal consumption of good z_i, we can set a component w_j of $\mathbf{w}$ to equal $\int_{\mathbf{N}} z_i(\mathbf{n}) dF(\mathbf{n})$.

We will illustrate the inclusion of externalities in inverse welfare functionals using a slightly more complex externality: inequality aversion. Suppose that in addition to having preferences over consumption and labor supply, agents also have an intrinsic distaste for inequality (Alesina and Giuliano, 2011). We assume that individuals n choose an income z to maximize utility but that individuals are also negatively impacted by the overall level of inequality (i.e., other individuals' incomes generate an externality by contributing to inequality), which we operationalize using the Gini coefficient, $G = \frac{100}{\int_Z z h(z) dz} \int_Z H(z)(1 - H(z)) dz$ where $H(z)$ is the CDF of z and $h(z)$ is the PDF of z:

$$\begin{aligned} U(n; T, G) &= \max_z \; c - \frac{z^{1+k}}{1+k} - \alpha G \\ \text{s.t. } \; c &= z - T(z) \end{aligned} \tag{49}$$

α represents the per-capita willingness to pay to decrease the Gini coefficient by 1. For reference, the Gini coefficient is ≈ 40 in the U.S. and is ≈ 30 in Sweden (OECD, 2024), so $\alpha = 100$ implies

[34] Note, this logic implicitly requires that the difference in GE welfare impacts between the true welfare weights and the inverse welfare weights is small compared to the direct welfare effect because Proposition 3 needs to be augmented when there are GE effects; see Proposition 3 GE in Appendix B.11.

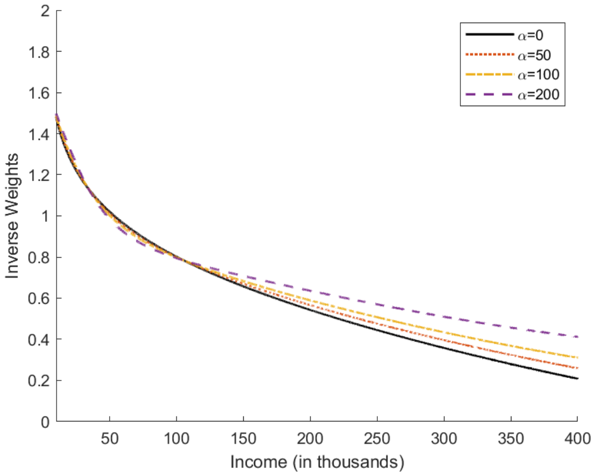


Figure 7: Inverse Welfare Weights with Inequality Aversion

Note: This figure shows the inverse welfare weights for a smooth approximation to the U.S. income tax schedule under various assumptions about inequality aversion α. α represents the per-capita willingness to pay to decrease the Gini coefficient (measured on scale from 0 to 100) by 1. Individuals have utility function 49 and the calibration is the same as in Section 5.1.

that the per-capita willingness-to-pay to live in an economy with Swedish inequality relative to U.S. inequality is $\approx \$1,000$ per year. Using the analysis from Section 3.1 (assuming that $h(\underline{z}) = h(\overline{z}) = 0$ for simplicity), we can calculate the Gateaux derivative of the government's Lagrangian as (where $\xi(z)$ is the substitution effect defined in Equation 12):

$$\int_Z \left\{ [1-\phi(n(z))]h(z) - \frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right] \right\} \tau(z)dz - \alpha \lim_{\epsilon\to 0} \frac{G(T+\epsilon\tau)-G(T)}{\epsilon} \int_Z \phi(n(z))h(z)dz = 0 \tag{50}$$

Changing variables and using the fact that by monotonicity of $n \mapsto z$, $H(z(n)) = F(n)$, so that the Jacobian equals $\frac{dz}{dn} = \frac{f(n)}{h(z(n))}$, we can express $G = \frac{100}{\int_N z(n)f(n)dn} \int_N H(z(n))(1 - H(z(n)))\frac{f(n)}{h(z(n))}dn$. We can use Equation 12 along with integration by parts to calculate the Gateaux derivative of G, $\lim_{\epsilon\to 0} \frac{G(T+\epsilon\tau)-G(T)}{\epsilon}$:

$$\frac{100 \int_Z \frac{\partial}{\partial z}\left[\frac{G(T)}{100}\xi(z)h(z) - \xi(z)(1-2H(z))h(z) + \xi(z)H(z)(1-H(z))h'(z)/h(z)\right]\tau(z)dz}{\int_Z zh(z)dz} \tag{51}$$

From here, we can solve the integral equation in Equation 50 via a fixed-point algorithm. Using the baseline calibration from Section 5.1, Figure 7 shows the inverse welfare weights that rationalize a smooth approximation to the U.S. tax schedule for various values of α: as inequality aversion increases, the inverse welfare weights that rationalize a given tax schedule increase for high income individuals. These findings are intuitive: in order to justify *not* wanting to increase taxes on high income individuals as inequality aversion increases, the inverse welfare weights must increase for these individuals. The presence of inequality aversion can have a meaningful impact on the inverse welfare functional; for instance, the baseline calibration yields that the inverse welfare weight on individuals earning \$400,000 is about 0.2 whereas if $\alpha = 200$ (i.e., per-capita willingness-to-pay to live in an economy with Swedish inequality relative to

U.S. inequality is $\approx \$2,000$), then the inverse welfare weight on individuals earning \$400,000 is about 0.4. Thus, without inequality aversion (with inequality aversion of $\alpha = 200$) society implicitly values giving a dollar to those earning \$400,000 20% (40%) as much as splitting a dollar evenly among the population.

7 Non-Existence of Inverse Welfare Functionals

We have shown that our framework allows us to construct inverse welfare functionals for a large class of tax schedules because Gateaux differentiability of government revenue is a relatively mild restriction. We now discuss one situation in which revenue fails to be Gateaux differentiable and hence inverse welfare functionals *do not* exist: when the dimension of the choice space $\mathbf{Z}$ is *larger* than the dimension of the type space $\mathbf{N}$. In Appendix B.13 we prove the following result:

Proposition 5. *Consider smooth $T(\mathbf{z})$ such that $R(T) = E$, $\mathbf{Z}$ is compact, and all $\mathbf{n}$ have a unique optimum. When the dimension of the choice space $\mathbf{Z}$ is larger than the dimension of the type space $\mathbf{N}$, then there are smooth tax schedules for which no inverse welfare functional exists.*

We prove Proposition 5 by showing non-existence of an inverse welfare functional in a simple model of income and savings taxation as in Atkinson and Stiglitz (1976). Mathematically, non-existence results due to government revenue failing to be Gateaux differentiable.[35] The intuition is as follows. Under some smoothness assumptions, there is a unique set of inverse welfare weights that ensures every perturbation to the *income* tax schedule leaves the government's Lagrangian unchanged. But there is also a unique set of inverse welfare weights that ensures every perturbation to the *savings* tax schedule leaves the government's Lagrangian unchanged. However, these two sets of potential inverse welfare weights do not necessarily need to coincide, leading to an overdetermined (infinite) system of equations that the inverse welfare weights must satisfy. This overdetermination then leads to non-existence of inverse welfare weights.

Practically, the choice space is likely smaller than the type space so that Proposition 5 does not apply. However, there is at least one important theoretical setting in which Proposition 5 holds: the setting of Atkinson and Stiglitz (1976). The Atkinson-Stiglitz Theorem is one of the most famous results in public economics, showing that when agents differ in terms of a unidimensional parameter $n \in N$, then multidimensional tax schedules which are a function of income and other choices (e.g., savings, commodities) are sub-optimal when the utility function is weakly separable between labor and all other goods. Proofs of the Atkinson-Stiglitz Theorem typically invoke the Pareto principle by showing that any multidimensional tax schedule is Pareto dominated by some non-linear income tax schedule (Kaplow, 2006). The non-existence result of Proposition 5 can therefore be viewed as a strengthening of the classic Atkinson-Stiglitz Theorem: in settings with unidimensional type heterogeneity and multidimensional choice spaces, many tax schedules are not just Pareto inefficient but are actually not supported

[35] Recall that both multidimensional integration by parts, Lemma 1, and Proposition 1 require that the set $\mathbf{Z}$ over which the functions are integrated is open (or the closure of an open set) in the ambient space (i.e., $\mathbf{Z}$ has non-empty interior). This condition fails when the choice space is larger than the type space. For instance, if individuals differ in terms of a unidimensional parameter n and have two choice variables, then the set $\mathbf{Z}$ of chosen (z_1, z_2) will be a curve in $\mathbb{R}^2$, which is *not* an open set (or the closure of an open set) in $\mathbb{R}^2$.

by *any* inverse welfare functionals, even those that allow for negative weights.[36] The economic takeaway is as follows: in the Atkinson-Stiglitz environment, it is often impossible to rationalize indirect taxes (e.g., savings or commodity taxes) even if the government wants to make some individuals as miserable as possible (via negative welfare weights).

8 Conclusion

This paper has developed a general theory to recover the inverse welfare functional that rationalizes a given tax schedule as optimal. The essential component required to construct such an inverse welfare functional is the Gateaux derivative of government revenue with respect to the tax schedule. Our theory allows for complex environments including the presence of multidimensional tax schedules, bunching/jumping behavior, optimization frictions, general equilibrium effects, and externalities. From a policy perspective, inverse welfare functionals are a simple tool that can be used to compare society's true preferences with the implicit preferences revealed by government policies. Moreover, we show that inverse welfare functionals can be used to assess the desirability of tax reforms as well as test for Pareto efficiency of a tax schedule. Finally, we have shown numerically how allowing for bunching/jumping behavior, sparsity-based frictions, multidimensional tax systems, general equilibrium wage effects, and inequality aversion can have large and meaningful impacts on the inverse welfare functional.

Moving forward, we believe there is still substantial scope for innovation in so-called "inverse optima" methods. First, all of the analysis in this paper has assumed that agents correctly perceive the tax schedule and their own utility function. While a general analysis of inverse welfare functionals in the presence of misperceptions, behavioral biases, and internalities seems challenging, we believe this is a useful area for future research (we present two examples of how our framework can be used as a starting point to show existence or non-existence of inverse welfare functionals in Appendix B.14 and B.15). Second, while this paper focuses on inverse welfare functions for tax schedules, much of the analysis can likely be extended to non-tax policy spaces, such as in-kind good provision, minimum wages, or social insurance.

References

Albouy, David, Gabriel Ehrlich, and Yingyi Liu. 2016. "Housing Demand, Cost-of-Living Inequality, and the Affordability Crisis." National Bureau of Economic Research Working Paper 22816.

Alesina, Alberto, and Paola Giuliano. 2011. "Chapter 4 - Preferences for Redistribution." In . Vol. 1 of *Handbook of Social Economics*, , ed. Jess Benhabib, Alberto Bisin and Matthew O. Jackson, 93–131. North-Holland.

Anagol, Santosh, Allan Davids, Benjamin B Lockwood, and Tarun Ramadorai. 2022. "Diffuse bunching with Lumpy incomes: Theory and estimation." *SSRN Electronic Journal.*

[36] Note, in Appendix B.13, we show that this non-existence does not rely on separability in any way: hence most tax schedules in this setting will not have associated inverse welfare functionals regardless of whether utility is weakly separable or not.

Atkinson, A. B., and J. E. Stiglitz. 1976. "The Design of Tax Structure: Direct versus Indirect Taxation." *Journal of Public Economics*, 6: 55–75.

Autor, David H., Lawrence F. Katz, and Melissa S. Kearney. 2008. "Trends in U.S. wage inequality: Revising the revisionists." *Review of Economics and Statistics*, 90(2): 300–323.

Bargain, Olivier, Mathias Dolls, Dirk Neumann, Andreas Peichl, and Sebastian Siegloch. 2013. "Comparing inequality aversion across countries when labor supply responses differ." *International Tax and Public Finance*, 21(5): 845–873.

Bergstrom, Katy, and William Dodds. 2021. "Optimal Taxation with Multiple Dimensions of Heterogeneity." *Journal of Public Economics*, 200: 104442.

Bierbrauer, Felix J., Pierre C. Boyer, and Emanuel Hansen. 2023. "Pareto-Improving Tax Reforms and the Earned Income Tax Credit." *Econometrica*, 91(3): 1077–1103.

Blundell, Richard, Mike Brewer, Peter Haan, and Andrew Shephard. 2009. "Optimal income taxation of lone mothers: An empirical comparison of the UK and Germany." *The Economic Journal*, 119(535).

Bourguignon, François, and Amedeo Spadaro. 2010. "Tax–benefit revealed social preferences." *The Journal of Economic Inequality*, 10(1): 75–108.

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

Chetty, Raj, Adam Guren, Day Manoli, and Andrea Weber. 2013. "Does Indivisible Labor Explain the Difference between Micro and Macro Elasticities? A Meta-Analysis of Extensive Margin Elasticities." *NBER Macroeconomics Annual*, 27: 1–56.

Choquet, Gustave. 1966. *Topology.* Academic Press.

Das, P. C. 1974. "Nonlinear integral equations in a measure space." *Proceedings of the American Mathematical Society*, 42(1): 181–185.

de Bartolome, Charles A.M. 1995. "Which tax rate do people use: Average or marginal?" *Journal of Public Economics*, 56(1): 79–96.

Ferey, Antoine, Benjamin Lockwood, and Dmitry Taubinsky. 2021. "Sufficient Statistics for Nonlinear Tax Systems with General Across-Income Heterogeneity." National Bureau of Economic Research Working Paper 29582.

Golosov, Mikhail, Aleh Tsyvinski, and Nicolas Werquin. 2014. "A Variational Approach to the Analysis of Tax Systems." *NBER Working Paper 20780*, 48.

Gruber, Jonathan, and Emmanuel Saez. 2002. "The elasticity of taxable income: evidence and implications." *Joural of Public Economics*, 84(2002): 1–32.

Havranek, Tomas, Zuzana Irsova, Lubica Laslopova, and Olesia Zeynalova. 2020. *The elasticity of substitution between skilled and unskilled labor: A meta-analysis.*

Hendren, Nathaniel. 2020. "Measuring economic efficiency using inverse-optimum weights." *Journal of Public Economics*, 187: 104198.

Jacobs, Bas, Egbert L.W. Jongen, and Floris T. Zoutman. 2017. "Revealed social preferences of Dutch political parties." *Journal of Public Economics*, 156: 81–100.

Kaplow, Louis. 2006. "On the undesirability of commodity taxation even when income taxation is not optimal." *Journal of Public Economics*, 90(6–7): 1235–1250.

Lorenz, Normann, and Dominik Sachs. 2016. "Identifying laffer bounds: A sufficient-statistics approach with an application to Germany." *The Scandinavian Journal of Economics*, 118(4): 646–665.

Milgrom, Paul, and Ilya Segal. 2002. "Envelope Theorems for Arbitrary Choice Sets." *Econometrica*, 70: 583–601.

Mirrlees, James. 1971. "An Exploration in the Theory of Optimal Income Taxation." *Review of Economic Studies*, 38: 175–208.

OECD. 2024. "Income Distribution Database." *https://stats.oecd.org/Index.aspx?DataSetCode=IDD#*, Accessed on 24 May 2024.

Rudin, Walter. 1974. *Real and complex analysis.* McGraw-Hill Book Company.

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

Sadka, Efraim. 1976. "On income distribution, incentive effects and optimal income taxation." *The Review of Economic Studies*, 43(2): 261.

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

Saez, Emmanuel. 2010. "Do Taxpayers Bunch at Kink Points?" *American Economic Journal: Economic Policy*, 2(3): 180–212.

Saez, Emmanuel, Joel Slemrod, and Seth H Giertz. 2012. "The elasticity of taxable income with respect to marginal tax rates: A critical review." *Journal of Economic Literature*, 50(1): 3–50.

Scheuer, Florian, and Ivan Werning. 2016. "Mirrlees meets Diamond-Mirrlees."

Seade, J.K. 1977. "On the shape of Optimal Tax Schedules." *Journal of Public Economics*, 7(2): 203–235.

Sharma, R. R. 1975. "Some problems of nonlinear integral equations in measure spaces." *Proceedings of the American Mathematical Society*, 51(2): 313–321.

SmartAsset. 2024. "Price-to-Rent Ratio in the 50 Largest U.S. Cities." https://smartasset.com/data-studies/price-to-rent-ratio-in-the-50-largest-us-cities-2022, Accessed on 24 May 2024.

Spiritus, Kevin, Etienne Lehmann, Sander Renes, and Floris Zoutman. 2022. "Optimal taxation with multiple incomes and types." *SSRN Electronic Journal*.

Sturm, John, and André Sztutman. 2023. "Income Taxation with Elasticity Heterogeneity."

U.S. Census Bureau. 2021. "MORTGAGE STATUS BY MEDIAN REAL ESTATE TAXES PAID (DOLLARS)." *U.S. Census Bureau*, Accessed on 24 May 2024.

Werning, Ivan. 2007. "Pareto Efficient Income Taxation."

A Appendix: Proofs

A.1 Proof of Theorem 1

Proof. First, $T(\mathbf{z})$, and hence $\tau(\mathbf{z})$, are assumed continuous so that if $R(T)$ is Gateaux differentiable then by the Riesz-Markov-Kakutani representation theorem, $\exists$ a Borel measure Γ (that is unique, regular, and countably additive) such that the Gateaux derivative (which is a continuous, linear functional by definition) can be written:

$$\lim_{\epsilon \to 0} \frac{R(T+\epsilon\tau) - R(T)}{\epsilon} = \int_{\mathbf{Z}} \tau(\mathbf{z}) d\Gamma(\mathbf{z}) \tag{52}$$

We aim to show that there exists a continuous linear functional $W(U(\mathbf{n};T))$ such that $T(\mathbf{z})$ is a stationary point for the Lagrangian $L(T;W)$.

Suppose for the moment that almost all $\mathbf{n}$ that choose each $\mathbf{z}$ have a unique optimum and that the distribution $F(\mathbf{n}|\mathbf{z})$ admits a density $f(\mathbf{n}|\mathbf{z})$ for all $\mathbf{z}$. Let us then find an inverse welfare functional of the following form where $\mathbf{N}(\mathbf{z})$ represents the set of $\mathbf{n}$ that optimally choose a given $\mathbf{z}$:

$$W(U(\mathbf{n};T)) = \int_{\mathbf{Z}} \int_{\mathbf{N}(\mathbf{z})} \frac{f(\mathbf{n}|\mathbf{z})}{u_c(\mathbf{n})} U(\mathbf{n};T) dS d\Phi(\mathbf{z}) \tag{53}$$

where dS is the hypersurface element of $\mathbf{N}(\mathbf{z})$. To take the Gateaux derivative of $W(U(\mathbf{n};T))$, we will appeal to the envelope theorem. Recalling that $U(\mathbf{n};T) \equiv u(y(\mathbf{z}(\mathbf{n})) - T(\mathbf{z}(\mathbf{n})), \mathbf{z}(\mathbf{n}); \mathbf{n})$, the envelope theorem implies that for all $\mathbf{n}$ with a unique optima:

$$\lim_{\epsilon \to 0} \frac{U(\mathbf{n};T+\epsilon\tau) - U(\mathbf{n};T)}{\epsilon} = -u_c(y(\mathbf{z}(\mathbf{n})) - T(\mathbf{z}(\mathbf{n})), \mathbf{z}(\mathbf{n}); \mathbf{n})\tau(\mathbf{z}(\mathbf{n})) \equiv -u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n})) \tag{54}$$

While application of the envelope theorem is standard in public economics, we nonetheless rigorously justify its use. Consider optimal utility for a given $\mathbf{n}$ as a function of ϵ where we explicitly note that $\mathbf{z}(\mathbf{n})$ is also a function of ϵ: $u(y(\mathbf{z}(\mathbf{n},\epsilon)) - T(\mathbf{z}(\mathbf{n},\epsilon)) - \epsilon\tau(\mathbf{z}(\mathbf{n},\epsilon)), \mathbf{z}(\mathbf{n},\epsilon); \mathbf{n})$. Note that by standard arguments, any $\mathbf{n}$ with a unique optimum will move continuously in response to a given tax perturbation for sufficiently small ϵ. Theorem 3 of Milgrom and Segal (2002) then implies that Equation 54 holds for any such $\mathbf{n}$ if we can show that $-\tau(\mathbf{z})u_c(y(\mathbf{z}) - T(\mathbf{z}), \mathbf{z}; \mathbf{n})$ is bounded as a function of $\mathbf{z}$ over $\mathbf{Z}$ (the chosen set of $\mathbf{z}$'s) and that

$$\frac{u(y(\mathbf{z}(\mathbf{n})) - T(\mathbf{z}(\mathbf{n})) - \epsilon\tau(\mathbf{z}(\mathbf{n})), \mathbf{z}(\mathbf{n}); \mathbf{n}) - u(y(\mathbf{z}(\mathbf{n})) - T(\mathbf{z}(\mathbf{n})), \mathbf{z}(\mathbf{n}); \mathbf{n})}{\epsilon} \tag{55}$$

converges uniformly in ϵ for all $\mathbf{z}$. Given that $T(\mathbf{z})$ and $\tau(\mathbf{z})$ are both continuous, we will make the technical assumptions that $\sup_{\mathbf{n},\mathbf{z}} u_c(y(\mathbf{z}) - T(\mathbf{z}), \mathbf{z}; \mathbf{n}) < \infty$ and that $\sup_{\mathbf{n},\mathbf{z}} |u_{cc}(y(\mathbf{z}) - T(\mathbf{z}), \mathbf{z}; \mathbf{n})| < \infty$ (this second condition ensures that Equation 55 converges uniformly in ϵ for all $\mathbf{z}$ via a Taylor series argument).

Because for every $\mathbf{z}$, almost all $\mathbf{n}$ have a unique optima, then we have by the envelope theorem:[37]

$$\frac{\partial W(U(\mathbf{n};T+\epsilon\tau))}{\partial\epsilon}\bigg|_{\epsilon=0}=\int_{\mathbf{Z}}\int_{\mathbf{N}(\mathbf{z})}-\frac{f(\mathbf{n}|\mathbf{z})}{u_c(\mathbf{n})}u_c(\mathbf{n})\tau(\mathbf{z})dSd\Phi(\mathbf{z})=-\int_{\mathbf{Z}}\tau(\mathbf{z})d\Phi(\mathbf{z})$$

Given a welfare functional as in Equation 53, we still need to show that there exists a measure Φ that makes the Gateaux derivative of the government's Lagrangian zero $\forall\tau$. Hence, $\forall\tau$ we must have:
$$\frac{\partial L(T+\epsilon\tau;W)}{\partial\epsilon}\bigg|_{\epsilon=0}=-\int_{\mathbf{Z}}\tau(\mathbf{z})d\Phi(\mathbf{z})+\lambda\int_{\mathbf{Z}}\tau(\mathbf{z})d\Gamma(\mathbf{z})=0$$
But from here, we can find an inverse welfare functional by normalizing λ to 1 and and choosing $\Phi=\Gamma$. Finally, we should show that our inverse welfare functional in Equation 53 is a continuous linear functional. First, note that $\mathcal{U}\subset C(\mathbf{N})$ because the utility function is continuous so any indirect profile consistent with individual optimization must be continuous by Berge's Maximum Theorem. Hence, we only need to show that Equation 53 yields a continuous linear functional on $C(\mathbf{N})$ (equipped with the supremum norm). Linearity follows immediately and continuity of $\int_{\mathbf{N}(\mathbf{z})}\frac{f(\mathbf{n}|\mathbf{z})}{u_c(\mathbf{n})}U(\mathbf{n};T)dS$ follows as for two utility profiles $U_1(\mathbf{n};T)$ and $U_2(\mathbf{n};T)$:

$$\left|\int_{\mathbf{N}(\mathbf{z})}\frac{f(\mathbf{n}|\mathbf{z})}{u_c(\mathbf{n})}U_2(\mathbf{n};T)dS-\int_{\mathbf{N}(\mathbf{z})}\frac{f(\mathbf{n}|\mathbf{z})}{u_c(\mathbf{n})}U_1(\mathbf{n};T)dS\right|\le\left\|\frac{1}{u_c(\mathbf{n})}\right\|_\infty||U_2(\mathbf{n};T)-U_1(\mathbf{n};T)||_\infty\le K||U_2(\mathbf{n};T)-U_1(\mathbf{n};T)||_\infty$$

where the final inequality follows assuming that marginal utility of consumption is bounded away from 0 (which is a standard assumption given that $\mathbf{N}$ and $\mathbf{Z}$ are compact). Hence, the inner integral is Lipschitz continuous[38] which implies that $\int_{\mathbf{Z}}\int_{\mathbf{N}(\mathbf{z})}\frac{f(\mathbf{n}|\mathbf{z})}{u_c(\mathbf{n})}U(\mathbf{n};T)dSd\Phi(\mathbf{z})$ is continuous as well given that $\Phi(\mathbf{z})$ defines a continuous functional (on the set of continuous functions over $\mathbf{Z}$) by equivalence with $\Gamma(\mathbf{z})$. □

A.2 Proof of Proposition 3

The total welfare impact of a small tax perturbation $\tau(\mathbf{z})=\tau_1(\mathbf{z})+\tau_2(\mathbf{z})$ where $\tau_1(\mathbf{z})>0$ raises taxes on those with choices in $\mathbf{V}$ and $\tau_2(\mathbf{z})=\tau_2$ is a lump sum transfer that makes the Gateaux derivative of revenue equal to zero is:

$$-\int_{\mathbf{V}}\tau_1(\mathbf{v})\int_{\mathbf{N}(\mathbf{v})}\phi(\mathbf{n})u_c(\mathbf{n})dF(\mathbf{n}|\mathbf{v})h(\mathbf{v})d\mathbf{v}-\tau_2\int_{\mathbf{N}}\phi(\mathbf{n})u_c(\mathbf{n})dF(\mathbf{n})=0$$

[37] If instead, we only have that for each $\mathbf{z}$ $\exists$ at least one $\hat{\mathbf{n}}(\mathbf{z})$ has a unique optimum at the given $\mathbf{z}$, then we can consider the following welfare functional:

$$W(U(\mathbf{n};T))=\int_{\mathbf{Z}}\int_{\mathbf{N}(\mathbf{z})}\frac{U(\mathbf{n};T)}{u_c(\hat{\mathbf{n}}(\mathbf{z}))}d\delta_{\hat{\mathbf{n}}(\mathbf{z})}(\mathbf{n})d\Phi(\mathbf{z})=\int_{\mathbf{Z}}\frac{U(\hat{\mathbf{n}}(\mathbf{z});T)}{u_c(\hat{\mathbf{n}}(\mathbf{z}))}d\Phi(\mathbf{z})$$

where $\delta_{\hat{\mathbf{n}}(\mathbf{z})}$ is the Dirac measure centered at $\hat{\mathbf{n}}(\mathbf{z})$. The Dirac measure centered at $\hat{\mathbf{n}}(\mathbf{z})$ satisfies:

$$\int_{\mathbf{N}(\mathbf{z})}U(\mathbf{n};T)d\delta_{\hat{\mathbf{n}}(\mathbf{z})}(\mathbf{n})=U(\hat{\mathbf{n}}(\mathbf{z});T)$$

This welfare functional also yields that $\frac{\partial W(U(\mathbf{n};T+\epsilon\tau))}{\partial\epsilon}=\int_{\mathbf{Z}}-\frac{1}{u_c(\hat{\mathbf{n}}(\mathbf{z}))}u_c(\hat{\mathbf{n}}(\mathbf{z}))\tau(\mathbf{z})d\Phi(\mathbf{z})=-\int_{\mathbf{Z}}\tau(\mathbf{z})d\Phi(\mathbf{z})$ because at each $\mathbf{z}$ we only care about the welfare of a type $\hat{\mathbf{n}}(\mathbf{z})$ with a unique optimum and we weight them according to the inverse of their marginal utility of consumption.

[38] In the case that we only have at least a single $\hat{\mathbf{n}}(\mathbf{z})$ with a unique optimum at each $\mathbf{z}$, the same arguments can be used to show that the term $\int_{\mathbf{N}(\mathbf{z})}\frac{U(\mathbf{n};T)}{u_c(\hat{\mathbf{n}}(\mathbf{z}))}d\delta_{\hat{\mathbf{n}}(\mathbf{z})}(\mathbf{n})$ is Lipschitz continuous.

The conditions of the proposition imply that:

$$-\int_{\mathbf{V}} \tau_1(\mathbf{v}) \int_{\mathbf{N}(\mathbf{v})} \phi^T(\mathbf{n}) u_c(\mathbf{n}) dF(\mathbf{n}|\mathbf{v}) h(\mathbf{v}) d\mathbf{v} - \tau_2 \int_{\mathbf{N}} \phi^T(\mathbf{n}) u_c(\mathbf{n}) dF(\mathbf{n}) > 0$$

If $\tau_1(\mathbf{v}) < 0$ and the inequality in Equation 34 is reversed, then the above inequality also holds, which completes the proof.

A.3 Proof of Proposition 4

Proof. We prove each statement below:

1. Suppose to the contrary that the Gateaux derivative of $R(T)$, which takes the form $\int_{\mathbf{Z}} \tau(\mathbf{z}) d\Gamma(\mathbf{z})$ for some Borel measure $\Gamma(\mathbf{z})$ by the Riesz-Markov-Kakutani representation theorem, is not positive. Hence, $\exists \tau(\mathbf{z}) \geq 0 \; \forall \mathbf{z}$ such that $\int_{\mathbf{Z}} \tau(\mathbf{z}) d\Gamma(\mathbf{z}) < 0$. Equivalently, by linearity, $\exists \tau(\mathbf{z}) \leq 0 \; \forall \mathbf{z}$ such that $\int_{\mathbf{Z}} \tau(\mathbf{z}) d\Gamma(\mathbf{z}) > 0$. In other words, we have found a way to (weakly) reduce taxes at all $\mathbf{z}$ yet increase revenue. Given that reducing taxes at a given $\mathbf{z}$ makes individuals who choose that $\mathbf{z}$ strictly better off and the fact that we assume marginal utility of consumption is strictly positive, we have found a Pareto improvement. If almost all $\mathbf{n}$ that choose each $\mathbf{z}$ have a unique optima, we can follow Appendix A.1 and consider the welfare functional:

$$W(U(\mathbf{n};T)) = \int_{\mathbf{Z}} \int_{\mathbf{N}(\mathbf{z})} \frac{1}{u_c(\mathbf{n})} U(\mathbf{n};T) f(\mathbf{n}|\mathbf{z}) dS d\Phi(\mathbf{z})$$

   where dS is the surface element of $\mathbf{N}(\mathbf{z})$. This is an inverse welfare functional if we choose $\Phi(\mathbf{z}) = \Gamma(\mathbf{z})$ by the logic of Appendix A.1. Finally, we show that this inverse welfare functional is positive. Suppose not so that $\exists \tilde{U}(\mathbf{n}) \in C(\mathbf{N})$ with $\tilde{U}(\mathbf{n}) > 0$ and $W(U(\mathbf{n};T)) < 0$:

$$\int_{\mathbf{Z}} \int_{\mathbf{N}(\mathbf{z})} \frac{\tilde{U}(\mathbf{n})}{u_c(\mathbf{n})} f(\mathbf{n}|\mathbf{z}) dS d\Gamma(\mathbf{z}) < 0$$

   But then consider $\tau(\mathbf{z}) = \int_{\mathbf{N}(\mathbf{z})} \frac{\tilde{U}(\mathbf{n})}{u_c(\mathbf{n})} f(\mathbf{n}|\mathbf{z}) dS$, yielding that for some $\tau(\mathbf{z}) \geq 0$:

$$\int_{\mathbf{Z}} \tau(\mathbf{z}) d\Gamma(\mathbf{z}) < 0$$

   which is a contradiction given that $\int_{\mathbf{Z}} \tau(\mathbf{z}) d\Gamma(\mathbf{z})$ is the Gateaux derivative of $R(T)$, which we previously established must be positive.

2. Suppose to the contrary that $T(\mathbf{z})$ was not Pareto optimal. Then $\exists T'(\mathbf{z})$ such that $U(\mathbf{n};T') \geq U(\mathbf{n};T) \; \forall \mathbf{n}$ with the inequality strict for some $\mathbf{n}$. Note that by continuity, if $U(\mathbf{n};T') > U(\mathbf{n};T)$ then this holds on some open ball around $\mathbf{n}$. Because $\phi_i(\mathbf{n}) > 0 \; \forall \mathbf{n}$, we have a contradiction because then:

$$\sum_i^M \int_{\mathbf{N}_i} \phi_i(\mathbf{n}) [U(\mathbf{n};T') - U(\mathbf{n};T)] d\mathbf{N}_i > 0$$

□

A.4 Proof to Theorem 2

Proof. First, $T(\mathbf{z})$, and hence $\tau(\mathbf{z})$, are assumed continuous so that if $R(T)$ is Gateaux differentiable then by the Riesz-Markov-Kakutani representation theorem, $\exists$ a Borel measure Γ

(that is unique, regular, and countably additive) such that the Gateaux derivative (which is a continuous, linear functional by definition) can be written:

$$\lim_{\epsilon \to 0} \frac{R(T + \epsilon\tau) - R(T)}{\epsilon} = \int_{\mathbf{Z}} \tau(\mathbf{z}) d\Gamma(\mathbf{z})$$

Similarly, for each $w_i \in \mathbf{w}$ (which is assumed Gateaux differentiable) there exists some Borel measure p_i such that:

$$\lim_{\epsilon \to 0} \frac{w_i(T + \epsilon\tau) - w_i(T)}{\epsilon} = \int_{\mathbf{Z}} \tau(\mathbf{z}) dp_i(\mathbf{z})$$

Next, let us form the government's Lagrangian under a welfare functional W:

$$L = W(U(\mathbf{n}; T, \mathbf{w})) + \lambda R(T)$$

We aim to show that there exists a linear functional $W(U(\mathbf{n}; T, \mathbf{w}))$ such that $T(\mathbf{z})$ is a stationary point for the Lagrangian $L(T; W)$. Let us suppose for simplicity that almost all $\mathbf{n}$ at each $\mathbf{z}$ have a unique optima. We will construct an inverse welfare functional that takes the following form for some Borel measure Φ where $\mathbf{N}(\mathbf{z})$ represents the set of types $\mathbf{n}$ that choose a given $\mathbf{z}$ and dS represents the hypersurface element of $\mathbf{N}(\mathbf{z})$:

$$W(U(\mathbf{n}; T, \mathbf{w})) = \int_{\mathbf{Z}} \int_{\mathbf{N}(\mathbf{z})} \frac{1}{u_c(\mathbf{n})} U(\mathbf{n}; T, \mathbf{w}) f(\mathbf{n}|\mathbf{z}) dS d\Phi(\mathbf{z}) \tag{56}$$

To take the Gateaux derivative of $W(U(\mathbf{n}; T, \mathbf{w}))$ we will appeal to the envelope theorem. Recalling that $U(\mathbf{n}; T, \mathbf{w}) \equiv u(y(\mathbf{z}(\mathbf{n}), \mathbf{w}) - T(\mathbf{z}(\mathbf{n})), \mathbf{z}(\mathbf{n}); \mathbf{n}, \mathbf{w})$ and that $\mathbf{w}$ is a function of the tax schedule, the envelope theorem implies that for individuals with a unique optima:[39]

$$\begin{aligned}
&\lim_{\epsilon \to 0} \frac{U(\mathbf{n}; T + \epsilon\tau, \mathbf{w}) - U(\mathbf{n}; T, \mathbf{w})}{\epsilon} \\
&= \lim_{\epsilon \to 0} \frac{u(y(\mathbf{z}(\mathbf{n})) - T(\mathbf{z}(\mathbf{n})) + \epsilon\tau(\mathbf{z}(\mathbf{n})), \mathbf{z}(\mathbf{n}); \mathbf{n}, \mathbf{w}(\epsilon)) - u(y(\mathbf{z}(\mathbf{n})) - T(\mathbf{z}(\mathbf{n})), \mathbf{z}(\mathbf{n}); \mathbf{n}, \mathbf{w}(\epsilon))}{\epsilon} \\
&= -u_c(y(\mathbf{z}(\mathbf{n})) - T(\mathbf{z}(\mathbf{n})), \mathbf{z}(\mathbf{n}); \mathbf{n})\tau(\mathbf{z}) + \sum_i u_{w_i}(\mathbf{n}) \frac{\partial w_i}{\partial \epsilon} \\
&= -u_c(y(\mathbf{z}(\mathbf{n})) - T(\mathbf{z}(\mathbf{n})), \mathbf{z}(\mathbf{n}); \mathbf{n})\tau(\mathbf{z}) + \sum_i u_{w_i}(\mathbf{n}) \int_{\mathbf{Z}} \tau(\mathbf{z}) dp_i(\mathbf{z})
\end{aligned}$$

Applying the envelope theorem to compute the Gateaux derivative of $W(U(\mathbf{n}; T, \mathbf{w}))$ we get the following expression for the Gateaux derivative of the Lagrangian:

$$\int_{\mathbf{Z}} \int_{\mathbf{N}(\mathbf{z})} \frac{-u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n})) + \sum_i u_{w_i}(\mathbf{n}) \int_{\mathbf{Z}} \tau(\mathbf{z}) dp_i(\mathbf{z})}{u_c(\mathbf{n})} f(\mathbf{n}|\mathbf{z}) dS d\Phi(\mathbf{z}) + \lambda \int_{\mathbf{Z}} \tau(\mathbf{z}) d\Gamma(\mathbf{z}) \tag{57}$$

Define $\overline{\frac{u_{w_i}}{u_c}}(\mathbf{z}) \equiv \int_{\mathbf{N}(\mathbf{z})} \frac{u_{w_i}(\mathbf{n})}{u_c(\mathbf{n})} f(\mathbf{n}|\mathbf{z}) dS$ so that Equation 57 can be rewritten:[40]

$$-\int_{\mathbf{Z}} \tau(\mathbf{z}) d\Phi(\mathbf{z}) + \sum_i \int_{\mathbf{Z}} \overline{\frac{u_{w_i}}{u_c}}(\mathbf{z}) d\Phi(\mathbf{z}) \int_{\mathbf{Z}} \tau(\mathbf{z}) dp_i(\mathbf{z}) + \lambda \int_{\mathbf{Z}} \tau(\mathbf{z}) d\Gamma(\mathbf{z}) \tag{58}$$

[39] See the proof to Theorem 1 for a rigorous justification of the envelope theorem application here.

[40] As in Theorem 1, if we only have that (at least) a single type $\hat{\mathbf{n}}(\mathbf{z})$ has a unique optimum at each $\mathbf{z}$ then we can use the welfare functional:

$$W(U(\mathbf{n}; T, \mathbf{w})) = \int_{\mathbf{Z}} \int_{\mathbf{N}(\mathbf{z})} \frac{1}{u_c(\hat{\mathbf{n}}(\mathbf{z}))} U(\mathbf{n}; T, \mathbf{w}) d\delta_{\hat{\mathbf{n}}(\mathbf{z})}(\mathbf{n}) d\Phi(\mathbf{z})$$

where $\delta_{\hat{\mathbf{n}}(\mathbf{z})}$ is the Dirac measure centered at $\hat{\mathbf{n}}(\mathbf{z})$. In this case, Equation 58 still holds but now $\overline{\frac{u_{w_i}}{u_c}}(\mathbf{z}) \equiv \frac{u_{w_i}(\hat{\mathbf{n}}(\mathbf{z}))}{u_c(\hat{\mathbf{n}}(\mathbf{z}))}$.

Or, changing the dummy variable of integration in $\int_{\mathbf{Z}} \frac{u_{w_i}}{u_c}(\mathbf{z})d\Phi(\mathbf{z})$ from $\mathbf{z}$ to $\tilde{\mathbf{z}}$, we have:

$$\int_{\mathbf{Z}} \tau(\mathbf{z}) \left(-d\Phi(\mathbf{z}) + \sum_i dp_i(\mathbf{z}) \int_{\mathbf{Z}} \frac{u_{w_i}}{u_c}(\tilde{\mathbf{z}})d\Phi(\tilde{\mathbf{z}}) + \lambda d\Gamma(\mathbf{z}) \right) \tag{59}$$

If the tax schedule $T(\mathbf{z})$ is a local extremum of the government's Lagrangian, then the Gateaux derivative of L is zero. A sufficient condition for this is that for all measurable $\mathbf{E} \subseteq \mathbf{Z}$ we have:

$$\int_{\mathbf{E}} \left(-d\Phi(\mathbf{z}) + \sum_i dp_i(\mathbf{z}) \int_{\mathbf{Z}} \frac{u_{w_i}}{u_c}(\tilde{\mathbf{z}})d\Phi(\tilde{\mathbf{z}}) + \lambda d\Gamma(\mathbf{z}) \right) = 0 \tag{60}$$

Or, expressing Equation 60 in terms of measures with $\Phi(\mathbf{E}) \equiv \int_{\mathbf{E}} d\Phi(\mathbf{z})$, $\Gamma(\mathbf{E}) \equiv \int_{\mathbf{E}} d\Gamma(\mathbf{z})$, and $p_i(\mathbf{E}) \equiv \int_{\mathbf{E}} dp_i(\mathbf{z})$, we have (normalizing $\lambda = 1$):

$$\Phi(\mathbf{E}) = \Gamma(\mathbf{E}) + \sum_i p_i(\mathbf{E}) \int_{\mathbf{Z}} \frac{u_{w_i}}{u_c}(\tilde{\mathbf{z}})d\Phi(\tilde{\mathbf{z}}) \tag{61}$$

which is an integral equation formulated in a measure space as in Das (1974) or Sharma (1975). As in Subsection 6.2, we are going to show that the map $(Q\Phi)(\mathbf{E}) = \Gamma(\mathbf{E}) + \sum_i p_i(\mathbf{E}) \int_{\mathbf{Z}} \frac{u_{w_i}}{u_c}(\tilde{\mathbf{z}})d\Phi(\tilde{\mathbf{z}})$ is a contraction mapping on the set of regular, countably additive Borel measures. Note that the space of regular, countably additive Borel measures on $\mathbf{Z}$, denoted $rca(\mathbf{Z})$, is a Banach space when equipped with the "total variation" norm (hence, we can apply the contraction mapping theorem):

$$||\mu||_{\text{TV}} = \sup_{||f||_\infty \leq 1} \int f d\mu \tag{62}$$

Thus, for two measures Φ and Φ', consider the total variation norm of $(Q\Phi') - (Q\Phi)$:

$$\begin{aligned} ||(Q\Phi') - (Q\Phi)||_{\text{TV}} &= \left|\left| \sum_i p_i \int_{\mathbf{Z}} \frac{u_{w_i}}{u_c}(\tilde{\mathbf{z}})d(\Phi'(\tilde{\mathbf{z}}) - \Phi(\tilde{\mathbf{z}})) \right|\right|_{\text{TV}} \\ &\leq \sum_i ||p_i||_{\text{TV}} \left| \int_{\mathbf{Z}} \frac{u_{w_i}}{u_c}(\tilde{\mathbf{z}})d(\Phi'(\tilde{\mathbf{z}}) - \Phi(\tilde{\mathbf{z}})) \right| \\ &= \sum_i ||p_i||_{\text{TV}} \left|\left| \frac{u_{w_i}}{u_c} \right|\right|_\infty \left| \int_{\mathbf{Z}} \frac{u_{w_i}}{u_c}(\tilde{\mathbf{z}}) \Big/ \left|\left| \frac{u_{w_i}}{u_c} \right|\right|_\infty d(\Phi'(\tilde{\mathbf{z}}) - \Phi(\tilde{\mathbf{z}})) \right| \\ &\leq \sum_i ||p_i||_{\text{TV}} \left|\left| \frac{u_{w_i}}{u_c} \right|\right|_\infty ||\Phi' - \Phi||_{\text{TV}} \\ &< ||\Phi' - \Phi||_{\text{TV}} \end{aligned} \tag{63}$$

Let us explain the steps detailed in Equation 63. The first line simply uses the definition of the measure $(Q\Phi') - (Q\Phi)$ from Equation 61. The second line uses the triangle inequality and the absolute homogeneity of the norm. The third line just multiplies and divides by $\left|\left|\frac{u_{w_i}}{u_c}\right|\right|_\infty$ (recognize that $\frac{u_{w_i}}{u_c}$ is a function of $\mathbf{z}$; hence, the supnorm is taken over $\mathbf{z}$). The fourth line uses the definition of the total variation norm in Equation 62. The final line uses our assumption on the size of $\sum_i ||p_i||_{\text{TV}} \left|\left|\frac{u_{w_i}}{u_c}\right|\right|_\infty$. Hence, $(Q\Phi)(E)$ is a contraction mapping, which implies the existience of a (unique) fixed point Φ which solves Equation 61. Hence, we have proved the existence of an inverse welfare functional taking the form of Equation 56. Finally, note that this inverse welfare functional can be shown to be continuous and linear using the arguments at the end of Appendix A.1. □

B Online Appendix: Additional Results

B.1 Existence of Global Inverse Welfare Functionals

We first need to introduce the multidimensional envelope theorem. Consider an allocation $(\tilde{T}(\mathbf{n}), \tilde{\mathbf{z}}(\mathbf{n}))$ (which is not necessarily generated by optimization under a tax schedule) which induces a utility profile $V(\mathbf{n}) = u(y(\tilde{\mathbf{z}}(\mathbf{n})) - \tilde{T}(\mathbf{n}), \tilde{z}(\mathbf{n}); \mathbf{n})$. We say that $V(\mathbf{n})$ satisfies the envelope condition if for any $\mathbf{n}_1$ and $\mathbf{n}_2$ and any path between these two points:

$$V(\mathbf{n}_1) - V(\mathbf{n}_2) = \int_{\mathbf{n}_2}^{\mathbf{n}_1} \nabla_{\mathbf{n}} u(y(\mathbf{z}) - T, \mathbf{z}; \mathbf{n})|_{T=\tilde{T}(\mathbf{n}), \mathbf{z}=\tilde{\mathbf{z}}(\mathbf{n})} \cdot d\mathbf{n} \tag{64}$$

Alternatively, for a.e. $\mathbf{n}$, we can consider the following "derivative version" of the envelope theorem:

$$\nabla_{\mathbf{n}} V(\mathbf{n}) = \nabla_{\mathbf{n}} u(y(\mathbf{z}) - T, \mathbf{z}; \mathbf{n})|_{T=\tilde{T}(\mathbf{n}), \mathbf{z}=\tilde{\mathbf{z}}(\mathbf{n})} \tag{65}$$

Equation 65 and $V(\mathbf{n}) = u(y(\tilde{\mathbf{z}}(\mathbf{n})) - \tilde{T}(\mathbf{n}), \tilde{z}(\mathbf{n}); \mathbf{n})$ define an a.e. correspondence $(V(\mathbf{n}), \nabla_{\mathbf{n}} V(\mathbf{n})) \mapsto (\tilde{T}(\mathbf{n}), \tilde{\mathbf{z}}(\mathbf{n}))$. Let us then define the object $\tilde{T}(V(\mathbf{n}), \nabla_{\mathbf{n}} V(\mathbf{n}))$ as a selection from this correspondence. Finally, let us define the set $\mathcal{V} \equiv \{V(\mathbf{n}) \text{ s.t. } \int_{\mathbf{N}} \tilde{T}(V(\mathbf{n}), \nabla_{\mathbf{n}} V(\mathbf{n})) dF(\mathbf{n}) \geq E\}$. We can then state:

Proposition 6. *Suppose all selections $\tilde{T}(V(\mathbf{n}), \nabla_{\mathbf{n}} V(\mathbf{n}))$ are concave in $(V(\mathbf{n}), \nabla_{\mathbf{n}} V(\mathbf{n}))$. Consider a tax schedule T such that $\nabla_{\mathbf{n}} u(y(\mathbf{z}) - T(\mathbf{z}), \mathbf{z}; \mathbf{n})$ is bounded on $\mathbf{Z} \times \mathbf{N}$. If T induces a utility profile $U(\mathbf{n}; T)$ on the boundary of the set $\mathcal{V}$ then T has an associated global inverse welfare functional.*

Proof. We are first going to show that the set $\mathcal{V}$ is convex. Consider $V_1(\mathbf{n}), V_2(\mathbf{n}) \in \mathcal{V}$. Now, for all $\mathbf{n}$ we have that:

$$\tilde{T}(\alpha V_1(\mathbf{n}) + (1-\alpha) V_2(\mathbf{n}), \alpha \nabla_{\mathbf{n}} V_1(\mathbf{n}) + (1-\alpha) \nabla_{\mathbf{n}} V_2(\mathbf{n})) \geq \alpha \tilde{T}(V_1(\mathbf{n}), \nabla_{\mathbf{n}} V_1(\mathbf{n})) + (1-\alpha) \tilde{T}(V_2(\mathbf{n}), \nabla_{\mathbf{n}} V_2(\mathbf{n}))$$

Hence:

$$\int_{\mathbf{N}} \tilde{T}(\alpha V_1(\mathbf{n}) + (1-\alpha) V_2(\mathbf{n}), \alpha \nabla_{\mathbf{n}} V_1(\mathbf{n}) + (1-\alpha) \nabla_{\mathbf{n}} V_2(\mathbf{n})) dF(\mathbf{n}) \geq E$$

Thus, we know that $\alpha V_1(\mathbf{n}) + (1-\alpha) V_2(\mathbf{n}) \in \mathcal{V}$, so that $\mathcal{V}$ is convex, as claimed.

By the geometric version of the Hahn-Banach Theorem (i.e., the infinite dimensional supporting hyperplane theorem), we know that for a convex set $\mathcal{V} \subset C(\mathbf{N})$ and $V \in \mathcal{V} \backslash \mathrm{Int}(\mathcal{V})$, there exists a continuous linear functional W that supports V:

$$W(V) = \sup_{V' \in \mathcal{V}} W(V')$$

Finally, we note that all feasible utility profiles generated by a optimization under a tax schedule $U(\mathbf{n}; T) \in \mathcal{U}$ must be within $\mathcal{V}$. This results by Corollary 1 of Milgrom and Segal (2002) which yields that any utility profile $U(\mathbf{n}; T)$ generated by a tax schedule $T(\mathbf{z})$ must satisfy the envelope condition 64 as long as $\nabla_{\mathbf{n}} u(y(\mathbf{z}) - T(\mathbf{z}), \mathbf{z}; \mathbf{n})$ is bounded on $\mathbf{Z} \times \mathbf{N}$ (recall we assume $\nabla_{\mathbf{n}} u(y(\mathbf{z}) - T(\mathbf{z}), \mathbf{z}; \mathbf{n})$ is continuous). Thus, if $U \in \mathcal{U}$ is on the boundary of $\mathcal{V} \supset \mathcal{U}$ then we clearly have:

$$W(U) = \sup_{V' \in \mathcal{V}} W(V') \geq \sup_{U' \in \mathcal{U}} W(U')$$

Given that $U \in \mathcal{U}$, we trivially have that $W(U) \leq \sup_{U' \in \mathcal{U}} W(U')$ so that $W(U) = \sup_{U' \in \mathcal{U}} W(U')$.

Thus, W is a global inverse optimal welfare functional for $U(\mathbf{n};T)$. □

Proposition 6 ensures that if we find a tax schedule generating an indirect utility profile on the boundary of the set of indirect utility profiles satisfying the envelope condition and the budget constraint, then we can find a global inverse optimal functional which supports that profile relative to all other feasible indirect utility profiles.[41] In practice, determining whether an indirect utility profile is on the boundary of $\mathcal{V}$ is relatively simple: it is sufficient to find an indirect utility profile arbitrarily close by that satisfies the envelope condition yet does not satisfy the budget constraint (typically, any indirect utility profile that satisfies the budget constraint with equality and also satisfies the envelope condition will be on the boundary of $\mathcal{V}$).

Remark 5. *As an example application of Proposition 6, suppose that* $\mathbf{N}$ *is a compact subset of* $(-\infty,0)^K$ *and*

$$u(y(z)-T,\mathbf{z};\mathbf{n}) = \frac{\left(\sum_{i=1}^K z_i - T\right)^{1-\sigma}}{1-\sigma} + \sum_{i=1}^K n_i \frac{z_i^{1+\theta_i}}{1+\theta_i}$$

with $z_1, z_2, ..., z_K \geq 0$ *and* $\theta_1, \theta_2, ..., \theta_K \geq 0$. *Then we have:*

$$\tilde{T}(V,\nabla_{\mathbf{n}}V) = \sum_{i=1}^K \left((1+\theta_i)\frac{\partial V}{\partial n_i}\right)^{\frac{1}{1+\theta_i}} - \left((1-\sigma)\left[V - \sum_{i=1}^K n_i \frac{\partial V}{\partial n_i}\right]\right)^{\frac{1}{1-\sigma}}$$

It is then straight-forward to establish then that $\tilde{T}(V,\nabla_{\mathbf{n}}V)$ *is concave as long as* $\sigma < 1$ *(noting that* $(1-\sigma)\left[V - \sum_{i=1}^K n_i \frac{\partial V}{\partial n_i}\right]$ *is always positive).*

B.2 Continuity of Tax Schedules

We establish conditions under which the tax schedule can be assumed continuous WLOG:

Lemma 2. *Suppose that given a* $T(\mathbf{z})$, *the set of choices,* $\mathbf{Z} = \{\mathbf{z}(\mathbf{n})|\mathbf{n} \in \mathbf{N}\}$, *is bounded. Further, suppose that indifference surfaces have bounded gradients:*

$$\left\|\frac{\nabla_{\mathbf{z}} u(c,\mathbf{z};\mathbf{n})}{u_c(c,\mathbf{z};\mathbf{n})}\right\| < M \ \forall \mathbf{n} \in \mathbf{N}, (c,\mathbf{z}) \text{ s.t. } \mathbf{z} \in \mathbf{Z} \text{ and } u(c,\mathbf{z};\mathbf{n}) = u(c(\mathbf{z}(\mathbf{n})),\mathbf{z}(\mathbf{n});\mathbf{n})$$

Then $\exists$ *(Lipschitz) continuous* $\tilde{T}(\mathbf{z})$ *that generates the same indirect utility profile as* $T(\mathbf{z})$: $U(\mathbf{n};T) = U(\mathbf{n};\tilde{T})$.

Proof. Under $T(\mathbf{z})$, for each type $\mathbf{n}$ consider the indifference surface, $\hat{c}(\mathbf{z};\mathbf{n})$, that goes through each of their (potentially multiple) optimal $\mathbf{z}(\mathbf{n})$. Note that each such indifference surface is implicitly defined by:

$$u(\hat{c}(\mathbf{z};\mathbf{n}),\mathbf{z};\mathbf{n}) = u(c(\mathbf{z}(\mathbf{n})),\mathbf{z}(\mathbf{n});\mathbf{n})$$

where $\mathbf{z}(\mathbf{n})$ denotes optimal choices for type $\mathbf{n}$ under tax schedule $T(\mathbf{z})$. Implicitly differentiating:

$$u_c(\hat{c}(\mathbf{z};\mathbf{n}),\mathbf{z};\mathbf{n})\nabla_{\mathbf{z}}\hat{c}(\mathbf{z};\mathbf{n}) + \nabla_{\mathbf{z}} u(\hat{c}(\mathbf{z};\mathbf{n}),\mathbf{z};\mathbf{n}) = 0$$

Equivalently:

$$\nabla_{\mathbf{z}}\hat{c}(\mathbf{z};\mathbf{n}) = -\frac{\nabla_{\mathbf{z}} u(\hat{c}(\mathbf{z};\mathbf{n}),\mathbf{z};\mathbf{n})}{u_c(\hat{c}(\mathbf{z};\mathbf{n}),\mathbf{z};\mathbf{n})}$$

By assumption then, the norm of the gradient of the indifference surface that goes through the optimal $\mathbf{z}(\mathbf{n})$ is bounded by M for every $\mathbf{n}$. Therefore the function $\hat{c}(\mathbf{z};\mathbf{n})$ is Lipschitz continuous

[41] In general, the associated global inverse functional need not be unique because the supporting hyperplane of a given point on the boundary of a convex set need not be unique.

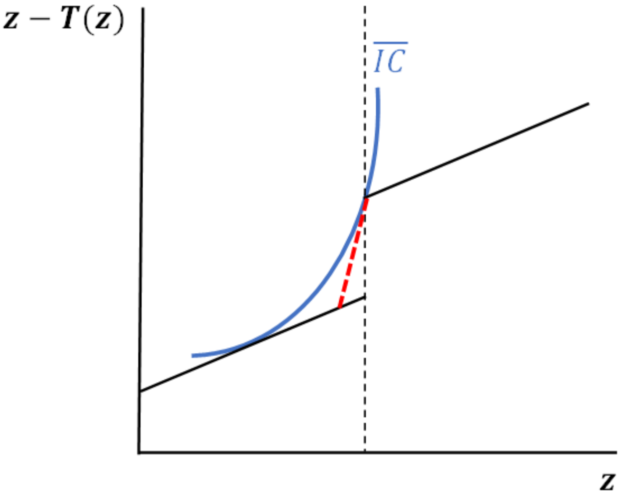


Figure 8: Alternative Continuous Tax Schedule for a Discontinuous Tax Schedule

Note: This figure shows a consumption schedule, $c(z) = z - T(z)$, in solid black corresponding to a discontinuous tax schedule along with the steepest indifference curve of an individual with multiple optima, shown in blue and labeled $\overline{IC}$, along with an alternative continuous tax schedule that yields equivalent welfare by replacing the relevant portion of the original tax schedule with the red dashed line.

(with constant M) for each $\mathbf{n}$.

Next, consider the consumption function, $\underline{c}(\mathbf{z})$, defined as the lower envelope of the family of functions $\{\hat{c}(\mathbf{z}; \mathbf{n})\}$. The lower envelope of a family of Lipschitz continuous functions with Lipschitz constant M is also Lipschitz continuous with Lipschitz constant M (see, for example, Proposition 6.3 of Choquet (1966)).

Now, under $\underline{c}(\mathbf{z})$ everyone (weakly) prefers his/her original optimal $\mathbf{z}(\mathbf{n})$ (and associated consumption level) to any of the points on this new consumption schedule defined by the lower envelope of indifference surfaces (by construction). Thus, we have constructed a Lipschitz continuous consumption schedule that yields the same welfare as our original discontinuous consumption schedule. Given that $c(\mathbf{z}) = y(\mathbf{z}) - T(\mathbf{z})$, any Lipschitz continuous consumption schedule defines a Lipschitz continuous tax schedule (as $y(\mathbf{z})$ is presumed smooth, hence Lipschitz). Hence, we have shown how to construct a Lipschitz continuous tax schedule that generates the same indirect utility profile as our original optimal tax schedule; thus, WLOG we can restrict attention to Lipschitz continuous optimal tax schedules. □

The intuition of Lemma 2 is that as long as indifference surfaces have bounded gradients for all types, then wherever the optimal tax schedule is discontinuous, we can always construct an alternative continuous tax schedule that coincides with the discontinuous tax schedule at all incomes which are chosen in equilibrium by some type yet also lies everywhere below everyone's indifference curves and hence leads to the same allocation. Figure 8 illustrates this alternative continuous tax schedule that yields equivalent welfare by replacing the relevant portion of the original tax schedule with the red dashed line, which lies everywhere below the steepest indifference curve labeled $\overline{IC}$.

B.3 Multiple Optima and Unidimensional Heterogeneity Example

We work through a similar example as in Section 3.2, but now only have a single dimension of heterogeneity. We consider a unidimensional tax schedule $T(z)$ with utility given by $u(c, z/n)$ where $n \in [\underline{n}, \overline{n}]$. Suppose that $u(c, z/n) = c - (z/n)^{1+k}/(1+k)$, which satisfies the Mirrlees (1971) single crossing property ensuring that $z(n)$ is monotonic in n and $c = z - T(z)$. Suppose that we want to find an inverse welfare functional for a piecewise linear tax schedule with two brackets for which the budget constraint is satisfied with equality; the marginal tax rates in the three brackets are denoted T_1, T_2 with $T_1 > T_2$ (so that we have one kink point with decreasing marginal rates). In other words, we want to find a welfare function such that this piecewise linear schedule is the optimal *non-linear* tax schedule.

Let us calculate the Gateaux derivative of $R(T)$. First, note that:

$$R(T) = \int_N T(z(n))f(n)dn$$

Let us consider the impacts of a tax perturbation from $T(z)$ to $T(z) + \epsilon\tau(z)$. First, recognize that no individual will locate at the kink point K_1 where marginal tax rates decrease, this should be immediate from an indifference curve diagram. By the single crossing property, $z(n)$ is monotonic in n so that there must be some individual n_1 who is indifferent between locating in the first bracket and in the second tax bracket. Thus, we split up the domain N into two regions: $[\underline{n}, n_1]$: the set of individuals locating in the first tax bracket and $(n_1, \overline{n}]$: the set of individuals locating in the second tax bracket. We can write tax revenue as:

$$\int_{\underline{n}}^{n_1} T(z(n))f(n)dn + \int_{n_1}^{\overline{n}} T(z(n))f(n)dn \tag{66}$$

We have the individual first order condition:

$$\big(1 - T'(z) - \epsilon\tau'(z)\big) - \frac{1}{n}\left(\frac{z}{n}\right)^k = 0 \tag{67}$$

For all individuals with a unique optimum where the tax schedule is twice continuously differentiable the second order condition holds strictly (see Lemma 3 of Bergstrom and Dodds (2021)), hence we can apply the implicit function theorem to determine the impact of a tax perturbation (note that $T''(z) = 0$ everywhere that $T'(z)$ exists):

$$\frac{\partial z}{\partial \epsilon}(n) = -\frac{\tau'(z)}{\frac{1}{n^2}\left(\frac{z}{n}\right)^{k-1}} \equiv \xi(n)\tau'(z(n)) \tag{68}$$

where $\xi(n) \equiv -\frac{1}{\frac{1}{n^2}\left(\frac{z}{n}\right)^{k-1}}$. Next, we consider the behavioral responses of the type n_1 with multiple optima who is indifferent between locating in the first and second tax brackets. Denoting z^- and z^+ the upper and lower optimal incomes for type n_1 we have:

$$z^- T(z^-) - \epsilon\tau(z^-) - (z^-/n_1)^{1+k}/(1+k) = z^+ T(z^+) - \epsilon\tau(z^+) - (z^+/n_1)^{1+k}/(1+k) \tag{69}$$

We can also calculate how the indifferent individual changes with the tax schedule by applying the implicit function theorem to Equation 69:

$$\frac{\partial n_1}{\partial \epsilon} = \frac{\tau(z^+) - \tau(z^-)}{\frac{1}{n_1}\left(\frac{z^+}{n_1}\right)^{1+k} - \frac{1}{n_1}\left(\frac{z^-}{n_1}\right)^{1+k}} \tag{70}$$

Let us then calculate the Gateaux derivative of R in the direction of $\tau(z)$:

$$\begin{aligned} &\lim_{\epsilon \to 0} \frac{R(T+\epsilon\tau) - R(T)}{\epsilon} \\ &= \int_{\underline{n}}^{n_1} \left[\frac{T(z(n))}{\partial \epsilon} + \tau(z(n))\right] f(n)dn + \int_{n_1}^{\overline{n}} \left[\frac{T(z(n))}{\partial \epsilon} + \tau(z(n))\right] f(n)dn + \left(T(z^-) - T(z^+)\right) f(n_1)\frac{\partial n_1}{\partial \epsilon} \end{aligned} \tag{71}$$

Note, the last term of Equation 71 results from applying Leibniz integral rule. Plugging in the value of $\frac{\partial z}{\partial \epsilon}(n)$ from the implicit function theorem (Equation 68) and changing the variable of integration from n to z we can rewrite Equation 71 as:[42]

$$\begin{aligned} &\int_{z(\underline{n})}^{z^-} \left(T'(z)\xi(z)\tau'(z) + \tau(z)\right) h(z)dz + \int_{z^+}^{z(\overline{n})} \left(T'(z)\xi(z)\tau'(z) + \tau(z)\right) h(z)dz \\ &+ \left(T(z^-) - T(z^+)\right) f(n_1)\frac{\tau(z^+) - \tau(z^-)}{\frac{1}{n_1}\left(\frac{z^+}{n_1}\right)^{1+k} - \frac{1}{n_1}\left(\frac{z^-}{n_1}\right)^{1+k}} \end{aligned} \tag{72}$$

Next, let us apply integration by parts to get rid of the $\tau'(z)$ terms in Equation 72:

$$\begin{aligned} &\lim_{\epsilon \to 0} \frac{R(T+\epsilon\tau) - R(T)}{\epsilon} \\ &= \int_{z(\underline{n})}^{z^-} \left(h(z) - \frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right]\right)\tau(z)dz + \int_{z^+}^{z(\overline{n})} \left(h(z) - \frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right]\right)\tau(z)dz \\ &+ T'(z)\xi(z)\tau(z)|_{z(\underline{n})}^{z^-} + T'(z)\xi(z)\tau(z)|_{z^+}^{z(\overline{n})} \\ &+ \left(T(z^-) - T(z^+)\right) f(n_1)\frac{\tau(z^+) - \tau(z^-)}{\frac{1}{n_1}\left(\frac{z^+}{n_1}\right)^{1+k} - \frac{1}{n_1}\left(\frac{z^-}{n_1}\right)^{1+k}} \end{aligned} \tag{73}$$

Note that all $\tau(z)$ terms enter Equation 73 linearly so that Equation 73 is a linear functional of $\tau(z)$ which means that $R(T)$ is Gateaux differentiable (assuming that all terms in Equation 73 are bounded so that Equation 73 is a linear, bounded (hence continuous) functional of $\tau(z)$). However, let us consider the welfare impacts of a tax perturbation, assuming that welfare is a linear functional of indirect utility $W(U(n;T))$. By the envelope theorem, the derivative of indirect utility with respect to ϵ of a tax perturbation from $T(z)$ to $T(z) + \epsilon\tau(z)$ for any type $n \neq n_1$ just equals $\tau(z(n))$. However, the utility impact of *any* tax perturbation $T(z)$ to $T(z) + \epsilon\tau(z)$ for type n_1 is *not* a linear function of $\tau(z^-)$ and $\tau(z^+)$. For instance, a tax perturbation that changes tax rates around z^- (but leaves taxes around z^+ unchanged) will have a utility impact proportional to $\tau(z^-)$ for type n_1 whereas a tax perturbation that changes tax rates around z^+ (but leaves taxes around z^- unchanged) will have a utility impact proportional

[42] Note by monotonicity that $H(z(n)) = F(n)$ so that $h(z(n)) = f(n)\left(\frac{\partial z(n)}{\partial n}\right)^{-1}$ so that $h(z)$ accounts for the Jacobian of the change of variables.

to $\tau(z^{+})$ for type n_1. Fundamentally, indirect utility for type n_1 is *not* differentiable in ϵ because $z(n)$ is not continuous at n_1 (so that we cannot apply the envelope theorem, e.g., Theorem 3 of Milgrom and Segal (2002)); this non-differentiability implies non-existence of a local inverse welfare functional. If such a local inverse welfare functional $W(U(n;T))$ existed then $W(U(n;T))$ must put positive "mass" on the utility of type n_1 in order for Equation 73 plus $\frac{\partial W(U(n;T))}{\partial \epsilon}$ to equal zero. But because indirect utility for type n_1 is *not* differentiable in ϵ, this means that $\frac{\partial W(U(n;T))}{\partial \epsilon}$ will not, in general, depend linearly on $\tau(z^{-})$ and $\tau(z^{+})$. The one exception is the knife-edge case wherein:

$$\frac{-\left(T(z^{-})-T(z^{+})\right)f(n_1)}{\frac{1}{n_1}\left(\frac{z^{+}}{n_1}\right)^{1+k}-\frac{1}{n_1}\left(\frac{z^{-}}{n_1}\right)^{1+k}}+T'(z^{-})\xi(z^{-})=0,\ \frac{\left(T(z^{-})-T(z^{+})\right)f(n_1)}{\frac{1}{n_1}\left(\frac{z^{+}}{n_1}\right)^{1+k}-\frac{1}{n_1}\left(\frac{z^{-}}{n_1}\right)^{1+k}}-T'(z^{+})\xi(z^{+})=0 \tag{74}$$

so that Equation 73 does not depend on $\tau(z^{-})$ and $\tau(z^{+})$. In this case, $W(U(n;T))$ does not need to put positive mass on the utility of type n_1 and the fact that indirect utility for type n_1 is *not* differentiable in ϵ is unimportant because $\{n_1\}$ is measure zero within $[\underline{n},\overline{n}]$. However, for arbitrary tax schedules, Equation 74 typically does not hold. For example, we compute the values in Equation 74 for a two bracket tax system with marginal tax rates of 60% and 40% (and a kink at \$25,000) using a value of $k = 1/0.3$ and $f(n)$ calibrated to the U.S. income distribution from the 2019 ACS, finding that:

$$\frac{-\left(T(z^{-})-T(z^{+})\right)f(n_1)}{\frac{1}{n_1}\left(\frac{z^{+}}{n_1}\right)^{1+k}-\frac{1}{n_1}\left(\frac{z^{-}}{n_1}\right)^{1+k}}+T'(z^{-})\xi(z^{-})=-0.0534$$

$$\frac{\left(T(z^{-})-T(z^{+})\right)f(n_1)}{\frac{1}{n_1}\left(\frac{z^{+}}{n_1}\right)^{1+k}-\frac{1}{n_1}\left(\frac{z^{-}}{n_1}\right)^{1+k}}-T'(z^{+})\xi(z^{+})=-0.0391$$

B.4 Proof of Equation 20

Proof. We assume that that conditional on each v, $u(c, z/n; v)$ satisfies the Mirrlees (1971) single crossing property which ensures that $z(n,v)$ is monotonic in n $\forall v$. We also assume that $z(n,v)$ is monotonic in v. Let us calculate the Gateaux derivative of $R(T)$. First, note that:

$$R(T)=\int_V\int_N T(z(n,v))f(n,v)dndv$$

Let us consider the impacts of a tax perturbation from $T(z)$ to $T(z)+\epsilon\tau(z)$. First, let us consider the impacts of such a perturbation on all of the types n for a fixed v. First, split up the domain N into four regions: $[\underline{n}, n_1]$: the set of individuals locating in the first tax bracket, $(n_1, n_2]$: the set of individuals bunching at the first kink, K_1, $(n_2, n_3]$: the set of individuals locating in the second tax bracket, and $(n_3, \overline{n}]$: the set of individuals locating in the third tax bracket (note that marginal tax rates decrease at K_2 so this generates an individual n_3 with multiple optima rather than bunching as in Bergstrom and Dodds (2021)). We can write tax revenue as:

$$\begin{aligned}&\int_V \left\{ \int_{\underline{n}}^{n_1(v)} T(z(n,v))f(n|v)dn + \int_{n_1(v)}^{n_2(v)} T(z(n,v))f(n|v)dn \right. \\ &\left. + \int_{n_2(v)}^{n_3(v)} T(z(n,v))f(n|v)dn + \int_{n_3(v)}^{\overline{n}} T(z(n,v))f(n|v)dn \right\} f(v)dv\end{aligned} \tag{75}$$

We have the individual first order condition:

$$u_1(z - T(z) - \epsilon\tau(z), z/n; v)\left(1 - T'(z) - \epsilon\tau'(z)\right) + \frac{1}{n}u_2(z - T(z) - \epsilon\tau(z), z/n; v) = 0$$

For all individuals with a unique optimum where the tax schedule is twice continuously differentiable the second order condition holds strictly (see Lemma 3 of Bergstrom and Dodds (2021)), hence we can apply the implicit function theorem to determine the impact of a tax perturbation (note that $T''(z) = 0$ everywhere that $T'(z)$ exists):

$$\begin{aligned}\frac{\partial z}{\partial \epsilon}(n,v) &= \frac{u_1\tau'(z) + \left[u_{11}(1 - T'(z)) + \frac{1}{n}u_{12}\right]\tau(z)}{u_{11}(1 - T'(z))^2 + \frac{2(1-T'(z))}{n}u_{12} + \frac{1}{n^2}u_{22}} \\ &\equiv \xi(n,v)\tau'(z(n,v)) + \eta(n,v)\tau(z(n,v))\end{aligned} \tag{76}$$

where $\xi(n,v) \equiv \frac{u_1}{u_{11}(1-T'(z))^2 + \frac{2(1-T'(z))}{n}u_{12} + \frac{1}{n^2}u_{22}}$ and $\eta(n,v) \equiv \frac{\left[u_{11}(1-T'(z)) + \frac{1}{n}u_{12}\right]}{u_{11}(1-T'(z))^2 + \frac{2(1-T'(z))}{n}u_{12} + \frac{1}{n^2}u_{22}}$.

For each v, almost all individuals $(n_1(v), n_2(v))$ that bunch at the kink point K_2 do not change their income in response to small tax perturbations because they are at a corner solution to begin with so that they strictly prefer this income level to all others; hence, $\frac{\partial T(z(n,v))}{\partial \epsilon} = 0$ for these individuals.[43] Next, we consider the behavioral responses of the types with multiple optima who are indifferent between locating in the second and third tax brackets. Let us denote $z^-(v)$ and $z^+(v)$ the upper and lower optimal incomes for type $n_3(v)$. Dropping the v argument, $z^-(v)$ and $z^+(v)$ satisfy the following indifference condition:

$$u(z^+ - T(z^+) - \epsilon\tau(z^+), z^+/n_3; v) = u(z^- - T(z^-) - \epsilon\tau(z^-), z^-/n_3; v) \tag{77}$$

We can also calculate how the indifferent individual (for each v) changes with the tax schedule by applying the implicit function theorem to Equation 77:[44]

$$\frac{\partial n_3}{\partial \epsilon} = \frac{u_1(z^+ - T(z^+), z^+/n_3; v)\tau(z^+) - u_1(z^- - T(z^-), z^-/n_3; v)\tau(z^-)}{u_2(z^- - T(z^-), z^-/n_3; v)z^-/(n_3)^2 - u_2(z^+ - T(z^+), z^+/n_3; v)z^+/(n_3)^2} \equiv \frac{u_1^+\tau(z^+) - u_1^-\tau(z^-)}{u_2^- z^+/(n_3)^2 - u_2^+ z^+/(n_3)^2} \tag{78}$$

[43] Footnote 21 of Bergstrom and Dodds (2021) discusses this point in more detail. Note, we have assumed that for each v, $n_2(v) < n_3(v)$ so that all bunching individuals have a unique optima.

[44] Note, we have implicitly assumed that the individual first order condition holds for $n_3(v)$ at both $z^+(v)$ and $z^-(v)$ in deriving Equation 78. However, this assumption can be dropped without changing Equation 78; see Appendix A.6 of Bergstrom and Dodds (2021).

Let us then calculate the Gateaux derivative of R in the direction of $\tau(z)$:

$$
\begin{aligned}
&\lim_{\epsilon\to 0}\frac{R(T+\epsilon\tau)-R(T)}{\epsilon}\\
&=\int_V\left\{\int_{\underline{n}}^{n_1(v)}\left[\frac{T(z(n,v))}{\partial\epsilon}+\tau(z(n,v))\right]f(n|v)dn+\int_{n_1(v)}^{n_2(v)}\left[\frac{T(z(n,v))}{\partial\epsilon}+\tau(z(n,v))\right]f(n|v)dn\right.\\
&+\int_{n_2(v)}^{n_3(v)}\left[\frac{T(z(n,v))}{\partial\epsilon}+\tau(z(n,v))\right]f(n|v)dn+\int_{n_3(v)}^{\overline{n}}\left[\frac{T(z(n,v))}{\partial\epsilon}+\tau(z(n,v))\right]f(n|v)dn\\
&\left.+\left(T(z^-(v))-T(z^+(v))\right)f(n_3(v)|v)\frac{\partial n_3}{\partial\epsilon}(v)\right\}f(v)dv
\end{aligned}
\tag{79}
$$

Note, the last term of Equation 79 results from applying the Leibniz integral rule (recognizing that this is the only such term arising from differentiating the limits of integration via the Leibniz integral rule because $T(z(n,v))$ is continuous as a function of n at all n other than $n_3(v)$). As argued previously, $\int_{n_1(v)}^{n_2(v)}\frac{T(z(n,v))}{\partial\epsilon}f(n|v)dn=0$. Plugging in the value of $\frac{\partial z(n,v)}{\partial\epsilon}$ from the implicit function theorem (Equation 76) and changing the variable of integration from n to z we find that:[45]

$$
\begin{aligned}
&\lim_{\epsilon\to 0}\frac{R(T+\epsilon\tau)-R(T)}{\epsilon}\\
&=\int\limits_V\left\{\int\limits_{z(\underline{n};v)}^{z(n_1(v);v)}\left(T'(z)\xi(z,v)\tau'(z)+\left[1+T'(z)\eta(z,v)\right]\tau(z)\right)h(z|v)dz+\int\limits_{n_1}^{n_2}\tau(K_1)f(n|v)dn\right.\\
&+\int\limits_{z(n_2(v);v)}^{z^-(n_3(v);v)}\left(T'(z)\xi(z,v)\tau'(z)+\left[1+T'(z)\eta(z,v)\right]\tau(z)\right)h(z|v)dz\\
&+\int\limits_{z^+(n_3(v);v)}^{z(\overline{n};v)}\left(T'(z)\xi(z,v)\tau'(z)+\left[1+T'(z)\eta(z,v)\right]\tau(z)\right)h(z|v)dz\\
&\left.+\left(T(z^-(v))-T(z^+(v))\right)f(n_3(v)|v)\frac{u_1^+\tau(z^+(v))-u_1^-\tau(z^-(v))}{u_2^-z^-(v)/(n_3(v))^2-u_2^+z^+(v)/(n_3(v))^2}\right\}f(v)dv
\end{aligned}
\tag{80}
$$

Next, let us switch the order of integration again and average out the various behavioral effects over v for each z. Let us denote $\underline{z}$ as the lowest z chosen by any type, $\overline{z}$ as the highest z chosen by any type, $\overline{z^-}$ as the highest $z^-(v)$ for any v, and $\underline{z^+}$ as the lowest $z^+(v)$ for any v. Furthermore, let us use $\overline{\xi}(z)$ to denote average $\xi(z,v)$ at a given z and define $\overline{\eta}(z)$ to denote average $\eta(z,v)$ at a given z. Let $M(K_1)=\int_{n_1}^{n_2}f(n|v)dn$ denote the mass of types bunching at

[45] Note by monotonicity that $H(z(n;v)|v)=F(n|v)$ so that $h(z(n;v)|v)=f(n|v)\left(\frac{\partial z(n;v)}{\partial n}\right)^{-1}$ so that $h(z|v)$ accounts for the Jacobian of the change of variables.

K_1. Finally, note $f(n_3(v)|v)f(v) = f(n_3(v), v)$.

$$\begin{aligned}
&\lim_{\epsilon\to 0}\frac{R(T+\epsilon\tau)-R(T)}{\epsilon}\\
&\int_{\underline{z}}^{K_1}\left(T'(z)\overline{\xi}(z)\tau'(z)+\left[1+T'(z)\overline{\eta}(z)\right]\tau(z)\right)h(z)dz + M(K_1)\tau(K_1)\\
&+\int_{K_1}^{\overline{z^-}}\left(T'(z)\overline{\xi}(z)\tau'(z)+\left[1+T'(z)\overline{\eta}(z)\right]\tau(z)\right)h(z)dz\\
&+\int_{\underline{z^+}}^{\overline{z}}\left(T'(z)\overline{\xi}(z)\tau'(z)+\left[1+T'(z)\overline{\eta}(z)\right]\tau(z)\right)h(z)dz\\
&+\int_V\left(T(z^-(v))-T(z^+(v))\right)\frac{u_1^+\tau(z^+(v))-u_1^-\tau(z^-(v))}{u_2^- z^-(v)/(n_3(v))^2-u_2^+ z^+(v)/(n_3(v))^2}f(n_3(v),v)dv
\end{aligned}\tag{81}$$

Next, let us apply integration by parts to get rid of the $\tau'(z)$ terms in Equation 81, supposing that $z(\underline{n}, v)$, $z^-(v)$, $z^+(v)$, and $z(\overline{n}, v)$ are all strictly monotonic in v. This ensures that $h(\underline{z}) = h(\overline{z^-}) = h(\underline{z^+}) = h(\overline{z}) = 0$ as long as $\left(\frac{\partial z(n;v)}{\partial n}\right)^{-1}$ is bounded away from infinity.[46] Denoting $T_1\overline{\xi}(K_1^-)h(K_1^-)$ as $\lim_{z\to K_1^-} T'(z)\overline{\xi}(z)h(z)$ and $T_2\overline{\xi}(K_1^+)h(K_1^+)$ as $\lim_{z\to K_1^+} T'(z)\overline{\xi}(z)h(z)$ (recall T_1 denotes the marginal tax rate in the first tax bracket and T_2 denotes the marginal tax rate in the second tax bracket):

$$\begin{aligned}
&\lim_{\epsilon\to 0}\frac{R(T+\epsilon\tau)-R(T)}{\epsilon}\\
&\int_{\underline{z}}^{K_1}\left(-\frac{\partial}{\partial z}\left[T'(z)\overline{\xi}(z)h(z)\right]+\left[1+T'(z)\overline{\eta}(z)\right]h(z)\right)\tau(z)dz + T_1\overline{\xi}(K_1^-)h(K_1^-)\tau(K_1) + M(K_1)\tau(K_1)\\
&-T_2\overline{\xi}(K_1^+)h(K_1^+)\tau(K_1)+\int_{K_1}^{\overline{z^-}}\left(-\frac{\partial}{\partial z}\left[T'(z)\overline{\xi}(z)h(z)\right]+\left[1+T'(z)\overline{\eta}(z)\right]h(z)\right)\tau(z)dz\\
&+\int_{\underline{z^+}}^{\overline{z}}\left(-\frac{\partial}{\partial z}\left[T'(z)\overline{\xi}(z)h(z)\right]+\left[1+T'(z)\overline{\eta}(z)\right]h(z)\right)\tau(z)dz\\
&+\int_V\left(T(z^-(v))-T(z^+(v))\right)\frac{u_1^+\tau(z^+(v))-u_1^-\tau(z^-(v))}{u_2^- z^-(v)/(n_3(v))^2-u_2^+ z^+(v)/(n_3(v))^2}f(n_3(v),v)dv
\end{aligned}\tag{82}$$

Finally, note that all $\tau(z)$ terms enter Equation 82 linearly so that Equation 82 is a linear functional of $\tau(z)$ which means that $R(T)$ is Gateaux differentiable (assuming that all terms in Equation 82 are bounded so that Equation 82 is a linear bounded - hence continuous - functional of $\tau(z)$). To recover the inverse welfare functional from Equation 14, we simply collect all of the terms in Equation 14 that involve a $\tau(z)$ at each income level z. Assuming that $z^-(v)$ and $z^+(v)$ are monotonic in v so that we can change the variable of integration in the final term of Equation 82 where Z^- is the set of all $z^-(v)$, Z^+ is the set of all $z^+(v)$, $\hat{h}^-(n_3(z^-), z^-) = f(n_3(v), v)\left(\frac{\partial z^-}{\partial v}\right)^{-1}$ (i.e., this new density just incorporates the Jacobian of

[46] If $z(\underline{n}, v)$ is strictly monotonic in v then $h(\underline{z}) = \int_V h(\underline{z}|v)dv = \int_V f(n(\underline{z}, v)|v)\left(\frac{\partial z(n;v)}{\partial n}\right)^{-1} = 0$ because $f(n(\underline{z}, v)|v) \neq 0$ only for a single type v. Similary, if the lower multiple optima income $z^-(v)$ is strictly monotonic in v then $h(\overline{z^-}) = 0$; identical logic holds for $h(\underline{z^+})$ and $h(\overline{z})$.

the transformation), and $\hat{h}^+(n_3(z^+), z^+) = f(n_3(v), v)\left(\frac{\partial z^-}{\partial v}\right)^{-1}$:[47]

$$\begin{aligned}
&\lim_{\epsilon\to 0}\frac{R(T+\epsilon\tau)-R(T)}{\epsilon}\\
&\int_{\underline{z}}^{K_1}\left(-\frac{\partial}{\partial z}\left[T'(z)\overline{\xi}(z)h(z)\right]+\left[1+T'(z)\overline{\eta}(z)\right]h(z)\right)\tau(z)dz+T_1\overline{\xi}(K_1^-)h(K_1^-)\tau(K_1)+M(K_1)\tau(K_1)\\
&-T_2\overline{\xi}(K_1^+)h(K_1^+)\tau(K_1)+\int_{K_1}^{\overline{z^-}}\left(-\frac{\partial}{\partial z}\left[T'(z)\overline{\xi}(z)h(z)\right]+\left[1+T'(z)\overline{\eta}(z)\right]h(z)\right)\tau(z)dz\\
&+\int_{\underline{z^+}}^{\overline{z}}\left(-\frac{\partial}{\partial z}\left[T'(z)\overline{\xi}(z)h(z)\right]+\left[1+T'(z)\overline{\eta}(z)\right]h(z)\right)\tau(z)dz\\
&+\int_{Z^-}\frac{-\left[T(z^-)-T(z^+)\right]u_1^-\tau(z^-)}{u_2^-z^-/(n_3)^2-u_2^+z^+/(n_3)^2}\hat{h}^-(n_3(z^-),z^-)dz^-+\int_{Z^+}\frac{\left[T(z^-)-T(z^+)\right]u_1^+\tau(z^+)}{u_2^-z^-/(n_3)^2-u_2^+z^+/(n_3)^2}\hat{h}^+(n_3(z^+),z^+)dz^+
\end{aligned}\tag{83}$$

From here we can just collect terms to incorporate the terms of the last two integrals in Equation 83 into the other integrals noting that $\hat{h}^+(n_3(z), z) \neq 0 \iff \mathbb{1}(z \in Z^+)$ and $\hat{h}^-(n_3(z), z) \neq 0 \iff \mathbb{1}(z \in Z^-)$:

$$\begin{aligned}
&\lim_{\epsilon\to 0}\frac{R(T+\epsilon\tau)-R(T)}{\epsilon}=\\
&\underbrace{\int_{\underline{z}}^{K_1}\left(-\frac{\partial}{\partial z}\left[T'(z)\overline{\xi}(z)h(z)\right]+\left[1+T'(z)\overline{\eta}(z)\right]h(z)\right)\tau(z)dz}_{\text{Perturbations in First Bracket}}\\
&+\underbrace{T_1\overline{\xi}(K_1^-)h(K_1^-)\tau(K_1)+M(K_1)\tau(K_1)-T_2\overline{\xi}(K_1^+)h(K_1^+)\tau(K_1)}_{\text{Perturbation at Kink}}\\
&+\underbrace{\int_{K_1}^{\overline{z^-}}\left(-\frac{\partial}{\partial z}\left[T'(z)\overline{\xi}(z)h(z)\right]+\left[1+T'(z)\overline{\eta}(z)\right]h(z)-\frac{\left(T(z)-T(z^+(z))\right)u_1^-(z)}{u_2^-(z)\frac{z}{(n_3(z))^2}-u_2^+(z)\frac{z}{(n_3(z))^2}}\hat{h}^-(n_3(z),z)\right)\tau(z)dz}_{\text{Perturbations in Second Bracket}}\\
&+\underbrace{\int_{\underline{z^+}}^{\overline{z}}\left(-\frac{\partial}{\partial z}\left[T'(z)\overline{\xi}(z)h(z)\right]+\left[1+T'(z)\overline{\eta}(z)\right]h(z)+\frac{\left(T(z^-(z))-T(z)\right)u_1^+(z)}{u_2^-(z)\frac{z}{(n_3(z))^2}-u_2^+(z)\frac{z}{(n_3(z))^2}}\hat{h}^+(n_3(z),z)\right)\tau(z)dz}_{\text{Perturbations in Third Bracket}}
\end{aligned}\tag{84}$$

From here, we recover Equation 20 by: (1) recognizing that $T'(z)$ equals T_1 in the first bracket, T_2 in the second bracket, and T_3 in the third bracket (2) defining $Z_1 = [\underline{z}, K_1]$, $Z_2 = [K_1, \overline{z^-}]$, and $Z_3 = [\underline{z^+}, \overline{z}]$ and (3) defining:

$$\begin{aligned}
J_2(z) &\equiv \frac{\left(T(z)-T(z^+(z))\right)u_1^-(z)}{u_2^-(z)\frac{z}{(n_3(z))^2}-u_2^+(z)\frac{z}{(n_3(z))^2}}\hat{h}^-(n_3(z),z)\\
J_3(z) &\equiv \frac{\left(T(z^-(z))-T(z)\right)u_1^+(z)}{u_2^-(z)\frac{z}{(n_3(z))^2}-u_2^+(z)\frac{z}{(n_3(z))^2}}\hat{h}^+(n_3(z),z)
\end{aligned}$$

□

[47] Note that in the first integral in the last line of Equation 83, everything (e.g., z^+, n_3, u_1^-, u_2^-) is a function of z^-; similarly, everything in the second integral in the last line is a function of z^+.

B.5 Proof of Proposition 1

Proof. Recall that tax revenue is given by:

$$R(T) = \int_{\mathbf{N}} T(\mathbf{z}(\mathbf{n}))dF(\mathbf{n})$$

Our goal is to show that we can find a continuous linear functional that represents:

$$\lim_{\epsilon \to 0} \frac{R(T + \epsilon\tau) - R(T)}{\epsilon}$$

We organize the proof up by first discussing the impact of a tax perturbation on individuals with a single optimum where the tax schedule is smooth, next discussing the impact of a tax perturbation on individuals with multiple optima, and finally discussing the impact of a tax perturbation on individuals for whom the tax schedule is not differentiable at their chosen $\mathbf{z}$. As mentioned, there are five additional regularity conditions that we assume will hold throughout:

1. The tax schedule everywhere is semi-differentiable in all directions (i.e., one way directional derivatives exist everywhere).
2. The set of individuals locating along the surfaces where the tax schedule is not differentiable and whose first order conditions are satisfied in some direction is measure zero.
3. The income distribution admits a density $h(\mathbf{z})$ at all $\mathbf{z}$ where $T(\mathbf{z})$ is differentiable. On non-differentiable hypersurfaces $\hat{\mathbf{Z}}$ of dimension ≥ 1, the income distribution also admits a "density" $\hat{h}(\mathbf{z})$ so that the mass of people locating on any $E \subset \hat{\mathbf{Z}}$ equals $\int_E \hat{h}(\mathbf{z})dS$, where dS is the hypersurface element.
4. The set of individuals with more than two optima is measure zero restricted to the set of surfaces of those who have multiple optima (i.e., almost all individuals with multiple optima just have two optima).[48]
5. Average behavioral effects of taxation are sufficiently smooth so as to apply integration by parts (which follows assuming the behavioral responses are in an appropriate Sobolev space defined within the proof)

B.5.1 Single Optimum Individuals and a Smooth Tax Schedule

First, let us consider the set of individuals who have a single optimum $\mathbf{z}$ and at which $T(\mathbf{z})$ is twice differentiable. These individuals satisfy first order conditions given by System 85 (which is just System 25 reproduced for clarity):

$$\begin{aligned} u_c(y(\mathbf{z}) - T(\mathbf{z}) - \epsilon\tau(\mathbf{z}), \mathbf{z}; \mathbf{n}) \left(y_{z_1}(\mathbf{z}) - T_{z_1}(\mathbf{z}) - \epsilon\tau_{z_1}(\mathbf{z})\right) + u_{z_1}(y(\mathbf{z}) - T(\mathbf{z}) - \epsilon\tau(\mathbf{z}), \mathbf{z}; \mathbf{n}) &= 0 \\ &\vdots \\ u_c(y(\mathbf{z}) - T(\mathbf{z}) - \epsilon\tau(\mathbf{z}), \mathbf{z}; \mathbf{n}) \left(y_{z_J}(\mathbf{z}) - T_{z_J}(\mathbf{z}) - \epsilon\tau_{z_J}(\mathbf{z})\right) + u_{z_J}(y(\mathbf{z}) - T(\mathbf{z}) - \epsilon\tau(\mathbf{z}), \mathbf{z}; \mathbf{n}) &= 0 \end{aligned} \tag{85}$$

For any such agent $\mathbf{n}$ with a unique optimal income $\mathbf{z}(\mathbf{n})$, compactness arguments imply that

[48] We require that almost all individuals have only two optima because if they had three or more optimal choices $\mathbf{z}$, then their decision over which choice to jump to depends in a non-linear way on the tax perturbation.

$\exists \upsilon$ such that for any δ:

$$u(c(\mathbf{z}(\mathbf{n})), \mathbf{z}(\mathbf{n}); \mathbf{n}) > u(c(\mathbf{z}), \mathbf{z}; \mathbf{n}) + \upsilon \ \forall \mathbf{z} \notin B_\delta(\mathbf{z}(\mathbf{n}))$$

Thus, for sufficiently small ϵ, all such individuals prefer some $\mathbf{z} \in B_\delta(\mathbf{z}(\mathbf{n}))$ to all $\mathbf{z} \notin B_\delta(\mathbf{z}(\mathbf{n}))$. Hence, these individuals must move continuously in response to sufficiently small tax perturbations.

By assumption, for all but some measure zero set of these individuals, the second order condition holds strictly so that the Hessian matrix of second derivatives $\mathbf{H}(\mathbf{n})$ is negative definite (and therefore invertible) so that we can apply the implicit function theorem to derive Equation 86 (which is just System 26 reproduced for clarity):

$$\begin{aligned} \frac{\partial \mathbf{z}(\mathbf{n})}{\partial \epsilon} &= \mathbf{H}^{-1}(\mathbf{n}) FOC(\mathbf{n})_\epsilon|_{\epsilon=0} = \mathbf{H}^{-1}(\mathbf{n})[\mathbf{a}(\mathbf{n})\tau(\mathbf{z}) + \mathbf{B}(\mathbf{n}) \cdot \nabla_{\mathbf{z}}\tau(\mathbf{z})] \\ &\equiv \vec{\eta}(\mathbf{n})\tau(\mathbf{z}) + \mathbf{X}(\mathbf{n}) \cdot \nabla_{\mathbf{z}}\tau(\mathbf{z}) \end{aligned} \tag{86}$$

where $FOC_\epsilon|_{\epsilon=0}$ is the vector of derivatives of the first order conditions 85 with respect to ϵ. The second equality in Equation 86 follows for some vector $\mathbf{a}$ and a matrix $\mathbf{B}$ (which depend on $\mathbf{n}$) given that the derivative of each first order condition with respect to ϵ (evaluated at $\epsilon = 0$) is linear in τ and each component of $\nabla_{\mathbf{z}}\tau(\mathbf{z}) = (\tau_{\mathbf{z}_1}, \tau_{\mathbf{z}_2}, ..., \tau_{\mathbf{z}_J})$. The third equality in Equation 86 simply follows by defining $\vec{\eta}(\mathbf{n}) \equiv \mathbf{H}^{-1}(\mathbf{n})\mathbf{a}(\mathbf{n})$ and $\mathbf{X}(\mathbf{n}) \equiv \mathbf{H}^{-1}(\mathbf{n})\mathbf{B}(\mathbf{n})$. $\vec{\eta}(\mathbf{n})$ represents the vector of income effects (how each component of $\mathbf{z}$ changes with the tax level, τ) and $\mathbf{X}(\mathbf{n})$ represents the matrix of substitution effects (how each component of $\mathbf{z}$ changes with each marginal tax rate).

Thus, for the set of individuals who have a unique optimum and the tax schedule is twice continuously differentiable, we know that for all but some measure zero set of agents:

$$\frac{\partial}{\partial \epsilon}\left[T(\mathbf{z}(\mathbf{n})) + \epsilon\tau(\mathbf{z}(\mathbf{n}))\right]|_{\epsilon=0} = \tau(\mathbf{z}(\mathbf{n})) + \nabla_{\mathbf{z}}T(\mathbf{z})\tau(\mathbf{z}(\mathbf{n}))\vec{\eta}(\mathbf{n}) + \nabla_{\mathbf{z}}T(\mathbf{z})\mathbf{X}(\mathbf{n}) \cdot \nabla_{\mathbf{z}}\tau(\mathbf{z}) \tag{87}$$

Note, the measure zero set of individuals for whom the second order conditions hold only weakly move in a continuous way (because they have a unique optimum to begin with); hence, they have a negligible impact on the Gateaux derivative of $R(T)$.

B.5.2 Individuals with Multiple Optima

Next, let us move on to the set of individuals who have multiple optima. For this set of agents, we assume everyone has two optima (other than potentially some measure zero set), which we will denote $\mathbf{z}_1(\mathbf{n})$ and $\mathbf{z}_2(\mathbf{n})$. For a given tax perturbation from $T(\mathbf{z})$ to $T(\mathbf{z}) + \epsilon\tau(\mathbf{z})$, the set of agents who initially had two optima will, in general, now strictly prefer one of their two optima, leading them to "jump" from one optimum to another. Moreover, some other agents who were close to indifferent will also jump to a point close to the initially indifferent agent's new optima. So the question becomes, what can we say about how the set of individuals with multiple optima changes as a result of the tax perturbation? Towards this purpose, let us note that for each $\tilde{\mathbf{n}}$ with two optima:

$$\max_{\mathbf{z} \in \mathbf{Z}_1} u(c(\mathbf{z}), \mathbf{z}; \tilde{\mathbf{n}}) = \max_{\mathbf{z} \in \mathbf{Z}_2} u(c(\mathbf{z}), \mathbf{z}; \tilde{\mathbf{n}}) \tag{88}$$

where $\mathbf{Z}_1, \mathbf{Z}_2$ are two disjoint compact sets which contain $\mathbf{z}_1(\tilde{\mathbf{n}})$ and $\mathbf{z}_2(\tilde{\mathbf{n}})$ on the interior, respectively. Now, because type $\tilde{\mathbf{n}}$ has a unique optimum on both $\mathbf{Z}_1$ and $\mathbf{Z}_2$ (and the utility function is smooth), we can apply the envelope theorem separately to restricted choice sets $\mathbf{Z}_1$ and $\mathbf{Z}_2$ for type $\tilde{\mathbf{n}}$ (Corollary 4 of Milgrom and Segal (2002)) to infer that:

$$\left(\frac{\partial u(c(\mathbf{z}_1), \mathbf{z}_1; \tilde{\mathbf{n}})}{\partial \epsilon} - \frac{\partial u(c(\mathbf{z}_2), \mathbf{z}_2; \tilde{\mathbf{n}})}{\partial \epsilon}\right) = (\nabla_{\mathbf{n}} u(c(\mathbf{z}_2), \mathbf{z}_2; \tilde{\mathbf{n}}) - \nabla_{\mathbf{n}} u(c(\mathbf{z}_1), \mathbf{z}_1; \tilde{\mathbf{n}})) \cdot \nabla_\epsilon \tilde{\mathbf{n}} \tag{89}$$

Given that ϵ only has a direct impact on consumption, we can rewrite Equation 89 as:

$$u_c(c(\mathbf{z}_2), \mathbf{z}_2; \tilde{\mathbf{n}})\tau(\mathbf{z}_2) - u_c(c(\mathbf{z}_1), \mathbf{z}_1; \tilde{\mathbf{n}})\tau(\mathbf{z}_1) = (\nabla_{\mathbf{n}} u(c(\mathbf{z}_2), \mathbf{z}_2; \tilde{\mathbf{n}}) - \nabla_{\mathbf{n}} u(c(\mathbf{z}_1), \mathbf{z}_1; \tilde{\mathbf{n}})) \cdot \nabla_\epsilon \tilde{\mathbf{n}} \tag{90}$$

Equation 90 tells us how the surface of indifferent types changes with ϵ: $\nabla_\epsilon \tilde{\mathbf{n}}$. By our assumption, there exists (at most) some finite set of surfaces across which individuals have multiple optima, allowing us to partition the space of $\mathbf{N}$ so that agents on the interior of each partition have a unique optimum and agents on the boundary surfaces have multiple optima. For simplicity, let us suppose that there is just one such surface - the argument is easy to adapt if there are a a finite set of surfaces. In this case, suppose that we have $\mathbf{N} = \mathbf{N}_1 \cup \mathbf{N}_2$ and all individuals on the interior of $\mathbf{N}_1$ and $\mathbf{N}_2$ have a single optimum whereas individuals on the (shared) boundary of these two regions have multiple optima. We have:

$$R(T + \epsilon\tau) = \int_{\mathbf{N}_1} [T(\mathbf{z}(\mathbf{n})) + \epsilon\tau(\mathbf{z}(\mathbf{n}))] dF(\mathbf{n}) + \int_{\mathbf{N}_2} [T(\mathbf{z}(\mathbf{n})) + \epsilon\tau(\mathbf{z}(\mathbf{n}))] dF(\mathbf{n})$$

Taking the Gateaux variation of $R(T + \epsilon\tau)$, appealing to the Reynold's Transport Theorem, we get:[49]

$$\int_{\mathbf{N}} \frac{\partial}{\partial \epsilon} [T(\mathbf{z}(\mathbf{n})) + \epsilon\tau(\mathbf{z}(\mathbf{n}))] dF(\mathbf{n}) + \int_{\partial \mathbf{N}_1} T(\mathbf{z}(\tilde{\mathbf{n}})) \nabla_\epsilon \tilde{\mathbf{n}} \cdot \rho_1 f(\tilde{\mathbf{n}}) dS + \int_{\partial \mathbf{N}_2} T(\mathbf{z}(\tilde{\mathbf{n}})) \nabla_\epsilon \tilde{\mathbf{n}} \cdot \rho_2 f(\tilde{\mathbf{n}}) dS$$

where ρ_i is the outward pointing unit normal to the boundary $\partial \mathbf{N}_i$ of the given region $\mathbf{N}_i$, $\nabla_\epsilon \tilde{\mathbf{n}}$ describes the "velocity" that the boundary is changing as we change ϵ, and dS is the hypersurface element. Next, note that $\partial \mathbf{N}_1 = \partial \mathbf{N}_2$ and that the outward pointing normals satisfy $\rho_1 = -\rho_2$. Hence, we simplify the Gateaux variation of $R(T + \epsilon\tau)$ to:

$$\int_{\mathbf{N}} \frac{\partial}{\partial \epsilon} [T(\mathbf{z}(\mathbf{n})) + \epsilon\tau(\mathbf{z}(\mathbf{n}))] dF(\mathbf{n}) + \int_{\partial \mathbf{N}_1} [T(\mathbf{z}_1(\tilde{\mathbf{n}})) - T(\mathbf{z}_2(\tilde{\mathbf{n}}))] \nabla_\epsilon \tilde{\mathbf{n}} \cdot \rho_1 f(\tilde{\mathbf{n}}) dS \tag{91}$$

Economically, the second term captures the total "jumping effects" of an infinitesimal set of individuals changing their choices from $\mathbf{z}_2$ to $\mathbf{z}_1$. This changes tax revenue by $[T(\mathbf{z}_1(\tilde{\mathbf{n}})) - T(\mathbf{z}_2(\tilde{\mathbf{n}}))]$ for each jumping individual multiplied by the rate of change of the boundary, $\nabla_\epsilon \tilde{\mathbf{n}} \cdot \rho_1$, integrated along the surface $\partial \mathbf{N}_1$. The key remaining question is: how do we determine the rate of change of the boundary $[\nabla_\epsilon \tilde{\mathbf{n}}] \cdot \rho_1$? The idea is to recognize that the surface $\partial \mathbf{N}_1$ is the level set of $\mathbf{n}$ such that:

$$\max_{\mathbf{z} \in \mathbf{Z}_1} u(c(\mathbf{z}), \mathbf{z}; \mathbf{n}) - \max_{\mathbf{z} \in \mathbf{Z}_2} u(c(\mathbf{z}), \mathbf{z}; \mathbf{n}) = 0$$

Thus, the normal vector ρ_1 to this surface is just the gradient of the LHS of the above equation, which by the envelope theorem is just $(\nabla_{\mathbf{n}} u(c(\mathbf{z}_1), \mathbf{z}_1; \tilde{\mathbf{n}}) - \nabla_{\mathbf{n}} u(c(\mathbf{z}_2), \mathbf{z}_2; \tilde{\mathbf{n}}))$. Thus, by

[49] The Reynold's Transport Theorem is simply the Leibniz integral rule for multivariable functions.

Equation 90 we have:[50]

$$\nabla_\epsilon \tilde{\mathbf{n}} \cdot \rho_1 = \frac{\nabla_{\mathbf{n}} u(c(\mathbf{z}_1), \mathbf{z}_1; \tilde{\mathbf{n}}) - \nabla_{\mathbf{n}} u(c(\mathbf{z}_2), \mathbf{z}_2; \tilde{\mathbf{n}})}{||\nabla_{\mathbf{n}} u(c(\mathbf{z}_1), \mathbf{z}_1; \tilde{\mathbf{n}}) - \nabla_{\mathbf{n}} u(c(\mathbf{z}_2), \mathbf{z}_2; \tilde{\mathbf{n}})||} \cdot \nabla_\epsilon \tilde{\mathbf{n}} = \frac{u_c(c(\mathbf{z}_1), \mathbf{z}_1; \tilde{\mathbf{n}})\tau(\mathbf{z}_1) - u_c(c(\mathbf{z}_2), \mathbf{z}_2; \tilde{\mathbf{n}})\tau(\mathbf{z}_2)}{||\nabla_{\mathbf{n}} u(c(\mathbf{z}_1), \mathbf{z}_1; \tilde{\mathbf{n}}) - \nabla_{\mathbf{n}} u(c(\mathbf{z}_2), \mathbf{z}_2; \tilde{\mathbf{n}})||}$$

Hence, we get that:

$$\begin{aligned} &\int_{\partial \mathbf{N}_1} [T(\mathbf{z}_1(\tilde{\mathbf{n}})) - T(\mathbf{z}_2(\tilde{\mathbf{n}}))] \nabla_\epsilon \tilde{\mathbf{n}} \cdot \rho_1 f(\tilde{\mathbf{n}}) dS \\ &= \int_{\partial \mathbf{N}_1} [T(\mathbf{z}_1(\tilde{\mathbf{n}})) - T(\mathbf{z}_2(\tilde{\mathbf{n}}))] \frac{u_c(c(\mathbf{z}_1), \mathbf{z}_1; \tilde{\mathbf{n}})\tau(\mathbf{z}_1) - u_c(c(\mathbf{z}_2), \mathbf{z}_2; \tilde{\mathbf{n}})\tau(\mathbf{z}_2)}{||\nabla_{\mathbf{n}} u(c(\mathbf{z}_1), \mathbf{z}_1; \tilde{\mathbf{n}}) - \nabla_{\mathbf{n}} u(c(\mathbf{z}_2), \mathbf{z}_2; \tilde{\mathbf{n}})||} f(\tilde{\mathbf{n}}) dS \end{aligned} \tag{92}$$

Importantly, note that Equation 92 is *linear* in the tax perturbation $\tau(\mathbf{z})$; this is the key property we require in order to ensure that $R(T)$ is Gateaux differentiable.

Note that if there is some measure zero set of individuals along the surface $\partial \mathbf{N}_1$ with more than two optima, then those individuals may not move from $\mathbf{z}_1$ to $\mathbf{z}_2$ (or from $\mathbf{z}_2$ to $\mathbf{z}_1$) according to Equation 90; however, by assumption there is only a measure zero set of these individuals when the domain is restricted to $\partial \mathbf{N}_1$ so that the presence of such individuals does not impact Equation 92.[51]

B.5.3 Individuals who Choose $\mathbf{z}$ with Non-smooth $T(\mathbf{z})$

Finally, we discuss individuals with a unique optimum who choose $\mathbf{z}$ where $T(\mathbf{z})$ is not differentiable.[52] [53] Note that by the same arguments as for individuals with a single optimum locating at $\mathbf{z}$ where $T(\mathbf{z})$ is differentiable, individuals with a single optimum locating at $\mathbf{z}$ where $T(\mathbf{z})$ is not differentiable must move locally in response to sufficiently small tax perturbations. Next, it is useful to point out that if $\mathbf{z}$ is unidimensional and the single crossing property holds, then it is obvious that the derivative of revenue for bunching individuals is linear in τ for such individuals because (1) bunching can only occur when the tax schedule is non-differentiable and (2) almost all individuals who locate at kinks in the tax schedule strictly prefer the kink point to all other possible income choices. Hence, there are (essentially) no behavioral responses for individuals locating at $\mathbf{z}$ with non-differentiable $T(\mathbf{z})$ so that the derivative of revenue at these income levels is just the mechanical effect.

However, in the multidimensional case, behavioral responses of individuals locating where $T(\mathbf{z})$ is non-differentiable are more complex because the tax schedule can be non-differentiable in some directions but differentiable in others (e.g., a three dimensional ridge). In particular, let us suppose that there is a single differentiable surface $\hat{\mathbf{Z}}$ such that $T(\cdot)$ is not differentiable

[50] We divide by the norm to transform $(\nabla_{\mathbf{n}} u(c(\mathbf{z}_1), \mathbf{z}_1; \tilde{\mathbf{n}}) - \nabla_{\mathbf{n}} u(c(\mathbf{z}_2), \mathbf{z}_2; \tilde{\mathbf{n}}))$ into a *unit* normal vector.

[51] More specifically, if we denote $E \subset \partial \mathbf{N}_1$ as the set of individuals along $\partial \mathbf{N}_1$ with more than two optima, then $\int_E f(\mathbf{n}) dS = 0$ where S is the surface element of $\partial \mathbf{N}_1$.

[52] Note, we already showed that we can express the behavioral effects of individuals with multiple optima as a linear functional of the tax schedule; this includes individuals who choose $\mathbf{z}$ where $T(\mathbf{z})$ is not differentiable. Hence, we can restrict attention to individuals with a unique optimum who choose $\mathbf{z}$ where $T(\mathbf{z})$ is not differentiable.

[53] We also could have surfaces where $T(\mathbf{z})$ is differentiable but not twice differentiable so that we cannot apply the implicit function theorem. We assume that the set of individuals locating on such surfaces is measure zero (these individuals do not have "strict second order conditions"). If these individuals have multiple optima, then their behavioral responses are covered by Section B.5.2; if these individuals have a unique optimum then they must move smoothly in response to the tax perturbation, at which point the total impact of of such individuals on the derivative of $R(T)$ is negligible.

across this surface (the argument is easily adapted when there are more such non-differentiable surfaces). We assume that $T(\cdot)$ is semi-differentiable all directions (i.e., that one-way directional derivatives exist everywhere) but that in directions ρ normal to the surface $\hat{\mathbf{Z}}$, $T(\cdot)$ is not directionally differentiable:

$$\lim_{h\to 0^+}\frac{T(\mathbf{z}+h\rho)-T(\mathbf{z})}{h}\neq\lim_{h\to 0^-}\frac{T(\mathbf{z}+h\rho)-T(\mathbf{z})}{h}$$

Along the surface $\hat{\mathbf{Z}}$, $T(\cdot)$ is assumed twice directionally differentiable. Let us denote a maximal linearly independent set of normal vectors to the given surface as $\vec{\rho}$ and a maximal linearly independent set of tangent vectors to the given surface as $\vec{\nu}$. Hence, we have the following set of first order conditions for individuals choosing incomes along $\hat{\mathbf{Z}}$:

$$u_c(y(\mathbf{z})-T(\mathbf{z})-\epsilon\tau(\mathbf{z}),\mathbf{z};\mathbf{n})\,(y_\nu(\mathbf{z})-T_\nu(\mathbf{z})-\epsilon\tau_\nu(\mathbf{z}))+u_\nu(y(\mathbf{z})-T(\mathbf{z})-\epsilon\tau(\mathbf{z}),\mathbf{z};\mathbf{n})=0\ \forall\nu\in\vec{\nu} \quad (93)$$

$$u_c(y(\mathbf{z})-T(\mathbf{z})-\epsilon\tau(\mathbf{z}),\mathbf{z};\mathbf{n})\,\big(y_{\rho^+}(\mathbf{z})-T_{\rho^+}(\mathbf{z})-\epsilon\tau_{\rho^+}(\mathbf{z})\big)+u_{\rho^+}(y(\mathbf{z})-T(\mathbf{z})-\epsilon\tau(\mathbf{z}),\mathbf{z};\mathbf{n})\le 0\ \forall\rho\in\vec{\rho} \quad (94)$$

$$u_c(y(\mathbf{z})-T(\mathbf{z})-\epsilon\tau(\mathbf{z}),\mathbf{z};\mathbf{n})\,\big(y_{\rho^-}(\mathbf{z})-T_{\rho^-}(\mathbf{z})-\epsilon\tau_{\rho^-}(\mathbf{z})\big)+u_{\rho^-}(y(\mathbf{z})-T(\mathbf{z})-\epsilon\tau(\mathbf{z}),\mathbf{z};\mathbf{n})\ge 0\ \forall\rho\in\vec{\rho} \quad (95)$$

Equations 93, 94, and 95 simply say that first order conditions are satisfied in the directions of differentiability, ν, and are negative in the "positive" direction ρ^+ and positive in the "negative" direction ρ^- along the directions of non-differentiability. By assumption, there are only a measure zero set of individuals for whom either Equations 93 are satisfied and either 94 or Equation 95 are satisfied with equality. Because these individuals move continuously in response to tax perturbations, we can ignore them when computing the impact on $R(T)$. Moreover, we note that for all individuals locating at a $\mathbf{z}$ where $T(\mathbf{z})$ is non-differentiable and Equations 94 and 95 hold only weakly, Equations 94 and 95 still hold weakly for a sufficiently small perturbation ϵ. In other words, almost all individuals do not move in the directions $\rho\in\vec{\rho}$ normal to the surface of non-differentiability in response to small tax perturbations. Thus, we only need to determine how these individuals move in the directions tangent to the surface of non-differentiability.

Let us parametrize the surface $\hat{\mathbf{Z}}$ with a set of curvilinear coordinates (as is done when taking a line integral in $\mathbb{R}^2$ or a surface integral in $\mathbb{R}^3$). Hence, let us consider $\hat{\mathbf{z}}(\mathbf{t})$ for some vector of coordinates $\mathbf{t}$ contained in some region of $\mathbb{R}^m$. Under such a parametrization, we can consider the following set of first order conditions written in vector form:

$$\nabla_{\mathbf{t}}u(y(\mathbf{t})-T(\mathbf{t})-\epsilon\tau(\mathbf{t}),\mathbf{t};\mathbf{n})=0 \quad (96)$$

We assume that for all but a measure zero set of individuals locating at $\mathbf{z}$ where $T(\mathbf{z})$ is not differentiable, the second order conditions hold strictly along the surface of non-differentiability so that the Hessian matrix $\mathbf{H_t}(\mathbf{n})$ of second derivatives with respect to $\mathbf{t}$ is negative definite so that we can apply the implicit function theorem to Equation 96 to derive:

$$\begin{aligned}\frac{\partial\mathbf{t}(\mathbf{n})}{\partial\epsilon}&=\mathbf{H}_{\mathbf{t}}^{-1}(\mathbf{n})FOC(\mathbf{n})_\epsilon|_{\epsilon=0}=\mathbf{H}_{\mathbf{t}}^{-1}(\mathbf{n})[\mathbf{a_t}(\mathbf{n})\tau(\mathbf{t})+\mathbf{B_t}(\mathbf{n})\cdot\nabla_{\mathbf{t}}\tau(\mathbf{t})]\\&=\vec{\eta}_{\mathbf{t}}(\mathbf{n})\tau(\mathbf{t})+\mathbf{X_t}(\mathbf{n})\cdot\nabla_{\mathbf{t}}\tau(\mathbf{t})\end{aligned} \quad (97)$$

where $FOC(\mathbf{n})_\epsilon|_{\epsilon=0}$ is the vector of derivatives of the first order conditions 96 with respect to ϵ.

$\nabla_{\mathbf{t}}\tau(\mathbf{t})$ denotes the gradient of τ with respect to $\mathbf{t}$ and the first equality in Equation 97 follows for some vector $\mathbf{a_t}$ and a matrix $\mathbf{B_t}$ (which depend on $\mathbf{n}$) given that the derivative of each first order condition with respect to ϵ (evaluated at $\epsilon = 0$) is linear in $\tau(\mathbf{z})$ and $\nabla_{\mathbf{t}}\tau(\mathbf{z})$. The second equality in Equation 97 simply follows by defining $\vec{\eta}_{\mathbf{t}} \equiv \mathbf{H}_{\mathbf{t}}^{-1}(\mathbf{n})\mathbf{a_t}(\mathbf{n})$ and $\mathbf{X_t} \equiv \mathbf{H}_{\mathbf{t}}^{-1}(\mathbf{n})\mathbf{B_t}(\mathbf{n})$. Thus, for the set of individuals who choose a $\mathbf{t}$ where $T(\mathbf{t})$ is not differentiable, we know that for all but some measure zero set of agents:

$$\frac{\partial}{\partial\epsilon}\left[T(\mathbf{t}(\mathbf{n})) + \epsilon\tau(\mathbf{t}(\mathbf{n}))\right]|_{\epsilon=0} = \tau(\mathbf{t}(\mathbf{n})) + \nabla_{\mathbf{t}}T(\mathbf{t})\vec{\eta}_{\mathbf{t}}\tau(\mathbf{t}(\mathbf{n})) + \nabla_{\mathbf{t}}T(\mathbf{t})\mathbf{X_t}\cdot\nabla_{\mathbf{t}}\tau(\mathbf{t}(\mathbf{n})) \tag{98}$$

B.5.4 Gateaux Differentiability of $R(T)$

Putting all of this together, we need to plug the expressions from Equations 87, 92, and 98 into Equation 91. Then splitting up $\mathbf{N}$ into $\mathbf{N}\setminus\hat{\mathbf{N}}$ and $\hat{\mathbf{N}}$ (where $\hat{\mathbf{N}}$ denotes the set of individuals choosing to locate at the non-differentiable surface $\hat{\mathbf{Z}}$), we get that the Gateaux variation of $R(T)$ for a tax schedule with a non-differentiable surface $\hat{\mathbf{Z}}$ and a surface $\partial\mathbf{N}_1$ of individuals with multiple optima equals:

$$\begin{aligned}&\int_{\mathbf{N}\setminus\hat{\mathbf{N}}}\left(\tau(\mathbf{z}) + \nabla_{\mathbf{z}}T(\mathbf{z}(\mathbf{n}))\vec{\eta}(\mathbf{n})\tau(\mathbf{z}) + \nabla_{\mathbf{z}}T(\mathbf{z}(\mathbf{n}))\mathbf{X}(\mathbf{n})\cdot\nabla_{\mathbf{z}}\tau(\mathbf{z})\right)dF(\mathbf{n})\\ &+\int_{\partial\mathbf{N}_1}\left[T(\mathbf{z}_1(\tilde{\mathbf{n}})) - T(\mathbf{z}_2(\tilde{\mathbf{n}}))\right]\frac{u_c(c(\mathbf{z}_1),\mathbf{z}_1;\tilde{\mathbf{n}})\tau(\mathbf{z}_1) - u_c(c(\mathbf{z}_2),\mathbf{z}_2;\tilde{\mathbf{n}})\tau(\mathbf{z}_2)}{||\nabla_{\mathbf{n}}u(c(\mathbf{z}_1),\mathbf{z}_1;\tilde{\mathbf{n}}) - \nabla_{\mathbf{n}}u(c(\mathbf{z}_2),\mathbf{z}_2;\tilde{\mathbf{n}})||}f(\tilde{\mathbf{n}})dS\\ &+\int_{\hat{\mathbf{N}}}\left(\tau(\mathbf{t}(\mathbf{n})) + \nabla_{\mathbf{t}}T(\mathbf{t})\vec{\eta}_{\mathbf{t}}\tau(\mathbf{t}(\mathbf{n})) + \nabla_{\mathbf{t}}T(\mathbf{t})\mathbf{X_t}\cdot\nabla_{\mathbf{t}}\tau(\mathbf{t}(\mathbf{n}))\right)dF(\mathbf{n})\end{aligned}$$

Integrating over $\mathbf{Z}$ we can write this as:[54]

$$\begin{aligned}&\int_{\mathbf{Z}\setminus\hat{\mathbf{Z}}}\int_{\mathbf{N}(\mathbf{z})}\left(\tau(\mathbf{z}) + \nabla_{\mathbf{z}}T(\mathbf{z})\vec{\eta}(\mathbf{n})\tau(\mathbf{z}) + \nabla_{\mathbf{z}}T(\mathbf{z})\mathbf{X}(\mathbf{n})\cdot\nabla_{\mathbf{z}}\tau(\mathbf{z})\right)f(\mathbf{n}|\mathbf{z})dSh(\mathbf{z})d\mathbf{z}\\ &+\int_{\partial\mathbf{N}_1}\left[T(\mathbf{z}_1(\tilde{\mathbf{n}})) - T(\mathbf{z}_2(\tilde{\mathbf{n}}))\right]\frac{u_c(c(\mathbf{z}_1),\mathbf{z}_1;\tilde{\mathbf{n}})\tau(\mathbf{z}_1) - u_c(c(\mathbf{z}_2),\mathbf{z}_2;\tilde{\mathbf{n}})\tau(\mathbf{z}_2)}{||\nabla_{\mathbf{n}}u(c(\mathbf{z}_1),\mathbf{z}_1;\tilde{\mathbf{n}}) - \nabla_{\mathbf{n}}u(c(\mathbf{z}_2),\mathbf{z}_2;\tilde{\mathbf{n}})||}f(\tilde{\mathbf{n}})dS\\ &+\int_{\hat{\mathbf{Z}}}\int_{\mathbf{N}(\mathbf{z})}\left(\tau(\mathbf{t}) + \nabla_{\mathbf{t}}T(\mathbf{t})\vec{\eta}_{\mathbf{t}}(\mathbf{n})\tau(\mathbf{t}) + \nabla_{\mathbf{t}}T(\mathbf{t})\mathbf{X_t}(\mathbf{n})\cdot\nabla_{\mathbf{t}}\tau(\mathbf{t})\right)f(\mathbf{n}|\mathbf{t})dS\hat{h}(\mathbf{t})d\mathbf{t}\end{aligned}$$

where $\mathbf{N}(\mathbf{z})$ denotes the set of $\mathbf{n}$ who choose a given $\mathbf{z}$, dS represents a hypersurface element, and $\hat{h}(\mathbf{t})$ is the density of households choosing to locate at coordinates $\mathbf{t}$ on $\hat{\mathbf{Z}}$.[55] Now, as long as $q(\mathbf{z}) \equiv \int_{\mathbf{N}(\mathbf{z})}\nabla_{\mathbf{z}}T(\mathbf{z}(\mathbf{n}))\mathbf{X}(\mathbf{n})f(\mathbf{n}|\mathbf{z})dSh(\mathbf{z})$ is sufficiently smooth (specifically, if each component of the vector valued function is in the Sobolev space $H^1(\mathbf{Z})$, see https://encyclopediaofmath.org/wiki/Integration_by_parts) then we can apply integration by parts:

$$\int_{\mathbf{Z}\setminus\hat{\mathbf{Z}}}\int_{\mathbf{N}(\mathbf{z})}\nabla_{\mathbf{z}}T(\mathbf{z}(\mathbf{n}))\mathbf{X}(\mathbf{n})\cdot\nabla_{\mathbf{z}}\tau(\mathbf{z})f(\mathbf{n}|\mathbf{z})dSh(\mathbf{z})d\mathbf{z} = -\int_{\mathbf{Z}\setminus\hat{\mathbf{Z}}}\mathrm{div}(q(\mathbf{z}))\tau(\mathbf{z})d\mathbf{z} + \int_{\partial(\mathbf{Z}\setminus\hat{\mathbf{Z}})}q(\mathbf{z})\tau(\mathbf{z})dS$$

Note, we have used the assumption that $\mathbf{Z}$ is the closure of an open set and that $\hat{\mathbf{Z}}$ is a closed set. Hence, $\mathbf{Z}\setminus\hat{\mathbf{Z}}\setminus\partial\mathbf{Z}$ is an open set in the ambient space, allowing us to perform integration by parts over the region $\mathbf{Z}\setminus\hat{\mathbf{Z}}\setminus\partial\mathbf{Z}$ or, equivalently (because inclusion of the boundary does not

[54] We have just integrated over $\mathbf{Z}$ first and then integrated these terms over the set of $\mathbf{n}$ who choose a given $\mathbf{z}$.

[55] Note, we assumed the existence of $h(\mathbf{z})$ on $\mathbf{Z}\setminus\hat{\mathbf{Z}}$ and $\hat{h}(\mathbf{z})$ along $\hat{\mathbf{Z}}$. Given the parametrization of $\hat{\mathbf{Z}}$ using some curvilinear coordinates in $\mathbf{t}$, $\hat{h}(\mathbf{t})\sqrt{g(\mathbf{t})} = \hat{h}(\mathbf{z})$ where $g(\mathbf{t})$ is the Riemannian metric of the hypersurface $\hat{\mathbf{Z}}$ (e.g., the line element for a curve in $\mathbb{R}^2$).

impact the integral) $\mathbf{Z} \setminus \hat{\mathbf{Z}}$. For example, if $\mathbf{z} = (z_1, z_2)$, we have assumed that $\mathbf{Z} \setminus \hat{\mathbf{Z}} \setminus \partial\mathbf{Z}$ has non-zero area in $\mathbb{R}^2$. If not (which will occur if the dimension of $\mathbf{N}$ is less than the dimension of $\mathbf{Z}$) then $R(T)$ will typically *not* be Gateaux differentiable as discussed in Section 7.

Finally, suppose that on $\hat{\mathbf{Z}}$, $\hat{q}(\mathbf{t}) \equiv \int_{\mathbf{N}(\mathbf{z})} \nabla_{\mathbf{t}} T(\mathbf{t}) \mathbf{X}_{\mathbf{t}}(\mathbf{n}) f(\mathbf{n}|\mathbf{t}) dS \hat{h}(\mathbf{t})$ is sufficiently smooth (specifically, if each component of the vector valued function is in the Sobolev space $H^1(\hat{\mathbf{Z}})$).[56] Then again we have that:

$$\int_{\hat{\mathbf{Z}}} \int_{\mathbf{N}(\mathbf{z})} \nabla_{\mathbf{t}} T(\mathbf{t}) \mathbf{X}_{\mathbf{t}}(\mathbf{n}) \cdot \nabla_{\mathbf{t}} \tau(\mathbf{t}) f(\mathbf{n}|\mathbf{t}) dS \hat{h}(\mathbf{t}) d\mathbf{t} = - \int_{\hat{\mathbf{Z}}} \operatorname{div}(\hat{q}(\mathbf{t})) \tau(\mathbf{t}) d\mathbf{t} + \int_{\partial\hat{\mathbf{Z}}} \hat{q}(\mathbf{t}) \tau(\mathbf{t}) dS$$

Note, we have to split the $\mathbf{Z} \setminus \hat{\mathbf{Z}}$ and $\hat{\mathbf{Z}}$ domains to perform integration by parts because $\hat{\mathbf{Z}}$ is measure zero and hence not open in the ambient space. Thus, we have to treat $\hat{\mathbf{Z}}$ (after a suitable parametrization) as the closure of an open subset of $\mathbb{R}^m$ for $m < \dim(\mathbf{Z})$.

Thus, we can write the Gateaux derivative of $R(T)$ as:

$$\int_{\mathbf{Z}\setminus\hat{\mathbf{Z}}} \int_{\mathbf{N}(\mathbf{z})} [\tau(\mathbf{z}) + \nabla_{\mathbf{z}} T(\mathbf{z}(\mathbf{n})) \vec{\eta}(\mathbf{n}) \tau(\mathbf{z})] f(\mathbf{n}|\mathbf{z}) dS d\mathbf{z} - \int_{\mathbf{Z}\setminus\hat{\mathbf{Z}}} \operatorname{div}(q(\mathbf{z})) \tau(\mathbf{z}) d\mathbf{z} + \int_{\partial(\mathbf{Z}\setminus\hat{\mathbf{Z}})} q(\mathbf{z}) \tau(\mathbf{z}) dS$$
$$+ \int_{\hat{\mathbf{Z}}} \int_{\mathbf{N}(\mathbf{z})} [\tau(\mathbf{t}) + \nabla_{\mathbf{t}} T(\mathbf{t}) \vec{\eta}_{\mathbf{t}}(\mathbf{n}) \tau(\mathbf{t})] f(\mathbf{n}|\mathbf{t}) dS \hat{h}(\mathbf{t}) d\mathbf{t} - \int_{\hat{\mathbf{Z}}} \operatorname{div}(\hat{q}(\mathbf{t})) \tau(\mathbf{t}) d\mathbf{t} + \int_{\partial\hat{\mathbf{Z}}} \hat{q}(\mathbf{t}) \tau(\mathbf{t}) dS$$
$$+ \int_{\partial\mathbf{N}_1} [T(\mathbf{z}_1(\tilde{\mathbf{n}})) - T(\mathbf{z}_2(\tilde{\mathbf{n}}))] \frac{u_c(c(\mathbf{z}_1), \mathbf{z}_1; \tilde{\mathbf{n}}) \tau(\mathbf{z}_1) - u_c(c(\mathbf{z}_2), \mathbf{z}_2; \tilde{\mathbf{n}}) \tau(\mathbf{z}_2)}{||\nabla_{\mathbf{n}} u(c(\mathbf{z}_1), \mathbf{z}_1; \tilde{\mathbf{n}}) - \nabla_{\mathbf{n}} u(c(\mathbf{z}_2), \mathbf{z}_2; \tilde{\mathbf{n}})||} f(\tilde{\mathbf{n}}) dS$$

If the behavioral effects of taxation are sufficiently smooth (i.e., all terms in the above expression are bounded), then the above expression is a bounded linear functional (and therefore a continuous linear functional): hence, $R(T)$ is Gateaux differentiable. □

B.6 Extensive Margin Responses

Let us consider another example with a smooth unidimensional tax schedule $T(z)$ but with two dimensions of heterogeneity $(n, v) \in [\underline{n}, \overline{n}] \times [\underline{v}, \overline{v}]$. As before n denotes productivity. v now denotes a fixed cost of working so that utility is given by:

$$u(c, z/n) - v\mathbb{1}[z > 0]$$

with $c = z - T(z)$ and some smooth $u(c, z/n)$ satisfying the Mirrlees (1971) single crossing property which ensures that $z(n)$ is monotonic in n $\forall v$. Let us calculate the Gateaux derivative of $R(T)$. First, note that by monotonicity of $z(n)$ in n $\forall v$, for every v $\exists \hat{n}(v) \in [\underline{n}, \overline{n}]$ such that $n > \hat{n}(v)$ choose $z > 0$ and $n \leq \hat{n}(v)$ choose $z = 0$ (suppose for simplicity that $\hat{n}(v) \in (\underline{n}, \overline{n})$ $\forall v$). $\hat{n}(v)$ satisfies the following indifference condition where $z(n, v)$ denotes the optimal income conditional on working for type (n, v):

$$u(z(\hat{n}(v)) - T(z(\hat{n}(v))) - \epsilon\tau(z(\hat{n}(v))), z(\hat{n}(v))/\hat{n}(v)) - v = u(-T(0) - \epsilon\tau(0), 0) \qquad (99)$$

We have that (note we have dropped the v argument from $z(n, v)$ for those who choose to work

[56] Note that this automatically holds in dimension 1 because $\mathbf{X}_{\mathbf{t}}(\mathbf{n}) = 0$ as there are no substitution effects for individuals locating at $\mathbf{z}$ with $T(\mathbf{z})$ non-differentiable in this case, see Bergstrom and Dodds (2021).

because their choice of z is not dependent on v conditional on working a positive amount):

$$R(T)=\int_V\int_N T(z(n,v))f(n,v)dndv=\int_V\int_{\underline{n}}^{\hat{n}(v)} T(0)f(n,v)dndv+\int_V\int_{\hat{n}(v)}^{\overline{n}} T(z(n))f(n,v)dndv$$

Let us consider the impacts of a tax perturbation from $T(z)$ to $T(z)+\epsilon\tau(z)$. We have the individual first order condition, which holds with $\epsilon=0$ for all types that choose to work:

$$\left(1-T'(z)-\epsilon\tau'(z)\right)u_1(z-T(z)-\epsilon\tau(z),z/n)+\frac{1}{n}u_2(z-T(z)-\epsilon\tau(z),z/n)=0 \tag{100}$$

For all individuals with a unique optimum where the tax schedule is twice continuously differentiable the second order condition holds strictly (see Lemma 3 of Bergstrom and Dodds (2021)), hence we can apply the implicit function theorem to determine the impact of a tax perturbation:

$$\frac{\partial z}{\partial\epsilon}(n,v)=\frac{u_1\tau'(z)+\left[u_{11}(1-T'(z))+\frac{1}{n}u_{12}\right]\tau(z)}{u_{11}(1-T'(z))^2+\frac{2}{n}u_{12}(1-T'(z))+\frac{1}{n^2}u_{22}-T''(z)u_1}\equiv\xi(n,v)\tau'(z(n,v))+\eta(n,v)\tau(z(n,v)) \tag{101}$$

Taking the derivative of $R(T)$ via the Leibniz integral rule recognizing that almost all individuals who choose not to work are at a corner solution and hence do not change incomes in response to small tax perturbations we have:

$$\begin{aligned}&\lim_{\epsilon\to 0}\frac{R(T+\epsilon\tau)-R(T)}{\epsilon}\\&=\int_V\int_{\hat{n}(v)}^{\overline{n}}\left[\frac{T(z(n))}{\partial\epsilon}+\tau(z(n,v))\right]f(n,v)dndv+\int_V\int_{\underline{n}}^{\hat{n}(v)}\tau(0)f(n,v)dndv\\&+\int_V\left[T(0)-T(z(\hat{n}(v)))\right]f(\hat{n}(v)|v)\frac{\partial\hat{n}(v)}{\partial\epsilon}f(v)dv\end{aligned} \tag{102}$$

We can also calculate how the indifferent individual changes with the tax schedule by applying the implicit function theorem to Equation 99 (and evaluating at $\epsilon=0$):

$$\frac{\partial\hat{n}(v)}{\partial\epsilon}=\frac{u_1(-T(0),0)\tau(0)-u_1(z(\hat{n}(v))-T(z(\hat{n}(v))),z(\hat{n}(v))/\hat{n}(v))\tau(z(\hat{n}(v)))}{u_2(z(\hat{n}(v))-T(z(\hat{n}(v))),z(\hat{n}(v))/\hat{n}(v))\frac{z(\hat{n}(v))}{\hat{n}(v)^2}} \tag{103}$$

Plugging in Equations 101 and 103 into Equation 102 and denoting $M(0)\equiv\int_V\int_{\underline{n}}^{\hat{n}(v)}f(n,v)dndv$, we have (note we have dropped some of the arguments from the derivatives of utility functions in Equation 103 for readability):

$$\begin{aligned}&\lim_{\epsilon\to 0}\frac{R(T+\epsilon\tau)-R(T)}{\epsilon}\\&=\int_V\int_{\hat{n}(v)}^{\overline{n}}\left[T'(z(n,v))\xi(n,v)\tau'(z(n,v))+\left(1+T'(z(n,v))\eta(n,v)\right)\tau(z(n,v))\right]f(n,v)dndv\\&+M(0)\tau(0)+\int_V\left[T(0)-T(z(\hat{n}(v),v))\right]f(\hat{n}(v)|v)\frac{u_1(0)\tau(0)-u_1(z(\hat{n}(v)))\tau(z(\hat{n}(v)))}{u_2(z(\hat{n}(v)))\frac{z(\hat{n}(v))}{\hat{n}(v)^2}}f(v)dv\end{aligned} \tag{104}$$

Changing the variable of integration for the first integral on the RHS of Equation 104 from n to z (and defining $h(z,v)=f(n,v)\left(\frac{\partial z}{\partial n}\right)^{-1}$ to take into account the Jacobian of the transformation), swapping the order of integration and taking averages over the V dimension as in Section 3.2

we have (where $\underline{z}$ and $\overline{z}$ are the lowest and highest incomes chosen by any type choosing $z > 0$), and then applying integration by parts to get rid of the $\tau'(z)$ term:

$$
\begin{aligned}
&\int_V \int_{\hat{n}(v)}^{\overline{n}} \left[T'(z(n,v))\xi(n,v)\tau'(z(n,v)) + \left(1 + T'(z(n,v))\eta(n,v)\right)\tau(z(n,v))\right] f(n,v)dndv \\
&= \int_V \int_{z(\hat{n}(v))}^{z(\overline{n})} \left[T'(z)\xi(z,v)\tau'(z) + \left(1 + T'(z)\eta(z,v)\right)\tau(z)\right] h(z,v)dzdv \\
&= \int_{\underline{z}}^{\overline{z}} \left[T'(z)\overline{\xi}(z)\tau'(z) + \left(1 + T'(z)\overline{\eta}(z)\right)\tau(z)\right] h(z)dz \\
&= \int_{\underline{z}}^{\overline{z}} \left(-\frac{\partial}{\partial z}\left[T'(z)\overline{\xi}(z)h(z)\right] + \left[1 + T'(z)\overline{\eta}(z)\right] h(z)\right)\tau(z)dz + T'(z)\overline{\xi}(z)h(z)\tau(z)\Big|_{\underline{z}}^{\overline{z}}
\end{aligned}
\tag{105}
$$

For simplicity, suppose that there is a monotonic relationship $v \to z(\hat{n}(v))$. Changing the variable of integration from v to z in the second integral on the RHS of Equation 104 and denoting $h(\hat{n}(z), z) \equiv f(\hat{n}(v)|v)f(v)\left(\frac{\partial z}{\partial v}\right)^{-1}$ to incorporate the Jacobian of the transformation:[57]

$$
\begin{aligned}
&\int_V \left[T(0) - T(z(\hat{n}(v), v))\right] f(\hat{n}(v)|v)\frac{u_1(0)\tau(0) - u_1(z(\hat{n}(v)))\tau(z(\hat{n}(v)))}{u_2(z(\hat{n}(v)))\frac{z(\hat{n}(v))}{\hat{n}(v)^2}} f(v)dv \\
&= \int_Z \left[T(0) - T(z)\right] \frac{u_1(0)\tau(0) - u_1(z)\tau(z)}{u_2(z)\frac{z}{\hat{n}(z)^2}} h(\hat{n}(z), z)dz
\end{aligned}
\tag{106}
$$

If $v \to z(\hat{n}(v))$ is monotonic then $h(z) \to 0$ as $z \to \underline{z}$ because $h(z|v) \to 0$ as $z \to \underline{z}$ for all $v > \underline{v}$. Thus $T'(\underline{z})\overline{\xi}(\underline{z})h(\underline{z}) = 0$, yielding:

$$
\begin{aligned}
&\lim_{\epsilon \to 0} \frac{R(T + \epsilon\tau) - R(T)}{\epsilon} \\
&= \int_{\underline{z}}^{\overline{z}} \left(-\frac{\partial}{\partial z}\left[T'(z)\overline{\xi}(z)h(z)\right] + \left[1 + T'(z)\overline{\eta}(z)\right] h(z)\right)\tau(z)dz + T'(\overline{z})\overline{\xi}(\overline{z})h(\overline{z})\tau(\overline{z}) \\
&+ M(0)\tau(0) + \int_Z \left[T(0) - T(z)\right] \frac{u_1(0)\tau(0) - u_1(z)\tau(z)}{u_2(z)\frac{z}{\hat{n}(z)^2}} h(\hat{n}(z), z)dz
\end{aligned}
\tag{107}
$$

From here, suppose welfare is given by:

$$
\begin{aligned}
&\iint_{N \times V} \phi(n,v)U(n,v;T)f(n,v)dndv + \int_V \overline{\phi}(v)U(n(\overline{z}),v;T)f(v|\overline{z})dv \\
&= \int_Z \iint_{N \times V} \phi(n,v)U(n,v;T)f(n,v|z)dndvdH(z) + \int_V \overline{\phi}(v)U(n(\overline{z}),v;T)f(v|\overline{z})dv
\end{aligned}
\tag{108}
$$

where $n(\overline{z})$ is the type n that chooses $\overline{z}$ given tax schedule $T(z)$ and $f(v|\overline{z})$ is the conditional density of type v at $\overline{z}$ under $T(z)$. By the envelope theorem, the Gateaux derivative of Equation 108 equals:

$$
-\int_Z \iint_{N \times V} \phi(n,v)u_c(n,v;T)\tau(z)f(n,v|z)dndvdH(z) - \int_V \overline{\phi}(v)\tau(\overline{z})u_c(n(\overline{z}),v;T)f(v|\overline{z})dv
\tag{109}
$$

We want to choose welfare weights such that for all $\tau(z)$, Equation 109 plus Equation 107 equals

[57] Note that we are slightly abusing notation here for brevity so that $u_1(z) = u_1(z - T(z), z/\hat{n}(z))$ and $u_2(z) = u_2(z - T(z), z/\hat{n}(z))$.

0. Hence, we pick $\phi(n,v)$ for all of those who do not work to satisfy:

$$\iint_{N\times V}\phi(n,v)u_c(n,v;T)\tau(z)f(n,v|0)dndvM(0)=M(0)\tau(0)+\int_Z[T(0)-T(z)]\frac{u_1(0)}{u_2(z)\frac{z}{\hat{n}(z)^2}}h(\hat{n}(z),z)dz \tag{110}$$

We pick $\phi(n,v)$ for those who earn less than the maximum income, $\overline{z}$, to satisfy at each z:

$$\begin{aligned}&\iint_{N\times V}\phi(n,v)u_c(n,v;T)\tau(z)f(n,v|z)dndvh(z)\\&=-\frac{\partial}{\partial z}\left[T'(z)\overline{\xi}(z)h(z)\right]+\left[1+T'(z)\overline{\eta}(z)\right]h(z)+[T(0)-T(z)]\frac{-u_1(z)}{u_2(z)\frac{z}{\hat{n}(z)^2}}h(\hat{n}(z),z)\end{aligned} \tag{111}$$

And we choose $\overline{\phi}(v)$ to satisfy:

$$\int_V\overline{\phi}(v)U(n(\overline{z}),v;T)f(v|\overline{z})dv=T'(\overline{z})\overline{\xi}(\overline{z})h(\overline{z}) \tag{112}$$

Choosing $\phi(n,v)$ and $\overline{\phi}(v)$ to satisfy the previous three equations ensures that any perturbation to the tax schedule leaves the government's Lagrangian unchanged; hence, we have shown how to construct a local inverse welfare functional in the presence of extensive margin effects.

B.7 Proof to Proposition 2

All individuals with a unique optima do not respond to infinitesimal tax perturbations so that for these individuals $\frac{\partial}{\partial\epsilon}[T(\mathbf{z}(\mathbf{n}))+\epsilon\tau(\mathbf{z}(\mathbf{n}))]=\tau(\mathbf{z}(\mathbf{n}))$. Thus, for these individuals the effect of a tax perturbation is a continuous linear functional given by $\int_{\mathbf{N}}\tau(\mathbf{z}(\mathbf{n}))f(\mathbf{n})d\mathbf{n}$.[58]

Individuals with multiple optima respond to tax perturbations by "jumping" between their optima. Under the two conditions stated in Proposition 2, we show in Appendix B.5.2 that these jumping effects can always be represented by a continuous, linear functional. Thus, revenue is Gateaux differentiable under the conditions stated in the proposition.

B.8 Sparsity-Based Frictions: Additional Discussion

First, we form the Lagrangian, recognizing that the Gateaux derivative of the Lagrangian is given by Equation 10. The Gateaux derivative of welfare, $\lim_{\epsilon\to 0}\frac{W(U(n;T+\epsilon\tau))-W(U(n;T))}{\epsilon}$ equals:

$$\begin{aligned}&-\int_A\left\{\int_{\underline{n}}^{n_1(a)}\phi(n,a)u_1(-T(0),0;n)\tau(0)f(n|a)dn+\int_{n_1(a)}^{n_2(a)}\phi(n,a)u_1(a/2-T(a/2),a/2;n)\tau(a/2)f(n|a)dn\right.\\&\left.+\int_{n_2(a)}^{\overline{n}}\phi(n,a)u_1(a-T(a),a;n)\tau(a)f(n|a)dn\right\}f(a)da\end{aligned} \tag{113}$$

We showed in Section 3.5 that $\lim_{\epsilon\to 0}\frac{R(T+\epsilon\tau)-R(T)}{\epsilon}$ is equal to Equation 33. First, plugging in the tax schedule $T(z)+\epsilon\tau(z)$ and differentiating Equations 31 and 32 w.r.t. ϵ yields the following expressions:

$$\frac{\partial n_1(a)}{\partial\epsilon}=\frac{u_1(-T(0),0;n_1(a))\tau(0)-u_1(a/2-T(a/2),a/2;n_1(a))\tau(a/2)}{u_n(-T(0),0;n_1(a))-u_n(a/2-T(a/2),a/2;n_1(a))} \tag{114}$$

[58] We can integrate over $\mathbf{N}$ ignoring those with multiple optima because they are measure zero within the set of $\mathbf{N}$ and therefore do not impact the integral.

$$\frac{\partial n_2(a)}{\partial \epsilon} = \frac{u_1(a/2 - T(a/2), a/2; n_2(a))\tau(a/2) - u_1(a - T(a), a; n_2(a))\tau(a)}{u_n(a/2 - T(a/2), a/2; n_2(a)) - u_n(a - T(a), a; n_2(a))} \tag{115}$$

Plugging these expressions into Equation 33, we see that the Gateaux variation is linear in $\tau(z)$ so that revenue is Gateaux differentiable as claimed. Next, let us collect terms that multiply a given $\tau(z)$ for each z in Equations 33 and 113. Beginning with Equation 33, for each value of z, $\tau(z)$ enters Equation 33 "twice": $\tau(z)$ impacts full-time individuals with choice set $\{0, z/2, z\}$ and also impacts part-time individuals with choice set $\{0, z, 2z\}$. First, plugging in the expressions for $\frac{\partial n_1(a)}{\partial \epsilon}$ and $\frac{\partial n_2(a)}{\partial \epsilon}$ let us collect all terms that multiply $\tau(a)$:

$$\begin{aligned}
&\int_A \left\{ \int_{n_2(a)}^{\overline{n}} f(n|a)dn + \frac{-u_1(a - T(a), a; n_2(a))[T(a/2) - T(a)]f(n_2(a)|a)}{u_n(a/2 - T(a/2), a/2; n_2(a)) - u_n(a - T(a), a; n_2(a))} \right\} \tau(a)f(a)da \\
&= \int_{\underline{a}}^{\overline{a}} \left\{ \int_{n_2(z)}^{\overline{n}} f(n|z)dn + \frac{-u_1(z - T(z), z; n_2(z))[T(z/2) - T(z)]f(n_2(z)|z)}{u_n(z/2 - T(z/2), z/2; n_2(z)) - u_n(z - T(z), z; n_2(z))} \right\} \tau(z)f(z)dz
\end{aligned} \tag{116}$$

where the equality simply changes the dummy variable of integration from a to z, noting that $A = [\underline{a}, \overline{a}]$. Next, let us collect terms that multiply $\tau(a/2)$:

$$\begin{aligned}
&\int_A \left\{ \int_{\underline{n}}^{n_1(a)} f(n|a)dn + \frac{-u_1(a/2 - T(a/2), a/2; n_1(a))[T(0) - T(a/2)]f(n_1(a)|a)}{u_n(-T(0), 0; n_1(a)) - u_n(a/2 - T(a/2), a/2; n_1(a))} \right. \\
&\left. + \frac{u_1(a/2 - T(a/2), a/2; n_2(a))[T(a/2) - T(a)]f(n_2(a)|a)}{u_n(a/2 - T(a/2), a/2; n_2(a)) - u_n(a - T(a), a; n_2(a))} \right\} \tau(a/2)f(a)da \\
&= \int_{\underline{a}/2}^{\overline{a}/2} \left\{ \int_{\underline{n}}^{n_1(2z)} f(n|2z)dn + \frac{-u_1(z - T(z), z; n_1(2z))[T(0) - T(z)]f(n_1(2z)|2z)}{u_n(-T(0), 0; n_1(2z)) - u_n(z - T(z), z; n_1(2z))} \right. \\
&\left. + \frac{u_1(z - T(z), z; n_2(2z))[T(z) - T(2z)]f(n_2(2z)|2z)}{u_n(z - T(z), z; n_2(2z)) - u_n(2z - T(2z), 2z; n_2(2z))} \right\} \tau(z)2f(2z)dz
\end{aligned} \tag{117}$$

where the equality does a change of variables from a to $2z$. There are also terms that multiply $\tau(0)$:

$$\begin{aligned}
&\int_A \int_{\underline{n}}^{n_1(a)} \tau(0)f(n|a)dnf(a)da \\
&+ \int_A \frac{u_1(-T(0), 0; n_1(a))[T(0) - T(a/2)]\tau(0)f(n_1(a)|a)}{u_n(-T(0), 0; n_1(a)) - u_n(a/2 - T(a/2), a/2; n_1(a))} f(a)da
\end{aligned} \tag{118}$$

Similar changes of variables show that the Gateaux derivative of government welfare, Equation 113, can be rewritten as:

$$\begin{aligned}
&- \int_A \int_{\underline{n}}^{n_1(a)} \phi(n, a)u_1(-T(0), 0; n)\tau(0)f(n|a)dnf(a)da \\
&- \int_{\underline{a}/2}^{\overline{a}/2} \int_{n_1(2z)}^{n_2(2z)} 2\phi(n, 2z)u_1(2z - T(2z), 2z; n)f(n|2z)f(2z)dn\tau(z)dz \\
&- \int_{\underline{a}}^{\overline{a}} \int_{n_2(z)}^{\overline{n}} \phi(n, z)u_1(z - T(z), z; n)f(n|z)f(z)dn\tau(z)dz
\end{aligned} \tag{119}$$

The Gateaux derivative of the government's Lagrangian equals Equation 119 plus the sum of Equations 116, 117, and 118 multiplied by the Lagrange multiplier λ. If $\lambda = 1$, which again just normalizes the inverse welfare functional, then the Gateaux derivative equals zero as long as $\forall z > 0$ (noting that $f(n|z) = 0$ when no types n choose a given value of z):

$$\begin{aligned}
&\int_{n_1(2z)}^{n_2(2z)} 2\phi(n,2z)u_1(2z-T(2z),2z;n)f(n|2z)f(2z)dn + \int_{n_2(z)}^{\overline{n}} \phi(n,z)u_1(z-T(z),z;n)f(n|z)f(z)dn \\
&= \int_{n_2(z)}^{\overline{n}} f(n|z)f(z)dn + \frac{-u_1(z-T(z),z;n_2(z))[T(z/2)-T(z)]f(n_2(z)|z)f(z)}{u_n(z/2-T(z/2),z/2;n_2(z))-u_n(z-T(z),z;n_2(z))} \\
&+ 2\Bigg\{\int_{\underline{n}}^{n_1(2z)} f(n|2z)dn + \frac{-u_1(z-T(z),z;n_1(2z))[T(0)-T(z)]f(n_1(2z)|2z)}{u_n(-T(0),0;n_1(2z))-u_n(z-T(z),z;n_1(2z))} \\
&+ \frac{u_1(z-T(z),z;n_2(2z))[T(z)-T(2z)]f(n_2(2z)|2z)}{u_n(z-T(z),z;n_2(2z))-u_n(2z-T(2z),2z;n_2(2z))}\Bigg\}f(2z)
\end{aligned} \tag{120}$$

And for $z = 0$:

$$\begin{aligned}
&\int_A\int_{\underline{n}}^{n_1(a)} \phi(n,a)u_1(-T(0),0;n)f(n|a)dn = \int_A\int_{\underline{n}}^{n_1(a)} f(n|a)dnf(a)da \\
&+ \int_A \frac{u_1(-T(0),0;n_1(a))[T(0)-T(a/2)]f(n_1(a)|a)}{u_n(-T(0),0;n_1(a))-u_n(a/2-T(a/2),a/2;n_1(a))}f(a)da
\end{aligned} \tag{121}$$

Equations 120 and 121 thus implicitly define an inverse welfare functional in the model of sparsity based frictions presented in Section 3.5.

B.9 Discussion of Inverse Weights Computed With Frictions

In this Appendix we illustrate heuristically how inverse weights at the top of the income distribution compare in the frictions model versus the baseline model. In the baseline model, inverse weights are given by:

$$\phi(z) = 1 - \frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right]/h(z) + \text{Extensive Margin Effects} \tag{122}$$

We calibrate the model so that the taxable income elasticity equals 0.3. In other words: $0.3 = \frac{\partial z}{\partial \epsilon}|_{\tau(z)=-z}\frac{1-T'}{z} = -\xi(z)\frac{1-T'}{z}$. For simplicity, let's assume that marginal tax rates are constant equal to T' and $T(0) = 0$:

$$\phi(z) = 1 + 0.3\frac{T'}{1-T'} \times \frac{\partial}{\partial z}\left[zh(z)\right]/h(z) + \text{Extensive Margin Effects} \tag{123}$$

Let us now compare this expression to the corresponding expression in the frictions model. We calibrate the "intensive margin elasticities" in the frictions model so that a tax perturbation that lowers the keep rate by 1% leads to an average income change of 0.3% for individuals of type a ($\overline{z}(a)$ represents average income for individuals of type a and $n_2(a)$ is the individual with type a indifferent between working part-time and full-time as in Equation 32):

$$\begin{aligned}
&-\frac{\partial n_2(a)}{\partial \epsilon}|_{\tau(z)=-z}f(n_2(a)|a)(a-a/2)\frac{1-T'}{\overline{z}(a)} \\
&= -\frac{(a-a/2)^2 f(n_2(a)|a)}{u_2(a/2-T(a/2),a/2;n_2(a))\frac{a/2}{n_2(a)^2} - u_2(a-T(a),a;n_2(a))\frac{a}{n_2(a)^2}}\frac{1-T'}{\overline{z}(a)} = 0.3
\end{aligned} \tag{124}$$

If we define $C(z) \equiv \int_{n_1(2z)}^{n_2(2z)} 2f(n|2z)f(2z)dn + \int_{n_2(z)}^{\overline{n}} f(n|z)f(z)dn$ (which is equal to the density of types choosing a given income z) and assume that all individuals chosing a given income are given the same welfare weight then this weight is given by (from Equation 120, using the fact that there are no income effects so that $u_1 = 1$):

$$\phi(z) = \frac{1}{C(z)}\left\{C(z) - \frac{[T(a/2) - T(a)]f(n_2(a)|a)f(a)}{u_n(a/2 - T(a/2), a/2; n_2(a)) - u_n(a - T(a), a; n_2(a))}\bigg|_{a=z} \right.$$
$$\left. + 2\frac{[T(a) - T(2a)]f(n_2(2a)|2a)f(2a)}{u_n(a - T(a), a; n_2(2a)) - u_n(2a - T(2a), 2a; n_2(2a))}\bigg|_{a=z}\right\} + \text{Extensive Margin Effects} \tag{125}$$

Using our calibration of the "intensive margin" effects in Equation 124, we can simplify this expression to read (also using the fact that marginal tax rates are assumed to be constant so that, for example, $T(z/2) - T(z) = T'[z/2 - z]$):

$$\phi(z) = \frac{1}{C(z)}\left\{C(z) - 0.3\frac{T'}{1-T'}\frac{\overline{z}(a)}{a - a/2}f(a)\bigg|_{a=z} + 2 \times 0.3\frac{T'}{1-T'}\frac{\overline{z}(2a)}{2a - a}f(2a)\bigg|_{a=z}\right\} + \text{Extensive Margin Effects} \tag{126}$$

Or equivalently:

$$\phi(z) = \frac{1}{C(z)}\left\{C(z) + 0.3\frac{T'}{1-T'}\frac{2\overline{z}(2a)f(2a) - 2\overline{z}(a)f(a)}{a}\bigg|_{a=z}\right\} + \text{Extensive Margin Effects} \tag{127}$$

Note that we calibrate the extensive margin effects between the two models to be essentially identical so these terms are equivalent between Equations 123 and 127. As in the baseline model, the substitition effects in Equation 127 depend on how fast income multiplied by a density is changing. The key difference is that the derivative in Equation 123 is replaced by a finite difference taken between z and $2z$: $\frac{2\overline{z}(2a)f(2a)-2\overline{z}(a)f(a)}{a}|_{a=z}$.[59] Hence, in regions like the top of the income distribution where the income density is decreasing and convex, the density is changing faster locally than it is over a large region so that $\frac{\partial}{\partial z}[zh(z)]$ will be more negative than $\frac{2\overline{z}(2a)f(2a)-2\overline{z}(a)f(a)}{a}|_{a=z}$ so that the behavioral effects of raising taxes are larger in the baseline model than the frictions model. This heuristic analysis suggests that, for instance, at the top of the income distribution where marginal tax rates are basically constant, that inverse welfare weights should be higher in the frictions model than in the baseline model.

B.10 Details on Derivations from Section 6.2

First, we need to determine how to express $\frac{\partial w}{\partial \epsilon}$ in terms of $\tau(z)$. Multiplying the market clearning condition, Equation 43, by w and implicitly differentiating with respect to ϵ (recognizing that labor demand does not react directly to a change in $\tau(z)$, only indirectly via the changing wage and that $wnl(n) = z(n)$):

$$\frac{\partial w}{\partial \epsilon}L + w\frac{\partial L}{\partial w}\frac{\partial w}{\partial \epsilon} - \int_{\mathbf{N}}\left(\frac{\partial z(n)}{\partial w}\bigg|_{\epsilon}\frac{\partial w}{\partial \epsilon} + \frac{\partial z(n)}{\partial \epsilon}\bigg|_{w}\right)dF(n) = 0 \tag{128}$$

We can recover $\frac{\partial z(n)}{\partial w}|_{\epsilon}$ by implicitly differentiating the individual first order condition with

[59] There is also an extra factor of 2, which results due to a change of variables. Note that if 1/3 of individuals with type a work full-time, 1/3 work part-time, and 1/3 do not work then $2\overline{z}(a) = a$.

respect to w:

$$\left.\frac{\partial z(n)}{\partial w}\right|_{\epsilon} = \frac{-(1+k)\left(\frac{z(n)}{nw}\right)^{k}\frac{1}{nw^2}}{-k\left(\frac{z(n)}{nw}\right)^{k-1}\frac{1}{n^2w^2} - T''(z(n))}$$

Similarly, by implicitly differentiating the individual first order condition with respect to ϵ, we find that $\frac{\partial z(n)}{\partial \epsilon}|_w = \tau'(z(n))\xi(n)$ for some function $\xi(n)$:

$$\left.\frac{\partial z(n)}{\partial \epsilon}\right|_{w} = \frac{\tau'(z(n))}{-k\left(\frac{z(n)}{nw}\right)^{k-1}\frac{1}{n^2w^2} - T''(z(n))} \equiv \tau'(z(n))\xi(n)$$

The firm's first order condition is that $Y'(L) - w = 0$. Thus, $\frac{\partial L}{\partial w} = \frac{\partial Y'^{-1}(w)}{\partial w}$. Plugging in $\frac{\partial z(n)}{\partial \epsilon}|_w \equiv \tau'(z(n))\xi(n)$ to Equaton 128 we have that:

$$\frac{\partial w}{\partial \epsilon}\left[L + w\frac{\partial Y'^{-1}(w)}{\partial w} - \int_N \left(\left.\frac{\partial z(n)}{\partial w}\right|_{\epsilon}\right) dF(n)\right] = \int_N \tau'(z(n))\xi(n)dF(n) \qquad (129)$$

Doing a change of variables from n to z (where $h(z)$ represents the density of z) and applying integration by parts, we find that (denoting $\overline{\frac{\partial z}{\partial w}}\big|_{\epsilon} \equiv \int_N \left(\frac{\partial z(n)}{\partial w}\big|_{\epsilon}\right) dF(n)$):

$$\frac{\partial w}{\partial \epsilon} = \frac{-\int_Z \frac{\partial[\xi(z)h(z)]}{\partial z}\tau(z)dz + \xi(z)h(z)\tau(z)\big|_{\underline{z}}^{\overline{z}}}{L + w\frac{\partial Y'^{-1}(w)}{\partial w} - \overline{\frac{\partial z}{\partial w}}\big|_{\epsilon}} \qquad (130)$$

Thus, $\frac{\partial w}{\partial \epsilon}$ exists and is a linear functional of $\tau(z)$; hence w is Gateaux differentiable in $T(z)$. If $h(z) = 0$ at the top and bottom of the distribution (this holds as long as $f(n) = 0$ at the top and bottom and $\frac{\partial z}{\partial n} \not\to 0$ as $n \to \underline{n}$ or $n \to \overline{n}$), then $\frac{\partial w}{\partial \epsilon} = \int_Z p(z)\tau(z)dz$ for $p(z) = \frac{-\frac{\partial[\xi(z)h(z)]}{\partial z}}{L + w\frac{\partial L}{\partial w} - \overline{\frac{\partial z}{\partial w}}\big|_{\epsilon}}$.

Next, let us consider the budgetary impact from Equation 45. Using a change of variables (recalling $n \mapsto z$ was assumed bijective and differentiable) and integration by parts we see that the government's budget is Gateaux differentiable in $T(z)$:

$$\begin{aligned}
&\int_N \left(\tau(z) + T'(z(n))\xi(n)\tau'(z(n)) + T'(z(n))\left.\frac{\partial z(n)}{\partial w}\right|_{\epsilon}\frac{\partial w}{\partial \epsilon}\right) dF(n) \\
&= \int_Z \left(h(z) - \frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right]\right)\tau(z)dz + \int_N \left(T'(z(n))\left.\frac{\partial z(n)}{\partial w}\right|_{\epsilon}\frac{\partial w}{\partial \epsilon}\right) dF(n) \\
&= \int_Z \left(h(z) - \frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right]\right)\tau(z)dz + \int_Z \int_N \left(T'(z(n))\left.\frac{\partial z(n)}{\partial w}\right|_{\epsilon}\right) dF(n)p(z)\tau(z)dz \\
&= \int_Z \left(h(z) - \frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right] + p(z)\int_N \left(T'(z(n))\left.\frac{\partial z(n)}{\partial w}\right|_{\epsilon}\right) dF(n)\right)\tau(z)dz
\end{aligned} \qquad (131)$$

Intuitively, Equation 131 captures two separate budgetary impacts: the direct budgetary impact of individuals responding to tax changes and the indirect budgetary impacts of individuals responding to wage changes that result from changes in labor supply as a result of tax changes. Turning to the welfare component of Equation 45, let us again do a change of variables and integration by parts:

$$
\begin{aligned}
&W\left(-\tau(z(n)) + \left(\frac{z(n)}{nw}\right)^{1+k}\frac{1}{w}\frac{\partial w}{\partial \epsilon} + s(n)\pi'(w)\frac{\partial w}{\partial \epsilon}\right)\\
&= -\int_Z \phi(n(z))\tau(z)h(z)dz + \int_N \phi(n)\left[\left(\frac{z(n)}{nw}\right)^{1+k}\frac{1}{w} + s(n)\pi'(w)\right]\frac{\partial w}{\partial \epsilon}f(n)dn\\
&= -\int_Z \phi(n(z))\tau(z)h(z)dz + \int_Z p(z)\tau(z)\left(\int_N \phi(n)\left[\left(\frac{z(n)}{nw}\right)^{1+k}\frac{1}{w} + s(n)\pi'(w)\right]f(n)dn\right)dz\\
&= -\int_Z \left[\phi(n(z))h(z) - p(z)\left(\int_Z \phi(n(\tilde{z}))\left[\left(\frac{\tilde{z}}{n(\tilde{z})w}\right)^{1+k}\frac{1}{w} + s(n(\tilde{z}))\pi'(w)\right]h(\tilde{z})d\tilde{z}\right)\right]\tau(z)dz
\end{aligned}
\tag{132}
$$

Combining Equations 131 and 132 yields Equation 46.

B.11 Proposition 3 with GE Effects

Proposition 3 GE. *Suppose that the conditions of Theorem 2 hold and suppose that the inverse welfare function can be written as:*[60]

$$\int_{\mathbf{Z}}\int_{\mathbf{N(z)}} \phi(\mathbf{n})U(\mathbf{n};T)dF(\mathbf{n}|\mathbf{z})h(\mathbf{z})d\mathbf{z}$$

Consider a tax perturbation $\tau(\mathbf{z}) = \tau_1(\mathbf{z}) + \tau_2(\mathbf{z})$ where $\tau_1(\mathbf{z})$ raises taxes on those with choices in $\mathbf{V}$ and $\tau_2(\mathbf{z}) = \tau_2$ is a lump sum transfer that balances the budget. Let $\frac{\partial \mathbf{w}}{\partial \epsilon}$ denote the vector of Gateaux variations of $\mathbf{w}$ in the direction of $\tau(\mathbf{z})$. Then if for some compact set $\mathbf{V} \in \mathbf{Z}$, we have that true societal welfare weights $\phi^T(\mathbf{n})$ satisfy:

$$\int_{\mathbf{N(v)}} \phi^T(\mathbf{n})u_c(\mathbf{n})dF(\mathbf{n}|\mathbf{v}) < \int_{\mathbf{N(v)}} \phi(\mathbf{n})u_c(\mathbf{n})dF(\mathbf{n}|\mathbf{v}) \;\forall \mathbf{v} \in \mathbf{V} \tag{133}$$

And:

$$\int_{\mathbf{Z}}\int_{\mathbf{N(z)}} \phi^T(\mathbf{n})u_{\mathbf{w}}(\mathbf{n})\frac{\partial \mathbf{w}}{\partial \epsilon}dF(\mathbf{n}|\mathbf{z})h(\mathbf{z})d\mathbf{z} \approx \int_{\mathbf{Z}}\int_{\mathbf{N(z)}} \phi(\mathbf{n})u_{\mathbf{w}}(\mathbf{n})\frac{\partial \mathbf{w}}{\partial \epsilon}dF(\mathbf{n}|\mathbf{z})h(\mathbf{z})d\mathbf{z} \tag{134}$$

and the true weights and inverse weights have the same normalization:

$$\int_{\mathbf{N}} \phi^T(\mathbf{n})u_c(\mathbf{n})dF(\mathbf{n}) = \int_{\mathbf{N}} \phi(\mathbf{n})u_c(\mathbf{n})dF(\mathbf{n}) \tag{135}$$

then $\tau(\mathbf{z})$ is a welfare improving tax reform direction. If the inequality in Equation 133 is reversed then $-\tau(\mathbf{z})$ is a welfare improving tax reform direction.

Proof. The total welfare impact of a small tax perturbation $\tau(\mathbf{z}) = \tau_1(\mathbf{z}) + \tau_2(\mathbf{z})$ where $\tau_1(\mathbf{z}) > 0$ raises taxes on those with choices in $\mathbf{V}$ and $\tau_2(\mathbf{z}) = \tau_2$ is a lump sum transfer that makes the Gateaux derivative of revenue equal to zero is:

$$
\begin{aligned}
&-\int_{\mathbf{V}} \tau_1(\mathbf{v})\int_{\mathbf{N(v)}} \phi(\mathbf{n})u_c(\mathbf{n})dF(\mathbf{n}|\mathbf{v})h(\mathbf{v})d\mathbf{v}\\
&+\int_{\mathbf{Z}}\int_{\mathbf{N(z)}} \phi(\mathbf{n})u_{\mathbf{w}}(\mathbf{n})\frac{\partial \mathbf{w}}{\partial \epsilon}dF(\mathbf{n}|\mathbf{z})h(\mathbf{z})d\mathbf{z} - \tau_2\int_{\mathbf{N}} \phi(\mathbf{n})u_c(\mathbf{n})dF(\mathbf{n}) = 0
\end{aligned}
\tag{136}
$$

[60] The proposition holds more generally. Under a continuous, linear welfare function, we can always write welfare as follows by the Disintegration Theorem: $\int_{\mathbf{Z}}\int_{\mathbf{N(z)}} U(\mathbf{n};T)d\Phi_1(\mathbf{n}|\mathbf{z})d\Phi_2(\mathbf{z})$. Then the theorem can be shown to hold if we replace Equation 133 with the following "statewise dominance" condition such that for all positive functions $\tau(\mathbf{v})$: $\int_{\mathbf{V}} \tau(\mathbf{v})\int_{\mathbf{N(v)}} u_c(\mathbf{n})d\Phi_1^T(\mathbf{n}|\mathbf{v})d\Phi_2^T(\mathbf{v}) < \int_{\mathbf{V}} \tau(\mathbf{v})\int_{\mathbf{N(v)}} u_c(\mathbf{n})d\Phi_1(\mathbf{n}|\mathbf{v})d\Phi_2(\mathbf{v})$

The conditions of the proposition imply that:

$$\begin{aligned}&-\int_{\mathbf{V}}\tau_1(\mathbf{v})\int_{\mathbf{N}(\mathbf{v})}\phi^T(\mathbf{n})u_c(\mathbf{n})dF(\mathbf{n}|\mathbf{v})h(\mathbf{v})d\mathbf{v}\\&+\int_{\mathbf{Z}}\int_{\mathbf{N}(\mathbf{z})}\phi^T(\mathbf{n})u_{\mathbf{w}}(\mathbf{n})\frac{\partial\mathbf{w}}{\partial\epsilon}dF(\mathbf{n}|\mathbf{z})h(\mathbf{z})d\mathbf{z}-\tau_2\int_{\mathbf{N}}\phi^T(\mathbf{n})u_c(\mathbf{n})dF(\mathbf{n})>0\end{aligned} \tag{137}$$

If Equation 133 is reversed, then the above inequality also holds replacing $\tau(\mathbf{z})$ with $-\tau(\mathbf{z})$ which completes the proof. □

B.12 Labor Demand with High and Low Skilled Labor

We again consider a government that chooses a tax schedule to maximize welfare for a given population of individuals indexed by a uni-dimensional type n. Individuals choose an income $z = w_i nl$ where l is labor supply and $w_i \in \{w_l, w_h\}$ is a wage paid on effective effort, nl, that varies with whether an individual is low-skilled or high-skilled (for simplicity, whether a worker is low-skilled or high-skilled is taken as exogenous). Furthemore, suppose for simplicity that all types with $n < \text{med}(n) \equiv \text{median}(n)$ are low-skilled and those with $n \geq \text{med}(n)$ are high-skilled; hence w is a function of n with $w(n)$ denoting the wage faced by a given individual with productivity n. Individuals choose z to maximize a quasi-linear iso-elastic utility function:

$$\begin{aligned}U(n;T,w_l,w_h) = \max_z\ c - \frac{[z/(nw(n))]^{1+k}}{1+k}\\ \text{s.t. } c = z - T(z) + s(n)\pi^*(w_l,w_h)\end{aligned} \tag{138}$$

where c is again numeraire consumption, $\pi^*(w_l, w_h)$ represents optimal firm profits given wages (w_l, w_h), and $s(n)$ represents the share of profits owned by a given type n with $\int_N s(n)f(n)dn = 1$. There is also a single firm that produces the consumption good c by hiring labor to maximize profits. Firm output depends on total hired effective effort of both low-skilled and high-skilled types, L_l and L_h. Firm profits are given by:

$$\pi = Y(L_l, L_h) - w_l L_l - w_h L_h$$

where $Y(L_l, L_h)$ is the firm's production function. Market clearing requires that:[61]

$$L_l = \int_{\underline{n}}^{\text{med}(n)} nl(n)dF(n) \tag{139}$$

$$L_h = \int_{\text{med}(n)}^{\overline{n}} nl(n)dF(n) \tag{140}$$

The firm first order conditions are given by:

$$Y_1(L_l, L_h) - w_l = 0 \tag{141}$$
$$Y_2(L_l, L_h) - w_h = 0 \tag{142}$$

Suppose that we are interested in calculating an inverse welfare functional in this setting for a smooth tax schedule under which all individuals have a unique optima. The government's

[61] Recognize that L_l, L_h, and $l(n)$ all depend on the wage, w_l and w_h. $l(n)$ also depends on the tax schedule, $T(\mathbf{z})$.

Lagrangian is given by:

$$W(U(n;T,w_l,w_h)) + \lambda\left[\int_N T(z(n))dF(n) - E\right] \tag{143}$$

Now, let us take the (Gateaux) derivative of Equation 143 in the direction of $\tau(z)$ (i.e., as we move from $T(z)$ to $T(z)+\epsilon\tau(z)$), assuming that $n \mapsto z$ is a smooth bijective function, individual second order conditions hold strictly, and that $\frac{\partial w_l}{\partial \epsilon}$, $\frac{\partial w_h}{\partial \epsilon}$ exist:

$$\begin{aligned} &W\left(-\tau(z(n)) + \left(\frac{z(n)}{nw(n)}\right)^{1+k}\frac{1}{w(n)}\frac{\partial w(n)}{\partial \epsilon} + s(n)\nabla_{\mathbf{w}}\pi(w_l,w_h)\nabla_\epsilon \mathbf{w}\right) \\ &+\lambda\int_N\left(\tau(z) + T'(z(n))\left.\frac{\partial z(n)}{\partial \epsilon}\right|_{\mathbf{w}} + T'(z(n))\left.\frac{\partial z(n)}{\partial w(n)}\right|_{\epsilon}\frac{\partial w(n)}{\partial \epsilon}\right)dF(n) \end{aligned} \tag{144}$$

where $\mathbf{w} = (w_l, w_h)$ and noting that labor supply decisions of low-skilled types do not depend on high-skilled wages (and vice-versa) due to the assumption of no income effects. Next, we need to determine how to express $\nabla_\epsilon \mathbf{w}$ in terms of $\tau(z)$. Multiplying Equations 139 and 140 by w_l and w_h, respectively, and implicitly differentiating with respect to ϵ (recognizing that labor supply only responds to changes in the own wage and that labor demand does not react directly to a change in $\tau(z)$, only indirectly via the changing wage and that $w(n)nl(n) = z(n)$):

$$\frac{\partial w_l}{\partial \epsilon}L_l + w_l\frac{\partial L_l}{\partial w_l}\frac{\partial w_l}{\partial \epsilon} + w_l\frac{\partial L_l}{\partial w_h}\frac{\partial w_h}{\partial \epsilon} - \int_{\underline{n}}^{\text{med}(n)}\left(\left.\frac{\partial z(n)}{\partial w_l}\right|_\epsilon\frac{\partial w_l}{\partial \epsilon} + \left.\frac{\partial z(n)}{\partial \epsilon}\right|_{\mathbf{w}}\right)dF(n) = 0 \tag{145}$$

$$\frac{\partial w_h}{\partial \epsilon}L_h + w_h\frac{\partial L_h}{\partial w_h}\frac{\partial w_h}{\partial \epsilon} + w_h\frac{\partial L_h}{\partial w_l}\frac{\partial w_l}{\partial \epsilon} - \int_{\text{med}(n)}^{\overline{n}}\left(\left.\frac{\partial z(n)}{\partial w_h}\right|_\epsilon\frac{\partial w_h}{\partial \epsilon} + \left.\frac{\partial z(n)}{\partial \epsilon}\right|_{\mathbf{w}}\right)dF(n) = 0 \tag{146}$$

Applying the implicit function theorem to the individual first order condition $(1 - T'(z(n)) - \epsilon\tau'(z(n))) - \left(\frac{z(n)}{nw(n)}\right)^k\frac{1}{nw(n)}$ we get an analogue to Equation 12 which says that:

$$\left.\frac{\partial z(n)}{\partial \epsilon}\right|_{\mathbf{w}} = \frac{\tau'(z(n))}{-k\left(\frac{z(n)}{nw(n)}\right)^{k-1}\frac{1}{n^2w(n)^2} - T''(z(n))} \equiv \tau'(z(n))\xi(n)$$

Implicitly differentiating the individual first order condition with respect to $w(n)$, we get:

$$\left.\frac{\partial z(n)}{\partial w(n)}\right|_\epsilon = \frac{-(1+k)\left(\frac{z(n)}{nw(n)}\right)^k\frac{1}{nw(n)^2}}{-k\left(\frac{z(n)}{nw(n)}\right)^{k-1}\frac{1}{n^2w(n)^2} - T''(z(n))}$$

We can recover $\frac{\partial L_l}{\partial w_l}$, $\frac{\partial L_l}{\partial w_h}$, $\frac{\partial L_h}{\partial w_l}$, and $\frac{\partial L_h}{\partial w_h}$ from the implicit function theorem applied to the firm first order conditions, Equations 141 and 142. The expressions are:

$$\begin{aligned} \frac{\partial L_l}{\partial w_l} &= \frac{Y_{22}(L_l,L_h)}{Y_{11}(L_l,L_h)Y_{22}(L_l,L_h) - Y_{12}(L_l,L_h)^2} \\ \frac{\partial L_h}{\partial w_l} &= \frac{-Y_{12}(L_l,L_h)}{Y_{11}(L_l,L_h)Y_{22}(L_l,L_h) - Y_{12}(L_l,L_h)^2} \\ \frac{\partial L_l}{\partial w_h} &= \frac{-Y_{12}(L_l,L_h)}{Y_{11}(L_l,L_h)Y_{22}(L_l,L_h) - Y_{12}(L_l,L_h)^2} \\ \frac{\partial L_h}{\partial w_h} &= \frac{Y_{11}(L_l,L_h)}{Y_{11}(L_l,L_h)Y_{22}(L_l,L_h) - Y_{12}(L_l,L_h)^2} \end{aligned}$$

Next, let us apply a change of variables from n to z (where $h(z)$ represents the density of z) and apply integration by parts to the terms involving $\tau'(z)$ in Equations 145 and 146:[62]

$$\begin{aligned}
\int\limits_{\underline{n}}^{\text{med}(n)} \left.\frac{\partial z(n)}{\partial \epsilon}\right|_{\mathbf{w}} f(n)dn &= \int\limits_{\underline{n}}^{\text{med}(n)} \tau'(z(n))\xi(n)f(n)dn = \int\limits_{\underline{z}}^{\text{med}(z)} \tau'(z)\xi(z)h(z)dz \\
&= \int\limits_{\underline{z}}^{\text{med}(z)} -\tau(z)\frac{\partial}{\partial z}\left(\xi(z)h(z)\right)dz + \tau(z)\xi(z)h(z)\Big|_{\underline{z}}^{\text{med}(z)}
\end{aligned} \tag{147}$$

Identical steps yield:

$$\int\limits_{\text{med}(n)}^{\overline{n}} \left.\frac{\partial z(n)}{\partial \epsilon}\right|_{\mathbf{w}} f(n)dn = \int\limits_{\text{med}(z)}^{\overline{z}} -\tau(z)\frac{\partial}{\partial z}\left(\xi(z)h(z)\right)dz + \tau(z)\xi(z)h(z)\Big|_{\text{med}(z)}^{\overline{z}} \tag{148}$$

We are now ready to express $\frac{\partial w_l}{\partial \epsilon}$ and $\frac{\partial w_h}{\partial \epsilon}$ as linear functionals of $\tau(z)$. Let us assume for simplicity that $f(n) = 0$ for $n \in \{\underline{n}, \text{med}(n), \overline{n}\}$ and that $\frac{\partial z}{\partial n} \not\to 0$ as $n \to \{\underline{n}, \text{med}(n), \overline{n}\}$ so that $h(z) = 0$ at $\{\underline{z}, \text{med}(z), \overline{z}\}$.[63] One can then solve Equations 145 and 146 to yield that $\frac{\partial w_l}{\partial \epsilon}$ and $\frac{\partial w_h}{\partial \epsilon}$ are linear functionals of $\tau(z)$:

$$\frac{\partial w_l}{\partial \epsilon} = \frac{w_l \frac{\partial L_l}{\partial w_h} \int_{\text{med}(z)}^{\overline{z}} \frac{\partial}{\partial z}\left[\xi(z)h(z)\right]\tau(z)dz - C_2 \int_{\underline{z}}^{\text{med}(z)} \frac{\partial}{\partial z}\left[\xi(z)h(z)\right]\tau(z)dz}{C_1C_2 - w_l \frac{\partial L_l}{\partial w_h} w_h \frac{\partial L_h}{\partial w_l}} \tag{149}$$

$$\frac{\partial w_h}{\partial \epsilon} = \frac{w_h \frac{\partial L_h}{\partial w_l} \int_{\underline{z}}^{\text{med}(z)} \frac{\partial}{\partial z}\left[\xi(z)h(z)\right]\tau(z)dz - C_1 \int_{\text{med}(z)}^{\overline{z}} \frac{\partial}{\partial z}\left[\xi(z)h(z)\right]\tau(z)dz}{C_1C_2 - w_l \frac{\partial L_l}{\partial w_h} w_h \frac{\partial L_h}{\partial w_l}} \tag{150}$$

with

$$\begin{aligned}
C_1 &= L_l + w_l \frac{\partial L_l}{\partial w_l} - \int_{\underline{n}}^{\text{med}(n)} \left.\frac{\partial z(n)}{\partial w_l}\right|_{\epsilon} dF(n) \\
C_2 &= L_h + w_h \frac{\partial L_h}{\partial w_h} - \int_{\text{med}(n)}^{\overline{n}} \left.\frac{\partial z(n)}{\partial w_h}\right|_{\epsilon} dF(n)
\end{aligned}$$

Let us condense notation and say that $\frac{\partial w_l}{\partial \epsilon} \equiv \int_Z p_l(z)\tau(z)dz$ and $\frac{\partial w_h}{\partial \epsilon} \equiv \int_Z p_h(z)\tau(z)dz$. Let us normalize equilibrium wages for both skill types to be equal to 1 (i.e., we are loading all of the equilibrium pay differences into the distribution of types n). Using a change of variables

[62] Note that $\underline{z} \equiv z(\underline{n})$, $\overline{z} \equiv z(\underline{n})$, and $\text{med}(z) \equiv z(\text{med}(n))$.

[63] Note that we can have $f(n) \to 0$ arbitrarily quickly at $\text{med}(n)$, which will lead to arbitrarily large weights right around median n, but this will in general have little impact on total welfare because it the large weights apply to a very small measure of types. Alternatively, we can still find an inverse welfare functional if $h(z) \neq 0$ at $\{\underline{z}, \text{med}(z), \overline{z}\}$, we just have to formulate the resulting integral equation in a measure space as in the proof to Theorem 2.

and integration by parts we see that the government's budget is Gateaux differentiable in $T(z)$:

$$
\begin{aligned}
&\int_N \left(\tau(z) + T'(z(n)) \frac{\partial z(n)}{\partial \epsilon}\bigg|_{\mathbf{w}} + T'(z(n)) \frac{\partial z(n)}{\partial w(n)}\bigg|_{\epsilon} \frac{\partial w(n)}{\partial \epsilon} \right) dF(n) \\
&= \int_Z \left(h(z) - \frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right] \right) \tau(z) dz + \int_N T'(z(n)) \frac{\partial z(n)}{\partial w(n)}\bigg|_{\epsilon} \frac{\partial w(n)}{\partial \epsilon} dF(n) \\
&= \int_Z \left(h(z) - \frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right] \right) \tau(z) dz \\
&+ \int_Z \int_{\underline{n}}^{\text{med}(n)} T'(z(n)) \frac{\partial z(n)}{\partial w}\bigg|_{\epsilon} dF(n) p_l(z) \tau(z) dz + \int_Z \int_{\text{med}(n)}^{\overline{n}} T'(z(n)) \frac{\partial z(n)}{\partial w}\bigg|_{\epsilon} dF(n) p_h(z) \tau(z) dz \\
&= \int_Z \left(h(z) - \frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right] \right. \\
&+ p_l(z) \int_{\underline{n}}^{\text{med}(n)} T'(z(n)) \frac{\partial z(n)}{\partial w}\bigg|_{\epsilon} dF(n) + p_h(z) \int_{\text{med}(n)}^{\overline{n}} T'(z(n)) \frac{\partial z(n)}{\partial w}\bigg|_{\epsilon} dF(n) \left. \vphantom{\frac{\partial}{\partial z}} \right) \tau(z) dz
\end{aligned}
\tag{151}
$$

Equation 151 captures two separate budgetary impacts: the direct budgetary impact of individuals responding to tax changes and the indirect budgetary impacts of households responding to wage changes that result from changes in labor supply as a result of tax changes. Using similar logic, doing a change of variables from n to z (recalling $n \mapsto z$ was assumed bijective and differentiable):

$$
\begin{aligned}
&W\left(-\tau(z(n)) + \left(\frac{z(n)}{nw(n)} \right)^{1+k} \frac{1}{w(n)} \frac{\partial w(n)}{\partial \epsilon} + s(n) \nabla_{\mathbf{w}} \pi(w_l, w_h) \nabla_{\epsilon} \mathbf{w} \right) \\
&= -\int_Z \phi(n(z)) \tau(z) h(z) dz + \int_N \phi(n) \left[\left(\frac{z(n)}{n} \right)^{1+k} \frac{\partial w(n)}{\partial \epsilon} + s(n) \nabla_{\mathbf{w}} \pi(1,1) \nabla_{\epsilon} \mathbf{w} \right] f(n) dn \\
&= -\int_Z \phi(n(z)) \tau(z) h(z) dz + \int_Z p_l(z) \tau(z) \left(\int_{\underline{n}}^{\text{med}(n)} \phi(n) \left(\frac{z(n)}{n} \right)^{1+k} f(n) dn \right) dz \\
&+ \int_Z p_h(z) \tau(z) \left(\int_{\text{med}(n)}^{\overline{n}} \phi(n) \left(\frac{z(n)}{n} \right)^{1+k} f(n) dn \right) dz + \sum_{i=l,h} \int_Z p_i(z) \tau(z) \left(\int_N \phi(n) s(n) \frac{\partial \pi}{\partial w_i} f(n) dn \right) dz \\
&= -\int_Z \left[\phi(n(z)) h(z) - p_l(z) \left(\int_{\underline{z}}^{\text{med}(z)} \phi(n(\tilde{z})) \left(\frac{\tilde{z}}{n(\tilde{z})} \right)^{1+k} dH(\tilde{z}) \right) \right. \\
&- p_h(z) \left(\int_{\text{med}(z)}^{\overline{z}} \phi(n(\tilde{z})) \left(\frac{\tilde{z}}{n(\tilde{z})} \right)^{1+k} dH(\tilde{z}) \right) - \sum_{i=l,h} p_i(z) \left(\int_Z \phi(n(\tilde{z})) s(n(\tilde{z})) \frac{\partial \pi}{\partial w_i} dH(\tilde{z}) \right) \left. \vphantom{\int} \right] \tau(z) dz
\end{aligned}
\tag{152}
$$

Equation 152 captures two types of welfare impacts: direct welfare impacts of tax changes along with the indirect welfare impacts of tax changes that result from general equilibrium wage changes. Finally, we can construst the inverse welfare functional. We want a set of welfare weights such that Equation 144 equals zero. Normalizing the Lagrange multiplier λ to 1 this requires that Equation 151 plus Equation 152 must equal zero for all perturbations $\tau(\mathbf{z})$. To construct a set of welfare weights that satisfy this condition, let us define:

$$
\chi(z) \equiv \frac{h(z) - \frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right] + p_l(z) \int_{\underline{n}}^{\text{med}(n)} T'(z(n)) \frac{\partial z(n)}{\partial w}\big|_{\epsilon} dF(n) + p_h(z) \int_{\text{med}(n)}^{\overline{n}} T'(z(n)) \frac{\partial z(n)}{\partial w}\big|_{\epsilon} dF(n)}{h(z)}
$$

$$K(z) \equiv \frac{p_l(z)\int_{\underline{z}}^{\text{med}(z)} \phi(n(\tilde{z}))\left(\frac{\tilde{z}}{n(\tilde{z})}\right)^{1+k} dH(\tilde{z}) + p_h(z)\int_{\text{med}(z)}^{\overline{z}} \phi(n(\tilde{z}))\left(\frac{\tilde{z}}{n(\tilde{z})}\right)^{1+k} dH(\tilde{z}) + \sum_{i=l,h} p_i(z)\int_Z \phi(n(\tilde{z}))s(n(\tilde{z}))\frac{\partial \pi}{\partial w_i} dH(\tilde{z})}{h(z)}$$

From here, we can simply match terms pointwise in Equations 151 and 152 to see that Equation 151 plus Equation 152 equals zero for all $\tau(\mathbf{z})$ (i.e., the Gateaux derivative of the Lagrangian is zero) as long as the following equation holds:

$$\phi(n(z)) = \chi(z) + K(z) \tag{153}$$

As long as $\chi(z) + K(z)$ defines a contraction mapping on the set of functions $\phi(n(z))$, then Equation 153 has a solution that can be computed via standard fixed point algorithms as discussed in Section 6.2. This solution to Equation 153 then defines an inverse welfare functional for the given arbitrary tax schedule.

B.13 Proof to Proposition 5

To prove Proposition 5, we first set up a simple model of joint savings and income taxation. We assume households choose how much to work in the first period and then choose how much to save for the second period given an interest rate r. Taxes are a function of both income and savings. Utility is given by $u(c, s, z/n)$ where $c = z - \frac{1}{1+r}s - T(z, s)$ where s represents your net-of-interest savings (i.e., if you save x dollars in the first period, in the second period you get to consume $s = (1 + r)x$).[64] Suppose the tax schedule $T(z, s)$ is smooth and the mappings $n \mapsto z$ and $n \mapsto s$ are both bijective, all types have a unique optima, and that second order conditions hold strictly for all n. Also, suppose that the density $f(n)$ is zero at the top and bottom (this simplification just allows us to ignore the boundary terms and does not impact the argument).

Next, consider the impact of tax perturbations from $T(z, s)$ to $T(z, s) + \epsilon\tau(z, s)$. Implicit function theorem arguments as in Section 3 can be used to show that the behavioral impacts of a tax change can be expressed as:

$$\frac{\partial z}{\partial \epsilon}(n) = \eta_z(n)\tau(z, s) + \xi_z^z(n)\tau_z(z, s) + \xi_s^z(n)\tau_s(z, s)$$

$$\frac{\partial s}{\partial \epsilon}(n) = \eta_s(n)\tau(z, s) + \xi_z^s(n)\tau_z(z, s) + \xi_s^s(n)\tau_s(z, s)$$

for some functions $\eta_z, \xi_z^z, \xi_s^z, \eta_s, \xi_z^s, \xi_s^s$. Note that η_i represents the income effect for variable i and ξ_i^j represents the substitution effect of variable j with respect to the marginal tax rate on variable i.

Next, consider two different perturbations $\tau(z, s)$: we will consider perturbing the tax schedule in the direction of an arbitrary income tax change $\tau(z)$ and in the direction of an arbitrary savings tax change $\tau(s)$. Lemma 3 provides first order conditions that must be satisfied by a set of inverse welfare weights $\phi(n)$ (recall that there are one-to-one relationships between n, s, and z by assumption):

[64] Note, in practice taxes are typically a function of savings *income* rather than savings directly; however, any tax on savings income can be translated into a tax on savings given a constant interest rate r. For instance, if there is a 10% tax on savings income at an interest rate of 5%, then this is equivalent to a 0.05% tax on savings.

Lemma 3. *Under the smoothness and regularity assumptions discussed above, inverse welfare weights must satisfy:*

$$\phi(n(z)) = \frac{\left(1 + T_z(z, s(z))\eta_z(z) + T_s(z, s(z))\eta_s(z)\right) h(z) - \frac{\partial}{\partial z}\left(\left[T_z(z, s(z))\xi_z^z(z) + T_s(z, s(z))\xi_z^s(z)\right] h(z)\right)}{u_c(n(z))} \tag{154}$$

$$\phi(n(z)) = \frac{\left(1 + T_z(z, s(z))\eta_z(z) + T_s(z, s(z))\eta_s(z)\right) h(z) - \frac{\partial}{\partial z}\left(\left[T_z(z, s(z))\xi_s^z(z) + T_s(z, s(z))\xi_s^s(z)\right] \left(\frac{ds}{dz}\right)^{-1} h(z)\right)}{u_c(n(z))} \tag{155}$$

Proof. We have:

$$\begin{aligned}
&\frac{\partial R(T(z,s) + \epsilon\tau(z))}{\partial \epsilon}|_{\epsilon=0} = \int_N \frac{\partial}{\partial \epsilon}\left[T(z(n), s(n)) + \epsilon\tau(z(n))\right] f(n)dn \\
&= \int\limits_N \Bigg(\left(1 + T_z(z(n), s(n))\eta_z(n) + T_s(z(n), s(n))\eta_s(n)\right) \tau(z(n)) \\
&+ \left(T_z(z(n), s(n))\xi_z^z(n) + T_s(z(n), s(n))\xi_z^s(n)\right) \tau_z(z(n))\Bigg) f(n)dn \\
&= \int\limits_Z \Bigg(\left(1 + T_z(z, s(z))\eta_z(z) + T_s(z, s(z))\eta_s(z)\right) \tau(z) + \left(T_z(z, s(z))\xi_z^z(z) + T_s(z, s(z))\xi_z^s(z)\right) \tau_z(z)\Bigg) h(z)dz \\
&= \int\limits_Z \Bigg[\left(1 + T_z(z, s(z))\eta_z(z) + T_s(z, s(z))\eta_s(z)\right) h(z)\Bigg] - \frac{\partial}{\partial z}\left(\left[T_z(z, s(z))\xi_z^z(z) + T_s(z, s(z))\xi_z^s(z)\right] h(z)\right) \tau(z)dz
\end{aligned} \tag{156}$$

The first equality is just the definition of $R(T(z,s) + \epsilon\tau(z))$; the second equality uses the chain rule to evaluate $\frac{\partial T(z(n),s(n))}{\partial \epsilon}$; the third just does a change of variables from n to z noting that we assumed $n \mapsto z$ is bijective and using the fact that $h(z) = f(n(z))\frac{dz}{dn}$ so that $h(z)$ incorporates the Jacobian of the transformation; the final equality just applies integration by parts using the fact that the boundary terms are 0 as we assume $f(n) = 0$ on the boundary (and assuming that $\frac{dz}{dn} \not\to 0$ as $n \to \underline{n}$ or as $n \to \overline{n}$). Similarly, we have:

$$\begin{aligned}
&\frac{\partial R(T(z,s) + \epsilon\tau(s))}{\partial \epsilon}|_{\epsilon=0} = \int_N \frac{\partial}{\partial \epsilon}\left[T(z(n), s(n)) + \epsilon\tau(s(n))\right] f(n)dn \\
&= \int\limits_N \Bigg(\left(1 + T_z(z(n), s(n))\eta_z(n) + T_s(z(n), s(n))\eta_s(n)\right) \tau(s(n)) \\
&+ \left(T_z(z(n), s(n))\xi_s^z(n) + T_s(z(n), s(n))\xi_s^s(n)\right) \tau_s(s(n))\Bigg) f(n)dn \\
&= \int\limits_Z \Bigg(\left(1 + T_z(z, s(z))\eta_z(z) + T_s(z, s(z))\eta_s(z)\right) \tau(z) + \left(T_z(z, s(z))\xi_s^z(z) + T_s(z, s(z))\xi_s^s(z)\right) \tau_s(s(z))\Bigg) h(z)dz \\
&= \int\limits_Z \Bigg(\left(1 + T_z(z, s(z))\eta_z(z) + T_s(z, s(z))\eta_s(z)\right) \tau(z) + \left(T_z(z, s(z))\xi_s^z(z) + T_s(z, s(z))\xi_s^s(z)\right) \tau_s(s(z))\frac{ds}{dz}\left(\frac{ds}{dz}\right)^{-1}\Bigg) h(z)dz \\
&= \int_Z \Bigg[\left[1 + T_z(z, s(z))\eta_z(z) + T_s(z, s(z))\eta_s(z)\right] h(z) - \frac{\partial}{\partial z}\left(\left[T_z(z, s(z))\xi_s^z(z) + T_s(z, s(z))\xi_s^s(z)\right]\left(\frac{ds}{dz}\right)^{-1} h(z)\right)\Bigg]\tau(z)dz
\end{aligned} \tag{157}$$

The first equality is just the definition of $R(T(z,s) + \epsilon\tau(s))$; the second equality uses the chain rule to evaluate $\frac{\partial T(z(n),s(n))}{\partial \epsilon}$; the third just does a change of variables from n to z noting that we assumed $n \mapsto z$ is bijective and using the fact that $h(z) = f(n(z))\frac{dz}{dn}$ so that $h(z)$ incorporates the Jacobian of the transformation; the fourth equality multiplies and divides by $\frac{ds}{dz}$ (note,

$\frac{ds}{dz}$ varies with z); the final equality just applies integration by parts using the fact that the boundary terms are 0 as we assume $f(n) = 0$ on the boundary (and assuming that $\frac{dz}{dn} \not\to 0$ as $n \to \underline{n}$ or as $n \to \overline{n}$) and the fact that $\frac{d\tau(s(z))}{dz} = \tau_s(s(z))\frac{ds}{dz}$.

Finally suppose that welfare is given by $W(U(n;T)) = \int_N \phi(n)U(n;T)dn$ (if the welfare functional has mass points at particular n then T cannot be a stationary point of the government's Lagrangian because the Gateaux variations 156 and 157 do not have mass points) and note that we have by the envelope theorem (given our assumption that all types have a unique optima):

$$\frac{\partial W(U(n;T(z,s)+\epsilon\tau(z)))}{\partial\epsilon}|_{\epsilon=0} = -\int_N \phi(n)u_c(n)\tau(z(n))dn$$

$$\frac{\partial W(U(n;T(z,s)+\epsilon\tau(s)))}{\partial\epsilon}|_{\epsilon=0} = -\int_N \phi(n)u_c(n)\tau(s(n))dn$$

Hence, in order for $\frac{\partial W(U(n;T(z,s)+\epsilon\tau(z)))+\lambda R(T(z,s)+\epsilon\tau(z))}{\partial\epsilon}|_{\epsilon=0} = 0$ and $\frac{\partial W(U(n;T(z,s)+\epsilon\tau(s)))+\lambda R(T(z,s)+\epsilon\tau(s))}{\partial\epsilon}|_{\epsilon=0} = 0$, Equations 154 and 155 must be satisfied (recognizing that we can normalize $\lambda = 1$).

□

Lemma 3 essentially says that for a set of inverse welfare weights to ensure that an arbitrary income tax perturbation leaves the government Lagrangian unchanged, Equation 154 must be satisfied. Similarly, to ensure that an arbitrary savings tax perturbation leaves the government Lagrangian unchanged, Equation 155 must be satisfied. The key point is that $\phi(n)$ is overdetermined and that Equation 154 often does not equal Equation 155 at all z. When Equation 154 does not equal Equation 155, we can find either an income tax perturbation or a savings tax perturbation that improves welfare under any given welfare weights (i.e., a local inverse welfare functional does not exist).

We can now prove Proposition 5 simply by providing a number of examples in Figure 9 of the inverse welfare weights that satisfy Equation 154 along with the inverse welfare weights that satisfy Equation 155. In all but the top left panel of Figure 9 where savings taxes are zero, the given tax schedules do not have any inverse welfare functional because the inverse weights that satisfy Equation 154 are different than the inverse weights that satisfy Equation 155. In general, Equation 154 equals Equation 155 only in knife-edge cases so that most arbitrary tax schedules will not satisfy this property. Note, this argument did not rely on separability in any way: hence most tax schedules will not have associated inverse welfare functionals regardless of whether utility is weakly separable or not (Figure 10 shows similar findings for a non-separable utility function as well as for non-linear tax schedules).

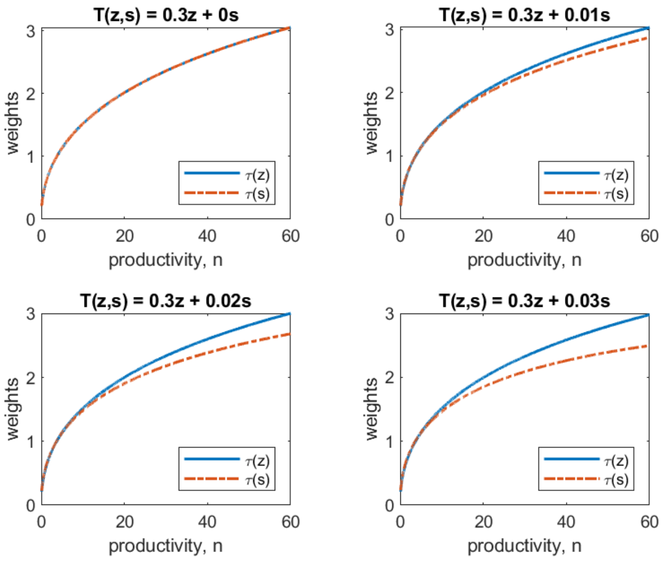


Figure 9: Inverse Weights for $\tau(z)$ and $\tau(s)$ Perturbations

Note: This figure shows the inverse welfare weights that satisfy Equation 154 in blue solid lines (i.e., ensure that the Gateaux variation of any income tax perturbation $\tau(z)$ is zero) and shows the inverse welfare weights that satisfy Equation 155 in orange dashed lines (i.e., ensure that the Gateaux variation of any savings tax perturbation $\tau(s)$ is zero). Each of the four panels is labeled with the tax schedule $T(z,s)$ for which we are finding inverse welfare weights. Utility is given by $u(c,s,z/n) = \frac{c^{1-\alpha}}{1-\alpha} + \beta\frac{s^{1-\alpha}}{1-\alpha} + \frac{(z/n)^{1+k}}{1+k}$ where $c = z - T(z,s) - \frac{s}{1+r}$ and $\{\alpha, \beta, k, r\} = \{0.5, 1/1.03, 1/0.3, 0.05\}$. $f(n)$ is calibrated to match the observed distribution of incomes from the 2019 ACS. At the assumed interest rate of 5%, a 1% (2%, 3%, respectively) savings tax is equivalent to a 20% (40%, 60%, respectively) tax on interest income.

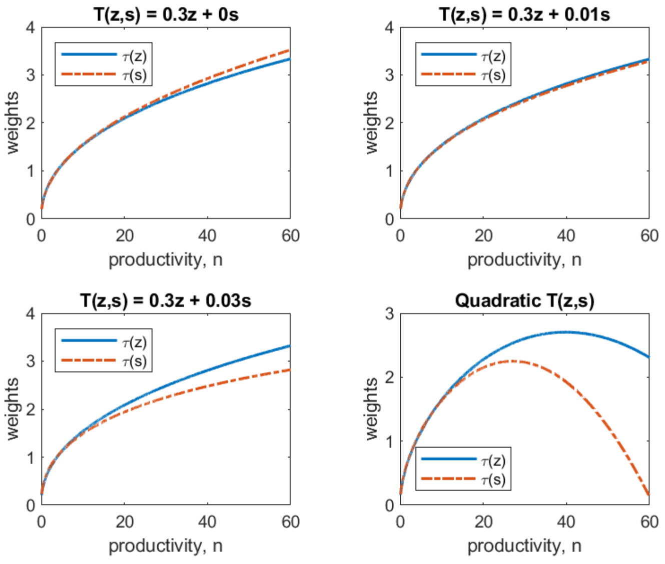


Figure 10: Inverse Weights for $\tau(z)$ and $\tau(s)$ Perturbations: Non-Separable Utility Function

Note: This figure shows the inverse welfare weights that satisfy Equation 154 in blue solid lines (i.e., ensure that the Gateaux variation of any income tax perturbation $\tau(z)$ is zero) and shows the inverse welfare weights that satisfy Equation 155 in orange dashed lines (i.e., ensure that the Gateaux variation of any savings tax perturbation $\tau(s)$ is zero). Each of the four panels is labeled with the tax schedule $T(z,s)$ for which we are finding inverse welfare weights. The parameters of the quadratic tax schedule are chosen so that marginal tax rates on both income and savings are 10% for the lowest type $\underline{n}$ and are 50% for the highest type $\overline{n}$. Utility is given by $u(c,s,z/n) = \frac{c^{1-\alpha}}{1-\alpha} + \beta(n)\frac{s^{1-\alpha}}{1-\alpha} + \frac{(z/n)^{1+k}}{1+k}$ where $c = z - T(z,s) - \frac{s}{1+r}$ and $\{\alpha, k, r\} = \{0.5, 1/0.3, 0.05\}$. $f(n)$ is calibrated as in Figure 9 and $\beta(n)$ is an increasing linear function of n that ranges from 0.7 for the lowest n to 0.99 for the highest n. At the assumed interest rate of 5%, a 1% (3%, respectively) savings tax is equivalent to a 20% (60%, respectively) tax on interest income.

Finally, it is worth noting that when utility is weakly separable in (c, s) and z (so that $u(c, s, z/n) = u(v(c, s), z/n)$ for some sub-utility function v) and taxes are only a function of income $T(z)$, then Equation 154 *will* equal Equation 155; this is why the associated inverse welfare weights satisfying Equation 154 and Equation 155 in the top left panel of Figure 9 *do* coincide. To show Equation 154 equals Equation 155 when $T_s = 0$ under weak separability it suffices to show that:

$$\xi_z^z(z) = \xi_s^z(z)\left(\frac{ds}{dz}\right)^{-1} \tag{158}$$

In other words, we require that the behavioral impact on z of a marginal tax change on z is equal to the behavioral impact on z of a marginal tax change on s scaled by $\left(\frac{ds}{dz}\right)^{-1}$. This follows almost immediately by Lemma 1 of Ferey, Lockwood and Taubinsky (2021) who prove that, more generally, $\xi_z^z(z) = \xi_s^z(z)\left(\frac{\partial s(n,z)}{\partial z}\right)^{-1}$. While we require instead that $\xi_z^z(z) = \xi_s^z(z)\left(\frac{ds(n(z),z)}{dz}\right)^{-1}$, weak separability ensures that $\frac{\partial s(n,z)}{\partial n} = 0$ because s is not a function of n conditional on a value of z (intuitively, this is because optimal s is determined by the first order condition $v_c(c, s)\frac{\partial c}{\partial s} + v_s = 0$, which does not depend on n). Thus, Equation 158 holds under weak separability. Hence:

Remark 6. *Under the assumptions listed at the start of this appendix and if utility is weakly separable in (c, s) and z, any $T(z)$ satisfying the budget constraint strictly yields a Gateaux differentiable $R(T)$.*[65] *By Theorem 1, any such schedule therefore has an inverse welfare functional that rationalizes this schedule (locally) within all tax schedules $T(z, s)$.*

B.14 Misperceptions Example: Existence

There is evidence in the empirical literature that individuals often misunderstand the difference between marginal tax rates and average tax rates (de Bartolome, 1995). We now show how to construct an inverse welfare functional if individuals misperceive their average tax rate as if it was their marginal tax rate.

Suppose that individuals vary in a unidimensional type n and face a smooth non-linear tax schedule $T(z)$. However, suppose that they misperceive their marginal tax rate as $T(z)/z$, which is, in actuality, their average tax rate. Suppose utility is quasi-linear and isoelastic: $u(c, z; n) = c - \frac{(z/n)^{1+k}}{1+k}$ (hence, $n \mapsto z$ is monotonic by the single crossing property). First order conditions are $(1 - T(z)/z) - (z/n)^k/n = 0$. Consider a tax change $T(z) \to T(z) + \epsilon\tau(z)$ where agents also incorrectly perceive their after-perturbation marginal tax rate as $(T(z) + \tau(z))/z$; hence, they perceive that their marginal rate has changed by $\tau(z)/z$. The impact on indirect

[65] We have actually shown that Gateaux variations in the directions $\tau(z)$ and $\tau(s)$ are described by the same continuous linear functional under weak separability and any $T(z)$ whereas Gateaux differentiability requires that Gateaux variations *in all directions* $\tau(z, s)$ are described by the same continuous linear functional. One can show using Equation 158 that, under weak separability and any $T(z)$, revenue is in fact Gateaux differentiable with Gateaux derivative:

$$\int_Z \left[(1 + T_z(z)\eta_z(z))\,h(z) - \frac{\partial}{\partial z}\left(T_z(z)\xi_z^z(z)h(z)\right)\right]\tau(z, s(z))dz$$

utility $U(n;T)$ of such a tax change equals:

$$\begin{aligned}\frac{\partial U(n;T)}{\partial\epsilon} &= -\tau(z(n)) + [(1-T'(z(n))) - (z(n)/n)^k/n]\kappa(n)\frac{\tau(z(n))}{z(n)} \\ &= -\tau(z(n)) + [(1-T'(z(n))) - (1-T(z(n))/z(n)))]\kappa(n)\frac{\tau(z(n))}{z(n)}\end{aligned}$$

where $\kappa(n)$ comes from applying the implicit function theorem to $(1 - T(z)/z - \epsilon\tau(z)/(z) - (z/n)^k/n = 0$ to determine $\frac{\partial z}{\partial\epsilon}(n) \equiv \kappa(n)\tau(z)$. We want to know whether we can find a linear functional, $W(U(n;T))$, that rationalizes this tax schedule as optimal? We show how to find an inverse welfare functional of the form:

$$W(U(n;T)) = \int_N \phi(n)U(n;T)f(n)dn$$

The Gateaux variation of $W(U(n;T))$ in the direction of a tax perturbation $\tau(z)$ is:

$$\begin{aligned}&\lim_{\epsilon\to 0}\frac{W(U(n;T+\epsilon\tau)) - W(U(n;T))}{\epsilon} \\ &= \int_N \phi(n)\left[-\tau(z(n)) + [(1-T'(z(n))) - (1-T(z(n))/z(n)))]\kappa(n)\frac{\tau(z(n))}{z(n)}\right]f(n)dn \qquad (159) \\ &= \int_Z \phi(n(z))\left[-\tau(z) + [(1-T'(z)) - (1-T(z)/z))]\kappa(z)\frac{\tau(z)}{z}\right]h(z)dz\end{aligned}$$

The budgetary impacts of a tax perturbation in the direction $\tau(z)$ are:

$$\lim_{\epsilon\to 0}\frac{R(T+\epsilon\tau) - R(T)}{\epsilon} = \int_N \left[\tau(z(n)) + T'(z)\kappa(n)\frac{\tau(z(n))}{z(n)}\right]f(n)dn = \int_Z \left[\tau(z) + T'(z)\kappa(z)\frac{\tau(z)}{z}\right]h(z)dz \tag{160}$$

where the second integral is just a change of variables assuming that $n \mapsto z$ is monotonic (e.g., the Mirrlees (1971) single crossing property is satisfied). By Equation 163, $R(T)$ *is* Gateaux differentiable. Hence, the Gateaux derivative of the government's Lagrangian $W(U(n;T)) + \lambda R(T)$ (with λ normalized to 1) equals:

$$\int_Z \left\{\phi(n(z))\left[-h(z) + [(1-T'(z)) - (1-T(z)/z))]\kappa(z)\frac{h(z)}{z}\right] + \left[h(z) + T'(z)\kappa(z)\frac{h(z)}{z}\right]\right\}\tau(z)dz \tag{161}$$

Hence, we can find an inverse welfare functional that sets Equation 161 equal to zero for all possible $\tau(z)$ as long as we choose $\phi(n(z))$ to satisfy:

$$\phi(n(z)) = \frac{h(z) + T'(z)\kappa(z)\frac{h(z)}{z}}{h(z) - [(1-T'(z)) - (1-T(z)/z))]\kappa(z)\frac{h(z)}{z}}$$

B.15 Misperceptions Example: Non-Existence

We present another example with misperceptions where government revenue is Gateaux differentiable yet an inverse welfare functional still does not exist. Suppose that individuals vary in a unidimensional type n and face a constant marginal tax rate; however, suppose that they misperceive this tax rate, causing them to optimize incorrectly. Let T_z represent the actual marginal tax rate and $\hat{T}_z$ represent the misperceived tax rate. Suppose utility is quasi-linear and isoelastic: $u(c,z;n) = c - \frac{(z/n)^{1+k}}{1+k}$ (hence, $n \mapsto z$ is monotonic by the single crossing property). First order conditions are $(1-\hat{T}_z)-(z/n)^k/n=0$. Consider a tax change $T(z)\to T(z)+\epsilon\tau(z)$ where agents correctly perceive the tax change $\tau(z)$. The impact on indirect utility $\frac{\partial U(n;T)}{\partial \epsilon}$ of such a tax change equals:

$$-\tau(z(n))+[(1-T_z)-(z(n)/n)^k/n]\xi(n)\tau'(z(n))=-\tau(z(n))+[(1-T_z)-(1-\hat{T}_z)]\xi(n)\tau'(z(n))$$

where $\xi(n)$ comes from applying the implicit function theorem to $(1-\hat{T}_z-\epsilon\tau'(z))-(z/n)^k/n=0$ to determine $\frac{\partial z}{\partial \epsilon}(n)\equiv\xi(n)\tau'(z(n))$ as in Equation 12. We want to know whether we can find a linear functional, $W(U(n;T))$, that rationalizes this tax schedule as optimal? The Riesz-Markov-Kakutani representation theorem ensures that every continuous linear functional $W(U(n;T))$ can be written as follows for some function of bounded variation $\Phi(n)$:

$$W(U(n;T))=\int_N U(n;T)d\Phi(n)$$

The Gateaux variation of $W(U(n;T))$ in the direction of a tax perturbation $\tau(z)$ is:

$$\int_N\left[-\tau(z(n))+[(1-T_z)-(1-\hat{T}_z)]\xi(n)\tau'(z(n))\right]d\Phi(n) \tag{162}$$

The budgetary impacts of a tax perturbation in the direction $\tau(z)$ are:

$$\begin{aligned}&\lim_{\epsilon\to 0}\frac{R(T+\epsilon\tau)-R(T)}{\epsilon}=\int_N\left[\tau(z(n))+T_z\xi(n)\tau'(z(n))\right]f(n)dn\\&=\int_Z\left[\tau(z)+T_z\xi(z)\tau'(z)\right]h(z)dz=\int_Z\left[h(z)-\frac{\partial}{\partial z}\left(T_z\xi(z)h(z)\right)\right]\tau(z)dz+T_z\xi(z)h(z)\tau(z)\Big|_{\underline{z}}^{\overline{z}}\end{aligned} \tag{163}$$

where the equality follows from a change of variables assuming that $n\mapsto z$ is monotonic (e.g., the Mirrlees (1971) single crossing property is satisfied) and the third equality applies integration by parts. By Equation 163, $R(T)$ *is* Gateaux differentiable. Returning to Equation 162, we claim $\Phi(n)$ cannot have mass points (which is equivalent to continuity of $\Phi(n)$). If $\Phi(n)$ has mass points for $n\in\text{Int}(N)$ (who choose $z\in\text{Int}(Z)$ by monotonicity), Equation 162 cannot equal Equation 163 because Equation 163 does not have mass points on the interior. Alternatively, if $\Phi(n)$ has mass points for $n\in\{\underline{n},\overline{n}\}$ (who choose $z\in\{\underline{z},\overline{z}\}$), then the welfare impacts of a tax change will depend on the value of $\tau'(z)$ for $z\in\{\underline{z},\overline{z}\}$, which again means Equation 162 cannot equal Equation 163. Hence, $\Phi(n)$ is continuous; given that $\Phi(n)$ is a function of bounded variation, $\Phi(n)$ is therefore absolutely continuous (bounded variation functions are always the difference of two monotone functions). Hence, $\Phi(n)$ is differentiable a.e. with $\frac{d\Phi(n)}{dn}\equiv\psi(n)f(n)$. Then using integration by parts to get rid of the $\tau'(z)$ terms from Equation 162:

$$\begin{aligned}&\int_N\left[-\tau(z(n))+[(1-T_z)-(1-\hat{T}_z)]\xi(n)\tau'(z(n))\right]\psi(n)f(n)dn\\&=\int_Z\left[-\tau(z)+[(1-T_z)-(1-\hat{T}_z)]\xi(z)\tau'(z)\right]\psi(n(z))h(z)dz\\&=\int_Z\left[-\psi(n(z))h(z)-\frac{\partial}{\partial z}\left\{[(1-T_z)-(1-\hat{T}_z)]\xi(z)\psi(n(z))h(z)\right\}\right]\tau(z)dz\\&+\left\{[(1-T_z)-(1-\hat{T}_z)]\xi(z)h(z)\psi(n(z))\right\}\tau(z)\Big|_{\underline{z}}^{\overline{z}}\end{aligned} \tag{164}$$

If $W(U(n;T))$ is an inverse welfare functional then the Gateaux derivative of the government

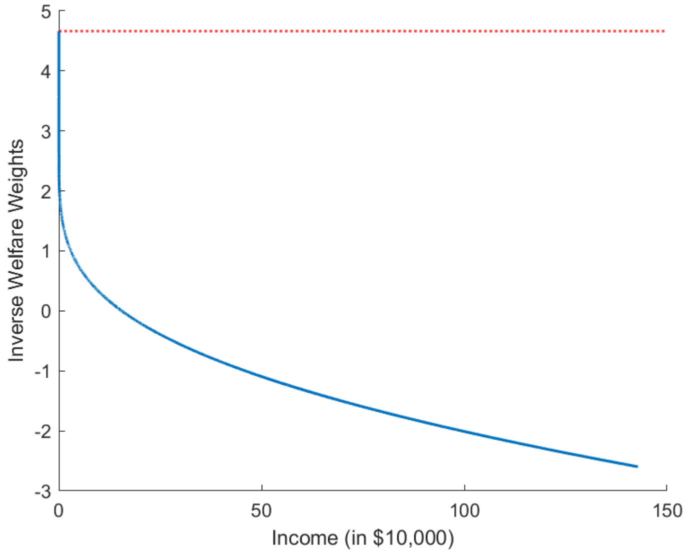


Figure 11: Inverse Welfare Weights that Solve System 166 Given Initial Value at $\underline{z}$

Note: This figure shows the inverse welfare weights that satisfy the differential equation in System 166 given the initial condition $\psi(n(\underline{z})) = -\frac{T_z}{[(1-T_z)-(1-\hat{T}_z)]}$ with the solid blue line. We also plot $-\frac{T_z}{[(1-T_z)-(1-\hat{T}_z)]}$ with a dashed red line. We set $k = 1/0.3$, $f(n)$ is calibrated to match the U.S. income distribution from the 2019 ACS, $(1-T_z) = 0.7$, and $(1-\hat{T}_z) = 0.85$. The solution to this differential equation with this given initial condition does not satisfy System 166 because $\psi(n(\overline{z})) \neq -\frac{T_z}{[(1-T_z)-(1-\hat{T}_z)]}$.

Lagrangian $W(U(n;T)) + \lambda R(T)$ must equal zero (where we can normalize $\lambda = 1$ as usual):

$$\int_Z \left[-\psi(n(z))h(z) - \frac{\partial}{\partial z}\left\{[(1-T_z)-(1-\hat{T}_z)]\xi(z)\psi(n(z))h(z)\right\} + h(z) - \frac{\partial}{\partial z}\left(T_z\xi(z)h(z)\right)\right]\tau(z)dz$$
$$+\left\{[(1-T_z)-(1-\hat{T}_z)]\xi(z)h(z)\psi(n(z)) + T_z\xi(z)h(z)\right\}\tau(z)\Big|_{\underline{z}}^{\overline{z}} = 0 \tag{165}$$

Thus, we get a boundary value problem wherein:

$$\begin{aligned}&\psi(n(z))h(z) + \frac{\partial}{\partial z}\left\{[(1-T_z)-(1-\hat{T}_z)]\xi(z)h(z)\psi(n(z))\right\} = h(z) - \frac{\partial}{\partial z}\left(T_z\xi(z)h(z)\right)\\&\text{and } [(1-T_z)-(1-\hat{T}_z)]\xi(z)h(z)\psi(n(z)) + T_z\xi(z)h(z) = 0 \text{ for } z \in \{\underline{z}, \overline{z}\}\end{aligned} \tag{166}$$

In most cases the boundary value problem given by System 166 will have no solution because it is overdetermined (i.e., there is a solution for this ODE with the initial value specified at $\underline{z}$, but in general the value of the solution will not coincide with the prescribed value at $\overline{z}$); we provide an explicit example of the unsolvability of System 166 in Figure 11.

The careful reader may notice that in Equation 164 we implicitly assumed that $\Phi(n)$ was not just differentiable but in fact *twice* differentiable. However, we can derive an integrated version of Equation 166 if we instead use integration by parts in Equations 162 and 163 to express the integrals involving $\tau(z)$ as integrals involving $\tau'(z)$ rather than the reverse. We omit this for brevity but are happy to provide the proof if requested.

C Online Appendix: Simulations

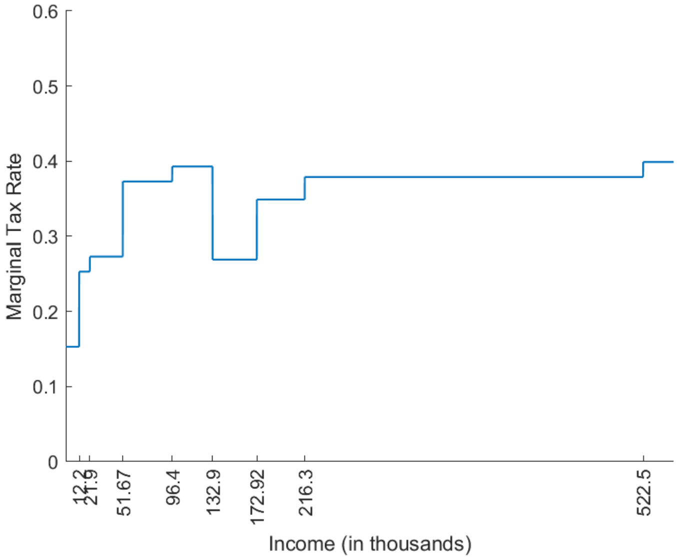


Figure 12: U.S. Income Tax Schedule 2019

Note: This figure shows the income tax schedule in the United States in 2019 for single adult households accounting for federal income taxes and payroll taxes (i.e., Social Security and Medicare taxes levied on wages). This tax schedule assumes all households take the standard deduction. This tax schedule also assumes that labor demand is perfectly elastic (as in the standard Mirrlees (1971) model) so that all taxes are perfectly passed through to workers.

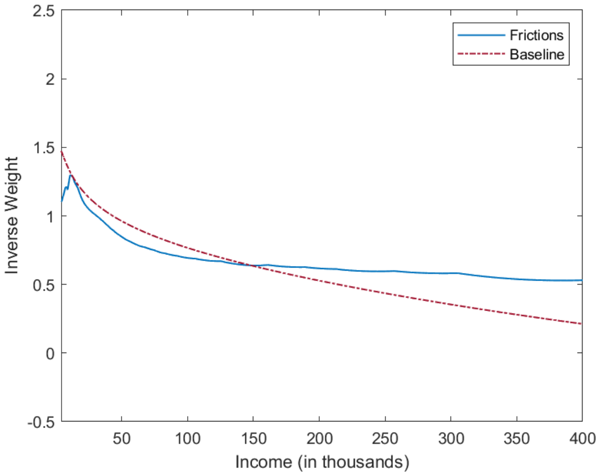


Figure 13: Inverse Welfare Weights with Frictions for Smooth Approximation to U.S. Income Tax Schedule

Note: This figure shows inverse welfare weights across the income distribution for a smooth approximation to the U.S. combined tax schedule of income and payroll taxes shown in Figure 12. The "Baseline" curve is identical to Figure 3. For the "Frictions" curve, agents are assumed to face sparsity-based frictions by solving Equation 39 with the calibration described in Section 5.3.

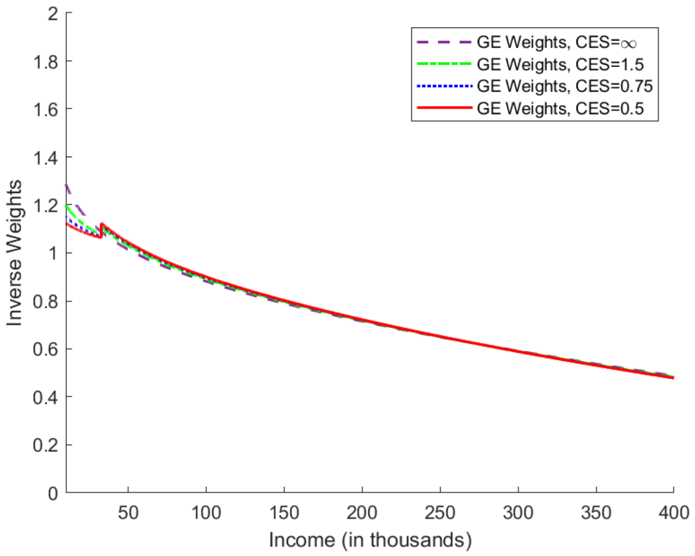


Figure 14: Inverse Welfare Weights with Finite Labor Demand Elasticity and CES Production

Note: This figure shows the inverse welfare weights for a smooth approximation to the U.S. income tax schedule computed under various assumptions about the degree of complementarity between high- and low-skilled labor. We assume the production function equals $Y(L_l, L_h) = (a_l L_l^\sigma + a_h L_h^\sigma)^{\frac{v}{\sigma}}$ where a_l and a_h are calibrated so that equilibrium wages for both high- and low-skilled types are normalized to 1 (i.e., we load all equilibrium income differences into the productivity distribution). Low-skilled types (those below median productivity) are paid wage w_l and high-skilled types (those above median productivity) are paid wage w_h. The productivity distribution is calibrated as in Section 5.1. The CES $= \infty$ line is computed assuming $v = 1/2$ so that the labor demand elasticity is -2 and $\sigma = 1$ so that high- and low-skilled labor are perfect substitutes; this line matches the $E^D = -2$ line in Figure 6. We then plot inverse welfare weights that support this tax schedule with $v = 1/2$ under various assumptions about the value of the elasticity of substitution between L_l and L_h: $\frac{1}{1-\sigma}$. Whenever the elasticity of substitution, $\frac{1}{1-\sigma}$, is less than 1, L_l and L_h are gross complements. The typical value used in macro studies is 1.5 (Autor, Katz and Kearney, 2008), although Havranek et al. (2020) argue that estimates from the literature are more consistent with a value of 0.6-0.9.