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 2308
December 2023

Abstract

Optimal taxation problems typically involve finding a tax schedule to maximize a welfare function. This paper considers the reverse problem of finding an inverse welfare function that rationalizes a given tax schedule as optimal. Inverse welfare functions encode the implicit interpersonal comparisons a society must make in order to justify a tax schedule. We develop a general theory to recover the inverse social welfare function not only for income tax schedules, but also for substantially more complex tax systems that incorporate many different forms of taxation and multidimensional agent heterogeneity. The key insight is that even in complex tax environments, the (Gateaux) derivative of government revenue with respect to the tax schedule is the key empirical object required to construct the inverse welfare function. Additionally, our framework allows us to characterize Pareto efficient schedules in complex environments and extend the Atkinson-Stiglitz result. Our framework can also be augmented to construct inverse welfare functions when there are general equilibrium effects of taxation and when agents make optimization errors. We provide a number of example inverse welfare function constructions related to the taxation of couples, income taxation with labor demand and endogenous wages, piecewise linear income taxation, and joint taxation of income and housing rent.

Keywords: inverse optimal, inverse welfare, multidimensional taxation, multidimensional heterogeneity
JEL codes: D82, D86, H21

A General Theory of Inverse Welfare Functions*

Katy Bergstrom† William Dodds‡

December 17, 2023

Abstract

Optimal taxation problems typically involve finding a tax schedule to maximize a welfare function. This paper considers the reverse problem of finding an inverse welfare function that rationalizes a given tax schedule as optimal. Inverse welfare functions encode the implicit interpersonal comparisons a society must make in order to justify a tax schedule. We develop a general theory to recover the inverse social welfare function not only for income tax schedules, but also for substantially more complex tax systems that incorporate many different forms of taxation and multidimensional agent heterogeneity. The key insight is that even in complex tax environments, the (Gateaux) derivative of government revenue with respect to the tax schedule is the key empirical object required to construct the inverse welfare function. Additionally, our framework allows us to characterize Pareto efficient schedules in complex environments and extend the Atkinson-Stiglitz result. Our framework can also be augmented to construct inverse welfare functions when there are general equilibrium effects of taxation and when agents make optimization errors. We provide a number of example inverse welfare function constructions related to the taxation of couples, income taxation with labor demand and endogenous wages, piecewise linear income taxation, and joint taxation of income and housing rent.

Keywords: inverse optimal, inverse welfare, multidimensional taxation, multidimensional heterogeneity

JEL: D82, D86, H21

*The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors.

†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 social 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 preferences, this places an onus on the government to change tax policy. For instance, if a given tax system is rationalized by an inverse welfare function that values giving \$1 to a billionaire more than giving \$1 to someone earning \$10,000 per year, this is likely to be quite inconsistent with societal preferences for redistribution, suggesting a need for a more progressive tax schedule.

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) explores 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, typically only allow for unidimensional heterogeneity, and assume that individuals respond smoothly to tax reforms.

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 not only for income tax schedules, but also for substantially more complex tax systems that incorporate many different forms of taxation. The key insight is that 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. This is our first main result, Theorem 1: 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. Importantly, the theorem is constructive: we show explicitly how to construct the inverse welfare function for arbitrary multidimensional tax schedules. Roughly, we compute the inverse welfare weight at a given set of 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, where an "instantaneous" tax change corresponds to an infinitesimal "bump function" perturbation. Importantly, the Gateaux derivative of revenue is (at least in principle) estimable; hence, inverse welfare weights can be recovered from empirical objects.

The next section of the paper provides a number of example constructions of Gateaux derivatives of government revenue along with associated inverse welfare functions. To build intuition, we show how to compute the Gateaux derivative of revenue and construct an inverse welfare function in the context of the income tax model of Mirrlees (1971) when the tax schedule and behavioral responses to taxation are smooth. We then showcase this procedure for income taxation with non-smooth tax schedules and/or non-smooth behavioral responses (e.g., when the tax schedule is piecewise linear, generating bunching and/or individuals with multiple optima). Finally, we show more broadly how the theory can be applied to compute the Gateaux derivative and an inverse welfare function for an arbitrary multidimensional tax schedule when behavioral responses are smooth.

The next key result of the paper, Proposition 1, establishes that, more generally, 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. Hence, Proposition 1 combined with Theorem 1 establishes that inverse welfare functions often exist even if the tax schedule is "pathological" in various ways.

Next, we highlight how our theory of inverse welfare functions relates to various important results in public economics. First, we note that inverse welfare functions may require "mass points" on different individuals so that the welfare function does not solely consist of welfare weights in the traditional sense; these mass points are required to, for example, rationalize income tax schedules with non-zero marginal rates at the top or bottom. Second, we discuss how to characterize Pareto efficient multidimensional tax schedules using our theory of inverse welfare functions. 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; conversely, tax schedules with a positive inverse welfare function are Pareto efficient. Third, we provide a non-existence result in the setting of Atkinson and Stiglitz (1976). The Atkinson-Stiglitz Theorem establishes that, when agents differ in terms of a unidimensional parameter and the utility function satisfies weak separability, any multidimensional tax schedule (as a function of income and other decisions) is Pareto dominated by a tax schedule that is only a function of income (Kaplow, 2006). Strengthening this result, we show that in the Atkinson-Stiglitz setting, most multidimensional tax schedules are not only Pareto dominated, but they are in fact not supported by *any* inverse welfare function, even those

that feature negative weights or have mass points. Fundamentally, this non-existence results due to a sort of "dimensionality mismatch" wherein the type space has a smaller dimension than the choice space: in this case the inverse welfare function is characterized by an overdetermined system of equations that often has no solution.[1]

We then discuss how our theory of inverse welfare functions can be extended to even more complex taxation problems: we illustrate how our theory can be adapted to settings with general equilibrium effects and to settings where agents make optimization mistakes. When there are general equilibrium effects of taxation (e.g., endogenous wages or prices), we show that an inverse welfare function can be constructed from the Gateaux derivative of government revenue *and* the Gateaux derivatives of equilibrium objects; the inverse welfare function in this case is computed as the fixed point of an integral equation. When agents make optimization mistakes we show that inverse welfare functions typically *do not* exist; however, we show that we can still recover a "generalized marginal inverse function" a la Saez and Stantcheva (2016) whenever revenue is Gateaux differentiable.

The final section of the paper illustrates four applications of inverse welfare functions. First, we recover the inverse welfare function associated with the joint income tax schedule for couples in the United States. We find that inverse welfare weights are, on average, lower for high earning couples than low earning couples, consistent with a redistributive welfare function. We also find that inverse welfare weights are higher for couples with high earning men compared to couples with high earning females; this results because taxes are a function of combined earnings yet empirical estimates of labor supply elasticities are typically larger for females than males. Second, we show how to compute inverse welfare functions in a model with both labor supply and labor demand, highlighting that general equilibrium wage effects can have substantial impacts on the inverse welfare function: ignoring implicit redistribution via general equilibrium wage effects typically vastly overestimates implicit "redistributive tastes" (i.e., the inverse welfare function that supports a given tax schedule values redistribution far less once wage effects are taken into account).

Next, we show how to use inverse welfare functions to approximate solutions for complicated optimal tax problems. As optimal taxation problems get more realistic and complex, solving these problems becomes analytically and computationally intractable; at best, the solution is governed by a highly non-linear partial differential equation and, at worst, the solution features non-smoothness (typically in the form of bunching or types with multiple optima) that make computation of solutions exceedingly difficult (Dodds (2023) or Krasikov and Golosov (2022)). We thus propose a new strategy: constrain the solution to be within some relatively simple

[1] In contrast, in settings where the type space is larger than the choice space, we argue that inverse welfare functions typically *do* exist.

class of functions (e.g., piecewise linear functions or polynomials) and then compute the inverse welfare function that rationalizes the proposed schedule as the fully non-linear optimum, continuing to solve the problem for more and more flexible function classes (e.g., piecewise linear functions with more brackets or polynomials with higher order terms) until we reach a point in which the inverse welfare function is sufficiently "close" to the true welfare function. Our third application computes inverse welfare functions for (constrained) optimal piecewise linear income tax schedules with multidimensional heterogeneity, showcasing how to compute inverse welfare functions in the presence of bunching and individuals with multiple optima. Our fourth application computes inverse welfare functions for (constrained) optimal polynomial income and housing tax schedules in a model where individuals vary in terms of three dimensions: labor income productivity, taste for housing, and curvature over utility of consumption. In both our third and fourth applications, we find that relatively simple tax schedules (i.e., piecewise linear schedules with just a few brackets or relatively low order polynomials) do a reasonable job of approximating the (assumed) true welfare functions.

The rest of the paper is organized as follows: Section 2 first presents a highly simplified model to build intuition without all of the mathematical machinery present in the rest of the paper. Section 2 then discusses notation and presents our first main result on existence and construction of inverse welfare functions. Section 3 discusses a number of example inverse welfare function constructions as well as provides a set of general sufficient conditions for Gateaux differentiability of government revenue. Section 4 discusses how our theory relates to several important results in public economics. Section 5 shows how our results can be extended to taxation problems with general equilibrium effects and optimization failures. Section 6 presents empirical applications of our theory. Section 7 concludes.

2 Notation and Construction of Inverse Optimal 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 finite choices 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):

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

Hence, as long as we pick $\phi(n)$ to satisfy Equation 1, then the tax schedule is locally extremal (note that we can normalize $\lambda = 1$ as this simply scales $\phi(n)$ multiplicatively). Thus, 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. 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 $\mathbf{z}$:

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

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/other choices 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 utility profile that arises when agents optimize under tax schedule $T(\mathbf{z})$ according to Equation 2 (i.e., the indirect utility level as a function of type $\mathbf{n}$ given $T(\mathbf{z})$). 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.[2]

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 $\Phi(\mathbf{n})$:*

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

Hence, 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 function where $\Phi(\mathbf{n})$ is a distribution that puts all weight on the lowest type $\mathbf{n}$ in society.[3]

Finally, much of our analysis will require us to take *Gateaux derivatives* and *Gateaux varia-*

[2]Note that in the introduction we abused languge 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.

[3]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: $W(U(\mathbf{n};T)) = \int_{\mathbf{N}} U(\mathbf{n};T)f(\mathbf{n})d\mathbf{n}$.

tions 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 $\tau \in \mathcal{T}$:*[4]

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

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. In the infinite dimensional case, the Gateaux derivative exists if the Gateaux variation in every direction can be written as the Gateaux derivative integrated against the direction function τ.*

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 Existence and Construction of Inverse Welfare Functionals

Our goal in this section 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 identify such an inverse welfare functional?"

First, it is helpful to recast this problem slightly: we are searching for an (infinite dimensional) supporting hyperplane for the utility profile $U(\mathbf{n}; T)$ generated when agents optimize according to a tax schedule $T(\mathbf{z})$. However, it turns out to be difficult to guarantee the existence of an infinite dimensional supporting hyperplane for a given utility profile $U(\mathbf{n}; T)$ because the set of the set of admissible utility profiles $\mathcal{U}$ is not necessarily a convex set.[5] In Appendix A.1 we prove Proposition 4 establishing existence of an infinite dimensional supporting hyperplane for a given utility profile $U(\mathbf{n}; T)$ 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

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

[5] 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 $U_3(\mathbf{n}) = \alpha U(\mathbf{n}; T_1) + (1 - \alpha)U(\mathbf{n}; T_2)$ is not necessarily consistent with individual optimization under some tax schedule.

under a welfare functional W:

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

$L(T;W)$ is simply the value of welfare plus a Lagrange multiplier λ multiplied by the value of the societal budget constraint. We will call a welfare functional $W(\cdot)$ a "local inverse welfare functional" for a given tax schedule $T(\mathbf{z})$ if $T(\mathbf{z})$ is a stationary point for the Lagrangian $L(T;W)$. More precisely:

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 in the paper, 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 $\mathbf{Z}$ 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 main 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.2 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, to express the Gateaux derivative of revenue we do often need to split up the domain $\mathbf{Z}$, e.g., into the interior and boundary):

$$\lim_{\epsilon \to 0} \frac{R(T + \epsilon\tau) - R(T)}{\epsilon} = \sum_i \int_{\mathbf{Z}_i} \tau(\mathbf{z})\gamma(\mathbf{z})h(\mathbf{z})d\mathbf{Z}_i \tag{5}$$

where $h(\mathbf{z})$ is the density of income (i.e., we assume there is no bunching) and $d\mathbf{Z}_i$ is the volume element of $\mathbf{Z}_i$. In Equation 5, $\gamma(\mathbf{z})$ should be understood as the "instantaneous budgetary effect" of an infinitesimal "bump function" 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 $c = z - T(z)$).

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 $W(U(\mathbf{n};T)) = \sum_i \int_{\mathbf{Z}_i} \int_{\mathbf{N}(\mathbf{z})} \phi(\mathbf{n})U(\mathbf{n};T)dF(\mathbf{n}|\mathbf{z})h(\mathbf{z})d\mathbf{Z}_i$.[6] Hence, our objective is to find the

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

weights $\phi(\mathbf{n})$ such that $T(\mathbf{z})$ is a stationary point of the government's Lagrangian:[7]

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

Next, we want to 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. However, the envelope theorem implies that for all $\mathbf{n}$ with a unique optima, behavioral responses to tax changes have 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 optima, then we can apply the envelope theorem to write the Gateaux derivative of the government's welfare functional as:

$$\frac{\partial W(U(\mathbf{n}; T + \epsilon\tau))}{\partial \epsilon} = -\sum_i \int_{\mathbf{Z}_i} \int_{\mathbf{N}(\mathbf{z})} \phi(\mathbf{n}) u_c(\mathbf{n}) \tau(\mathbf{z}) dF(\mathbf{n}|\mathbf{z}) h(\mathbf{z}) d\mathbf{Z}_i \tag{6}$$

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

$$-\sum_i \int_{\mathbf{Z}_i} \int_{\mathbf{N}(\mathbf{z})} \phi(\mathbf{n}) u_c(\mathbf{n}) \tau(\mathbf{z}) dF(\mathbf{n}|\mathbf{z}) h(\mathbf{z}) d\mathbf{Z}_i + \lambda \sum_i \int_{\mathbf{Z}_i} \tau(\mathbf{z}) \gamma(\mathbf{z}) h(\mathbf{z}) d\mathbf{Z}_i \tag{7}$$

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

$$\int_{\mathbf{N}(\mathbf{z})} \phi(\mathbf{n}) u_c(\mathbf{n}) dF(\mathbf{n}|\mathbf{z}) h(\mathbf{z}) = \lambda \gamma(\mathbf{z}) h(\mathbf{z}) \tag{8}$$

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 8.[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 8 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 just 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

[7] 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.

[8] Intuitively, 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.

impact of changing taxes at a given $\mathbf{z}$ that is proportional to the tax change at that $\mathbf{z}$. 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 points to discuss. First, note then 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 because for each $\mathbf{z}$ we can pick any $\phi(\mathbf{n})$ for the $\mathbf{n}$ that choose $\mathbf{z}$ so as to satisfy Equation 7. Loosely, 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 weight functional that puts the requisite amount of weight on these types collectively will support the given tax schedule. Second, the assumption that $T(\mathbf{z})$ is continuous is mostly WLOG; Lemma 3 in Appendix C 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 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. Loosely, if an individual has two optima and no other types choose those that $\mathbf{z}$, then in order to rationalize a given tax schedule, the government may need to care differentially about raising utility for this individual depending on which of the two incomes they choose. In other words, we may require different inverse welfare weights for this individual depending upon which income they choose, which is not consistent with any linear welfare functional. We discuss this point more in Section 4.4.

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. 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?

3 Examples of Inverse Welfare Functional Construction

This Section first provides 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 (which simply captures how government revenue changes with the tax schedule) is, in principle, an empirically estimable object; thus, inverse welfare functionals can be estimated with sufficient data. Finally, we provide general sufficient conditions for Gateaux differentiability in Section 3.4.

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 $u(c, z/n)$ that satisfies the single crossing property of Mirrlees (1971) so that $n \mapsto z$ is weakly increasing for all n and is strictly increasing whenever $T'(z)$ exists (Lemma 1 of Bergstrom and Dodds (2021*a*)). Suppose that we want to find an inverse welfare functional for a smooth $T(z)$ for which all individuals have a unique optima. 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. Let us first calculate the Gateaux derivative of $R(T)$ in the direction of some $\tau(z)$ (importantly, recall that $z(n)$ is also a function of the tax schedule even though we omit it as an argument for brevity):

$$\lim_{\epsilon \to 0} \frac{R(T+\epsilon\tau) - R(T)}{\epsilon} = \int_N \frac{\partial}{\partial \epsilon} \left(T(z(n)) + \epsilon\tau(z(n))\right) f(n) dn = \int_N \left(T'(z(n)) \frac{\partial z}{\partial \epsilon}(n) + \tau(z(n))\right) f(n) dn \tag{9}$$

We have the individual first order condition:

$$u_1(z - T(z) - \epsilon\tau(z), z/n)\left(1 - T'(z) - \epsilon\tau'(z)\right) + \frac{1}{n} u_2(z - T(z) - \epsilon\tau(z), z/n) = 0 \tag{10}$$

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

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

Plugging in Equation 11 into Equation 9 and changing the variable of integration from n to z (with $h(z)$ denoting the income density) we find that:[9]

$$\lim_{\epsilon \to 0} \frac{R(T+\epsilon\tau) - R(T)}{\epsilon} = \int_{\underline{z}}^{\overline{z}} \left(T'(z)\xi(z)\tau'(z) + \left[1 + T'(z)\eta(z)\right]\tau(z)\right) h(z) dz \tag{12}$$

where $\overline{z} \equiv z(\overline{n})$ and $\underline{z} \equiv z(\underline{n})$. However, Equation 12 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* function of $\tau(z)$. Using integration by parts to get rid of the $\tau'(z)$ term, Equation 12 is equal to:

$$\int_{\underline{z}}^{\overline{z}} \left(-\frac{\partial}{\partial z}\left[T'(z)\xi(z)(z)h(z)\right] + \left[1 + T'(z)\eta(z)\right]h(z)\right)\tau(z)dz + T'(z)\xi(z)h(z)\tau(z)\Big|_{\underline{z}}^{\overline{z}} \tag{13}$$

Note that all $\tau(z)$ terms enter Equation 13 linearly so that Equation 13 is a linear functional of $\tau(z)$ which means that $R(T)$ is Gateaux differentiable (assuming that all terms in Equation 13 are bounded so that the Gateaux derivative is a bounded - hence continuous - linear functional

[9] 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.

of $\tau(z)$). Thus, an inverse welfare functional exists for $T(z)$ by Theorem 1. To see this, consider the following welfare functional:

$$W(U) = \int_N \phi(n)U(n;T)f(n)dn + \overline{\phi}U(\overline{n};T)h(z(\overline{n})) + \underline{\phi}U(\underline{n};T)h(z(\underline{n})) \tag{14}$$

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 $u_c(n)\tau(z(n)) \equiv u_c(z(n)-T(z(n)), z(n)/n)\tau(z(n))$ (recall $z(n)$ represents optimal income for each type n). Hence, the Gateaux derivative of $W(U)$ from Equation 14 equals:

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

From here, we can solve for $\phi(n(z))$ by simply equating terms for all $z \in \text{Int}(Z)$ (see Equation 15) and for $z \notin \text{Int}(Z)$ (see Equations 16 and 17), noting that we can normalize the Lagrange multiplier in the government's Lagrangian to equal 1, which simply scales the inverse welfare functional multiplicatively:

$$\phi(n(z))u_c(n(z))h(z) = -\frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right] + \left[1 + T'(z)\eta(z)\right]h(z) \tag{15}$$

$$\overline{\phi}u_c(\overline{n})h(\overline{z}) = T'(\overline{z})\xi(\overline{z})h(\overline{z}) \tag{16}$$

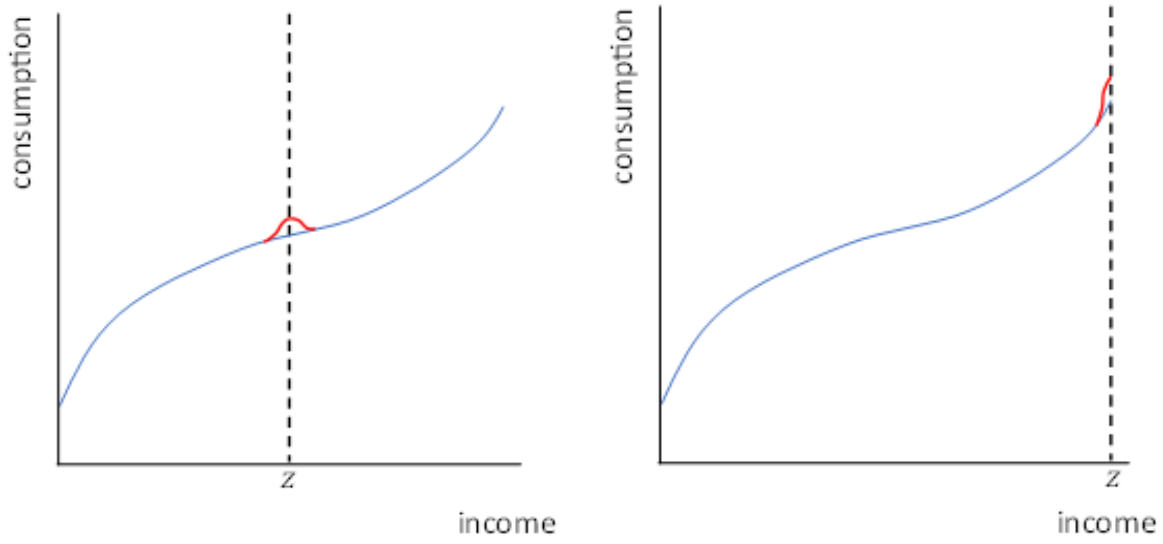
$$\underline{\phi}u_c(\underline{n})h(\underline{z}) = -T'(\underline{z})\xi(\underline{z})h(\underline{z}) \tag{17}$$

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. Note that this welfare functional has additional terms which allow the government to care a *discrete* amount about the welfare of the top and bottom individuals. This allows us to rationalize non-zero tax rates at the top and bottom, a point we discuss further in Subsection 4.1.

Intuitively, Equations 15, 16, and 17 ensure that the total impact on the government's Lagrangian of every possible "bump" perturbation is zero. Equation 15 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 15. Moreover, an interior bump perturbation leads to a mechanical budgetary impact along with an income effect, $[1 + T'(z)\eta(z)]$, 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, $-\frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right]$.[10] Equation 16

[10]Note that Equation 15 is just a differentiated version of Equation (19) from Saez (2001).

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 16. Identical logic explains the intuition behind Equation 17.



(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.

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 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 $\forall v$. 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

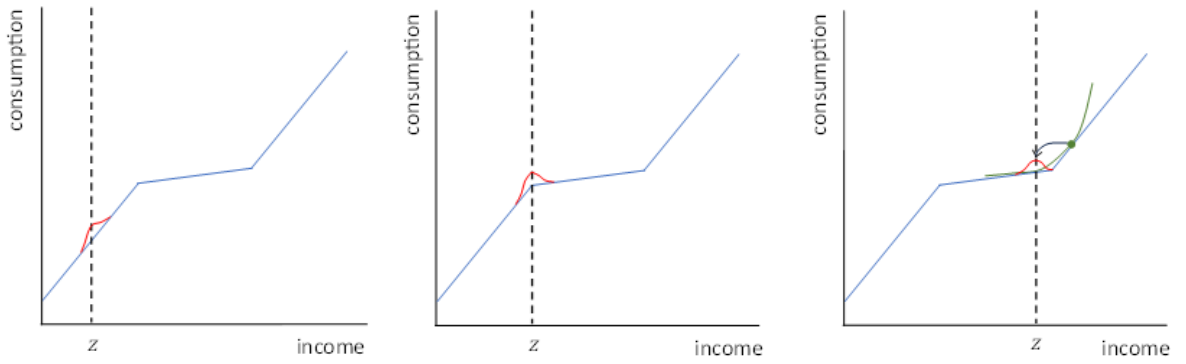
schedule is the optimal *non-linear* tax schedule.

There are two additional complexities relative to Section 3.1. 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 optima in the second tax bracket and one in the third tax bracket, see Figure 2c) around the second kink, K_2. For all individuals with a unique optima 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 v, we show in Appendix A.3 that the Gateaux derivative of $R(T)$ with three tax brackets is given by:

$$\begin{aligned}
&\lim_{\epsilon\to 0}\frac{R(T+\epsilon\tau)-R(T)}{\epsilon} = \\
&\underbrace{\int_{Z_1}\left(-\frac{\partial}{\partial z}\left[T_1\overline{\xi}(z)h(z)\right]+\left[1+T_1\overline{\eta}(z)\right]h(z)\right)\tau(z)dz}_{\text{Perturbations in First Bracket}} \\
&+\underbrace{T_1\overline{\xi}(K_1^-)h(K_1^-)\tau(K_1)+M(K_1)\tau(K_1)-T_2\overline{\xi}(K_1^+)h(K_1^+)\tau(K_1)}_{\text{Perturbation at Kink}} \\
&+\underbrace{\int_{Z_2}\left(-\frac{\partial}{\partial z}\left[T_2\overline{\xi}(z)h(z)\right]+\left[1+T_2\overline{\eta}(z)\right]h(z)-J_2(z)\right)\tau(z)dz}_{\text{Perturbations in Second Bracket}} \\
&+\underbrace{\int_{Z_3}\left(-\frac{\partial}{\partial z}\left[T_3\overline{\xi}(z)h(z)\right]+\left[1+T_3\overline{\eta}(z)\right]h(z)+J_3(z)\right)\tau(z)dz}_{\text{Perturbations in Third Bracket}}
\end{aligned} \tag{18}$$

where Z_1, Z_2, Z_3 represent the sets of incomes in the first bracket, second bracket, and third bracket, respectively, and $J_2(z)$ and $J_3(z)$ capture the budgetary impacts of individuals with multiple optima "jumping" (Bergstrom and Dodds, 2021*a*) in response to tax perurbations in the second and third brackets, respectively (these terms are defined in Appendix A.3). Importantly, Equation 18 is linear in $\tau(z)$ so that $R(T)$ is Gateaux differentiable (again assuming that all the terms in Equation 18 are bounded).

The intuition behind Equation 18 is fairly straight-forward. Essentially, Equation 18 collects the impacts of an infinite number of infinitesimal "bump" perturbations (as discussed in Section 3.1) on government revenue. Because there is bunching and there are individuals with multiple optima, the impacts of these bump perturbations are more complex, 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



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

Figure 2: Bump Function Perturbations for Piece-Wise Linear Schedule

Note: This figure shows three different "bump function" perturbations to a piece-wise 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.

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 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, if we consider a bump perturbation at an income where someone has multiple optima (as in Figure 2c), then there is 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 18.[11]

How can we use Equation 18 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 almost all types locating at each z with $h(z) > 0$ have a unique optima):

$$-\int_Z \iint_{N\times V} \phi(n,v)u_c(n,v)dF(n,v|z)\tau(z)dH(z) + \lambda \lim_{\epsilon\to 0} \frac{R(T+\epsilon\tau) - R(T)}{\epsilon} \tag{19}$$

From here, we note that if $\phi(n,v)$ are inverse welfare weights, then Equation 19 equals 0 for any

[11] Note, Equation 18 is a differentiated version of the budgetary effects from Equation (43) in Bergstrom and Dodds (2021*a*) that also allows for non-differentiable tax schedules. Also note that Equation 18 does not feature any "boundary terms" as in Equation 13; this is due to the monotonicity assumptions of $z(n,v)$ in v which ensure that the income density is zero at the top and bottom incomes.

$\tau(z)$. We can find such a set of weights by plugging in the Gateaux derivative from Equation 18 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:

$$\iint_{N\times V} \phi(n,v)u_c(n,v)dF(n,v|z)h(z) = -\frac{\partial}{\partial z}\left[T_1\overline{\xi}(z)h(z)\right] + \left[1+T_1\overline{\eta}(z)\right]h(z) \tag{20}$$

Or at the kink K_1:

$$\iint_{N\times V} \phi(n,v)u_c(n,v)dF(n,v|K_1)M(K_1) = T_1\overline{\xi}(K_1^-)h(K_1^-) + M(K_1) - T_2\overline{\xi}(K_1^+)h(K_1^+) \tag{21}$$

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

$$\iint_{N\times V} \phi(n,v)u_c(n,v)dF(n,v|z)h(z) = -\frac{\partial}{\partial z}\left[T_2\overline{\xi}(z)h(z)\right] + \left[1+T_2\overline{\eta}(z)\right]h(z) - J_2(z) \tag{22}$$

Equations 20, 21, and 22 ensure the 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 this given z, there are many choices of weights that satisfy the above equation: any of them will be an inverse optimal welfare functional.[12] 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 individuals solve Equation 2. We assume the tax schedule is smooth and all individuals move smoothly in response to tax perturbations. Consider individual choices under the perturbed tax schedule $T(\mathbf{z}) + \epsilon\tau(\mathbf{z})$. Because the tax schedule is assumed smooth, then individual choices over $\mathbf{z}$ satisfy the following first order conditions where J represents the number of choice variables:

$$\begin{aligned} &u_c(y(\mathbf{z}) - T(\mathbf{z}) - \epsilon\tau(\mathbf{z}), \mathbf{z}; \mathbf{n})\left(y_{z_1}(\mathbf{z}) - T_{z_1}(\mathbf{z}) - \epsilon\tau_{z_1}(\mathbf{z})\right) + u_{z_1}(y(\mathbf{z}) - T(\mathbf{z}) - \epsilon\tau(\mathbf{z}), \mathbf{z}; \mathbf{n}) = 0 \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ &u_c(y(\mathbf{z}) - T(\mathbf{z}) - \epsilon\tau(\mathbf{z}), \mathbf{z}; \mathbf{n})\left(y_{z_J}(\mathbf{z}) - T_{z_J}(\mathbf{z}) - \epsilon\tau_{z_J}(\mathbf{z})\right) + u_{z_J}(y(\mathbf{z}) - T(\mathbf{z}) - \epsilon\tau(\mathbf{z}), \mathbf{z}; \mathbf{n}) = 0 \end{aligned} \tag{23}$$

If 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 Gateaux variation of $\mathbf{z}(\mathbf{n})$ for each $\mathbf{n}$ in the direction of a given $\tau(\mathbf{z})$ via the implicit function theorem (importantly, recall that $\mathbf{z}(\mathbf{n})$ is also a function of the tax schedule even though we omit it as an argument

[12] 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 would be equal to $\frac{-\frac{\partial}{\partial z}\left[T_1\overline{\xi}(z)h(z)\right]+\left[1+T_1\overline{\eta}(z)\right]h(z)}{\iint_{N\times V} u_c(n,v)dF(n,v|z)h(z)}$. Weights at the kink, second, and third brackets would be defined analogously using Equations 21 and 22.

for brevity):

$$\begin{aligned}\frac{\partial \mathbf{z}}{\partial \epsilon}(\mathbf{n}) &= \mathbf{H}^{-1}(\mathbf{n}) FOC(\mathbf{n})_{\epsilon}|_{\epsilon=0} = \mathbf{H}^{-1}(\mathbf{n})[\mathbf{a}(\mathbf{n})\tau(\mathbf{z}) + \mathbf{B}(\mathbf{n}) \cdot \nabla_{\mathbf{z}}\tau(\mathbf{z})] \\ &\equiv \vec{\eta}(\mathbf{n})\tau(\mathbf{z}) + \mathbf{X}(\mathbf{n}) \cdot \nabla_{\mathbf{z}}\tau(\mathbf{z})\end{aligned} \tag{24}$$

where $FOC_{\epsilon}|_{\epsilon=0}$ is the vector of derivatives of the first order conditions 23 with respect to ϵ. The second equality in Equation 24 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 24 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).

The government's Lagrangian is:

$$\mathcal{L}(T; W) = W(U(\mathbf{n}; T)) + \lambda \int_{\mathbf{N}} [T(\mathbf{z}(\mathbf{n})) - E] dF(\mathbf{n})$$

The Gateaux derivative of the government's Lagrangian is therefore:

$$\begin{aligned}&W(-u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n}))) + \lambda \int_{\mathbf{N}} \left(\tau(\mathbf{z}) + \nabla_{\mathbf{z}} T(\mathbf{z}(\mathbf{n}))\left[\vec{\eta}(\mathbf{n})\tau(\mathbf{z}) + \mathbf{X}(\mathbf{n}) \cdot \nabla_{\mathbf{z}}\tau(\mathbf{z})\right]\right) dF(\mathbf{n}) \\ &= W(-u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n}))) + \lambda \int_{\mathbf{Z}} \int_{\mathbf{N}(\mathbf{z})} \left(\tau(\mathbf{z}) + \nabla_{\mathbf{z}} T(\mathbf{z})\left[\vec{\eta}(\mathbf{n})\tau(\mathbf{z}) + \mathbf{X}(\mathbf{n}) \cdot \nabla_{\mathbf{z}}\tau(\mathbf{z})\right]\right) dF(\mathbf{n}|\mathbf{z}) dH(\mathbf{z}) \\ &= W(-u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n}))) + \lambda \int_{\mathbf{Z}} \left(\tau(\mathbf{z}) + \nabla_{\mathbf{z}} T(\mathbf{z})\left[\overline{\vec{\eta}}(\mathbf{z})\tau(\mathbf{z}) + \overline{\mathbf{X}}(\mathbf{z}) \cdot \nabla_{\mathbf{z}}\tau(\mathbf{z})\right]\right) dH(\mathbf{z})\end{aligned} \tag{25}$$

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 25 first integrates the budget constraint over each $\mathbf{n}$ that chooses 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 25 by appealing to multi-dimensional integration by parts to get rid of the derivatives of $\tau(\mathbf{z})$:

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.

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 $\mathbf{Z}$):[13]

$$\begin{aligned} & W(-u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n}))) + \lambda \int_{\mathbf{Z}} \left(\tau(\mathbf{z}) + \nabla_{\mathbf{z}} T(\mathbf{z}) \left[\overline{\vec{\eta}}(\mathbf{z})\tau(\mathbf{z}) + \overline{\mathbf{X}}(\mathbf{z}) \cdot \nabla_{\mathbf{z}}\tau(\mathbf{z})\right]\right) h(\mathbf{z}) d\mathbf{z} \\ & = W(-u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n}))) + \lambda \int_{\mathbf{Z}} \left(\left[1 + \nabla_{\mathbf{z}} T(\mathbf{z})\overline{\vec{\eta}}(\mathbf{z})\right] h(\mathbf{z}) - \nabla_{\mathbf{z}} \cdot \left[\nabla_{\mathbf{z}} T(\mathbf{z})\overline{\mathbf{X}}(\mathbf{z})h(\mathbf{z})\right]\right) \tau(\mathbf{z}) d\mathbf{z} \\ & + \lambda \int_{\partial \mathbf{Z}} \nabla_{\mathbf{z}} T(\mathbf{z})\overline{\mathbf{X}}(\mathbf{z})\tau(\mathbf{z})h(\mathbf{z}) \cdot \rho dS \end{aligned} \tag{26}$$

Importantly, note that Equation 26 is *linear* in $\tau(\mathbf{z})$. Thus, if behavioral responses are sufficiently smooth, then revenue is Gateaux differentiable in the tax schedule (again assuming all terms in Equation 26 are bounded so that the Gateaux derivative is a bounded - hence, continuous - linear functional). To construct an inverse welfare functional, suppose that the welfare functional takes the following form:

$$W(U) = \int_{\mathbf{Z}} \int_{\mathbf{N}(\mathbf{z})} \phi(\mathbf{n}) U(\mathbf{n}; T) dF(\mathbf{n}|\mathbf{z}) h(\mathbf{z}) d\mathbf{z} + \int_{\partial \mathbf{Z}} \int_{\mathbf{N}(\mathbf{z})} \phi(\mathbf{n}) U(\mathbf{n}; T) dF(\mathbf{n}|\mathbf{z}) h(\mathbf{z}) dS \tag{27}$$

Then we can write the Gateaux derivative of the inverse optimal welfare functional as follows:

$$-\int_{\mathbf{Z}} \int_{\mathbf{N}(\mathbf{z})} \phi(\mathbf{n}) u_c(\mathbf{n}) \tau(\mathbf{z}) dF(\mathbf{n}|\mathbf{z}) h(\mathbf{z}) d\mathbf{z} - \int_{\partial \mathbf{Z}} \int_{\mathbf{N}(\mathbf{z})} \phi(\mathbf{n}) u_c(\mathbf{n}) dF(\mathbf{n}|\mathbf{z}) \tau(\mathbf{z}) h(\mathbf{z}) dS \tag{28}$$

To satisfy Equation 26, we choose $\phi(\mathbf{n})$ for those locating at each $\mathbf{z} \in \text{Int}(\mathbf{Z})$ such that:

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

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 26 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 observables $\mathbf{z}$. In principle, all of these objects can be estimated given sufficient tax variation. In practice, it is difficult to estimate heterogeneous behavioral responses to different tax reforms; hence, practitioners typically will need to make simplifying assumptions such as assuming elasticities are constant (thereby reducing empirical requirements to a more manageable set of sufficient statistics), or making structural assumptions on utility and calibrating. We employ both of these approaches in applications in Section 6.

[13] Note, to apply Lemma 1 we also require that the set $\mathbf{Z}$ is an open set in the ambient space $\mathbb{R}^J$ so that the chosen set of $\mathbf{Z}$ has non-empty interior in $\mathbb{R}^J$. This requires that that $\dim(\mathbf{N}) \geq \dim(\mathbf{Z})$. 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 $\int_{\mathbf{Z}} [\nabla_{\mathbf{z}} \cdot \mathbf{v}(\mathbf{z})]\tau(\mathbf{z}) d\mathbf{z}$.

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 the context of unidimensional piecewise linear taxation. Section 3.3 shows that $R(T)$ can be Gateaux differentiable even if the tax schedule is a function of multiple choices and features multidimensional heterogeneity. 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 A.4. □

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 (which also allows for individuals responding on the extensive margin by entering/exiting the workforce, see Appendix B.1 for further discussion). Even in spite of these pathologies, government revenue can still be Gateaux differentiable in the tax schedule.

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 ensure that at most $\mathbf{z}$, the tax schedule is smooth and individuals have a unique optima 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.[14] There may be surfaces where individuals have multiple optima and thereby react to some tax changes by jumping to a different income level; 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 curve that creates a ridge in three dimensional space. Almost all individuals who choose $\mathbf{z}$ on this curve strictly prefer their chosen $\mathbf{z}$ to any $\mathbf{z}$ that is off the curve (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 curve but do not move off the curve. For small tax perturbations we can then recast the optimization problem for individuals on the curve as a choice over some parameter t which parameterizes the curve. Thus, for these individuals, we have reduced the problem to a unidimensional optimization problem wherein we can use integration by parts (over the curve) to express the revenue impact of a tax change as a linear functional just as in Section 3.1.[15]

As a result of Proposition 1, we believe that Gateaux differentiability of government revenue is a relatively mild restriction. While one could come up with examples of tax schedules which are not differentiable on positive measure sets of choices or that feature large fractions of the population with multiple optima, these cases are presumably not overly realistic and hence not practically relevant.

4 Relationship to Previous Optimal Taxation Results

Next, we discuss how our theory of inverse welfare functionals relates to a number of important existing theoretical results in public economics.

4.1 Boundary Tax Rates and the Zero Top/Bottom Result

Welfare functionals used in the optimal taxation literature are typically comprised of a set of welfare weights, such as:

$$W(U(\mathbf{n};T)) = \int_{\mathbf{N}} \phi(\mathbf{n})U(\mathbf{n};T)f(\mathbf{n})d\mathbf{n} \tag{31}$$

Suppose welfare weights are given by Equation 31 and the Assumptions in Section 3.3 hold. In order for a tax schedule to be locally optimal (i.e., all tax perturbations yields a zero effect on

[14] Note, to apply Lemma 1, we require that the set of choices $\mathbf{Z}$ is open (or the closure of an open set) in $\mathbb{R}^J$. For instance, if there are two choice variables, then the set $\mathbf{Z}$ must have positive area in $\mathbb{R}^2$.

[15] 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 which is open (or the closure of an open set) in the lower dimensional space.

the Lagrangian), Equation 26 mandates that we must have $\nabla_{\mathbf{z}} T(\mathbf{z}) \overline{\mathbf{X}}(\mathbf{z}) h(\mathbf{z}) \cdot \rho = 0$ along the boundary $\partial \mathbf{Z}$. This so-called "natural boundary condition" mandates that substitution effects are zero in the direction normal to the boundary: in the classic Mirrleesian unidimensional case, this condition ensures that marginal tax rates are zero at the top and bottom of the income distribution. However, Sections 3.1 and 3.3 show that we *can* rationalize non-zero top/bottom tax rates (or more, generally, tax schedules that do not satisfy the natural boundary condition) with a continuous, linear inverse welfare functional that puts positive "mass" on the boundary types (so that, in a sense, we care about each boundary individual more than an interior individual). More generally, as we will see later on, the inverse optimal functional may also require the government to put positive weight on measure zero sets of individuals that have multiple optima or where the tax schedule is non-differentiable.

4.2 Pareto Efficiency

There has been a good amount of recent work characterizing Pareto efficient tax schedules (e.g., Werning (2007) or Scheuer and Werning (2016)); hence, it seems prudent to comment on how inverse welfare functionals and the Gateaux derivative of government revenue can be used to characterize Pareto efficiency in multidimensional taxation settings.

Proposition 2. *The following necessary condition 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 with $W(U) > 0$ when $U(\mathbf{n}) > 0$ $\forall \mathbf{n}$.*
2. *$T(\mathbf{z})$ is Pareto efficient if the (global) inverse welfare functional for $T(\mathbf{z})$ 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.5. □

Proposition 2 follows mostly from the definition of Pareto efficiency: 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.

The most important point is that Pareto efficient schedules are associated with positive Gateaux derivatives of government revenue and positive local inverse welfare functionals; hence, 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. Thus, existence of a local inverse functional is *weaker* than Pareto efficiency. Moreover, there is an argument to made that negative welfare weights are reasonable: for instance, if one has an inherent distaste for the existence of billionaires, one may place negative welfare weights on these individuals. This violation of the Pareto principle is often ruled out ex ante in economic analysis but can be accommodated in the context of local inverse welfare functionals.

4.3 Relationship to the Atkinson-Stiglitz Theorem

The Atkinson-Stiglitz Theorem (Atkinson and Stiglitz, 1976) 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). Given the discussion in Subsection 4.2, this begs the question: "Are there (non-positive) inverse welfare functionals that support multidimensional tax schedules that feature non-zero taxes on choice variables other than income in the Atkinson-Stiglitz world?" In other words, if we allow negative and/or discrete welfare weights, can any multidimensional tax schedule be supported by some welfare weight functional?

Perhaps surprisingly, the answer turns out to be no: we can identify many such tax schedules that are not supported by *any* linear welfare functional in the Atkinson-Stiglitz environment, regardless of whether weak separability holds. Fundamentally, this ensues due to a "dimensionality mismatch" which results because the choice space is a higher dimensional space than the type space. As a result of this dimensionality mismatch, the system of equations that pins down the inverse welfare functional is typically overdetermined and hence does not have a solution. More fundamentally, when the choice space is a higher dimension than the type space, government revenue typically fails to be Gateaux differentiable because we cannot apply integration by parts: recall that multidimensional integration by parts, Lemma 1, requires 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). 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 $\mathbb{R}^2$. As a result, government revenue is generally *not* Gateaux differentiable, leading to non-existence of an inverse welfare

functional.

To see this, we work through a simple example in the context of joint savings and income taxation. We assume households work in the first period, save some amount at interest rate r that is consumed in the next period, and then consume the rest today. 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 $s = (1 + r)x$).[16] 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 individual 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, we 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:

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

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

for some functions $\eta_z, \xi^z_z, \xi^z_s, \eta_s, \xi^s_z, \xi^s_s$. Note that η_i represents the income effect for variable i and ξ^j_i represents the substitution effect of variable j with respect to the marginal tax rate on variable i.

Next, we 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 2 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):

Lemma 2. *Under the smoothness and regularity assumptions discussed in Section 4.3, inverse welfare weights $\phi(n)$ must satisfy:*

$$\phi(n(z)) = \frac{(1 + T_z(z, s(z))\eta_z(z) + T_s(z, s(z))\eta_s(z))\, h(z) - \frac{\partial}{\partial z}\left([T_z(z, s(z))\xi^z_z(z) + T_s(z, s(z))\xi^s_z(z)]\, h(z)\right)}{u_c(n(z))} \tag{32}$$

$$\phi(n(z)) = \frac{(1 + T_z(z, s(z))\eta_z(z) + T_s(z, s(z))\eta_s(z))\, h(z) - \frac{\partial}{\partial z}\left([T_z(z, s(z))\xi^z_s(z) + T_s(z, s(z))\xi^s_s(z)]\left(\frac{ds}{dz}\right)^{-1} h(z)\right)}{u_c(n(z))} \tag{33}$$

[16] 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.

Proof. See Appendix A.6. □

Lemma 2 essentially says that for a set of inverse welfare weights to ensure that an arbitrary income tax perturbation leaves the government Lagrangian unchanged, Equation 32 must be satisfied. Similarly, to ensure that an arbitrary savings tax perturbation leaves the government Lagrangian unchanged, Equation 33 must be satisfied. The key point is that $\phi(n)$ is overdetermined and that Equation 32 often does not equal Equation 33 at all z. When Equation 32 does not equal Equation 33, 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):[17]

Proposition 3. *There are tax schedules $T(z, s)$ for which Equation 32 does not equal Equation 33 so that no inverse welfare functional exists.*

We prove Proposition 3 simply by providing a number of examples in Figure 3 of the inverse welfare weights that satisfy Equation 32 along with the inverse welfare weights that satisfy Equation 33. In all but the top left panel of Figure 3, the given tax schedules do not have any inverse welfare functional because the inverse weights that satisfy Equation 32 are different than the inverse weights that satisfy Equation 33. In general, Equation 32 equals Equation 33 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 9 in Appendix D shows similar findings for a non-separable utility function as well as for non-linear tax schedules).

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 32 *will* equal Equation 33; this is why the associated inverse welfare weights satisfying Equation 32 and Equation 33 in the top left panel of Figure 3 *do* coincide. To see this more generally, note that because $T_s = 0$, to show Equation 32 equals Equation 33 under weak separability it suffices to show that:

$$\xi_z^z(z) = \xi_s^z(z) \left(\frac{ds}{dz}\right)^{-1} \tag{34}$$

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

[17] Technically, the Gateaux variation of government revenue in the direction of an arbitrary $\tau(z)$ is not the same linear functional as the Gateaux variation of government revenue in the direction of an arbitrary $\tau(s)$; hence, revenue is not, in general, Gateaux differentiable as there is no continuous linear functional that represents the revenue effects of every possible perturbation as required in Definition 2.

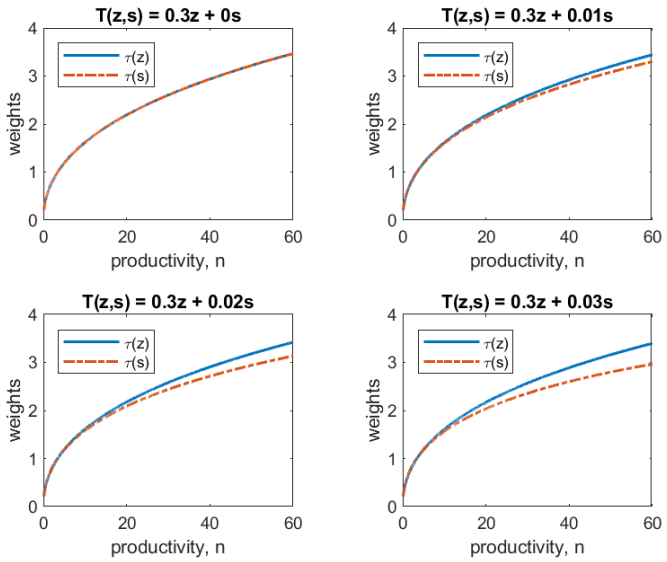


Figure 3: Inverse Weights for $\tau(z)$ and $\tau(s)$ Perturbations

Note: This figure shows the inverse welfare weights that satisfy Equation 32 in blue solid lines (i.e., ensure that the Gateaux variation of any income tax perturbations $\tau(z)$ is zero) and shows the inverse welfare weights that satisfy Equation 33 in orange dashed lines (i.e., ensure that the Gateaux variation of any savings tax perturbations $\tau(s)$ 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 0.01% (0.02%, 0.03%, respectively) savings tax is equivalent to a 20% (40%, 60%, respectively) tax on interest income.

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 34 holds under weak separability via Lemma 1 of Ferey, Lockwood and Taubinsky (2021). Hence:

Remark 4. *Under the assumptions listed at the start of this section and if utility is weakly separable in (c,s) and z, any $T(z)$ satisfying the budget constraint strictly yields a Gateaux differentiable $R(T)$.*[18] *By Theorem 1, any such schedule therefore has an inverse welfare functional that rationalizes this schedule (locally) within all tax schedules $T(z,s)$.*

Summing up, the non-existence result of Proposition 3 can 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 (Kaplow,

[18] 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 34 that, under weak separability and any $T(z)$, revenue is in fact Gateaux differentiable with Gateaux derivative:

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

2006) but are actually not supported by *any* inverse welfare functionals, even those that allow for negative and/or discrete weights; moreover, this result holds with or without weak separability. Fundamentally, this non-existence occurs because government revenue often fails to be Gateaux differentiable when the type space is a lower dimension than the choice space, resulting in an overdetermined system of equations characterizing the inverse welfare functional.

4.4 A Note on Dimensionality

Finally, it is useful to quickly discuss more broadly how the dimension of the spaces $\mathbf{N}$ and $\mathbf{Z}$ relate to Theorem 1 and Proposition 1. First, if $\dim(\mathbf{N}) < \dim(\mathbf{Z})$ then the chosen set of $\mathbf{z}$'s will not typically be an open set (or the closure of an open set) in the ambient space so that we cannot apply Proposition 1. This will often lead to non-existence of an inverse social welfare functional as discussed in Section 4.3 in the context of the Atkinson-Stiglitz Theorem. If $\dim(\mathbf{N}) = \dim(\mathbf{Z})$, then Theorem 1 shows that there exists an inverse welfare functional as long as all individuals have a unique optima and the conditions for Proposition 1 hold. If all individuals *do not* have a unique optima then even if the conditions for Proposition 1 hold so that revenue is Gateaux differentiable, an inverse welfare functional will typically not exist. The intuition is that if an individual has two optima, z^- and z^+, the weight on this type that ensures any local change to taxes around z^- leaves the Lagrangian unchanged is not the same weight that ensures any local change to taxes around z^+ leaves the Lagrangian unchanged; because there can only be one welfare weight on this type, no matter what welfare weight we choose we can therefore always improve welfare by changing taxes at either z^- or z^+.[19] On the other hand, when $\dim(\mathbf{N}) > \dim(\mathbf{Z})$, if the conditions of Proposition 1 hold, then we often *can* come up with an inverse welfare functional even if some individuals have multiple optima. The key idea is that the Gateaux derivative of government revenue pins down the (marginal utility of consumption weighted) *average* inverse welfare weight at each choice level $\mathbf{z}$ (e.g., Equation 8). Hence, as long as some types $\mathbf{n}$ at each $\mathbf{z}$ have a unique optima, we can construct an inverse welfare functional that renders a given tax schedule locally extremal by setting welfare weights to zero for individuals with multiple optima and augmenting the welfare weights on those with a unique optima in order to satisfy the requisite average inverse welfare weight at each $\mathbf{z}$. 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.

[19] We work through a numerical example of this in Appendix B.2.

5 More Complex Taxation Problems: General Equilibrium and Optimization Failures

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, there are at least two key restrictions that we have made: (1) we have only considered a "partial equilibrium" setting in which individual's decisions $\mathbf{z}$ do not impact the economy more broadly and (2) we assumed that individuals maximize their utility correctly. Extending the analysis to settings that allow for these issues is nonetheless possible. In this Section, we will first show how to compute inverse welfare functionals when there are general equilibrium effects by augmenting Theorem 1. We will state and prove this result in Theorem 2 in Subsection 5.2. However, the statement and proof of Theorem 2 are somewhat complex and technical; thus, to build intuition we will first illustrate the impact of general equilibrium effects on inverse welfare functionals via a simple labor demand/labor supply model with endogenous wages. This example contains all of the key intuition of the more general result and is substantially simpler. Finally, we will show how to recover a "generalized marginal inverse functional" a la Saez and Stantcheva (2016) when individuals do not correctly optimize utility.

5.1 Example: Labor Demand

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 $z = wnl$ 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:

$$\begin{aligned} U(n;T,w) = \max_{z} \; & c - \frac{[z/(nw)]^{1+k}}{1+k} \\ \text{s.t. } & c = z - T(z) + s(n)\pi(w) \end{aligned} \tag{35}$$

where c is again numeraire consumption, $\pi(w)$ represents firm profits, 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, $L = \int_{\mathbf{N}} nL(n)dF(n)$ where $L(n)$ is the hired labor from type n. Thus, firm profits are given by:

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

where $Y(L)$ is the firm's production function. Market clearing requires that:[20]

$$L = \int_N nL(n)dF(n) = \int_N nl(n)dF(n) \tag{36}$$

[20] Recognize that $L(n)$ and $l(n)$ both depend on the wage, w. $l(n)$ also depends on the tax schedule, $T(\mathbf{z})$.

The firm first order condition is given by:

$$Y'(L) - w = 0$$

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

$$W(U(n;T,w)) + \lambda \left[\int_N T(z(n)) dF(n) - E \right] \tag{37}$$

Now, let us take the Gateaux variation of Equation 37 in the direction of $\tau(z)$ (i.e., 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 $\frac{\partial w}{\partial \epsilon}$ 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):

$$\begin{aligned} &W\left(-\tau(z(n)) + \left(\frac{z(n)}{nw}\right)^{1+k} \frac{1}{w}\frac{\partial w}{\partial \epsilon} + s(n)\pi'(w)\frac{\partial w}{\partial \epsilon} \right) \\ &+ \lambda \int_N \left(\tau(z) + T'(z(n)) \frac{\partial z(n)}{\partial \epsilon}\bigg|_w + T'(z(n)) \frac{\partial z(n)}{\partial w}\bigg|_\epsilon \frac{\partial w}{\partial \epsilon} \right) dF(n) \end{aligned} \tag{38}$$

Next, we need to determine how to express $\frac{\partial w}{\partial \epsilon}$ in terms of $\tau(z)$. Multiplying Equation 36 by w and implicitly differentiating Equation 36 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 $wnl(n) = z(n)$):

$$\frac{\partial w}{\partial \epsilon} L + w \frac{\partial L}{\partial w}\frac{\partial w}{\partial \epsilon} - \int_{\mathbf{N}} \left(\frac{\partial z(n)}{\partial w}\bigg|_\epsilon \frac{\partial w}{\partial \epsilon} + \frac{\partial z(n)}{\partial \epsilon}\bigg|_w \right) dF(n) = 0 \tag{39}$$

We can recover $\frac{\partial z(n)}{\partial w}|_\epsilon$ by implicitly differentiating the individual first order condition with respect to w. 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 $\xi(n)$. The firm's first order condition yields $\frac{\partial L}{\partial w} = \frac{\partial Y'^{-1}(w)}{\partial w}$. Hence, we have all the components needed to determine $\frac{\partial w}{\partial \epsilon}$ from Equation 39. Applying integration by parts to the term involving $\frac{\partial z(n)}{\partial \epsilon}|_w$ in Equation 39, we find that $\frac{\partial w}{\partial \epsilon}$ is a linear functional of $\tau(z)$ (i.e., w is Gateaux differentiable in T). If $h(z) = 0$ at the top and bottom of the income distribution, then we can express $\frac{\partial w}{\partial \epsilon} = \int_Z p(z)\tau(z)dz$ for some function $p(z)$ (details provided in Appendix A.7).

Next, let us consider the budgetary impact. Again using a change of variables and integration

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

$$\begin{aligned}
&\int_N \left(\tau(z) + T'(z(n))\xi(n)\tau'(z(n)) + T'(z(n))\frac{\partial z(n)}{\partial w}\bigg|_\epsilon \frac{\partial w}{\partial \epsilon}\right) dF(n) \\
&= \int_Z \left(h(z) - \frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right]\right)\tau(z)dz + \int_N \left(T'(z(n))\frac{\partial z(n)}{\partial w}\bigg|_\epsilon \frac{\partial w}{\partial \epsilon}\right) dF(n) \\
&= \int_Z \left(h(z) - \frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right]\right)\tau(z)dz + \int_Z \int_N \left(T'(z(n))\frac{\partial z(n)}{\partial w}\bigg|_\epsilon\right) dF(n)p(z)\tau(z)dz \\
&= \int_Z \left(h(z) - \frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right] + p(z)\int_N \left(T'(z(n))\frac{\partial z(n)}{\partial w}\bigg|_\epsilon\right) dF(n)\right)\tau(z)dz
\end{aligned} \tag{40}$$

Intuitively, Equation 40 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. Using similar logic, let us assume that $W(U(n;T)) = \int_N \phi(n)U(n;T)dF(n)$ and then do a change of variables from n to z (recalling $n \mapsto z$ was assumed bijective and differentiable), finding that the Gateaux derivative of welfare is given by:

$$\begin{aligned}
&W\left(-\tau(z(n)) + \left(\frac{z(n)}{nw}\right)^{1+k}\frac{1}{w}\frac{\partial w}{\partial \epsilon} + s(n)\pi'(w)\frac{\partial w}{\partial \epsilon}\right) \\
&= -\int_Z \phi(n(z))\tau(z)h(z)dz + \int_N \phi(n)\left[\left(\frac{z(n)}{nw}\right)^{1+k}\frac{1}{w} + s(n)\pi'(w)\right]\frac{\partial w}{\partial \epsilon}f(n)dn \\
&= -\int_Z \phi(n(z))\tau(z)h(z)dz + \int_Z p(z)\tau(z)\left(\int_N \phi(n)\left[\left(\frac{z(n)}{nw}\right)^{1+k}\frac{1}{w} + s(n)\pi'(w)\right]f(n)dn\right)dz \\
&= -\int_Z \left[\phi(n(z))h(z) - p(z)\left(\int_Z \phi(n(\tilde{z}))\left[\left(\frac{\tilde{z}}{n(\tilde{z})w}\right)^{1+k}\frac{1}{w} + s(n(\tilde{z}))\pi'(w)\right]h(\tilde{z})d\tilde{z}\right)\right]\tau(z)dz
\end{aligned} \tag{41}$$

Intuitively, Equation 41 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. So, a local inverse welfare functional in this example is a set of weights $\phi(n)$ such that the budgetary effect exactly offsets the welfare impact; hence, Equation 40 plus Equation 41 equals zero. As in Section 3, let us simply match up Equations 40 and 41 pointwise, yielding:

$$\begin{aligned}
&\phi(n(z))h(z) - p(z)\left(\int_Z \phi(n(\tilde{z}))\left[\left(\frac{\tilde{z}}{n(\tilde{z})w}\right)^{1+k}\frac{1}{w} + s(n(\tilde{z}))\pi'(w)\right]h(\tilde{z})d\tilde{z}\right) \\
&= h(z) - \frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right] + p(z)\int_N \left(T'(z(n))\frac{\partial z(n)}{\partial w}\bigg|_\epsilon\right)dF(n)
\end{aligned} \tag{42}$$

Note, Equation 42 is more complex to solve for weights than Equation 15 in the analogous partial equilibrium case with a smooth bijective relationship between n and z. For a given z, Equation 15 is linear in $\phi(n(z))$ whereas Equation 42 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 42 can be expressed as:

$$\phi(n(z)) = \chi(z) + \int_Z K(z,\tilde{z})\phi(n(\tilde{z}))d\tilde{z} \tag{43}$$

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 43 (i.e., a fixed point of the contraction mapping) follows immediately from the contraction mapping theorem.[21] Economically, the condition that $\int_Z K(z,\tilde{z})d\tilde{z} < 1\ \forall z$ ensures that the welfare gain from the direct impact of a change in taxes, $\int_Z \phi(n(z))\tau(z)h(z)dz$, is larger than the welfare gain from the indirect wage impacts that result from the change in taxes, $\int_Z p(z)\left(\int_Z \phi(n(\tilde{z}))\left[\left(\frac{\tilde{z}}{n(\tilde{z})w}\right)^{1+k}\frac{1}{w}+s(n(\tilde{z}))\pi'(w)\right]h(\tilde{z})d\tilde{z}\right)\tau(z)dz$. 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.

5.2 Main Result with General Equilibrium Effects

We can generalize the previous example to allow for arbitrary 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 $\mathbf{z} = (z_1, z_2, ..., z_J) \in \mathbf{Z}$. However, individual utility also depends on a vector, $\mathbf{w}$, of "general equilibrium" parameters which are impacted by the tax schedule (either directly or indirectly via behavioral changes to $\mathbf{z}$). $\mathbf{w}$ might consist of prices, wages, or other quantities that are impacted in some way by individual decisions (and hence depend on taxes). Thus, individuals maximize the following utility function which is assumed smooth in all arguments:

$$\begin{aligned} U(n;T,\mathbf{w}) = \max_{\mathbf{z}}\ & u\left(c,\mathbf{z};\mathbf{n},\mathbf{w}\right) \\ \text{s.t. } & c = y(\mathbf{z},\mathbf{w}) - T(\mathbf{z}) \end{aligned} \tag{44}$$

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

1. *$R(T)$ is Gateaux differentiable*

[21] 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 43 until convergence.

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 total maximum average willingness-to-pay for an increase in each* w_i, $\left|\left|\frac{\overline{u_{w_i}}}{u_c}\right|\right|_\infty$ *(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:*[22]

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

Proof. See Appendix A.8 □

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. However, the key intuition is entirely unchanged from Subsection 5.1: if the "direct" impacts of a tax on utility are larger than the "indirect" effects that taxes have on general equilibrium objects $\mathbf{w}$ which in turn impact utility, 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 takes the form of an integral equation formulated in a measure space, but all of the intuition for Theorem 2 can be understood from the example in Subsection 5.1.

5.3 Optimization Failures

Theorem 1 relies crucially on the envelope theorem to recover an inverse welfare functional, implicitly assuming that agents choose $\mathbf{z}$ to optimize $u(c, \mathbf{z}; \mathbf{n})$. But what if agents have optimization failures or misunderstand the tax schedule or face frictions in optimization? Can we still recover inverse welfare functionals? In many cases, the answer is "no". Let us briefly discuss an example. 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 $\tau(\mathbf{z})$.

[22] $||p_i||_{\text{TV}}$ denotes the total variation norm which is defined in the proof.

The impact on indirect utility $U(n;T)$ 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 11. 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 $\Phi(n)$:

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

Thus, the welfare impact of a tax change is given by:

$$\int_N \left[-\tau(z(n)) + [(1-T_z) - (1-\hat{T}_z)]\xi(n)\tau'(z(n))\right] d\Phi(n) \tag{45}$$

Let us now turn to the budgetary impacts of this reform, which by the same logic are given by:

$$\begin{aligned}&\int_N \left[\tau(z(n)) + T_z\xi(n)\tau'(z(n))\right] f(n)dn = \int_Z \left[\tau(z) + T_z\xi(z)\tau'(z)\right] h(z)dz \\ &= \int_Z \left[h(z) - \frac{\partial}{\partial z}\left(T_z\xi(z)h(z)\right)\right]\tau(z)dz + T_z\xi(z)h(z)\tau(z)\Big|_{\underline{z}}^{\overline{z}}\end{aligned} \tag{46}$$

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 46, $R(T)$ *is* Gateaux differentiable. Returning to Equation 45, we claim $\Phi(n)$ cannot have mass points. If $\Phi(n)$ has mass points for $n \in \text{Int}(N)$ (who choose $z \in \text{Int}(Z)$ by monotonicity), Equation 45 cannot equal Equation 46 because Equation 46 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 45 cannot equal Equation 46. On the other hand, if $\Phi(n)$ is smooth, then we can do a change of variables to get rid of the $\tau'(z)$ terms using integration by parts after defining the derivative of $\Phi(n)$ to be $\frac{d\Phi(n)}{dn} \equiv \psi(n)f(n)$:

$$\begin{aligned}&\int_N \left[-\tau(z(n)) + [(1-T_z) - (1-\hat{T}_z)]\xi(n)\tau'(z(n))\right]\psi(n)f(n)dn \\ &= \int_Z \left[-\tau(z) + [(1-T_z) - (1-\hat{T}_z)]\xi(z)\tau'(z)\right]\psi(n(z))h(z)dz \\ &= \int_Z \left[-\psi(n(z))h(z) - \frac{\partial}{\partial z}\left\{[(1-T_z) - (1-\hat{T}_z)]\xi(z)\psi(n(z))h(z)\right\}\right]\tau(z)dz \\ &+ \left\{[(1-T_z) - (1-\hat{T}_z)]\xi(z)h(z)\psi(n(z))\right\}\tau(z)\Big|_{\underline{z}}^{\overline{z}}\end{aligned} \tag{47}$$

But from here we see that in order for an inverse welfare functional to exist, Equation 46 must equal Equation 47 for all $\tau(z)$. Matching terms for each z, we get a boundary value problem

wherein:

$$\begin{aligned} &-\psi(n(z))h(z) - \frac{\partial}{\partial z}\left\{[(1-T_z)-(1-\hat{T}_z)]\xi(z)h(z)\psi(n(z))\right\} = h(z) - \frac{\partial}{\partial z}\left(T_z\xi(z)h(z)\right) \\ &\text{and } [(1-T_z)-(1-\hat{T}_z)]\xi(z)h(z)\psi(n(z)) = T_z\xi(z)h(z) \text{ for } z \in \{\underline{z}, \overline{z}\} \end{aligned} \tag{48}$$

In most cases the boundary value problem given by System 48 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 48 in Figure 10 in Appendix D.

Given this example, how can we proceed when agents have optimization failures, frictions, and internalities? It turns out we can nonetheless recover a so-called "generalized marginal inverse functional" a la Saez and Stantcheva (2016):

Definition 4. *Suppose the government chooses T to maximize some objective function $O(T)$ with associated Lagrangian given by:*

$$O(T) + \lambda[R(T) - E] \tag{49}$$

We refer to a linear functional G_T as a "generalized marginal inverse functional" for a tax schedule T if the Gateaux derivative of the government's objective function, $O(T)$, equals $G_T(\tau)$ and T is a stationary point for the Lagrangian in Equation 49.

A generalized marginal inverse functional simply tells us how much the government's objective function must change in response to perturbing taxes at each choice level $\mathbf{z}$ in order to (locally) rationalize a given tax schedule. Importantly, the government's objective function no longer needs to be a weighted sum of utilities: the government is no longer assumed to maximize a continuous linear welfare functional as in Equation 4. Intuitively, if revenue is Gateaux differentiable, we can immediately recover a generalized marginal inverse functional:

Theorem 3. *Consider continuous $T(\mathbf{z})$ such that $R(T) = E$ and that $\mathbf{Z}$ is compact.*[23] *A generalized marginal inverse functional exists if $R(T)$ is Gateaux differentiable.*

Proof. Taking the Gateaux derivative of Equation 49, any generalized marginal inverse functional must satisfy:

$$G_T(\tau) + \lambda DR_T(\tau) = 0 \tag{50}$$

Thus, choosing $\lambda = 1$ and $G_T = -DR_T$ ensures Equation 50 is zero.

□

For a concrete example, if $DR_T(\tau(\mathbf{z})) = \int_{\mathbf{Z}} \tau(\mathbf{z})\gamma(\mathbf{z})dH(\mathbf{z})$ where where $H(\mathbf{z})$ is the distribution of choices $\mathbf{z}$ under a given tax schedule, then $G_T(\tau(\mathbf{z})) = -\int_{\mathbf{Z}} \tau(\mathbf{z})g(\mathbf{z})dH(\mathbf{z})$ where $g(\mathbf{z}) = \gamma(\mathbf{z})$.

[23] Note that we have dropped the assumption that for all $\mathbf{z}$ $\exists$ $\mathbf{n}$ with a unique optimal $\mathbf{z}$. This assumption is no longer necessary to recover a generalized marginal inverse functional because we no longer need the envelope theorem.

In this case, $g(\mathbf{z})$ represents the impact of an infinitesimal bump function perturbation at $\mathbf{z}$ on the government's objective function.

Theorem 3 is essentially definitional: the generalized marginal inverse functional is just the (negative) Gateaux derivative of revenue; intuitively, this generalized marginal inverse functional simply ensures that the impact of any given tax perturbation is zero. While Theorem 3 is exceedingly simple to prove given the right setup, the economic interpretation is also extremely general: even if the envelope theorem fails to hold and/or individuals do not correctly optimize (e.g., due to frictions or misperceptions) and/or the welfare functional is non-welfarist, the Gateaux derivative of the government's budget nonetheless pins down an generalized marginal inverse functional that informs us how much society implicitly cares about lowering taxes at a particular set of choices $\mathbf{z}_1$ relative to lowering taxes at any other set of choices $\mathbf{z}_2$. In other words, society's implicit valuations of raising taxes are always pinned down by the empirical impacts of tax changes.

However, it is also important to point out that generalized marginal inverse functionals are inherently difficult to interpret, making them, in general, substantially less useful than inverse welfare functionals as in Theorem 1. When given an inverse welfare functional for a given tax schedule, we can immediately infer how society must care about giving a dollar to individuals making decisions $\mathbf{z}_1$ vs. individuals making decisions $\mathbf{z}_2$, all else equal. On the other hand, when given a generalized marginal inverse functional, we can only infer how much the government's objective function must change by lowering taxes at $\mathbf{z}_1$ vs. lowering taxes at $\mathbf{z}_2$. Given that a generalized marginal inverse functional need not correspond to a welfarist objective (e.g., the government's objective could entail poverty alleviation or encode inequality aversion), we cannot infer anything about implicit interpersonal utility comparisons; we can only infer how the unknown government objective must be changing with tax perturbations at each choice level.

6 Examples of Inverse Welfare Functionals

In this Section, we illustrate two applications of inverse welfare functionals. First, we show how to recover the inverse welfare functional associated with two different tax schedules: (1) the joint income tax schedule for couples in the United States and (2) an income tax schedule in the presence of general equilibrium wage effects. Second, we provide an illustration of a new way in which inverse welfare functionals can be helpful: approximating solutions for complicated optimal taxation problems. Solving more realistic (and hence complex) optimal taxation problems quickly becomes computationally intractable; at best, the solution is governed by a highly non-linear partial differential equation and, at worst, the solution features

non-smoothness (typically in the form of bunching or types with multiple optima) that make computation of solutions exceedingly difficult (Dodds (2023) or Krasikov and Golosov (2022)). We thus propose a new strategy: constrain the solution to be within some relatively simple class of functions (e.g., piecewise linear functions or polynomials) and then compute the inverse welfare functional that rationalizes the proposed schedule as the fully non-linear optimum, continuing to solve the problem for more and more flexible function classes (e.g., piecewise linear functions with more brackets or polynomials with higher order terms) until we reach a point in which the inverse welfare functional is sufficiently "nearby" to the true welfare functional. In addition to highlighting a new path forward to solve complex taxation problems, this procedure highlights how to solve for inverse welfare functionals in a number of interesting, complex cases.

6.1 Example: Couples Taxation

First, we discuss how to find an inverse welfare functional for the observed couples tax schedule in the United States, which is a function of combined household income $T(z_1+z_2)$ where z_1 denotes male income and z_2 denotes female income (we restrict attention to heterosexual couples). We seek to find an inverse welfare functional that rationalizes the observed joint couples income tax schedule in the United States as the fully non-linear optima (Figure 12 in the Appendix shows the observed federal income tax schedule for couples in 2019). The first step is calculating the Gateaux derivative of government revenue. Given that the observed income distribution does not feature any appreciable bunching (see Figure 11 in Appendix D, recognizing that marginal tax rates change along lines $z_1 + z_2 = C$ for various values of C), we calculate the Gateaux derivative of government revenue by assuming that all individuals respond smoothly to tax perturbations. Hence, we require a few elements: (1) the joint income distribution for couples, taken from the American Community Survey (ACS); (2) the 2x2 matrix of substitution effects of male and female incomes with respect to the marginal tax rates on male and female incomes ($\overline{\mathbf{X}}(\mathbf{z})$, in Equation 25); and (3) the estimated income effects for male and female incomes ($\overline{\vec{\eta}}(\mathbf{z})$ in Equation 25). We assume the compensated taxable income elasticities for men and women w.r.t. their own tax rates are 0.2 and 1, respectively (taken from Blomquist and Selin (2010)), and that the compensated taxable income elasticities for men and women w.r.t. each others tax rates are 0.[24] These compensated elasticities of income with respect to marginal tax rates allow us to back out the substitution effects, $\overline{\mathbf{X}}(\mathbf{z})$, which are just the *derivatives* of income with respect to marginal tax rates. We assume income effects are zero.

From here, the Gateaux variation of government revenue can be computed as in Equation

[24]Note that in principle these elasticities could vary over the income distribution; we assume they are constant simply due to a lack of credible estimates of heterogeneous tax elasticities.

25 from Section 3.3 to equal:

$$\int_{\mathbf{Z}} \left(\tau(\mathbf{z}) + \nabla_{\mathbf{z}} T(\mathbf{z}) \left[\overline{\vec{\eta}}(\mathbf{z})\tau(\mathbf{z}) + \overline{\mathbf{X}}(\mathbf{z}) \cdot \nabla_{\mathbf{z}} \tau(\mathbf{z})\right]\right) dH(\mathbf{z}) \tag{51}$$

Next, we can apply integration by parts to Equation 51 *separately* on each of the two dimensional piecewise linear segments. Specifically, we split the domain $\mathbf{Z}$ into regions $\mathbf{Z}_1, \mathbf{Z}_2, ...$ on which marginal tax rates are constant. Performing integration by parts on Equation 51 separately for each region $\mathbf{Z}_i$ (noting that the behavioral revenue effects of taxation are continuous on each region but not across regions because marginal tax rates chase discontinuously) allows us to recover the Gateaux derivative of government revenue. If we assume that everyone is optimizing correctly and marginal utility of consumption is constant (consistent with no income effects), then Equation 29 pins down average inverse welfare weights at each interior income level $\mathbf{z}$. In particular, under these assumptions, average inverse welfare weights for any interior $\mathbf{z}$ are given by:[25]

$$\overline{\phi(\mathbf{z})} \equiv \int_{\mathbf{N}} \phi(\mathbf{n}) dF(\mathbf{n}|\mathbf{z}) = \left[1 + \nabla_{\mathbf{z}} T(\mathbf{z}) \overline{\vec{\eta}}(\mathbf{z})\right] - \frac{1}{h(\mathbf{z})} \nabla_{\mathbf{z}} \cdot \left[\nabla_{\mathbf{z}} T(\mathbf{z}) \overline{\mathbf{X}}(\mathbf{z}) h(\mathbf{z})\right] \tag{52}$$

Note that if only a single type $\mathbf{n}$ locates at each $\mathbf{z}$, then Equation 29 pins down inverse welfare weights for each type on the interior of $\mathbf{Z}$ where the marginal tax rate is constant.

Figure 4 plots these inverse welfare weights for all interior $\mathbf{z}$ (i.e., Figure 4 plots the right-hand-side of Equation 52 for all interior $\mathbf{z}$; Figure 4 does not include the inverse weights along the boundaries or ridges where tax rates are non-differentiable).[26] As to be expected, Figure 4 shows that the implicit welfare weights for the observed couples tax schedule are broadly higher for low income households than for high income households, consistent with a redistributive inverse welfare functional. Note also that inverse welfare weights change discretely across the surfaces where marginal taxes change discretely. Inverse welfare weights for couples with high earning men and low earning women are *higher* than for couples with high earning women with low earning men (e.g., couples with a sole male earner making $500,000 have a welfare weight of approximately 1.36 whereas couples with a sole female earner making $500,000 have a welfare weight of approximately 0.76). This is because observed taxes are a function of $z_1 + z_2$ so that couples with the same total income are taxed identically, yet women tend to have higher

[25] Assuming that everyone both optimizes utility correctly and moves smoothly in response to tax perturbations is arguably inconsistent with the observed lack of bunching in the observed income distribution. While this conceptual issue is not specific to our implementation, we discuss interpretations to resolve this apparent inconsistency more in Appendix D.1.

[26] Applying integration by parts to Equation 51 also leads to "boundary terms" for the boundaries of each region $\mathbf{Z}_i$, which consist of the boundary of the income space and the non-differentiable ridges in the tax schedule. Hence, the aggregate welfare of types along the boundary and non-differentiable ridges has a non-zero contribution to total welfare even though these lines have zero area in $\mathbb{R}^2$. In contrast, the contribution of any other curve in $\mathbf{Z} \subset \mathbb{R}^2$ to total welfare is zero because the area integral along these lines is zero. For example, the value of taking $1 from households with either a man and/or woman making the highest income in our income grid (a measure zero set) is equivalent to the value of giving $1 to the bottom 0.15% of households; the value of giving $1 to households on the second non-differentiable ridge (i.e., with a combined income of $78,950) is equal to the value of giving $1 to the bottom 1.8% of households.

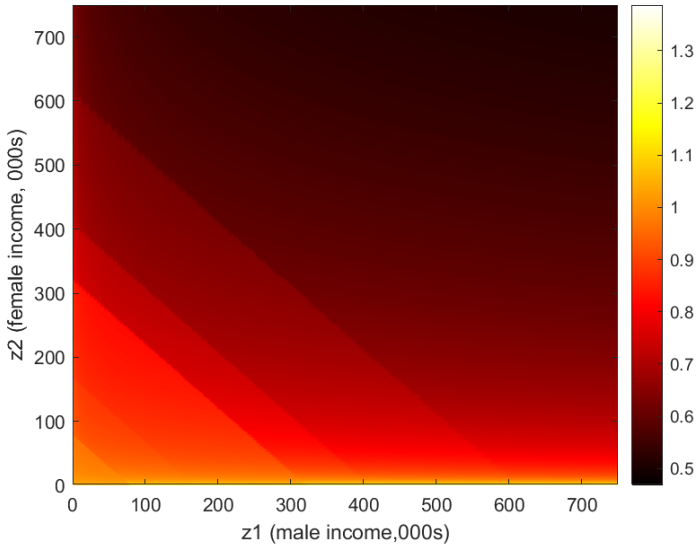


Figure 4: Interior inverse welfare weights based on couples' earnings

Note: The colorbar on the right represents the value of the inverse welfare weights. Lighter colors represent higher weights while darker colors represent lower weights.

elasticities than men.[27] Thus, in order to rationalize identical marginal tax rates on couples with high earning women as couples with high earning men, the government must have lower implicit weights on couples with a high earning women than with a high earning man; the inverse welfare function is thus "gender-biased" in the sense that couples with identical earnings have different inverse welfare weights depending on the earnings composition of the couple.

6.2 Example: Income Taxation with Finitely Elastic Labor Demand

Let us turn to our second example: income taxation with labor demand as in Section 5.1. The goal of this exercise is to show how general equilibrium wage effects impact inverse welfare weights. We suppose that individuals have quasi-linear iso-elastic utility as in Equation 35 from Section 5.1. For purposes of illustration, we calibrate the distribution of types $f(n)$ to match the U.S. income distribution in 2019 from the ACS given that $k = 0.3$, which implies the taxable income elasticity equals 0.3. The government has a set of smoothly decreasing welfare weights, $\phi^*(n)$, that equal the inverse of individual indirect utility under zero taxes, implying that the government cares approximately 10,000 times as much about giving a dollar to the lowest n types relative to giving a dollar to the highest n types. Finally, we 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 that the government assumes, as is common in the optimal taxation literature, that

[27] While in general women have higher labor supply elasticities than men, we do not yet have enough evidence to know for certain whether this holds or not at the very high end of the income distribution; this highlights that inverse welfare functionals are only as reliable as their elasticity parameter inputs.

labor demand is infinitely elastic (corresponding to a production function with $\beta = 1$). Given this assumption, the government's optimization problem collapses to the standard Mirrlees (1971) problem. Hence, the government proposes to change the tax schedule to the schedule that solves the optimal taxation problem as in Mirrlees (1971) or Saez (2001), denoted $T_{Mirrlees}$ (i.e., $T_{Mirrlees}$ satisfies Equation 15 with zero marginal rates at the top and bottom). Under the assumption of infinitely elastic labor demand, the inverse welfare weights that support $T_{Mirrlees}$ are trivially the weights that the government started with: $\phi^*(n)$. We then consider how the inverse weights that support the Mirrleesian tax schedule change if labor demand is, in fact, finitely elastic (the inverse welfare weights change with finitely elastic labor demand because tax perturbations have additional indirect redistributive effects via general equilibrium wage changes). We compute the inverse welfare functional that supports the proposed Mirrleesian tax schedule, $T_{Mirrlees}$, under various values of the labor demand elasticity. Specifically, we compute the inverse welfare weights that support $T_{Mirrlees}$ by finding the fixed point of integral equation 42. We show these inverse weights for various values of E^D:[28]

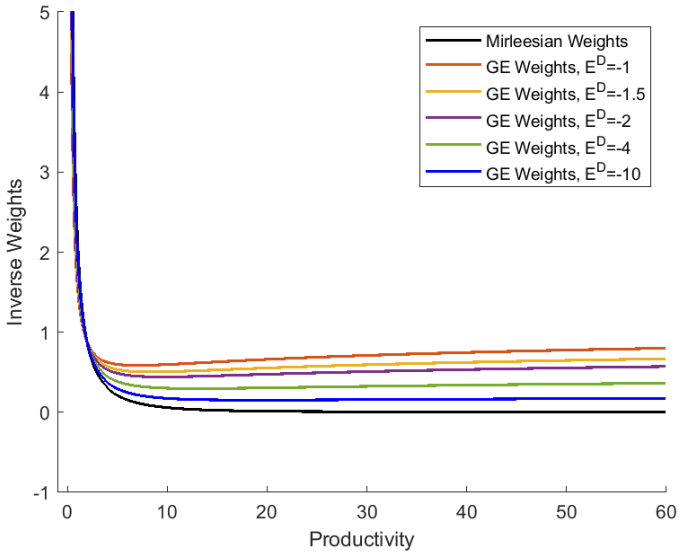


Figure 5: Inverse Welfare Weights with Finite Labor Demand Elasticity

Note: This figure shows the inverse welfare weights for a particular tax schedule computed under various assumptions about the size of the labor demand elasticity. We begin with a set of welfare weights, the "Mirrleesian Weights", and compute the optimal tax schedule assuming that labor demand is infinitely elastic as in Mirrlees (1971) or Saez (2001). We assume a skill distribution calibrated to the U.S. income distribution using the 2019 ACS and a labor supply elasticity of 0.3. Finally, we plot the inverse welfare weights that support this tax schedule under various assumptions about the value of β assuming 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 5 is that a larger labor demand elasticity (in absolute value) corresponds to inverse welfare weights that are much higher for high productivity types, implying

[28] We do not present the boundary weights in Figure 5; these weights are small and do not change the overall findings.

less aversion to inequality. More specifically, recall that we began with a government seeking to optimize welfare given some welfare weights wherein the value of giving a dollar to the lowest productivity type was approximately 10,000 times as much as giving a dollar to the highest productivity individual; assuming the labor demand elasticity is infinite, this government proposes implementing the Mirrlesian tax schedule. However, if the labor demand elasticity is in fact close to -1, then implicit welfare weights that support this Mirrlesian tax schedule only value giving a dollar to the lowest productivity type approximately 25 times as much as giving a dollar to the highest productivity individual. Intuitively, if labor demand is relatively inelastic, then this warrants higher tax rates on high income individuals due to general equilibrium wage effects: by raising taxes on high income individuals, the government not only raises money directly, but also redistributes indirectly via increased wages as a result of lower labor supply from high income individuals. Hence, for a fixed tax schedule $T_{Mirrlees}$, less elastic labor demand implies that inverse welfare weights are higher for high income individuals; in other words, with inelastic labor demand, $T_{Mirrlees}$ implies far weaker implicit redistributive preferences than the government intends.

One concern with the simplistic model of Section 5.1 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 5.1 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 D.2; we show in Figure 13 in Appendix D how the inverse welfare weights vary with the degree of complementarity between L_l and L_h.[29] The key takeaway is that even with a relatively high degree of complementarity between high- and low-skilled labor, accounting for GE wage effects has a substantial impact on inverse welfare weights. Further investigation using more sophisticated labor demand models is certainly warranted.

6.3 Approximate Optimal Taxation: Conceptual Framework

In this section, we show how to use the theory of inverse welfare functionals to approximate solutions to optimal tax problems. The key idea is that numerical computation of inverse welfare functionals is usually substantially simpler than computing the optimal tax schedule

[29] 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 (e.g., 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.

because multidimensional screening problems are notoriously difficult optimization problems whereas inverse welfare functionals are comparatively easy to construct because the can be computed pointwise (e.g., Equations 29 and 30). A typical optimal taxation problem takes as given a welfare functional and seeks to maximize welfare subject to a budget constraint given that individual choices depend on the tax schedule. Instead, we suppose that the government seeks to maximize a welfare functional $W^*(U(\mathbf{n};T)$ but recognizes that solving for the tax schedule T that maximizes this welfare functional may be exceedingly difficult. Alternatively, the government may have a preference for simplicity in the tax code and so may have a preference to maximize $W^*(U(\mathbf{n};T)$ within some restricted class of functions. Let W_T^{Inv} denote a (local) inverse welfare functional for a given tax schedule T. We suppose that the government solves the following problem:

$$\text{Find } T \text{ s.t. } W_T^{Inv} \text{ is sufficiently close to } W^* \text{ and } \int_{\mathbf{N}} T(\mathbf{z}(\mathbf{n}))dF(\mathbf{n}) \geq E \tag{53}$$

Thus, Problem 53 is a relaxation of a typical optimal taxation problem: the government is now content to implement any tax schedule such that the implicit inverse welfare functional rationalizing this tax schedule is "sufficiently close" to the original welfare functional. Now "sufficiently close" is ultimately a normative determination that depends on societal preferences as well as other (unmodeled) complexity constraints that the government may face; for instance, if the government would like to maximize a Rawlsian social welfare function, would an income tax schedule that implicitly puts equal weight on all those with incomes below \$20,000 and no weight on individuals above \$20,000 be "sufficiently close" so as to be acceptable? Setting this normative determination of "close enough" aside, we will simply show how to solve Problem 53 by computing inverse welfare functionals, thereby allowing the planner to determine whether a proposed tax schedule is "sufficiently close".

The next question then is: how can we find tax schedules that generate inverse welfare functionals that are "close" to the true welfare functional? There are a few technical results required, which we now outline, relegating the results and further discussion to Appendix C:

1. If indifference surfaces have bounded gradients, then for every tax schedule there exists a continuous tax schedule that generates the same indirect utility profile; hence, we can WLOG restrict attention to continuous tax schedules.
2. By Stone-Weierstrass, we can approximate the optimal tax schedule arbitrarily well with polynomials or piecewise linear tax schedules.
3. If the revenue effects of taxation are sufficiently smooth around the optimal tax schedule, then the inverse welfare functionals become arbitrarily close to the true welfare functional as the tax schedule becomes arbitrarily close to the optimal tax schedule.

6.4 Piecewise Linear Income Taxation with Multidimensional Heterogeneity

Let us now illustrate how to solve Problem 53 in the case of income taxation with multidimensional heterogeneity. Suppose that individuals have the following utility function, where individuals vary in terms of labor productivity n and their taxable income elasticity k according to some density function $f(n,k)$:

$$\begin{aligned} u(c,z;n,k) &= c - \frac{(z/n)^{1+k}}{1+k} \\ c &= z - T(z) \end{aligned} \tag{54}$$

Given this utility function, suppose that the government seeks to maximize a welfare functional W^*. However, suppose that the government wants to maximize this welfare functional using a piecewise linear tax schedule that is not overly complicated (perhaps due to complexity constraints or because in the presence of multidimensional heterogeneity, solving for the fully non-linear income tax schedule is challenging as one cannot typically ensure that bunching or types with multiple optima will not arise under the optimal tax schedule).[30] Recasting as in Problem 53, suppose the government wants to find a piecewise linear tax schedule T such that the inverse welfare functional that supports T is sufficiently close (i.e., within some tolerance) to the government's welfare functional W^* (recognizing that "sufficiently close" is normative).

For purposes of illustration, we calibrate the distribution of types $f(n,k)$ to match the U.S. income distribution in 2019 from the ACS under the assumption that k is uniformly distributed in $[1/0.35, 1/0.25]$, which implies that taxable income elasticities are between 0.25 and 0.35 with an average elasticity of 0.3 (Saez, Slemrod and Giertz, 2012). We consider a welfare functional W^* consisting of welfare weights that vary only with n, incorporating the concept of "preference neutrality" as in Lockwood and Weinzierl (2016) or Fleurbaey and Maniquet (2006). We choose weights that equal the inverse of individual indirect utility under zero taxes, implying that the government cares approximately 10,000 times as much about giving a dollar to the lowest n types relative to giving a dollar to the highest n types. We maximize this welfare functional using piecewise linear tax schedules with varying numbers of brackets (for simplicity, we choose the kink points of this tax schedule up front, but these could be chosen by the government as well). Figure 16 in Appendix D shows optimal piecewise linear schedules with 2, 4, and 6 brackets. Next, we calculate the inverse optimal welfare functional for each of these piecewise linear schedules (i.e., the welfare functional that rationalizes each piecewise linear schedule as the fully non-linear optima) using the formulas developed in Section 3.2; we present these inverse

[30] The logic of Proposition 4 from Bergstrom and Dodds (2021*a*) guarantees that for sufficient variation in k, some types will necessarily have multiple optima, generating jumping behavior, under the optimal non-linear tax schedule. Bergstrom and Dodds (2021*a*) show how to account for "jumping effects" when solving a simpler version of this sort of problem where the distribution of k is discrete; the case with continuously distributed k is much more complex.

optimal weights for the optimal two bracket, four bracket, and six bracket tax schedules in Figure 6 along with the government's true welfare weights, W^*.[31] There are two key takeaways

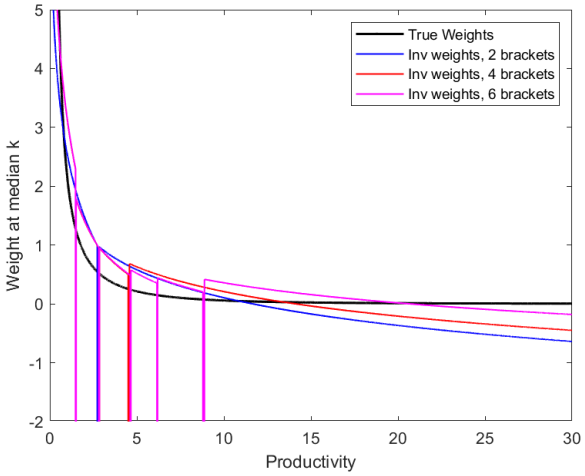


Figure 6: Inverse Welfare Weights for Piecewise Linear Tax Schedules

Note: This figure shows inverse welfare weights for the optimal two bracket tax system with a kink at \$25000, the optimal four bracket tax system with kink points in [\$10000, \$25000, \$50000] and the optimal six bracket tax system with kink points in [\$10000, \$25000, \$50000, \$75000, \$125000] to maximize the welfare functional depicted by the line "true weights". We plot inverse weights across the n distribution for median k when utility is given by Equation 54 and the distribution of n is calibrated to match the U.S. income distribution in 2019 from the ACS; k is uniformly distributed in $[1/0.35, 1/0.25]$, which implies that taxable income elasticities are between 0.25 and 0.35. Figure 14 in Appendix D shows the weights at each n averaged across the k distribution. Because many types choose each given z, inverse weights for each type (n, k) are computed as discussed in Appendix C to ensure pointwise convergence (precisely, we plot $q(\mathbf{z}; T_i)\phi^*(\mathbf{n})$ from Equation 121).

from Figure 6. First, right around each of the kink points, the inverse welfare weights become extremely negative, implying that the presence of the kink point is Pareto inefficient as discussed in Section 4.2.[32] In other words, the government can raise tax revenue by lowering tax rates around the kink point. Second, we see that the optimal six bracket tax system generates inverse weights that appear closer to the true welfare weights than the optimal four bracket tax system, which generates inverse weights that appear closer to the true welfare weights than the optimal two bracket tax system. These inverse welfare weights can then be used by the government to ascertain whether, for a given piecewise linear schedule, the associated inverse weights are "close enough" as in Problem 53.[33]

6.5 Nonlinear Taxation of Income and Housing Rent

Our final example of constructing inverse welfare functionals involves joint taxation of income and (implicit) housing rent. We consider a simple model with a government that wants to create

[31] We verify inverse weights for all tax schedules in Figure 6 by numerically checking that the derivative of the inverse welfare functional is zero in the direction of many small budget neutral perturbations.

[32] Note that these extreme negative weights only have a small impact on the welfare functional because these extreme weights only apply to a tiny set of individuals.

[33] The government may want to know inverse weights over the entire (n, k) distribution rather than just over the n distribution for median k as plotted in Figure 6. Figure 15 in Appendix D plots a heat map of weights over the entire distribution for the six bracket piecewise linear schedule depicted in Figure 6.

a joint tax schedule of both income, z_1, and housing rent, z_2. For households that rent, housing rent is simply equal to the amount of money spent on rent each year; for households that own their home, we assume implicit housing rent (i.e., rent paid to yourself) is equal to a fraction of the property value. With perfect pass-through of taxes onto renters and a constant rental rate of return, such a tax is equivalent to a property tax. Individuals differ in terms of three dimensions: labor productivity n_1, preferences over housing n_2, and curvature in consumption utility α.

$$\begin{aligned} u(c, z_1, z_2; n_1, n_2, \alpha) &= \frac{c^{1-\alpha}}{1-\alpha} - \frac{(z_1/n_1)^{1+k_1}}{1+k_1} + n_2 \frac{z_2^{1+k_2}}{1+k_2} \\ c &= z_1 - z_2 - T(z_1, z_2) \end{aligned} \tag{55}$$

Given this utility function, suppose that the government seeks to maximize a welfare functional W^*. However, this sort of multidimensional screening problem is exceedingly difficult to solve.[34] In contrast, solving for the inverse welfare functional is comparatively straightforward as we can compute inverse weights pointwise (e.g., via Equations 29 and 30). Thus, suppose the government decides to instead solve Problem 53 by finding a polynomial tax schedule T such that the inverse welfare functional that supports T is sufficiently close (i.e., within some tolerance) to the government's welfare functional W^* (recognizing that "sufficiently close" is normative).

We calibrate the distribution of $f(n_1, n_2, \alpha)$ to match the empirical joint distribution of labor income and implicit housing rents from the 2019 American Community Survey (ACS) where implicit rents for homeowners are assumed to be 5% of the property value. 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 the distribution of α is uniform between 0.5 and 0.75. This implies that income effects for taxable income vary between 0.01 and 0.15 (an average of approximately 0.1) across the joint distribution of (n_1, n_2, α), which is in line with estimates of income effects from Gruber and Saez (2002). We assume that the government seeks to maximize a "preference neutral" welfare function as in Bergstrom and Dodds (2021*b*) or Fleurbaey and Maniquet (2006): if there is no heterogeneity in productivity n_1, then optimal taxes are zero so that the government only seeks to redistribute based on productivity differences. Thus, the only motive for taxing housing expenditures is to target those with higher productivity via the correlation between productivity and tastes for housing.

There are two findings from this exercise. First, we find that optimal housing rent taxes are typically negative; for instance, the optimal linear housing tax is approximately -5% (for

[34] The type space is larger than the choice space so this problem does not satisfy the conditions of Rochet (1987) or the generalized single crossing property of Dodds (2023); even if the solution is smooth and characterized by Equation 26, this is a *highly* non-linear partial differential equation.

purposes of illustration, we plot marginal tax rates from the optimal third order polynomial in Figure 17 in Appendix D). Given a 5% rental rate of return, this corresponds to a 0.25% *subsidy* on property. It is optimal to subsidize housing as there is a negative correlation between labor productivity and taste for housing; this results because, empirically, higher income individuals typically spend a smaller fraction of their income on housing. This subsidy is relatively small, however, because the elasticity of housing with respect to the subsidy rate is large compared to the labor supply elasticity, making redistribution via the income tax schedule generally more efficient than redistribution via housing subsidies. The second finding is that the inverse welfare weights for the optimal polynomial tax schedules get closer to the true (assumed) welfare weights with the degree of the polynomial (see Figure 7). Moreover, the inverse welfare weights are reasonably close to the true welfare weights even for the relatively low order polynomials considered. Thus, in this example the planner can find a reasonably simple tax schedule that is rationalized by a welfare functional that is relatively similar to the welfare functional that he/she initially wanted to maximize.

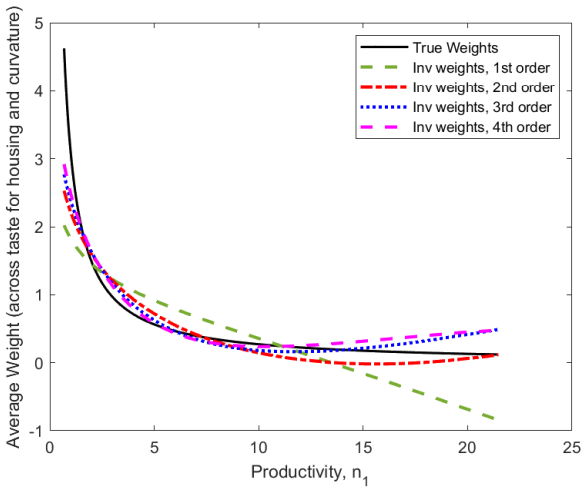


Figure 7: Average Inverse Weights over Labor Productivity

Note: This figure shows average inverse welfare weights across the labor productivity distribution n_1 for the optimal first, second, third, and fourth degree polynomial tax schedules which are a function of income and housing rent, z_1 and z_2. Utility is given by Equation 55 and the type distribution $f(n_1, n_2, \alpha)$ is calibrated to match the empirical joint distribution of labor income and implicit housing rents from the 2019 American Community Survey (ACS) where implicit rents for homeowners are assumed to be 5% of the property value. k_1 and k_2 are chosen 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). α is uniformly distributed between 0.5 and 0.75. The government maximizes a "preference neutral" welfare function as in Bergstrom and Dodds (2021*b*) or Fleurbaey and Maniquet (2006). Because many types choose each given (z_1, z_2), inverse weights for each type (n_1, n_2, α) are computed as discussed in Appendix C to ensure pointwise convergence (precisely, we plot $q(\mathbf{z}; T_i)\phi^*(\mathbf{n})$ from Equation 121).

7 Conclusion

This paper has developed a general theory to recover the inverse welfare functional that rationalizes a given multidimensional tax schedule as optimal. The essential component required to

construct such an inverse welfare functional is the Gateaux derivative of government revenue. We have shown how this theory can be used to characterize Pareto efficiency and can be used to strengthen the Atkinson-Stiglitz theorem. Furthermore, our theory can be even further generalized to allow for general equilibrium effects and optimization errors. Finally, we have provided a number of numerical examples, showcasing not only how to construct inverse welfare functionals, but also highlighting a number of interesting findings: the current couples tax schedule in the U.S. places higher implicit weights on couples with high earning males relative to high earning females, income tax schedules that are rationalized with high preferences for redistribution under a Mirrleesian model with exogenous wages may actually be rationalized with inverse welfare functions that encompass low preferences for redistribution once one accounts for wage endogeneity, piecewise linear income tax schedules with just 5 or 6 brackets seem to do a fairly good job of approximating the optimal tax schedule in the sense that the associated inverse welfare weights are reasonably close to the "true" (assumed) welfare weights, and low-order polynomials can do a decent job approximating the optimal joint income and housing rent tax schedule also in the sense that the associated inverse welfare weights are reasonably close to the "true" (assumed) welfare weights.

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. If the inverse welfare function for a given policy diverges sharply from society's preferences, this places an onus on the government to alter this policy. Further, inverse welfare functions are attractive because they are free from any normative considerations; hence, they can (at least in principle) be estimated purely from data. Moving forward, we believe there is still substantial scope for innovation in so-called "inverse optimal" methods. While this paper focuses on inverse welfare functions for tax schedules, much of the analysis can likely be extended in a straightforward manner to non-tax policy spaces, such as in-kind good provision, minimum wages, or social insurance. Furthermore, inverse optimal methods may be useful in the context of other multidimensional screening problems, such as nonlinear pricing (e.g., Mussa and Rosen (1978) or Armstrong (1996)), public procurement (e.g., Laffont and Tirole (1994)), or regulation of monopolies (e.g., Baron and Myerson (1982)).

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.

Armstrong, Mark. 1996. "Multiproduct nonlinear pricing." *Econometrica*, 64: 51–75.

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.

Baron, David P., and Roger B. Myerson. 1982. "Regulating a monopolist with unknown costs." *Econometrica*, 50(4): 911.

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

Bergstrom, Katy, and William Dodds. 2021*b*. "Using Labor Supply Elasticities to Learn about Income Inequality: The Role of Productivities versus Preferences." *American Economic Journal: Economic Policy*, 13(3): 28–62.

Blomquist, Sören, and Håkan Selin. 2010. "Hourly wage rate and taxable labor income responsiveness to changes in marginal tax rates." *Journal of Public Economics*, 94: 878–889.

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.

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.

Dodds, William. 2023. "Solving multidimensional screening problems using a generalized single crossing property." *Economic Theory.*

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.

Fleurbaey, Marc, and Francois Maniquet. 2006. "Fair Income Tax." *Review of Economic Studies*, 73: 55–83.

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.

Kinderlehrer, David, and Guido Stampacchia. 1980. *An introduction to variational inequalities and their applications.*

Krasikov, Ilia, and Mikhail Golosov. 2022. "Multidimensional Screening in Public Finance: The Optimal Taxation of Couples."

Laffont, Jean Jacques, and Jean Tirole. 1994. *A theory of incentives in procurement and Regulation.* MIT Press.

Lockwood, Benjamin B., and Matthew Weinzierl. 2016. "De Gustibus non est Taxandum: Heterogeneity in preferences and optimal redistribution." *Journal of Public Economics*, 124: 74–80.

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.

Mussa, Michael, and Sherwin Rosen. 1978. "Monopoly and product quality." *Journal of Economic Theory*, 18(2): 301–317.

Rochet, Jean-Charles. 1987. "A necessary and sufficient condition for rationalizability in a quasi-linear context." *Journal of Mathematical Economics*, 16(2): 191–200.

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

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

Saez, Emmanuel, and Stefanie Stantcheva. 2016. "Generalized Social Marginal Welfare Weights for Optimal Tax Theory." *American Economic Review*, 106(1): 24–45.

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."

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

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

A Appendix: Proofs

A.1 Existence of 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:

$$V(\mathbf{n}_1) - V(\mathbf{n}_2) = \int_{\mathbf{n}_2}^{\mathbf{n}_1} \nabla_{\mathbf{n}} u(y(\mathbf{z}) - T, \mathbf{z}; \mathbf{n})|_{T=\tilde{T}(\mathbf{n}), \mathbf{z}=\tilde{\mathbf{z}}(\mathbf{n})} \cdot d\mathbf{n} \tag{56}$$

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

$$\nabla_{\mathbf{n}} V(\mathbf{n}) = \nabla_{\mathbf{n}} u(y(\mathbf{z}) - T, \mathbf{z}; \mathbf{n})|_{T=\tilde{T}(\mathbf{n}), \mathbf{z}=\tilde{\mathbf{z}}(\mathbf{n})} \tag{57}$$

Equation 57 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 4. *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 $\mathbf{Z} \times \mathbf{N}$ (recall $\mathbf{Z}$ is the set of chosen $\mathbf{z}$ when agents optimize under T). If T induces a utility profile $U(\mathbf{n}; T)$ on the boundary of the set $\mathcal{V}$ then T has an associated 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 $V \in \mathcal{V} \backslash \text{Int}(\mathcal{V})$, 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 56 as long as $\nabla_{\mathbf{n}} u(y(\mathbf{z}) - T(\mathbf{z}), \mathbf{z}; \mathbf{n})$ is bounded on $\mathbf{Z} \times \mathbf{N}$. 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 an inverse optimal welfare functional for $U(\mathbf{n}; T)$. □

Proposition 4 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 an inverse optimal functional which support that profile relative to all other feasible indirect utility profiles.[35] 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 4, suppose that* $\mathbf{N}$ *is a compact subset of* $(-\infty, 0)^K$ *and*

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

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

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

It is then straight-forward to establish then that $\tilde{T}(V, \nabla_{\mathbf{n}} V)$ *is concave.*

A.2 Proof of Theorem 1

Proof. First, $T(\mathbf{z})$, and hence $\tau(\mathbf{z})$, are assumed continuous so that if $R(T(\mathbf{z})$ 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:

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

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 $L(T; W)$. 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, let us show that there exists some functional which is continuous and linear on $C(\mathbf{N})$ that satisfies the statement of the Theorem. In particular, we will show that there exists an inverse welfare functional of the following form

[35] In general, the associated 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.

for some regular, countably additive Borel measures Φ_1 and Φ_2:

$$W(U(\mathbf{n};T)) = \int_{\mathbf{Z}} \int_{\mathbf{N}(\mathbf{z})} U(\mathbf{n};T) d\Phi_1(\mathbf{n}|\mathbf{z}) d\Phi_2(\mathbf{z})$$

where $\mathbf{N}(\mathbf{z})$ again represents the set of $\mathbf{n}$ that optimally choose a given $\mathbf{z}$; if an individual has multiple optima we arbitrarily assign them to a single $\mathbf{z}$.

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:

$$\lim_{\epsilon \to 0} \frac{U(\mathbf{n};T+\epsilon\tau) - U(\mathbf{n};T)}{\epsilon} = -u_c(y(\mathbf{z}(\mathbf{n})) - T(\mathbf{z}(\mathbf{n})), \mathbf{z}(\mathbf{n}); \mathbf{n})\tau(\mathbf{z}(\mathbf{n})) \equiv -u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n})) \quad (58)$$

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 optima 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 58 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 $\{-\tau(\mathbf{z})u_c(y(\mathbf{z}) - T(\mathbf{z}), \mathbf{z}; \mathbf{n})\}_{\mathbf{z}\in\mathbf{Z}}$ is equicontinuous. Both properties follow by smoothness assumptions on utility, continuity of $T(\mathbf{z})$ and $\tau(\mathbf{z})$, and the fact that $\mathbf{Z}$ is compact.

Let us then pick Φ_1 as follows. By assumption, for each $\mathbf{z}$ $\exists$ at least one $\hat{\mathbf{n}}(\mathbf{z})$ with a unique optima at the given $\mathbf{z}$, so set $\Phi_1(\mathbf{n}|\mathbf{z}) = \frac{1}{u_c(\hat{\mathbf{n}})}\delta_{\hat{\mathbf{n}}(\mathbf{z})}(\mathbf{n})$, 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:

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

Thus, we have defined Φ_1 such that if $\mathcal{M}$ represents the set of individuals with multiple optima, then $\forall \mathbf{z}$, $\int_{\mathbf{N}(\mathbf{z})\cap\mathcal{M}} d\Phi_1(\mathbf{n}|\mathbf{z}) = 0$ (i.e., Φ_1-a.e. $\mathbf{n}$ have a unique optima). Let us then consider the derivative of $W(U(\mathbf{n};T+\epsilon\tau))$ with respect to ϵ. Because for every $\mathbf{z}$, Φ_1-a.e. $\mathbf{n}$ have a unique optima, then we have:[36]

$$\frac{\partial W(U(\mathbf{n};T+\epsilon\tau))}{\partial\epsilon} = \int_{\mathbf{Z}} \int_{\mathbf{N}(\mathbf{z})} -u_c(\mathbf{n})\tau(\mathbf{z}) d\Phi_1(\mathbf{n}|\mathbf{z}) d\Phi_2(\mathbf{z}) = -\int_{\mathbf{Z}} \tau(\mathbf{z}) d\Phi_2(\mathbf{z})$$

Given this choice of Φ_1, we still need to show that there exists a measure Φ_2 that makes the

[36] If instead almost all $\mathbf{n}$ that choose each $\mathbf{z}$ have a unique optima, we could avoid using the Dirac measure and instead maximize the more standard welfare functional:

$$W(U(\mathbf{n};T)) = \int_{\mathbf{Z}} \int_{\mathbf{N}(\mathbf{z})} \phi_1(\mathbf{n}) U(\mathbf{n};T) dF(\mathbf{n}|\mathbf{z}) d\Phi_2(\mathbf{z})$$

where we choose $\phi_1(\mathbf{n}) = \frac{1}{u_c(\mathbf{n})}$. This would also lead us to $\frac{\partial W(U(\mathbf{n};T+\epsilon\tau))}{\partial\epsilon} = -\int_{\mathbf{Z}} \tau(\mathbf{z}) d\Phi_2(\mathbf{z})$.

Gateaux derivative of the government's Lagrangian zero $\forall\tau$. Hence, $\forall\tau$ we must have:

$$\left.\frac{\partial L(T+\epsilon\tau;W)}{\partial\epsilon}\right|_{\epsilon=0} = -\int_{\mathbf{Z}}\tau(\mathbf{z})d\Phi_2(\mathbf{z}) + \lambda\int_{\mathbf{Z}}\tau(\mathbf{z})d\Gamma(\mathbf{z}) = 0$$

But from here, we can find an inverse welfare functional by normalizing λ to 1 and and choosing $\Phi_2 = \Gamma$. Finally, we should show that given this choice of Φ_1 and Φ_2 that $\int_{\mathbf{Z}}\int_{\mathbf{N(z)}}U(\mathbf{n};T)d\Phi_1(\mathbf{n}|\mathbf{z})d\Phi_2(\mathbf{z})$ is a continuous linear functional. Linearity follows immediately and continuity of $\int_{\mathbf{N(z)}}U(\mathbf{n};T)d\Phi_1(\mathbf{n}|\mathbf{z})$ follows as:

$$\left|\int_{\mathbf{N(z)}}U_2(\mathbf{n};T)d\Phi_1(\mathbf{n}|\mathbf{z}) - \int_{\mathbf{N(z)}}U_1(\mathbf{n};T)d\Phi_1(\mathbf{n}|\mathbf{z})\right| \leq \frac{||U_2(\mathbf{n};T)-U_1(\mathbf{n};T)||_\infty}{u_c(\hat{\mathbf{n}})} \leq K||U_2(\mathbf{n};T)-U_1(\mathbf{n};T)||_\infty$$

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 which implies, given that $\Phi_2(\mathbf{z})$ defines a continuous functional by equivalence with $\Gamma(\mathbf{z})$, that $\int_{\mathbf{Z}}\int_{\mathbf{N(z)}}U(\mathbf{n};T)d\Phi_1(\mathbf{n}|\mathbf{z})d\Phi_2(\mathbf{z})$ is continuous as well. □

A.3 Proof of Equation 18

Proof. 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*a*)). We can write tax revenue as:

$$\begin{aligned}\int_V\Bigg\{&\int_{\underline{n}}^{n_1(v)}T(z(n,v))f(n|v)dn + \int_{n_1(v)}^{n_2(v)}T(z(n,v))f(n|v)dn\\ &+\int_{n_2(v)}^{n_3(v)}T(z(n,v))f(n|v)dn + \int_{n_3(v)}^{\overline{n}}T(z(n,v))f(n|v)dn\Bigg\}f(v)dv\end{aligned} \tag{59}$$

We have the individual first order condition:

$$u_1(z-T(z)-\epsilon\tau(z), z/n; v)\left(1-T'(z)-\epsilon\tau'(z)\right) + \frac{1}{n}u_2(z-T(z)-\epsilon\tau(z), z/n; v) = 0 \tag{60}$$

For all individuals with a unique optima where the tax schedule is twice continuously differentiable the second order condition holds strictly (see Lemma 3 of Bergstrom and Dodds (2021*a*)), 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):

$$\begin{aligned}\frac{\partial z}{\partial \epsilon}(n,v) &= \frac{u_1\tau'(z) + \left[u_{11}(1-T'(z)) + \frac{1}{n}u_{12}\right]\tau(z)}{u_{11}(1-T'(z))^2 + \frac{2(1-T'(z))}{n}u_{12} + \frac{1}{n^2}u_{22}} \\ &\equiv \xi(n,v)\tau'(z(n,v)) + \eta(n,v)\tau(z(n,v))\end{aligned} \tag{61}$$

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.[37] 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:

$$u(z^+ - T(z^+) - \epsilon\tau(z^+), z^+/n_3; v) = u(z^- - T(z^-) - \epsilon\tau(z^-), z^-/n_3; v) \tag{62}$$

We can also calculate how the indifferent individual (for each v) changes with the tax schedule by applying the implicit function theorem to Equation 62:[38]

$$\frac{\partial n_3}{\partial \epsilon} = \frac{u_1(z^+ - T(z^+), z^+/n_3; v)\tau(z^+) - u_1(z^- - T(z^-), z^-/n_3; v)\tau(z^-)}{u_2(z^- - T(z^-), z^-/n_3; v)z^-/(n_3)^2 - u_2(z^+ - T(z^+), z^+/n_3; v)z^+/(n_3)^2} \equiv \frac{u_1^+\tau(z^+) - u_1^-\tau(z^-)}{u_2^- z^+/(n_3)^2 - u_2^+ z^+/(n_3)^2} \tag{63}$$

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

$$\begin{aligned}&\lim_{\epsilon\to 0}\frac{R(T+\epsilon\tau) - R(T)}{\epsilon} \\ &= \int_V \Bigg\{ \int_{\underline{n}}^{n_1(v)} \left[\frac{T(z(n,v))}{\partial\epsilon} + \tau(z(n,v))\right] f(n|v)dn + \int_{n_1(v)}^{n_2(v)} \left[\frac{T(z(n,v))}{\partial\epsilon} + \tau(z(n,v))\right] f(n|v)dn \\ &+ \int_{n_2(v)}^{n_3(v)} \left[\frac{T(z(n,v))}{\partial\epsilon} + \tau(z(n,v))\right] f(n|v)dn + \int_{n_3(v)}^{\overline{n}} \left[\frac{T(z(n,v))}{\partial\epsilon} + \tau(z(n,v))\right] f(n|v)dn \\ &+ \left(T(z^-(v)) - T(z^+(v))\right) f(n_3(v)|v)\frac{\partial n_3}{\partial\epsilon}(v)\Bigg\} f(v)dv\end{aligned} \tag{64}$$

Note, the last term of Equation 64 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 61) and changing the variable of integration from n to

[37] Footnote 21 of Bergstrom and Dodds (2021*a*) 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.

[38] 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 63. However, this assumption can be dropped without changing Equation 63; see Appendix A.6 of Bergstrom and Dodds (2021*a*).

z we find that:[39]

$$\begin{aligned}
&\lim_{\epsilon\to 0}\frac{R(T+\epsilon\tau)-R(T)}{\epsilon}\\
&=\int_V\Bigg\{\int_{z(\underline{n};v)}^{z(n_1(v);v)}\left(T'(z)\xi(z,v)\tau'(z)+\left[1+T'(z)\eta(z,v)\right]\tau(z)\right)h(z|v)dz+\int_{n_1}^{n_2}\tau(K_1)f(n|v)dn\\
&+\int_{z(n_2(v);v)}^{z^-(n_3(v);v)}\left(T'(z)\xi(z,v)\tau'(z)+\left[1+T'(z)\eta(z,v)\right]\tau(z)\right)h(z|v)dz\\
&+\int_{z^+(n_3(v);v)}^{z(\overline{n};v)}\left(T'(z)\xi(z,v)\tau'(z)+\left[1+T'(z)\eta(z,v)\right]\tau(z)\right)h(z|v)dz\\
&+\left(T(z^-(v))-T(z^+(v))\right)f(n_3(v)|v)\frac{u_1^+\tau(z^+(v))-u_1^-\tau(z^-(v))}{u_2^-z^-(v)/(n_3(v))^2-u_2^+z^+(v)/(n_3(v))^2}\Bigg\}f(v)dv
\end{aligned}\tag{65}$$

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

$$\begin{aligned}
&\lim_{\epsilon\to 0}\frac{R(T+\epsilon\tau)-R(T)}{\epsilon}\\
&\int_{\underline{z}}^{K_1}\left(T'(z)\overline{\xi}(z)\tau'(z)+\left[1+T'(z)\overline{\eta}(z)\right]\tau(z)\right)h(z)dz+M(K_1)\tau(K_1)\\
&+\int_{K_1}^{\overline{z^-}}\left(T'(z)\overline{\xi}(z)\tau'(z)+\left[1+T'(z)\overline{\eta}(z)\right]\tau(z)\right)h(z)dz\\
&+\int_{\underline{z^+}}^{\overline{z}}\left(T'(z)\overline{\xi}(z)\tau'(z)+\left[1+T'(z)\overline{\eta}(z)\right]\tau(z)\right)h(z)dz\\
&+\int_V\left(T(z^-(v))-T(z^+(v))\right)\frac{u_1^+\tau(z^+(v))-u_1^-\tau(z^-(v))}{u_2^-z^-(v)/(n_3(v))^2-u_2^+z^+(v)/(n_3(v))^2}f(n_3(v),v)dv
\end{aligned}\tag{66}$$

Next, let us apply integration by parts to get rid of the $\tau'(z)$ terms in Equation 66, 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.[40] 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

[39]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.

[40]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 $h(\overline{z})$.

in the second tax bracket):

$$\begin{aligned}
&\lim_{\epsilon\to 0}\frac{R(T+\epsilon\tau)-R(T)}{\epsilon}\\
&\int_{\underline{z}}^{K_1}\left(-\frac{\partial}{\partial z}\left[T'(z)\overline{\xi}(z)h(z)\right]+\left[1+T'(z)\overline{\eta}(z)\right]h(z)\right)\tau(z)dz+T_1\overline{\xi}(K_1^-)h(K_1^-)\tau(K_1)+M(K_1)\tau(K_1)\\
&-T_2\overline{\xi}(K_1^+)h(K_1^+)\tau(K_1)+\int_{K_1}^{\overline{z^-}}\left(-\frac{\partial}{\partial z}\left[T'(z)\overline{\xi}(z)h(z)\right]+\left[1+T'(z)\overline{\eta}(z)\right]h(z)\right)\tau(z)dz\\
&+\int_{\underline{z^+}}^{\overline{z}}\left(-\frac{\partial}{\partial z}\left[T'(z)\overline{\xi}(z)h(z)\right]+\left[1+T'(z)\overline{\eta}(z)\right]h(z)\right)\tau(z)dz\\
&+\int_V\left(T(z^-(v))-T(z^+(v))\right)\frac{u_1^+\tau(z^+(v))-u_1^-\tau(z^-(v))}{u_2^-z^-(v)/(n_3(v))^2-u_2^+z^+(v)/(n_3(v))^2}f(n_3(v),v)dv
\end{aligned}\tag{67}$$

Finally, note that all $\tau(z)$ terms enter Equation 67 linearly so that Equation 67 is a linear functional of $\tau(z)$ which means that $R(T)$ is Gateaux differentiable (assuming that all terms in Equation 67 are bounded so that Equation 67 is a linear bounded - hence continuous - functional of $\tau(z)$). To recover the inverse welfare functional from Equation 13, we simply collect all of the terms in Equation 13 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 67 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 the transformation), and $\hat{h}^+(n_3(z^+),z^+)=f(n_3(v),v)\left(\frac{\partial z^-}{\partial v}\right)^{-1}$:[41]

$$\begin{aligned}
&\lim_{\epsilon\to 0}\frac{R(T+\epsilon\tau)-R(T)}{\epsilon}\\
&\int_{\underline{z}}^{K_1}\left(-\frac{\partial}{\partial z}\left[T'(z)\overline{\xi}(z)h(z)\right]+\left[1+T'(z)\overline{\eta}(z)\right]h(z)\right)\tau(z)dz+T_1\overline{\xi}(K_1^-)h(K_1^-)\tau(K_1)+M(K_1)\tau(K_1)\\
&-T_2\overline{\xi}(K_1^+)h(K_1^+)\tau(K_1)+\int_{K_1}^{\overline{z^-}}\left(-\frac{\partial}{\partial z}\left[T'(z)\overline{\xi}(z)h(z)\right]+\left[1+T'(z)\overline{\eta}(z)\right]h(z)\right)\tau(z)dz\\
&+\int_{\underline{z^+}}^{\overline{z}}\left(-\frac{\partial}{\partial z}\left[T'(z)\overline{\xi}(z)h(z)\right]+\left[1+T'(z)\overline{\eta}(z)\right]h(z)\right)\tau(z)dz\\
&+\int_{Z^-}\frac{-\left[T(z^-)-T(z^+)\right]u_1^-\tau(z^-)}{u_2^-z^-/(n_3)^2-u_2^+z^+/(n_3)^2}\hat{h}^-(n_3(z^-),z^-)dz^-+\int_{Z^+}\frac{\left[T(z^-)-T(z^+)\right]u_1^+\tau(z^+)}{u_2^-z^-/(n_3)^2-u_2^+z^+/(n_3)^2}\hat{h}^+(n_3(z^+),z^+)dz^+
\end{aligned}\tag{68}$$

From here we can just collect terms to incorporate the terms of last two integrals in Equation 68 into the other integrals noting that $\hat{h}^+(n_3(z),z)\neq 0 \iff \mathbb{1}(z\in Z^+)$ and $\hat{h}^-(n_3(z),z)\neq$

[41] Note that in the first integral in the last line of Equation 68, 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^+.

$0 \iff \mathbb{1}(z \in Z^-)$:

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

$$\underbrace{\int_{\underline{z}}^{K_1} \left(-\frac{\partial}{\partial z}\left[T'(z)\overline{\xi}(z)h(z)\right] + \left[1 + T'(z)\overline{\eta}(z)\right]h(z) \right)\tau(z)dz}_{\text{Perturbations in First Bracket}}$$

$$+ \underbrace{T_1\overline{\xi}(K_1^-)h(K_1^-)\tau(K_1) + M(K_1)\tau(K_1) - T_2\overline{\xi}(K_1^+)h(K_1^+)\tau(K_1)}_{\text{Perturbation at Kink}}$$

$$+ \underbrace{\int_{K_1}^{\overline{z^-}} \left(-\frac{\partial}{\partial z}\left[T'(z)\overline{\xi}(z)h(z)\right] + \left[1 + T'(z)\overline{\eta}(z)\right]h(z) - \frac{\left(T(z) - T(z^+(z))\right)u_1^-(z)}{u_2^-(z)\frac{z}{(n_3(z))^2} - u_2^+(z)\frac{z}{(n_3(z))^2}}\hat{h}^-(n_3(z), z) \right)\tau(z)dz}_{\text{Perturbations in Second Bracket}}$$

$$+ \underbrace{\int_{\underline{z^+}}^{\overline{z}} \left(-\frac{\partial}{\partial z}\left[T'(z)\overline{\xi}(z)h(z)\right] + \left[1 + T'(z)\overline{\eta}(z)\right]h(z) + \frac{\left(T(z^-(z)) - T(z)\right)u_1^+(z)}{u_2^-(z)\frac{z}{(n_3(z))^2} - u_2^+(z)\frac{z}{(n_3(z))^2}}\hat{h}^+(n_3(z), z) \right)\tau(z)dz}_{\text{Perturbations in Third Bracket}} \tag{69}$$

From here, we recover Equation 18 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:

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

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

□

A.4 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 optima and 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 six 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).[42]
5. Average behavioral effects of taxation are sufficiently smooth

A.4.1 Single Optima Individuals and a Smooth Tax Schedule

First, let us consider the set of individuals who have a single optima $\mathbf{z}$ and at which $T(\mathbf{z})$ is twice differentiable. These individuals satisfy first order conditions given by System 70 (which is just System 23 reproduced for clarity):

$$\begin{aligned} &u_c(y(\mathbf{z}) - T(\mathbf{z}) - \epsilon\tau(\mathbf{z}), \mathbf{z}; \mathbf{n})\left(y_{z_1}(\mathbf{z}) - T_{z_1}(\mathbf{z}) - \epsilon\tau_{z_1}(\mathbf{z})\right) + u_{z_1}(y(\mathbf{z}) - T(\mathbf{z}) - \epsilon\tau(\mathbf{z}), \mathbf{z}; \mathbf{n}) = 0 \\ &\qquad\qquad \vdots \\ &u_c(y(\mathbf{z}) - T(\mathbf{z}) - \epsilon\tau(\mathbf{z}), \mathbf{z}; \mathbf{n})\left(y_{z_J}(\mathbf{z}) - T_{z_J}(\mathbf{z}) - \epsilon\tau_{z_J}(\mathbf{z})\right) + u_{z_J}(y(\mathbf{z}) - T(\mathbf{z}) - \epsilon\tau(\mathbf{z}), \mathbf{z}; \mathbf{n}) = 0 \end{aligned} \tag{70}$$

For any such agent $\mathbf{n}$ with a unique optimal income $\mathbf{z}(\mathbf{n})$, compactness arguments imply that $\exists \upsilon$ 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 71 (which is just System 24 reproduced for clarity):

$$\begin{aligned} &\frac{\partial \mathbf{z}(\mathbf{n})}{\partial \epsilon} = \mathbf{H}^{-1}(\mathbf{n}) FOC(\mathbf{n})_\epsilon|_{\epsilon=0} = \mathbf{H}^{-1}(\mathbf{n})[\mathbf{a}(\mathbf{n})\tau(\mathbf{z}) + \mathbf{B}(\mathbf{n}) \cdot \nabla_{\mathbf{z}}\tau(\mathbf{z})] \\ &\equiv \vec{\eta}(\mathbf{n})\tau(\mathbf{z}) + \mathbf{X}(\mathbf{n}) \cdot \nabla_{\mathbf{z}}\tau(\mathbf{z}) \end{aligned} \tag{71}$$

where $FOC_\epsilon|_{\epsilon=0}$ is the vector of derivatives of the first order conditions 70 with respect to ϵ. The second equality in Equation 71 follows for some vector $\mathbf{a}$ and a matrix $\mathbf{B}$ (which depend

[42] 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.

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 71 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 optima and the tax schedule is twice continuously differentiable, we know that for all but some measure zero set of agents:

$$\frac{\partial}{\partial \epsilon}\left[T(\mathbf{z}(\mathbf{n})) + \epsilon\tau(\mathbf{z}(\mathbf{n}))\right]|_{\epsilon=0} = \tau(\mathbf{z}(\mathbf{n})) + \nabla_{\mathbf{z}}T(\mathbf{z})\tau(\mathbf{z}(\mathbf{n}))\vec{\eta}(\mathbf{n}) + \nabla_{\mathbf{z}}T(\mathbf{z})\mathbf{X}(\mathbf{n}) \cdot \nabla_{\mathbf{z}}\tau(\mathbf{z}) \quad (72)$$

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 optima to begin with); hence, they have a negligible impact on the Gateaux derivative of $R(T)$.

A.4.2 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 optima 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:

$$\max_{\mathbf{z}\in\mathbf{Z}_1} u(c(\mathbf{z}), \mathbf{z}; \tilde{\mathbf{n}}) = \max_{\mathbf{z}\in\mathbf{Z}_2} u(c(\mathbf{z}), \mathbf{z}; \tilde{\mathbf{n}}) \quad (73)$$

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 optima 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:

$$\left(\frac{\partial u(c(\mathbf{z}_1), \mathbf{z}_1; \tilde{\mathbf{n}})}{\partial \epsilon} - \frac{\partial u(c(\mathbf{z}_2), \mathbf{z}_2; \tilde{\mathbf{n}})}{\partial \epsilon}\right) = (\nabla_{\mathbf{n}} u(c(\mathbf{z}_2), \mathbf{z}_2; \tilde{\mathbf{n}}) - \nabla_{\mathbf{n}} u(c(\mathbf{z}_1), \mathbf{z}_1; \tilde{\mathbf{n}})) \cdot \nabla_{\epsilon}\tilde{\mathbf{n}} \quad (74)$$

Given that ϵ only has a direct impact on consumption, we can rewrite Equation 74 as:

$$u_c(c(\mathbf{z}_2), \mathbf{z}_2; \tilde{\mathbf{n}})\tau(\mathbf{z}_2) - u_c(c(\mathbf{z}_1), \mathbf{z}_1; \tilde{\mathbf{n}})\tau(\mathbf{z}_1) = (\nabla_{\mathbf{n}} u(c(\mathbf{z}_2), \mathbf{z}_2; \tilde{\mathbf{n}}) - \nabla_{\mathbf{n}} u(c(\mathbf{z}_1), \mathbf{z}_1; \tilde{\mathbf{n}})) \cdot \nabla_{\epsilon}\tilde{\mathbf{n}} \quad (75)$$

Equation 75 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 optima 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 optima whereas individuals on the (shared) boundary of these two regions have multiple optima. We have:

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

Differentiating $R(T + \epsilon\tau)$, appealing to the Reynold's Transport Theorem, we get:[43]

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

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 surface 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:

$$\int_{\mathbf{N}} \frac{\partial}{\partial\epsilon}[T(\mathbf{z}(\mathbf{n})) + \epsilon\tau(\mathbf{z}(\mathbf{n}))]dF(\mathbf{n}) + \int_{\partial\mathbf{N}_1} [T(\mathbf{z}_1(\tilde{\mathbf{n}})) - T(\mathbf{z}_2(\tilde{\mathbf{n}}))]\nabla_\epsilon\tilde{\mathbf{n}} \cdot \rho_1 f(\tilde{\mathbf{n}})dS \tag{76}$$

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 Equation 75 we have:[44]

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

Hence, we get that:

$$\begin{aligned} &\int_{\partial\mathbf{N}_1} [T(\mathbf{z}_1(\tilde{\mathbf{n}})) - T(\mathbf{z}_2(\tilde{\mathbf{n}}))]\nabla_\epsilon\tilde{\mathbf{n}} \cdot \rho_1 f(\tilde{\mathbf{n}})d\tilde{\mathbf{n}} \\ &= \int_{\partial\mathbf{N}_1} [T(\mathbf{z}_1(\tilde{\mathbf{n}})) - T(\mathbf{z}_2(\tilde{\mathbf{n}}))] \frac{u_c(c(\mathbf{z}_1), \mathbf{z}_1; \tilde{\mathbf{n}})\tau(\mathbf{z}_1) - u_c(c(\mathbf{z}_2), \mathbf{z}_2; \tilde{\mathbf{n}})\tau(\mathbf{z}_2)}{||\nabla_{\mathbf{n}} u(c(\mathbf{z}_1), \mathbf{z}_1; \tilde{\mathbf{n}}) - \nabla_{\mathbf{n}} u(c(\mathbf{z}_2), \mathbf{z}_2; \tilde{\mathbf{n}})||} f(\tilde{\mathbf{n}})d\tilde{\mathbf{n}} \end{aligned} \tag{77}$$

Importantly, note that Equation 77 is *linear* in the tax perturbation $\tau(\mathbf{z})$; this is the key property

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

[44] 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.

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 75; 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 77.[45]

A.4.3 Individuals who Choose z with Non-smooth $T(\mathbf{z})$

Finally, we discuss individuals with a unique optima who choose $\mathbf{z}$ where $T(\mathbf{z})$ is not differentiable.[46] [47] Note that by the same arguments as for individuals with a single optima locating at $\mathbf{z}$ where $T(\mathbf{z})$ is differentiable, individuals with a single optima 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 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:

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

[45] 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 $\partial\mathbf{N}_1$.

[46] 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 optima who choose $\mathbf{z}$ where $T(\mathbf{z})$ is not differentiable have behavioral effects that are linear in $\tau(\mathbf{z})$.

[47] 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 A.4.2; if these individuals have a unique optima 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.

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 $\hat{\mathbf{Z}}$:

$$u_c(y(\mathbf{z}) - T(\mathbf{z}) - \epsilon\tau(\mathbf{z}), \mathbf{z}; \mathbf{n}) \left(y_\nu(\mathbf{z}) - T_\nu(\mathbf{z}) - \epsilon\tau_\nu(\mathbf{z})\right) + u_\nu(y(\mathbf{z}) - T(\mathbf{z}) - \epsilon\tau(\mathbf{z}), \mathbf{z}; \mathbf{n}) = 0 \ \forall \nu \in \vec{\nu} \quad (78)$$

$$u_c(y(\mathbf{z}) - T(\mathbf{z}) - \epsilon\tau(\mathbf{z}), \mathbf{z}; \mathbf{n}) \left(y_{\rho^+}(\mathbf{z}) - T_{\rho^+}(\mathbf{z}) - \epsilon\tau_{\rho^+}(\mathbf{z})\right) + u_{\rho^+}(y(\mathbf{z}) - T(\mathbf{z}) - \epsilon\tau(\mathbf{z}), \mathbf{z}; \mathbf{n}) \leq 0 \ \forall \rho \in \vec{\rho} \quad (79)$$

$$u_c(y(\mathbf{z}) - T(\mathbf{z}) - \epsilon\tau(\mathbf{z}), \mathbf{z}; \mathbf{n}) \left(y_{\rho^-}(\mathbf{z}) - T_{\rho^-}(\mathbf{z}) - \epsilon\tau_{\rho^-}(\mathbf{z})\right) + u_{\rho^-}(y(\mathbf{z}) - T(\mathbf{z}) - \epsilon\tau(\mathbf{z}), \mathbf{z}; \mathbf{n}) \geq 0 \ \forall \rho \in \vec{\rho} \quad (80)$$

Equations 78, 79, and 80 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 78 are satisfied and either 79 or Equation 80 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 $\mathbf{z}$ for whom $T(\mathbf{z})$ is non-differentiable and Equations 79 and 80 hold only weakly, Equations 79 and 80 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:

$$\nabla_{\mathbf{t}} u(y(\mathbf{t}) - T(\mathbf{t}) - \epsilon\tau(\mathbf{t}), \mathbf{t}; \mathbf{n}) = 0 \quad (81)$$

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 81 to derive:

$$\begin{aligned} \frac{\partial \mathbf{t}(\mathbf{n})}{\partial \epsilon} &= \mathbf{H_t}^{-1}(\mathbf{n}) FOC(\mathbf{n})_\epsilon|_{\epsilon=0} = \mathbf{H_t}^{-1}(\mathbf{n})[\mathbf{a_t}(\mathbf{n})\tau(\mathbf{t}) + \mathbf{B_t}(\mathbf{n}) \cdot \nabla_{\mathbf{t}}\tau(\mathbf{t})] \\ &= \vec{\eta}_{\mathbf{t}}(\mathbf{n})\tau(\mathbf{t}) + \mathbf{X_t}(\mathbf{n}) \cdot \nabla_{\mathbf{t}}\tau(\mathbf{t}) \end{aligned} \quad (82)$$

where $FOC(\mathbf{n})_\epsilon|_{\epsilon=0}$ is the vector of derivatives of the first order conditions 81 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 82 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 82 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:

$$\frac{\partial}{\partial \epsilon}\left[T(\mathbf{t}(\mathbf{n})) + \epsilon\tau(\mathbf{t}(\mathbf{n}))\right]|_{\epsilon=0} = \tau(\mathbf{t}(\mathbf{n})) + \nabla_{\mathbf{t}}T(\mathbf{t})\vec{\eta}_{\mathbf{t}}\tau(\mathbf{t}(\mathbf{n})) + \nabla_{\mathbf{t}}T(\mathbf{t})\mathbf{X_t}\cdot\nabla_{\mathbf{t}}\tau(\mathbf{t}(\mathbf{n})) \tag{83}$$

A.4.4 Gateaux Differentiability of $R(T)$

Putting all of this together, we need to plug the expressions from Equations 72, 77, and 83 into Equation 76. 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 derivative 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:

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

Integrating over $\mathbf{Z}$ we can write this as:[48]

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

where $\hat{h}(\mathbf{t})$ is the density of households choosing to locate at coordinates $\mathbf{t}$ on $\hat{\mathbf{Z}}$.[49] Now, as long as $q(\mathbf{z}) \equiv \int_{\mathbf{N}\setminus\hat{\mathbf{N}}}\nabla_{\mathbf{z}}T(\mathbf{z}(\mathbf{n}))\mathbf{X}(\mathbf{n})f(\mathbf{n}|\mathbf{z})d\mathbf{n}h(\mathbf{z})$ is sufficiently smooth (specifically, if each component of the vector valued function is in the Sobolev space $H^1(\mathbf{Z})$) then we can apply integration by parts:

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

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

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

[49] Note, we assumed the existence of $\hat{h}(\mathbf{z})$ 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 4.3.

Finally, suppose that on $\hat{\mathbf{Z}}$, $\hat{q}(\mathbf{t}) \equiv \int_{\hat{\mathbf{N}}} \nabla_{\mathbf{t}} T(\mathbf{t}) \mathbf{X}_{\mathbf{t}}(\mathbf{n}) f(\mathbf{n}|\mathbf{t}) d\mathbf{n} \hat{h}(\mathbf{t})$ is sufficiently smooth (specifically, if each component of the vector valued function is in the Sobolev space $H^1(\hat{\mathbf{Z}})$).[50] Then again we have that:

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

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 $m < J$.

Thus, we can write the Gateaux derivative of $R(T)$ as a linear functional of $\tau(\mathbf{z})$:

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

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. Hence, $R(T)$ is Gateaux differentiable.

□

A.5 Proof of Proposition 2

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 footnote 36 from

[50] 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*a*).

Appendix A.2 and consider the welfare functional:

$$W(U(\mathbf{n};T)) = \int_{\mathbf{Z}} \int_{\mathbf{N(z)}} \phi_1(\mathbf{n}) U(\mathbf{n};T) dF(\mathbf{n}|\mathbf{z}) d\Phi_2(\mathbf{z})$$

where we choose $\phi_1(\mathbf{n}) = \frac{1}{u_c(\mathbf{n})}$. This is an inverse welfare functional if we choose $\Phi_2(\mathbf{z}) = \Gamma(\mathbf{z})$ by the logic of Appendix A.2. 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 $W(U(\mathbf{n};T)) < 0$:

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

But then consider $\tau(\mathbf{z}) = \int_{\mathbf{N(z)}} \frac{\tilde{U}(\mathbf{n})}{u_c(\mathbf{n})} dF(\mathbf{n}|\mathbf{z})$, 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.6 Proof to Lemma 2

We have:

$$\begin{aligned}
&\frac{\partial R(T(z,s) + \epsilon\tau(z))}{\partial \epsilon}|_{\epsilon=0} = \int_N \frac{\partial}{\partial \epsilon} \left[T(z(n),s(n)) + \epsilon\tau(z(n))\right] f(n) dn \\
&= \int_N \Big(\left(1 + T_z(z(n),s(n))\eta_z(n) + T_s(z(n),s(n))\eta_s(n)\right) \tau(z(n)) \\
&+ \left(T_z(z(n),s(n))\xi_z^z(n) + T_s(z(n),s(n))\xi_z^s(n)\right) \tau_z(z(n)) \Big) f(n) dn \\
&= \int_Z \Big(\left(1 + T_z(z,s(z))\eta_z(z) + T_s(z,s(z))\eta_s(z)\right) \tau(z) + \left(T_z(z,s(z))\xi_z^z(z) + T_s(z,s(z))\xi_z^s(z)\right) \tau_z(z) \Big) h(z) dz \\
&= \int_Z \Big[\left(1 + T_z(z,s(z))\eta_z(z) + T_s(z,s(z))\eta_s(z)\right) h(z) \Big] - \frac{\partial}{\partial z} \left(\left[T_z(z,s(z))\xi_z^z(z) + T_s(z,s(z))\xi_z^s(z)\right] h(z) \right) \tau(z) dz
\end{aligned} \tag{84}$$

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:

$$\begin{aligned}
&\frac{\partial R(T(z,s)+\epsilon\tau(s))}{\partial \epsilon}|_{\epsilon=0} = \int_N \frac{\partial}{\partial \epsilon}\left[T(z(n),s(n))+\epsilon\tau(s(n))\right] f(n)dn \\
&= \int_N \Big(\left(1+T_z(z(n),s(n))\eta_z(n)+T_s(z(n),s(n))\eta_s(n)\right)\tau(s(n)) \\
&+ \left(T_z(z(n),s(n))\xi_s^z(n)+T_s(z(n),s(n))\xi_s^s(n)\right)\tau_s(s(n))\Big) f(n)dn \\
&= \int_Z \Big(\left(1+T_z(z,s(z))\eta_z(z)+T_s(z,s(z))\eta_s(z)\right)\tau(z) + \left(T_z(z,s(z))\xi_s^z(z)+T_s(z,s(z))\xi_s^s(z)\right)\tau_s(s(z))\Big) h(z)dz \\
&= \int_Z \Big(\left(1+T_z(z,s(z))\eta_z(z)+T_s(z,s(z))\eta_s(z)\right)\tau(z) + \left(T_z(z,s(z))\xi_s^z(z)+T_s(z,s(z))\xi_s^s(z)\right)\tau_s(s(z))\frac{ds}{dz}\left(\frac{ds}{dz}\right)^{-1}\Big) h(z)dz \\
&= \int_Z \left[\left[1+T_z(z,s(z))\eta_z(z)+T_s(z,s(z))\eta_s(z)\right]h(z) - \frac{\partial}{\partial z}\left(\left[T_z(z,s(z))\xi_s^z(z)+T_s(z,s(z))\xi_s^s(z)\right]\left(\frac{ds}{dz}\right)^{-1}h(z)\right)\right]\tau(z)dz
\end{aligned} \tag{85}$$

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 $W(U(n;T)) = \int_N \phi(n)U(n;T)dn$ (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 84 and 85 do not have mass points) and note that, given our assumption that all types have a unique optima, we can infer via the envelope theorem that:

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

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

Hence, in order to satisfy $\frac{\partial W(U(n;T(z,s)+\epsilon\tau(z)))+\lambda R(T(z,s)+\epsilon\tau(s))}{\partial \epsilon}|_{\epsilon=0} = 0$ (noting that we can normalize $\lambda = 1$), we must have that:

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

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

A.7 Details on Derivation of $\frac{\partial w}{\partial \epsilon}$ From Section 5.1

First, for completeness, the full expressions for $\frac{\partial z(n)}{\partial \epsilon}|_w$ and $\frac{\partial z(n)}{\partial w}|_\epsilon$ are provided below:

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

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

Next, plugging in $\frac{\partial z(n)}{\partial \epsilon}|_w \equiv \tau'(z(n))\xi(n)$ to Equaton 39 we have that:

$$\frac{\partial w}{\partial \epsilon}\left[L + w\frac{\partial L}{\partial w} - \int_N \left(\left.\frac{\partial z(n)}{\partial w}\right|_\epsilon\right) dF(n)\right] = \int_N \tau'(z(n))\xi(n)dF(n) \tag{86}$$

Doing a change of variables from n to z (where $h(z)$ represents the density of z) and applying integration by parts as in Equation 13, we find that (denoting $\overline{\frac{\partial z}{\partial w}}|_\epsilon \equiv \int_N \left(\frac{\partial z(n)}{\partial w}|_\epsilon\right) dF(n)$):

$$\frac{\partial w}{\partial \epsilon} = \frac{-\int_Z \frac{\partial[\xi(z)h(z)]}{\partial z}\tau(z)dz + \xi(z)h(z)\tau(z)\big|_{\underline{z}}^{\overline{z}}}{L + w\frac{\partial L}{\partial w} - \overline{\frac{\partial z}{\partial w}}|_\epsilon} \tag{87}$$

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 $p(z) = \frac{-\frac{\partial[\xi(z)h(z)]}{\partial z}}{L + w\frac{\partial L}{\partial w} - \overline{\frac{\partial z}{\partial w}}|_\epsilon}$.

A.8 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 positive linear functional $W(U(\mathbf{n}; T, \mathbf{w}))$ such that $T(\mathbf{z})$ is a stationary point for the Lagrangian $L(T; W)$. 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. Hence, let us show that there exists some functional which is continuous and linear on $C(\mathbf{N})$ that satisfies the statement of the Theorem. In particular, we will construct an inverse

welfare functional that takes the following form for some Borel measures Φ_1 and Φ_2:

$$W(U(\mathbf{n};T,\mathbf{w})) = \int_{\mathbf{Z}} \int_{\mathbf{N}(\mathbf{z})} U(\mathbf{n};T,\mathbf{w}) d\Phi_1(\mathbf{n}|\mathbf{z}) d\Phi_2(\mathbf{z}) \tag{88}$$

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:[51]

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

Choose Φ_1 as in the proof to Theorem 1 so that if $\mathcal{M}$ represents the set of individuals with multiple optima, then $\forall \mathbf{z}$, $\int_{\mathbf{N}(\mathbf{z})\cap\mathcal{M}} d\Phi_1(\mathbf{n}|\mathbf{z}) = 0$ (i.e., Φ_1-a.e. $\mathbf{n}$ have a unique optima). Hence, we can apply the envelope theorem to compute the Gateaux derivative of $W(U(\mathbf{n};T,\mathbf{w}))$, yielding the following expression for the Gateaux derivative of the Lagrangian:

$$\int_{\mathbf{Z}} \int_{\mathbf{N}} \left[-u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n})) + \sum_i u_{w_i}(\mathbf{n}) \int_{\mathbf{Z}} \tau(\mathbf{z}) dp_i(\mathbf{z}) \right] d\Phi_1(\mathbf{n}|\mathbf{z}) d\Phi_2(\mathbf{z}) + \lambda \int_{\mathbf{Z}} \tau(\mathbf{z}) d\Gamma(\mathbf{z}) \tag{89}$$

Given that we choose Φ_1 as in the proof to Theorem 1 we have that $\int_{\mathbf{N}(\mathbf{z})} u_c(\mathbf{n}) d\Phi_1(\mathbf{n}|\mathbf{z}) = \int_{\mathbf{N}(\mathbf{z})} \frac{u_c(\mathbf{n})}{u_c(\hat{\mathbf{n}})} d\delta_{\hat{\mathbf{n}}(\mathbf{z})}(\mathbf{n}) = \frac{u_c(\hat{\mathbf{n}})}{u_c(\hat{\mathbf{n}})} = 1$ (where $\hat{\mathbf{n}}(\mathbf{z})$ is a type with a unique optima at $\mathbf{z}$). Given this Φ_1, define $\overline{\frac{u_{w_i}}{u_c}}(\mathbf{z}) \equiv \int_{\mathbf{N}(\mathbf{z})} u_{w_i}(\mathbf{n}) d\Phi_1(\mathbf{n}|\mathbf{z}) = \frac{u_{w_i}(\hat{\mathbf{n}}(\mathbf{z}))}{u_c(\hat{\mathbf{n}}(\mathbf{z}))}$ so that Equation 89 can be rewritten:[52]

$$-\int_{\mathbf{Z}} \tau(\mathbf{z}) d\Phi_2(\mathbf{z}) + \sum_i \int_{\mathbf{Z}} \overline{\frac{u_{w_i}}{u_c}}(\mathbf{z}) d\Phi_2(\mathbf{z}) \int_{\mathbf{Z}} \tau(\mathbf{z}) dp_i(\mathbf{z}) + \lambda \int_{\mathbf{Z}} \tau(\mathbf{z}) d\Gamma(\mathbf{z}) \tag{90}$$

Or, changing the dummy variable of integration in $\int_{\mathbf{Z}} \overline{\frac{u_{w_i}}{u_c}}(\mathbf{z}) d\Phi_2(\mathbf{z})$ from $\mathbf{z}$ to $\tilde{\mathbf{z}}$, we have:

$$\int_{\mathbf{Z}} \tau(\mathbf{z}) \left(-d\Phi_2(\mathbf{z}) + \sum_i dp_i(\mathbf{z}) \int_{\mathbf{Z}} \overline{\frac{u_{w_i}}{u_c}}(\tilde{\mathbf{z}}) d\Phi_2(\tilde{\mathbf{z}}) + \lambda d\Gamma(\mathbf{z}) \right) \tag{91}$$

If the tax schedule $T(\mathbf{z})$ is a local extremum of the government's Lagrangian, then the Gateaux

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

[52] As in Theorem 1, if instead almost all $\mathbf{n}$ that choose each $\mathbf{z}$ have a unique optima, we could instead maximize the welfare functional:

$$W(U(\mathbf{n};T,\mathbf{w})) = \int_{\mathbf{Z}} \int_{\mathbf{N}(\mathbf{z})} \phi_1(\mathbf{n}) U(\mathbf{n};T,\mathbf{w}) dF(\mathbf{n}|\mathbf{z}) d\Phi_2(\mathbf{z})$$

where we choose $\phi_1(\mathbf{n}) = \frac{1}{u_c(\mathbf{n})}$. This would also lead us to rewrite the Gateaux derivative of the Lagrangian as in Equation 90 except $\overline{\frac{u_{w_i}}{u_c}}(\mathbf{z})$ would equal $\int_{\mathbf{N}(\mathbf{z})} \frac{u_{w_i}(\mathbf{n})}{u_c(\mathbf{n})} dF(\mathbf{n}|\mathbf{z})$.

derivative of L is zero. A sufficient condition for this is that for all measurable $\mathbf{E} \subseteq \mathbf{Z}$ we have:[53]

$$\int_{\mathbf{E}} \left(-d\Phi_2(\mathbf{z}) + \sum_i dp_i(\mathbf{z}) \int_{\mathbf{Z}} \overline{\frac{u_{w_i}}{u_c}}(\tilde{\mathbf{z}}) d\Phi_2(\tilde{\mathbf{z}}) + \lambda d\Gamma(\mathbf{z}) \right) = 0 \tag{92}$$

Or, expressing Equation 92 in terms of measures with $\Phi_2(E) \equiv \int_{\mathbf{E}} d\Phi_2(\mathbf{z})$, $\Gamma(E) \equiv \int_{\mathbf{E}} d\Gamma(\mathbf{z})$, and $p_i(E) \equiv \int_{\mathbf{E}} dp_i(\mathbf{z})$, we have (normalizing $\lambda = 1$):

$$\Phi_2(E) = \Gamma(E) + \sum_i p_i(E) \int_{\mathbf{Z}} \overline{\frac{u_{w_i}}{u_c}}(\tilde{\mathbf{z}}) d\Phi_2(\tilde{\mathbf{z}}) \tag{93}$$

which is an integral equation formulated in a measure space as in Das (1974) or Sharma (1975). As in Subsection 5.1, we are going to show that the map $(Q\Phi_2)(E) = \Gamma(E) + \sum_i p_i(E) \int_{\mathbf{Z}} \overline{\frac{u_{w_i}}{u_c}}(\tilde{\mathbf{z}}) d\Phi_2(\tilde{\mathbf{z}})$ 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):

$$||\mu||_{\text{TV}} = \sup_{||f||_\infty \leq 1} \int f d\mu \tag{94}$$

Thus, for two measures Φ_2 and Φ_2', consider the total variation norm of $(Q\Phi_2') - (Q\Phi_2)$:

$$\begin{aligned}
||(Q\Phi_2') - (Q\Phi_2)||_{\text{TV}} &= \left|\left| \sum_i p_i \int_{\mathbf{Z}} \overline{\frac{u_{w_i}}{u_c}}(\tilde{\mathbf{z}}) d(\Phi_2'(\tilde{\mathbf{z}}) - \Phi_2'(\tilde{\mathbf{z}})) \right|\right|_{\text{TV}} \\
&\leq \sum_i ||p_i||_{\text{TV}} \left| \int_{\mathbf{Z}} \overline{\frac{u_{w_i}}{u_c}}(\tilde{\mathbf{z}}) d(\Phi_2'(\tilde{\mathbf{z}}) - \Phi_2(\tilde{\mathbf{z}})) \right| \\
&= \sum_i ||p_i||_{\text{TV}} \left|\left| \overline{\frac{u_{w_i}}{u_c}} \right|\right|_\infty \left| \int_{\mathbf{Z}} \overline{\frac{u_{w_i}}{u_c}}(\tilde{\mathbf{z}}) \Big/ \left|\left| \overline{\frac{u_{w_i}}{u_c}} \right|\right|_\infty d(\Phi_2'(\tilde{\mathbf{z}}) - \Phi_2(\tilde{\mathbf{z}})) \right| \\
&\leq \sum_i ||p_i||_{\text{TV}} \left|\left| \overline{\frac{u_{w_i}}{u_c}} \right|\right|_\infty ||\Phi_2' - \Phi_2||_{\text{TV}} \\
&< ||\Phi_2' - \Phi_2||_{\text{TV}}
\end{aligned} \tag{95}$$

Let us explain the steps detailed in Equation 95. The first line simply uses the definition of the measure $(Q\Phi_2') - (Q\Phi_2)$ from Equation 93. The second line uses the triangle inequality and the absolute homogeneity of the norm. The third line just multiplies and divides by $\left|\left| \overline{\frac{u_{w_i}}{u_c}} \right|\right|_\infty$ (recognize that $\overline{\frac{u_{w_i}}{u_c}}$ 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 94. The final line uses our assumption on the size of $\sum_i ||p_i||_{\text{TV}} \left|\left| \overline{\frac{u_{w_i}}{u_c}} \right|\right|_\infty$. Hence, $(Q\Phi_2)(E)$ is a contraction mapping, which implies the existience of a (unique) fixed point Φ_2 which solves Equation 93. Hence, we have proved the existence of an inverse welfare functional taking the form of Equation 88. Finally, note that the inverse welfare functional with the given Φ_1 and Φ_2 can be shown to be continuous and linear

[53] The condition in Equation 92 shows that $\left(-d\Phi_2(\mathbf{z}) + \int_{\mathbf{Z}} \sum_i \overline{\frac{u_{w_i}}{u_c}}(\tilde{\mathbf{z}}) d\Phi_2(\tilde{\mathbf{z}}) dp_i(\mathbf{z}) + \lambda d\Gamma(\mathbf{z}) \right)$ is zero a.e.; hence, integrating $\tau(\mathbf{z})$ against this measure over $\mathbf{Z}$ for any $\tau(\mathbf{z})$ is zero.

using the same arguments as at the Appendix A.2.

□

B Appendix: Additional Results

B.1 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):

$$u(z(\hat{n}(v)) - T(z(\hat{n}(v))) - \epsilon\tau(z(\hat{n}(v))), z(\hat{n}(v))/\hat{n}(v)) - v = u(-T(0) - \epsilon\tau(0), 0) \tag{96}$$

We have that (note we have dropped the v argument from $z(n, v)$ for those who choose to work because their choice of z is not dependent on v conditional on working a positive amount):

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

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 for all types that choose to work:

$$\left(1 - T'(z) - \epsilon\tau'(z)\right) u_1(z - T(z), z/n) - \frac{1}{n} u_2(z - T(z), z/n) = 0 \tag{97}$$

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

$$\frac{\partial z}{\partial \epsilon}(n, v) = \frac{-u_1\tau'(z) + \left[u_{11}(1 - T'(z)) + \frac{1}{n}u_{12}\right]\tau(z)}{u_{11}(1 - T'(z)^2) + \frac{2}{n}u_{12} + \frac{1}{n^2}u_{22} - T''(z)u_1} \equiv \xi(n, v)\tau'(z(n, v)) + \eta(n, v)\tau(z(n, v)) \tag{98}$$

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:

$$\begin{aligned}
&\lim_{\epsilon\to 0}\frac{R(T+\epsilon\tau)-R(T)}{\epsilon}\\
&=\int_V\int_{\hat{n}(v)}^{\overline{n}}\left[\frac{T(z(n))}{\partial\epsilon}+\tau(z(n,v))\right]f(n,v)dndv+\int_V\int_{\underline{n}}^{\hat{n}(v)}\tau(0)f(n,v)dndv\\
&+\int_V\left[T(0)-T(z(\hat{n}(v)))\right]f(\hat{n}(v)|v)\frac{\partial\hat{n}(v)}{\partial\epsilon}f(v)dv
\end{aligned}\tag{99}$$

We can also calculate how the indifferent individual changes with the tax schedule by applying the implicit function theorem to Equation 96 (and evaluating at $\epsilon=0$):

$$\frac{\partial\hat{n}(v)}{\partial\epsilon}=\frac{u_1(-T(0),0)\tau(0)-u_1(z(\hat{n}(v))-T(z(\hat{n}(v))),z(\hat{n}(v))/\hat{n}(v))\tau(z(\hat{n}(v)))}{u_2(z(\hat{n}(v))-T(z(\hat{n}(v))),z(\hat{n}(v))/\hat{n}(v))\frac{z(\hat{n}(v))}{\hat{n}(v)^2}}\tag{100}$$

Plugging in Equations 98 and 100 into Equation 99 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 100 for readability):

$$\begin{aligned}
&\lim_{\epsilon\to 0}\frac{R(T+\epsilon\tau)-R(T)}{\epsilon}\\
&=\int_V\int_{\hat{n}(v)}^{\overline{n}}\left[T'(z(n,v))\xi(n,v)\tau'(z(n,v))+\left(1+T'(z(n,v))\eta(n,v)\right)\tau(z(n,v))\right]f(n,v)dndv\\
&+M(0)\tau(0)+\int_V\left[T(0)-T(z(\hat{n}(v),v))\right]f(\hat{n}(v)|v)\frac{u_1(0)\tau(0)-u_1(z(\hat{n}(v)))\tau(z(\hat{n}(v)))}{u_2(z(\hat{n}(v)))\frac{z(\hat{n}(v))}{\hat{n}(v)^2}}f(v)dv
\end{aligned}\tag{101}$$

Changing the variable of integration for the first integral on the RHS of Equation 101 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:

$$\begin{aligned}
&\int_V\int_{\hat{n}(v)}^{\overline{n}}\left[T'(z(n,v))\xi(n,v)\tau'(z(n,v))+\left(1+T'(z(n,v))\eta(n,v)\right)\tau(z(n,v))\right]f(n,v)dndv\\
&=\int_V\int_{z(\hat{n}(v))}^{z(\overline{n})}\left[T'(z)\xi(z,v)\tau'(z)+\left(1+T'(z)\eta(z,v)\right)\tau(z)\right]h(z,v)dzdv\\
&=\int_{\underline{z}}^{\overline{z}}\left[T'(z)\overline{\xi}(z)\tau'(z)+\left(1+T'(z)\overline{\eta}(z)\right)\tau(z)\right]h(z)dz\\
&=\int_{\underline{z}}^{\overline{z}}\left(-\frac{\partial}{\partial z}\left[T'(z)\overline{\xi}(z)h(z)\right]+\left[1+T'(z)\overline{\eta}(z)\right]h(z)\right)\tau(z)dz+T'(z)\overline{\xi}(z)h(z)\tau(z)\Big|_{\underline{z}}^{\overline{z}}
\end{aligned}\tag{102}$$

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 101 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:[54]

$$\begin{aligned}&\int_V \left[T(0) - T(z(\hat{n}(v), v))\right] f(\hat{n}(v)|v) \frac{u_1(0)\tau(0) - u_1(z(\hat{n}(v)))\tau(z(\hat{n}(v)))}{u_2(z(\hat{n}(v)))\frac{z(\hat{n}(v))}{\hat{n}(v)^2}} f(v)dv \\ &= \int_Z \left[T(0) - T(z)\right] \frac{u_1(0)\tau(0) - u_1(z)\tau(z)}{u_2(z)\frac{z}{\hat{n}(z)^2}} h(\hat{n}(z), z)dz\end{aligned} \tag{103}$$

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:

$$\begin{aligned}&\lim_{\epsilon \to 0} \frac{R(T + \epsilon\tau) - R(T)}{\epsilon} \\ &= \int_{\underline{z}}^{\overline{z}} \left(-\frac{\partial}{\partial z}\left[T'(z)\overline{\xi}(z)h(z)\right] + \left[1 + T'(z)\overline{\eta}(z)\right] h(z) \right) \tau(z)dz + T'(\overline{z})\overline{\xi}(\overline{z})h(\overline{z})\tau(\overline{z}) \\ &+ M(0)\tau(0) + \int_Z \left[T(0) - T(z)\right] \frac{u_1(0)\tau(0) - u_1(z)\tau(z)}{u_2(z)\frac{z}{\hat{n}(z)^2}} h(\hat{n}(z), z)dz\end{aligned} \tag{104}$$

From here, suppose welfare is given by:

$$\begin{aligned}&\iint_{N \times V} \phi(n, v)U(n, v; T)f(n, v)dndv + \int_V \overline{\phi}(v)U(n(\overline{z}), v; T)f(v|\overline{z})dv \\ &= \int_Z \iint_{N \times V} \phi(n, v)U(n, v; T)f(n, v|z)dndvdH(z) + \int_V \overline{\phi}(v)U(n(\overline{z}), v; T)f(v|\overline{z})dv\end{aligned} \tag{105}$$

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 105 equals:

$$-\int_Z \iint_{N \times V} \phi(n, v)u_c(n, v; T)\tau(z)f(n, v|z)dndvdH(z) - \int_V \overline{\phi}(v)\tau(\overline{z})u_c(n(\overline{z}), v; T)f(v|\overline{z})dv \tag{106}$$

We want to choose welfare weights such that for all $\tau(z)$, Equation 106 plus Equation 104 equals 0. Hence, we pick $\phi(n, v)$ for all of those who do not work to satisfy:

$$\iint_{N \times V} \phi(n, v)u_c(n, v; T)\tau(z)f(n, v|0)dndvM(0) = M(0)\tau(0) + \int_Z \left[T(0) - T(z)\right] \frac{u_1(0)}{u_2(z)\frac{z}{\hat{n}(z)^2}} h(\hat{n}(z), z)dz \tag{107}$$

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

$$\begin{aligned}&\iint_{N \times V} \phi(n, v)u_c(n, v; T)\tau(z)f(n, v|z)dndvh(z) \\ &= -\frac{\partial}{\partial z}\left[T'(z)\overline{\xi}(z)h(z)\right] + \left[1 + T'(z)\overline{\eta}(z)\right] h(z) + \left[T(0) - T(z)\right] \frac{-u_1(z)}{u_2(z)\frac{z}{\hat{n}(z)^2}} h(\hat{n}(z), z)\end{aligned} \tag{108}$$

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

$$\int_V \overline{\phi}(v)U(n(\overline{z}), v; T)f(v|\overline{z})dv = T'(\overline{z})\overline{\xi}(\overline{z})h(\overline{z}) \tag{109}$$

Choosing $\phi(n, v)$ and $\overline{\phi}(v)$ to satisfy the previous three equations ensures that any perturbation

[54] 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 $u_2(z) = u_2(z - T(z), z/\hat{n}(z))$.

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.2 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:

$$\int_{\underline{n}}^{n_1} T(z(n))f(n)dn + \int_{n_1}^{\overline{n}} T(z(n))f(n)dn \tag{110}$$

We have the individual first order condition:

$$\left(1 - T'(z) - \epsilon\tau'(z)\right) - \frac{1}{n}\left(\frac{z}{n}\right)^k = 0 \tag{111}$$

For all individuals with a unique optima where the tax schedule is twice continuously differentiable the second order condition holds strictly (see Lemma 3 of Bergstrom and Dodds (2021*a*)), 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):

$$\frac{\partial z}{\partial \epsilon}(n) = -\frac{\tau'(z)}{\frac{1}{n^2}\left(\frac{z}{n}\right)^{k-1}} \equiv \xi(n)\tau'(z(n)) \tag{112}$$

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:

$$z^- T(z^-) - \epsilon\tau(z^-) - (z^-/n_1)^{1+k}/(1+k) = z^+ T(z^+) - \epsilon\tau(z^+) - (z^+/n_1)^{1+k}/(1+k) \quad (113)$$

We can also calculate how the indifferent individual changes with the tax schedule by applying the implicit function theorem to Equation 113:

$$\frac{\partial n_1}{\partial \epsilon} = \frac{\tau(z^+) - \tau(z^-)}{\frac{1}{n_1}\left(\frac{z^+}{n_1}\right)^{1+k} - \frac{1}{n_1}\left(\frac{z^-}{n_1}\right)^{1+k}} \quad (114)$$

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

$$\begin{aligned}
&\lim_{\epsilon\to 0} \frac{R(T+\epsilon\tau) - R(T)}{\epsilon} \\
&= \int_{\underline{n}}^{n_1} \left[\frac{T(z(n))}{\partial\epsilon} + \tau(z(n))\right] f(n)dn + \int_{n_1}^{\overline{n}} \left[\frac{T(z(n))}{\partial\epsilon} + \tau(z(n))\right] f(n)dn + \left(T(z^-) - T(z^+)\right) f(n_1)\frac{\partial n_1}{\partial\epsilon}
\end{aligned} \quad (115)$$

Note, the last term of Equation 115 results from applying Leibniz integral rule. Plugging in the value of $\frac{\partial z}{\partial\epsilon}(n)$ from the implicit function theorem (Equation 112) and changing the variable of integration from n to z we find that:[55]

$$\begin{aligned}
&\lim_{\epsilon\to 0} \frac{R(T+\epsilon\tau) - R(T)}{\epsilon} \\
&= \int_{z(\underline{n})}^{z^-} \left(T'(z)\xi(z)\tau'(z) + \tau(z)\right) h(z)dz + \int_{z^+}^{z(\overline{n})} \left(T'(z)\xi(z)\tau'(z) + \tau(z)\right) h(z)dz \\
&+ \left(T(z^-) - T(z^+)\right) f(n_1) \frac{\tau(z^+) - \tau(z^-)}{\frac{1}{n_1}\left(\frac{z^+}{n_1}\right)^{1+k} - \frac{1}{n_1}\left(\frac{z^-}{n_1}\right)^{1+k}}
\end{aligned} \quad (116)$$

Next, let us apply integration by parts to get rid of the $\tau'(z)$ terms in Equation 116:

$$\begin{aligned}
&\lim_{\epsilon\to 0} \frac{R(T+\epsilon\tau) - R(T)}{\epsilon} \\
&= \int_{z(\underline{n})}^{z^-} \left(h(z) - \frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right]\right)\tau(z)dz + \int_{z^+}^{z(\overline{n})} \left(h(z) - \frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right]\right)\tau(z)dz \\
&+ T'(z)\xi(z)\tau(z)|_{z(\underline{n})}^{z^-} + T'(z)\xi(z)\tau(z)|_{z^+}^{z(\overline{n})} \\
&+ \left(T(z^-) - T(z^+)\right) f(n_1) \frac{\tau(z^+) - \tau(z^-)}{\frac{1}{n_1}\left(\frac{z^+}{n_1}\right)^{1+k} - \frac{1}{n_1}\left(\frac{z^-}{n_1}\right)^{1+k}}
\end{aligned} \quad (117)$$

Note that all $\tau(z)$ terms enter Equation 117 linearly so that Equation 117 is a linear functional of $\tau(z)$ which means that $R(T)$ is Gateaux differentiable (assuming that all terms in Equation 117 are bounded so that Equation 117 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

[55] 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.

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 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 117 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:

$$\frac{-\left(T(z^-)-T(z^+)\right)f(n_1)}{\frac{1}{n_1}\left(\frac{z^+}{n_1}\right)^{1+k}-\frac{1}{n_1}\left(\frac{z^-}{n_1}\right)^{1+k}}+T'(z^-)\xi(z^-)=0,\ \frac{\left(T(z^-)-T(z^+)\right)f(n_1)}{\frac{1}{n_1}\left(\frac{z^+}{n_1}\right)^{1+k}-\frac{1}{n_1}\left(\frac{z^-}{n_1}\right)^{1+k}}-T'(z^+)\xi(z^+)=0 \tag{118}$$

so that Equation 117 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 118 typically does not hold. For example, we compute the values in Equation 118 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:

$$\frac{-\left(T(z^-)-T(z^+)\right)f(n_1)}{\frac{1}{n_1}\left(\frac{z^+}{n_1}\right)^{1+k}-\frac{1}{n_1}\left(\frac{z^-}{n_1}\right)^{1+k}}+T'(z^-)\xi(z^-)=-0.0534$$
$$\frac{\left(T(z^-)-T(z^+)\right)f(n_1)}{\frac{1}{n_1}\left(\frac{z^+}{n_1}\right)^{1+k}-\frac{1}{n_1}\left(\frac{z^-}{n_1}\right)^{1+k}}-T'(z^+)\xi(z^+)=-0.0391$$

C Appendix: Approximating Optimal Schedules

In this Appendix, we will discuss a number of technical results needed to solve Problem 53. In particular, how can we identify tax schedules T_i which have inverse welfare functionals $W^{Inv}_{T_i}$ that are "close enough" to W^*? More precisely, how can we find a sequence of increasingly flexible function classes, $\{\mathcal{T}_i\}$ to ensure $W^{Inv}_{T_i} \to W^*$ so that for any sufficiently flexible function class $\mathcal{T}_i$, the inverse weights associated to the optimal schedule within $\mathcal{T}_i$ are arbitrarily close to

W^*? First, let us suppose that some optimal tax schedule exists so that the following problem has a solution:[56]

$$\max_T W^*(U(\mathbf{n};T) \text{ s.t. } \int_{\mathbf{N}} T(\mathbf{z}(\mathbf{n}))dF(\mathbf{n}) \geq E \tag{119}$$

It will be helpful to establish conditions under which the tax schedule can be assumed continuous:

Lemma 3. *Suppose that given a $T(\mathbf{z})$, the set of choices, $\mathbf{Z} = \{\mathbf{z}(\mathbf{n})\}$, is bounded. Further, suppose that indifference surfaces have bounded gradients:*

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

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 (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

[56] Existence of solutions to optimal tax problems is typically not straight-forward to establish. Appendix B.2 of Dodds (2023) proves an existence result by leveraging coercivity of the budget constraint along the lines of classic existence results a la Kinderlehrer and Stampacchia (1980).

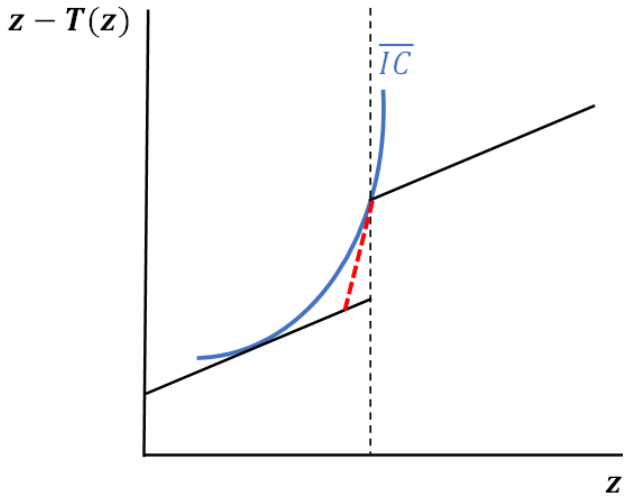


Figure 8: Alternative Continuous Tax Schedule for a Discontinuous Tax Schedule

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.

can restrict attention to Lipschitz continuous optimal tax schedules. □

The intuition of Lemma 3 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 $\overline{IC}$.

Remark 6. *Let $T^*(\mathbf{z})$ be the solution to Equation 119. Under the conditions of Lemma 3, $\exists$ a sequence of polynomial tax schedules, $\{T_i(\mathbf{z})\}$, such that $T_i(\mathbf{z}) \to T^*(\mathbf{z})$ uniformly. Alternatively, $\exists$ a sequence of piecewise linear tax schedules, $\{T_i(\mathbf{z})\}$, such that $T_i(\mathbf{z}) \to T^*(\mathbf{z})$ uniformly.*

Proof. The first statement follows immediately from the Stone-Weierstrass Theorem.

For the second statement, this is presumably a standard fact, but we struggled to find a proof to reference. First triangulate $\mathbf{Z}$ (the region of chosen $\mathbf{z}$'s). For each triangle in $\mathbf{Z}$ defined by three points, consider the plane connecting these points parametrized by $a + \mathbf{b} \cdot \mathbf{z}$ for some scalar a and a vector $\mathbf{b}$; this defines a piecewise linear tax schedule which agrees with $T^*(\mathbf{z})$ on the three vertices of each triangle region. Next, consider an arbitrary ϵ; there exists a sufficiently fine triangularization such that for any $\mathbf{z}, \mathbf{z}'$ within a given triangle, $||\mathbf{z} - \mathbf{z}'|| \leq \frac{\epsilon}{2M}$, where

M is the Lipschitz constant for $T^*(\mathbf{z})$. Moreover, by the Lipschitz continuity of $T^*(\mathbf{z})$, which implies uniform continuity, there exists a sufficiently small δ such that $||\mathbf{z}' - \mathbf{z}|| \leq \delta$ implies $||T^*(\mathbf{z}) - T^*(\mathbf{z}')|| \leq \epsilon/2$. Finally, for sufficiently fine triangularizations (i.e., for any $\mathbf{z}, \mathbf{z}'$ within a given triangle, $||\mathbf{z} - \mathbf{z}'|| \leq \min\{\frac{\epsilon}{2M}, \delta\}$) we thereby have:

$$\begin{aligned} &||T(\mathbf{z}) - (a + \mathbf{b} \cdot \mathbf{z})|| = ||T(\mathbf{z}) - T(\mathbf{z}_1) + (a + \mathbf{b} \cdot \mathbf{z}_1) - (a + \mathbf{b} \cdot \mathbf{z})|| \\ &\leq ||T(\mathbf{z}) - T(\mathbf{z}_1)|| + ||\mathbf{b} \cdot (\mathbf{z}_1 - \mathbf{z})|| \leq \epsilon/2 + ||\mathbf{b}||\frac{\epsilon}{2M} \leq \epsilon/2 + M\frac{\epsilon}{2M} = \epsilon \end{aligned} \tag{120}$$

where z_1 is a vertex of the triangle containing $\mathbf{z}$ and we have used the fact that $||\mathbf{b}|| \leq M$ because the norm of the gradient of the hyperplane going through any three points of a Lipschitz function must be (definitionally) bounded from above by the Lipschitz constant. □

The next step is to show conditions under which not only can we find a sequence of polynomial (or piecewise linear) tax schedules $\{T_i\} \to T^*(\mathbf{z})$, but that these tax schedules generate local inverse welfare functionals that converge in some sense to the welfare functional W^* that we want to maximize.

First, we should briefly mention how we construct inverse welfare functionals when there is *not* a unique local welfare functional that supports the optimal utility profile (this will almost always occur when $\dim(\mathbf{N}) > \dim(\mathbf{Z})$ so that $\mathbf{n} \mapsto \mathbf{z}$ is not bijective). In this case, if we want $W^{Inv}_{T_i}$ to converge to W^*, we need to carefully choose a sequence of inverse welfare weight functionals. Suppose that $W^* = \int_{\mathbf{N}} \phi^*(\mathbf{n})U(\mathbf{n})f(\mathbf{n})d\mathbf{n}$ and the Gateaux derivative of government revenue under each tax schedule T_i can be written as a sum over some partition $\{\mathbf{Z}_i\}$ with $\mathbf{Z}_1 \cup \mathbf{Z}_2 \cup \cdots \cup \mathbf{Z}_M = \mathbf{Z}$ as in Equation 5 (reproduced below):

$$\sum_j \int_{\mathbf{Z}_j} \tau(\mathbf{z})\gamma(\mathbf{z}; T_i)d\mathbf{Z}_j$$

If almost all $\mathbf{n}$ locating at each $\mathbf{z}$ have a unique optima under each T_i, then we can compute an inverse welfare functional via Theorem 1 of the form:

$$W^{Inv}_{T_i}(U(\mathbf{n}; T_i)) = \sum_j \int_{\mathbf{Z}_j} \int_{\mathbf{N}(\mathbf{z})} q(\mathbf{z}; T_i)\phi^*(\mathbf{n})U(\mathbf{n}; T_i)dF(\mathbf{n}|\mathbf{z})\tau(\mathbf{z})d\mathbf{Z}_j \tag{121}$$

The Gateaux derivative of the government's Lagrangian therefore equals:

$$-\sum_j \int_{\mathbf{Z}_j} \int_{\mathbf{N}(\mathbf{z})} q(\mathbf{z}; T_i)\phi^*(\mathbf{n})u_c(\mathbf{n})dF(\mathbf{n}|\mathbf{z})\tau(\mathbf{z})d\mathbf{Z}_j + \lambda \sum_j \int_{\mathbf{Z}_j} \tau(\mathbf{z})\gamma(\mathbf{z}; T_i)d\mathbf{Z}_j \tag{122}$$

Thus, we can recover the inverse welfare functional by choosing $q(\mathbf{z}; T_i)$ to satisfy for every $\mathbf{z}$:

$$\left[\int_{\mathbf{N}(\mathbf{z})} \phi^*(\mathbf{n})u_c(\mathbf{n})dF(\mathbf{n}|\mathbf{z})\right] q(\mathbf{z}; T_i) = \gamma(\mathbf{z}; T_i)$$

At any $\mathbf{z}$ where $\gamma(\mathbf{z}; T)$ is continuous in the tax schedule around T^*, then $\gamma(\mathbf{z}; T_i) \to \gamma(\mathbf{z}; T^*)$ so that as long as $\int_{\mathbf{N}(\mathbf{z})} \phi(\mathbf{n})u_c(\mathbf{n}; T_i)dF(\mathbf{n}|\mathbf{z}) \to \int_{\mathbf{N}(\mathbf{z})} \phi(\mathbf{n})u_c(\mathbf{n}; T^*)dF(\mathbf{n}|\mathbf{z})$ then $q(\mathbf{z}; T_i) \to 1$ as

$T_i \to T^*$. Hence, at all such $\mathbf{z}$, inverse welfare weights $q(\mathbf{z}; T_i)\phi^*(\mathbf{n})$ converge to $\phi^*(\mathbf{n})$. In this sense, we can guarantee a sort-of pointwise convergence of the inverse welfare functionals $W^{Inv}_{T_i}$ to W^* wherever $\gamma(\mathbf{z}; T)$ is continuous in the tax schedule even when $\mathbf{n} \mapsto \mathbf{z}$ is not bijective.

If we are satisfied instead with finding only a generalized marginal inverse functional, we can establish the following convergence result:

Proposition 5. *Suppose that T^* is continuous and that $R(T)$ is continuously Gateaux differentiable around T^* so that the mapping $T \mapsto DR_T$ is a continuous mapping from $C(\mathbf{Z}) \to \mathcal{L}(C(\mathbf{Z}))$ where $C(\mathbf{Z})$ is the set of continuous functions on $\mathbf{Z}$ and $\mathcal{L}(C(\mathbf{Z}))$ denotes the set of continuous linear functionals on $C(\mathbf{Z})$ equipped with the dual norm so that for $L \in \mathcal{L}$, $||L||_D = \sup\{|L(T)| : ||T||_\infty \leq 1\}$. Then if $T_i(\mathbf{z}) \to T^*(\mathbf{z})$ uniformly, $R(T_i) = E$ $\forall i$, and $\mathbf{Z}$ (the chosen set of $\mathbf{z}$) is compact $\forall i$, then the generalized marginal inverse functionals converge uniformly in the dual norm as well.*

Proof. By the definition of continuity, $\forall \delta$ $\exists \epsilon$ s.t. $||T_i(\mathbf{z}) - T^*(\mathbf{z})||_\infty < \epsilon \implies ||DR_{T_i} - DR_{T^*}||_D < \delta$. Thus, by Theorem 3, we know that $||T_i(\mathbf{z}) - T^*(\mathbf{z})||_\infty < \epsilon \implies ||G_{T_i} - G_{T^*}||_D < \delta$, where G_T is a generalized marginal inverse functional for tax schedule T. □

Note, Proposition 5 ensures that the generalized marginal inverse functional converge as long as $R(T)$ is continuously Gateaux differentiable around T^*. Loosely speaking, continuous Gateaux differentiability requires that average income and substitution effects, $\overline{\vec{\eta}}(\mathbf{z})$ and $\overline{\mathbf{X}}(\mathbf{z})$, are smooth as a function of the tax schedule.

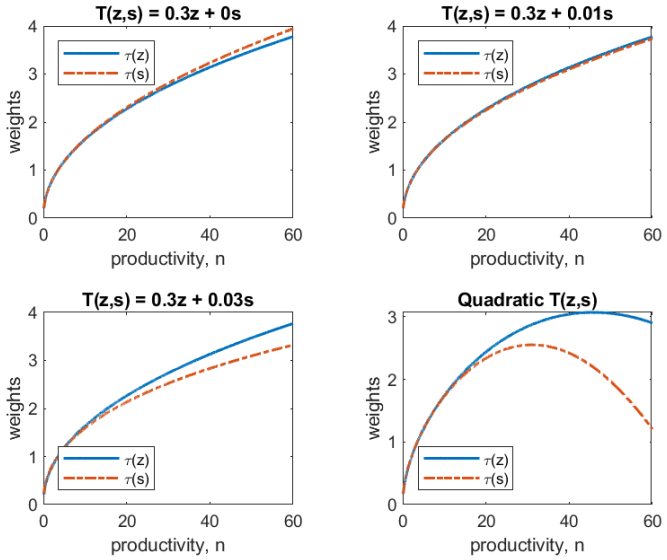


Figure 9: Inverse Weights for $\tau(z)$ and $\tau(s)$ Perturbations: Non-Separable Utility Function

Note: This figure shows the inverse welfare weights that satisfy Equation 32 in blue solid lines (i.e., ensure that the Gateaux variation of any income tax perturbations $\tau(z)$ is zero) and shows the inverse welfare weights that satisfy Equation 33 in orange dashed lines (i.e., ensure that the Gateaux variation of any savings tax perturbations $\tau(s)$ 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 $u(c,s,z/n) = \frac{c^{1-\alpha}}{1-\alpha} + \beta(n)\frac{s^{1-\alpha}}{1-\alpha} + \frac{(z/n)^{1+k}}{1+k}$ 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 3 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 0.01% (0.03%, respectively) savings tax is equivalent to a 20% (60%, respectively) tax on interest income.

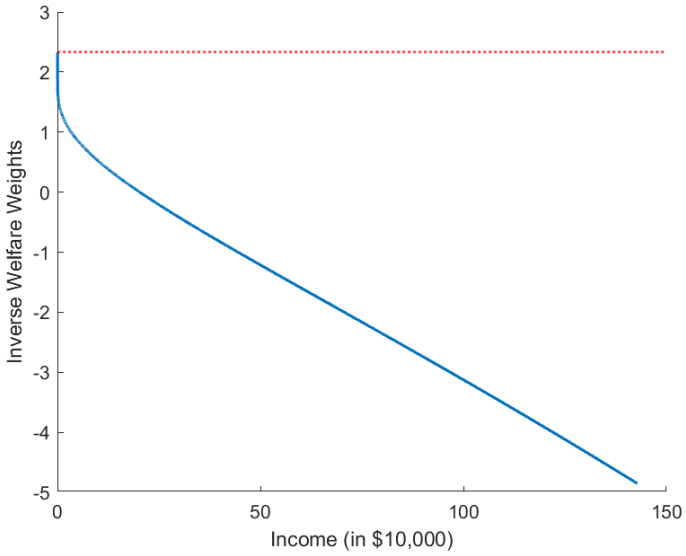


Figure 10: Inverse Welfare Weights that Solve System 48 Given Initial Value at $\underline{z}$

Note: This figure shows the inverse welfare weights that satisfy the differential equation in System 48 given the initial condition $\psi(n(\underline{z})) = \frac{T_z}{[(1-T_z)-(1-\hat{T}_z)]}$ with $k = 1/0.3$, $f(n)$ calibrated to match the U.S. income distribution from the 2019 ACS, and $(1-T_z) = 0.7$, $(1-\hat{T}_z) = 0.85$. The solution to this differential equation with this given initial condition does not satisfy System 48 because $\psi(n(\overline{z})) \neq \frac{T_z}{[(1-T_z)-(1-\hat{T}_z)]}$.

D Appendix: Simulations

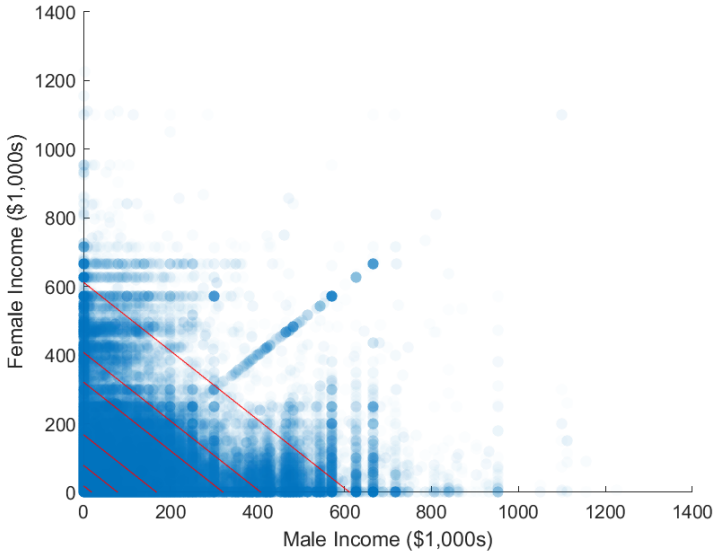


Figure 11: Couples Income Distribution from the 2019 ACS

Note: This figure shows a scatter plot of the joint distribution of incomes for heterosexual couples in the 2019 American Community Survey. Survey weights are indicated by the transparency of the data points. The red diagonal lines indicate joint income levels where marginal tax rates change.

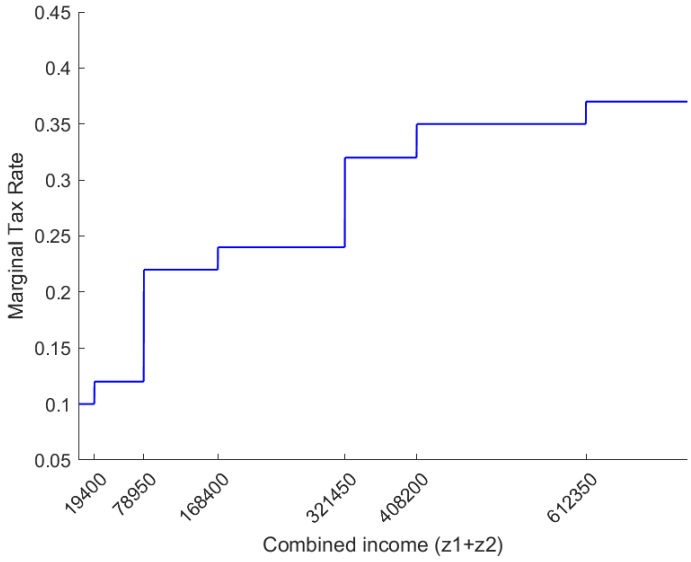


Figure 12: Federal Couple's Income Tax Schedule 2019

Note: This figure shows the federal income tax schedule for couples in 2019.

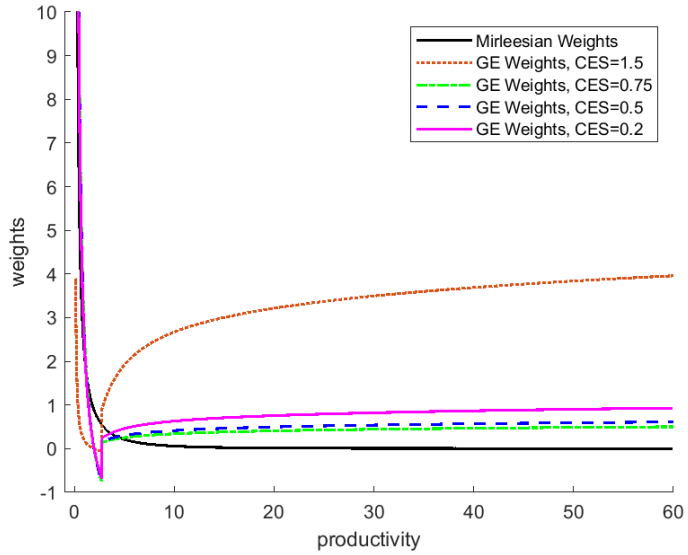


Figure 13: Inverse Welfare Weights with Finite Labor Demand Elasticity and CES Production

Note: This figure shows the inverse welfare weights for a particular 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 $v = 1/2$, and 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. We begin with a set of welfare weights, the "Mirrleesian Weights", and compute the optimal tax schedule assuming that labor demand is infinitely elastic as in Mirrlees (1971) or Saez (2001). We assume a skill distribution calibrated to the U.S. income distribution using the 2019 ACS and a labor supply elasticity of 0.3. Finally, we plot the inverse welfare weights that support this tax schedule under various assumptions about the value of the elasticity of substitution between L_l and L_h: $\frac{1}{1-\sigma}$.

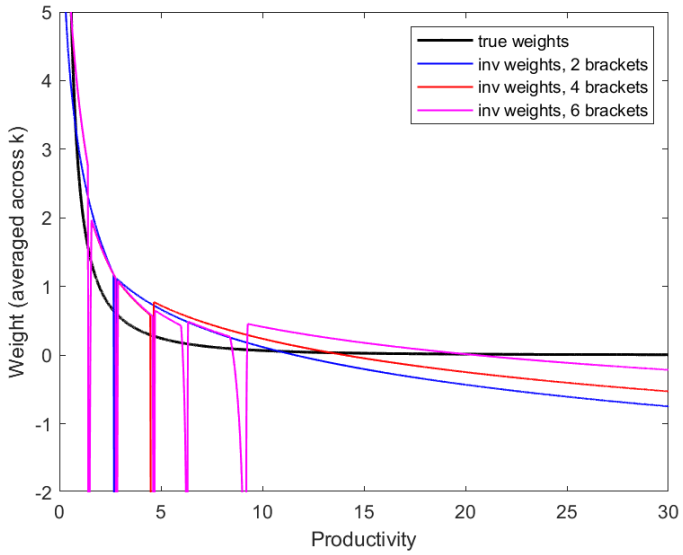


Figure 14: Average Inverse Welfare Weights for Piecewise Linear Tax Schedules

Note: This figure shows inverse welfare weights for the optimal two bracket tax system with a kink at \$25000, the optimal four bracket tax system with kink points in $[\$10000, \$25000, \$50000]$ and the optimal six bracket tax system with kink points in $[\$10000, \$25000, \$50000, \$75000, \$125000]$ to maximize the welfare functional depicted by the line "true weights". We plot inverse weights averaged over all k across the n distribution when utility is given by Equation 54 and the distribution of n is calibrated to match the U.S. income distribution in 2019 from the ACS; k is uniformly distributed in $[1/0.35, 1/0.25]$, which implies that taxable income elasticities are between 0.25 and 0.35. Because many types choose each given z, inverse weights for each type (n, k) are computed as discussed in Appendix C to ensure pointwise convergence (precisely, we plot $q(\mathbf{z}; T_i)\phi^*(\mathbf{n})$ from Equation 121).

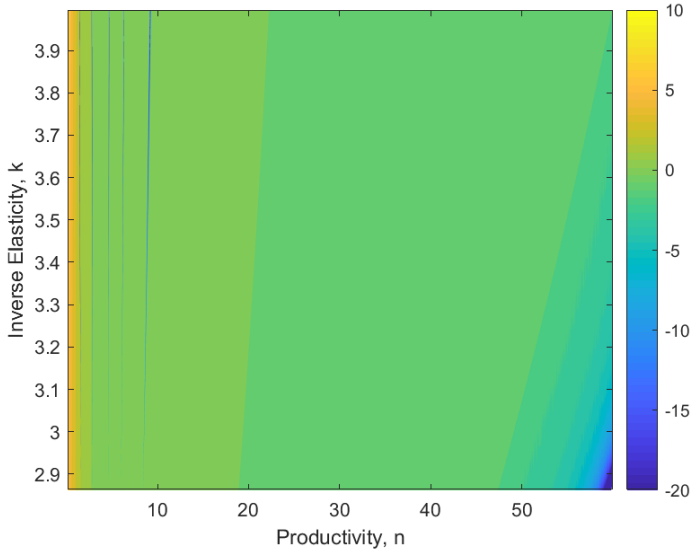
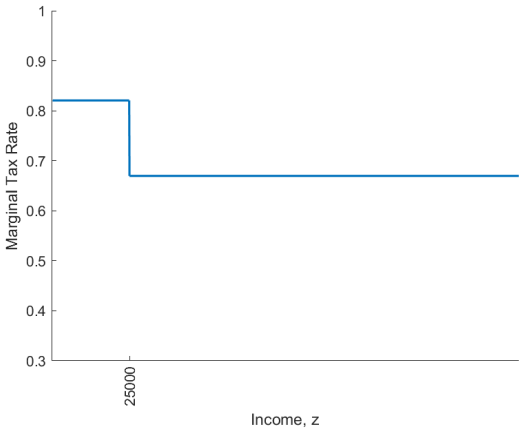
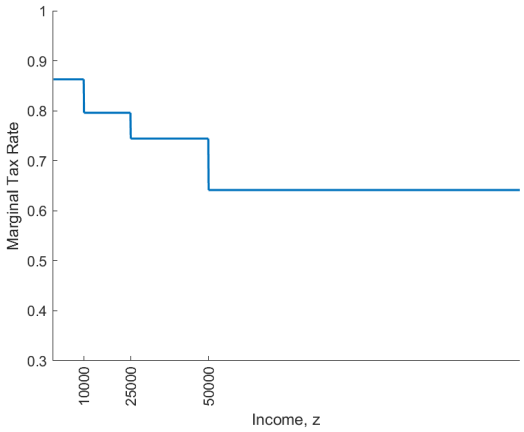


Figure 15: Inverse Welfare Weights for Optimal Six Bracket Piecewise Linear Tax Schedule

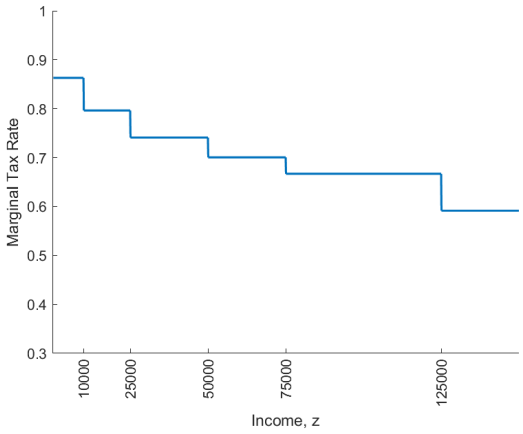
Note: This figure shows inverse welfare weights for the optimal six bracket tax system with kink points in $[\$10000, \$25000, \$50000, \$75000, \$125000]$ to maximize the welfare functional depicted by the line "true weights". We plot inverse weights over the (n, k) distribution when utility is given by Equation 54 and the distribution of n is calibrated to match the U.S. income distribution in 2019 from the ACS; k is uniformly distributed in $[1/0.35, 1/0.25]$, which implies that taxable income elasticities are between 0.25 and 0.35. Because many types choose each given z, inverse weights for each type (n, k) are computed as discussed in Appendix C to ensure pointwise convergence (precisely, we plot $q(\mathbf{z}; T_i)\phi^*(\mathbf{n})$ from Equation 121).



(a) Optimal Two Bracket



(b) Optimal Four Bracket



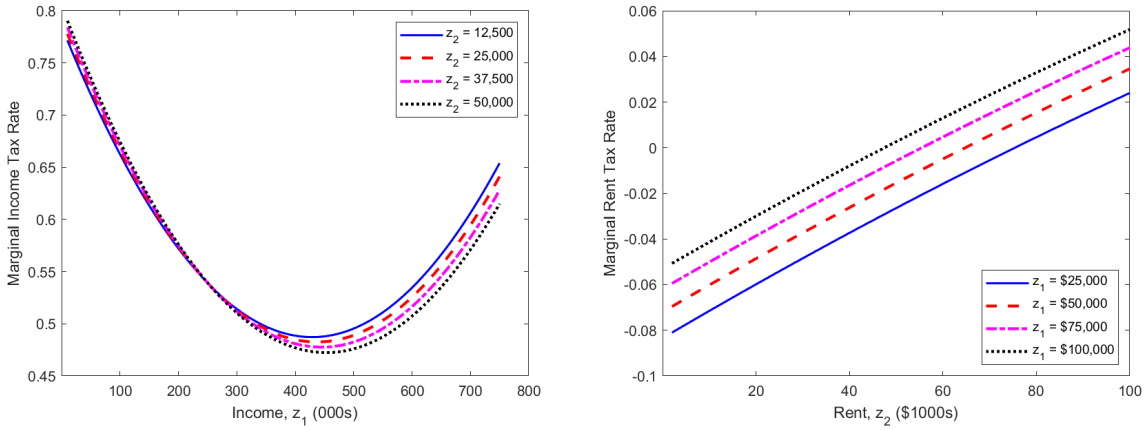
(c) Optimal Six Bracket

Figure 16: Optimal Piecewise Linear Tax Schedules

Note: This figure shows optimal piecewise linear schedules when we allow for two brackets with a kink at \$25,000, four brackets with kinks at $[\$10,000, \$25,000, \$50,000]$, and six brackets with kinks at $[\$10,000, \$25,000, \$50,000, \$75,000, \$125,000]$. To solve for the optimal piecewise linear schedules, we maximize the welfare functional under the "true weights" subject to the constraint that the tax schedule must be piecewise linear with the prescribed kink points.

D.1 Further Discussion of Exercise in Appendix 6.1

The exercise in Section 6.1 infers inverse welfare weights using Equation 29. To apply Equation 29 we assumed that individuals both respond smoothly to tax perturbations and also (via application of the envelope theorem) implicitly assumed individuals optimize utility. Both of these assumptions are standard in previous empirical calculations of inverse optimal welfare weights (Blundell et al. (2009); Bourguignon and Spadaro (2010); Bargain et al. (2013); Jacobs, Jongen and Zoutman (2017); and Hendren (2020)). However, smooth responses to tax perturbations and a lack of optimization errors is somewhat difficult to reconcile with the observed lack of bunching in the empirical income distribution. Given this difficulty, how can we properly interpret the results from Section 6.1 as well as previous empirical applications of inverse optimal theory?



(a) Marginal Income Tax Rates

(b) Marginal Housing Rent Tax Rates

Figure 17: Marginal Income and Housing Tax Rates: Optimal Third Order Polynomial

Note: This figure shows marginal tax rates on income and housing rent from the optimal third order polynomial from Section 6.5. Utility is given by Equation 55 and the type distribution $f(n_1, n_2, \alpha)$ is calibrated to match the empirical joint distribution of labor income and implicit housing rents from the 2019 American Community Survey (ACS) where implicit rents for homeowners are assumed to be 5% of the property value. k_1 and k_2 are chosen 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). α is uniformly distributed between 0.5 and 0.75. The government maximizes a "preference neutral" welfare function as in Bergstrom and Dodds (2021*b*) or Fleurbaey and Maniquet (2006).

There are at least three possible responses to this seeming inconsistency. The first (and almost surely least satisfactory) possibility is that the underlying type distribution $f(\mathbf{n})$ is discontinuous, generating an income distribution without bunching even if everyone responds smoothly to tax perturbations and optimizes correctly. While the type distribution is inherently unobservable without making assumptions on the individual optimization problem and the extent of optimization frictions, this seems unlikely. A second possibility is just to dispense with the assumption that individuals are optimizing correctly and simply interpret Figure 4 as depicting generalized marginal inverse weights as in Theorem 3, which does not require that individuals be correctly optimizing. This interpretation is logically consistent, but comes with the downsides discussed in Section 5.3 regarding generalized marginal inverse weights. Finally, we can potentially dispense with the assumption that individuals respond smoothly to tax perturbations. More specifically, the inverse welfare weights in Section 6.1 are inferred from the Gateaux derivative of revenue, which tells us how revenue changes in the direction of any possible tax perturbation and is calculated from Equation 29. The Gateaux derivative of revenue takes the form $\int_{\mathbf{Z}} \tau(\mathbf{z})\gamma(\mathbf{z})d\mathbf{z}$ where $\gamma(\mathbf{z})$ is just the RHS of Equation 29. As long as this Gateaux derivative of revenue is correct and individuals are optimizing utility, then we will correctly infer inverse welfare weights via Equation 8 from Theorem 1, *even if our model of individual behavior used to infer* $\gamma(\mathbf{z})$ *was incorrect.* In other words, even if individuals face substantial frictions (e.g., limited choice sets or difficulties changing jobs), as long as *aggregate*

revenue impacts of tax perturbations can be modeled *as if* all individuals respond smoothly, it is unimportant whether individuals all do in fact move smoothly. As a simple example, consider a population of 1,000 individuals, one of whom discretely decreases his tax liability by \$1,000 in response to a small tax perturbation and the rest of whom do not respond due to frictions; the revenue impacts of this tax perturbation can be equivalently *modeled* as if all 1,000 individuals smoothly decrease their tax liability by \$1 in response to a small tax change. In this way, it may be possible for the Gateaux derivative of government revenue to be correct even if the underlying model of behavior used to infer that Gateaux derivative is incorrect; with that said, the construction of an explicit model with frictions yielding identical aggregate revenue effects as a model with smooth behavioral responses is well beyond the scope of this paper. Nonetheless, as long as individuals optimize correctly and the Gateaux derivative of revenue is correct (i.e., *aggregate* revenue impacts of taxation can be modeled *as if* individuals are all responding smoothly), then the weights in Figure 4 can be interpreted as inverse welfare weights as opposed to simply generalized marginal inverse weights.

D.2 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 $z = w_i n l$ 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:

$$\begin{aligned} U(n; T, w_l, w_h) = \max_z \; & c - \frac{[z/(nw(n))]^{1+k}}{1+k} \\ \text{s.t. } & c = z - T(z) + s(n)\pi^*(w_l, w_h) \end{aligned} \tag{123}$$

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 each type, $L_l = \int_{\underline{n}}^{\text{med}(n)} nL(n)dF(n)$ and $L_h = \int_{\text{med}(n)}^{\overline{n}} nL(n)dF(n)$. 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:[57]

$$L_l = \int_{\underline{n}}^{\text{med}(n)} nL(n)dF(n) = \int_{\underline{n}}^{\text{med}(n)} nl(n)dF(n) \tag{124}$$

$$L_h = \int_{\text{med}(n)}^{\overline{n}} nL(n)dF(n) = \int_{\text{med}(n)}^{\overline{n}} nl(n)dF(n) \tag{125}$$

The firm first order conditions are given by:

$$Y_1(L_l, L_h) - w_l = 0 \tag{126}$$
$$Y_2(L_l, L_h) - w_h = 0 \tag{127}$$

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 Lagrangian is given by:

$$W(U(n; T, w_l, w_h)) + \lambda \left[\int_N T(z(n))dF(n) - E \right] \tag{128}$$

Now, let us take the (Gateaux) derivative of Equation 128 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:

$$\begin{aligned} &W\left(-\tau(z(n)) + \left(\frac{z(n)}{nw(n)}\right)^{1+k} \frac{1}{w(n)} \frac{\partial w(n)}{\partial \epsilon} + s(n)\nabla_{\mathbf{w}}\pi(w_l, w_h)\nabla_\epsilon \mathbf{w}\right) \\ &+ \lambda \int_N \left(\tau(z) + T'(z(n)) \frac{\partial z(n)}{\partial \epsilon}\bigg|_{\mathbf{w}} + T'(z(n)) \frac{\partial z(n)}{\partial w(n)}\bigg|_\epsilon \frac{\partial w(n)}{\partial \epsilon}\right) dF(n) \end{aligned} \tag{129}$$

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 124 and 125 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 $w(n)nl(n) = z(n)$):

$$\frac{\partial w_l}{\partial \epsilon} L_l + w_l \frac{\partial L_l}{\partial w_l}\frac{\partial w_l}{\partial \epsilon} + w_l \frac{\partial L_l}{\partial w_h}\frac{\partial w_h}{\partial \epsilon} - \int_{\underline{n}}^{\text{med}(n)} \left(\frac{\partial z(n)}{\partial w_l}\bigg|_\epsilon \frac{\partial w_l}{\partial \epsilon} + \frac{\partial z(n)}{\partial \epsilon}\bigg|_{\mathbf{w}}\right) dF(n) = 0 \tag{130}$$

$$\frac{\partial w_h}{\partial \epsilon} L_h + w_h \frac{\partial L_h}{\partial w_h}\frac{\partial w_h}{\partial \epsilon} + w_h \frac{\partial L_h}{\partial w_l}\frac{\partial w_l}{\partial \epsilon} - \int_{\text{med}(n)}^{\overline{n}} \left(\frac{\partial z(n)}{\partial w_h}\bigg|_\epsilon \frac{\partial w_h}{\partial \epsilon} + \frac{\partial z(n)}{\partial \epsilon}\bigg|_{\mathbf{w}}\right) dF(n) = 0 \tag{131}$$

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 11 which says that:

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

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

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 126 and 127. 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 130 and 131:[58]

$$\begin{aligned}
&\int_{\underline{n}}^{\text{med}(n)} \left.\frac{\partial z(n)}{\partial \epsilon}\right|_{\mathbf{w}} f(n)dn = \int_{\underline{n}}^{\text{med}(n)} \tau'(z(n))\xi(n)f(n)dn = \int_{\underline{z}}^{\text{med}(z)} \tau'(z)\xi(z)h(z)dz \\
&= \int_{\underline{z}}^{\text{med}(z)} -\tau(z)\frac{\partial}{\partial z}\left(\xi(z)h(z)\right)dz + \tau(z)\xi(z)h(z)\Big|_{\underline{z}}^{\text{med}(z)}
\end{aligned} \tag{132}$$

Identical steps yield:

$$\int_{\text{med}(n)}^{\overline{n}} \left.\frac{\partial z(n)}{\partial \epsilon}\right|_{\mathbf{w}} f(n)dn = \int_{\text{med}(z)}^{\overline{z}} -\tau(z)\frac{\partial}{\partial z}\left(\xi(z)h(z)\right)dz + \tau(z)\xi(z)h(z)\Big|_{\text{med}(z)}^{\overline{z}} \tag{133}$$

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 $\{\underline{z}, \text{med}(z), \overline{z}\}$.[59] One can then solve Equations 130 and 131 to yield that $\frac{\partial w_l}{\partial \epsilon}$ and $\frac{\partial w_h}{\partial \epsilon}$ are linear functionals of $\tau(z)$:

$$\frac{\partial w_l}{\partial \epsilon} = \frac{w_l\frac{\partial L_l}{\partial w_h}\int_{\text{med}(z)}^{\overline{z}}\frac{\partial}{\partial z}\left[\xi(z)h(z)\right]\tau(z)dz - C_2\int_{\underline{z}}^{\text{med}(z)}\frac{\partial}{\partial z}\left[\xi(z)h(z)\right]\tau(z)dz}{C_1C_2 - w_l\frac{\partial L_l}{\partial w_h}w_h\frac{\partial L_h}{\partial w_l}} \tag{134}$$

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

[59] 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.

$$\frac{\partial w_h}{\partial \epsilon} = \frac{w_h \frac{\partial L_h}{\partial w_l} \int_{\underline{z}}^{\text{med}(z)} \frac{\partial}{\partial z} \left[\xi(z)h(z)\right] \tau(z) dz - C_1 \int_{\text{med}(z)}^{\overline{z}} \frac{\partial}{\partial z} \left[\xi(z)h(z)\right] \tau(z) dz}{C_1 C_2 - w_l \frac{\partial L_l}{\partial w_h} w_h \frac{\partial L_h}{\partial w_l}} \tag{135}$$

with

$$C_1 = L_l + w_l \frac{\partial L_l}{\partial w_l} - \int_{\underline{n}}^{\text{med}(n)} \frac{\partial z(n)}{\partial w_l}\bigg|_{\epsilon} dF(n)$$

$$C_2 = L_h + w_h \frac{\partial L_h}{\partial w_h} - \int_{\text{med}(n)}^{\overline{n}} \frac{\partial z(n)}{\partial w_h}\bigg|_{\epsilon} dF(n)$$

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

$$\begin{aligned}
&\int_N \left(\tau(z) + T'(z(n)) \frac{\partial z(n)}{\partial \epsilon}\bigg|_{\mathbf{w}} + T'(z(n)) \frac{\partial z(n)}{\partial w(n)}\bigg|_{\epsilon} \frac{\partial w(n)}{\partial \epsilon}\right) dF(n) \\
&= \int_Z \left(h(z) - \frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right]\right)\tau(z)dz + \int_N T'(z(n)) \frac{\partial z(n)}{\partial w(n)}\bigg|_{\epsilon} \frac{\partial w(n)}{\partial \epsilon} dF(n) \\
&= \int_Z \left(h(z) - \frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right]\right)\tau(z)dz \\
&+ \int_Z \int_{\underline{n}}^{\text{med}(n)} T'(z(n)) \frac{\partial z(n)}{\partial w}\bigg|_{\epsilon} dF(n) p_l(z)\tau(z)dz + \int_Z \int_{\text{med}(n)}^{\overline{n}} T'(z(n)) \frac{\partial z(n)}{\partial w}\bigg|_{\epsilon} dF(n) p_h(z)\tau(z)dz \\
&= \int_Z \Bigg(h(z) - \frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right] \\
&+ p_l(z) \int_{\underline{n}}^{\text{med}(n)} T'(z(n)) \frac{\partial z(n)}{\partial w}\bigg|_{\epsilon} dF(n) + p_h(z) \int_{\text{med}(n)}^{\overline{n}} T'(z(n)) \frac{\partial z(n)}{\partial w}\bigg|_{\epsilon} dF(n)\Bigg)\tau(z)dz
\end{aligned} \tag{136}$$

Equation 136 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):

$$
\begin{aligned}
&W\left(-\tau(z(n))+\left(\frac{z(n)}{nw(n)}\right)^{1+k}\frac{1}{w(n)}\frac{\partial w(n)}{\partial \epsilon}+s(n)\nabla_{\mathbf{w}}\pi(w_l,w_h)\nabla_{\epsilon}\mathbf{w}\right)\\
&=-\int_Z \phi(n(z))\tau(z)h(z)dz+\int_N \phi(n)\left[\left(\frac{z(n)}{n}\right)^{1+k}\frac{\partial w(n)}{\partial \epsilon}+s(n)\nabla_{\mathbf{w}}\pi(1,1)\nabla_{\epsilon}\mathbf{w}\right]f(n)dn\\
&=-\int_Z \phi(n(z))\tau(z)h(z)dz+\int_Z p_l(z)\tau(z)\left(\int_{\underline{n}}^{\text{med}(n)}\phi(n)\left(\frac{z(n)}{n}\right)^{1+k}f(n)dn\right)dz\\
&+\int_Z p_h(z)\tau(z)\left(\int_{\text{med}(n)}^{\overline{n}}\phi(n)\left(\frac{z(n)}{n}\right)^{1+k}f(n)dn\right)dz+\sum_{i=l,h}\int_Z p_i(z)\tau(z)\left(\int_N \phi(n)s(n)\frac{\partial \pi}{\partial w_i}f(n)dn\right)dz\\
&=-\int_Z\left[\phi(n(z))h(z)-p_l(z)\left(\int_{\underline{z}}^{\text{med}(z)}\phi(n(\tilde{z}))\left(\frac{\tilde{z}}{n(\tilde{z})}\right)^{1+k}dH(\tilde{z})\right)\right.\\
&\left.-p_h(z)\left(\int_{\text{med}(z)}^{\overline{z}}\phi(n(\tilde{z}))\left(\frac{\tilde{z}}{n(\tilde{z})}\right)^{1+k}dH(\tilde{z})\right)-\sum_{i=l,h}\int_Z p_i(z)\left(\int_Z \phi(n(\tilde{z}))s(n(\tilde{z}))\frac{\partial \pi}{\partial w_i}dH(\tilde{z})\right)\right]\tau(z)dz
\end{aligned}
\tag{137}
$$

Equation 137 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 129 equals zero. Normalizing the Lagrange multiplier λ to 1 this requires that Equation 136 plus Equation 137 must equal zero for all perturbations $\tau(\mathbf{z})$. To construct a set of welfare weights that satisfy this condition, let us define:

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

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

From here, we can simply match terms pointwise in Equations 136 and 137 to see that Equation 136 plus Equation 137 equals zero for all $\tau(\mathbf{z})$ (i.e., the Gateaux derivative of the Lagrangian is zero) as long as the following equation holds:

$$
\phi(n(z))=\chi(z)+K(z) \tag{138}
$$

As long as $\chi(z)+K(z)$ defines a contraction mapping on the set of functions $\phi(n(z))$, then Equation 138 has a solution that can be computed via standard fixed point algorithms as discussed in Section 5.1. This solution to Equation 138 then defines an inverse welfare functional for the given arbitrary tax schedule.