A general theory of inverse welfare functions*

Katy Bergstrom† William Dodds‡

March 27, 2026

Abstract

This paper develops a general theory to recover the inverse welfare function that rationalizes a given tax schedule as optimal. Our theory allows for complex environments including the presence of multidimensional tax schedules, bunching/jumping behavior, optimization frictions, general equilibrium effects, and externalities. We show that inverse welfare functions can be used to test for Pareto efficiency, construct Pareto improving reforms, and characterize the set of welfare improving local reforms, thereby extending previous results to more general environments. We show numerically that allowing for such generalities can have meaningful impacts on inverse welfare functions.

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

JEL: H21, H31, D60

*The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. We would like to thank Nathan Hendren, Nicolas Werquin, Juan Rios, and conference participants at SAET, SEA, and Tulane University for their very helpful comments and suggestions. Finally, thanks to three anonymous referees and the editor, Rakesh Vohra, for their numerous helpful and constructive comments.

†Department of Economics, Tulane University. Email: kbergstrom@tulane.edu

‡Department of Economics, Tulane University. Email: wdodds@tulane.edu

1 Introduction

Since the seminal work of Mirrlees (1971), the vast majority of theoretical work on taxation involves a planner seeking to maximize a social welfare function 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 "inverse welfare function" that rationalizes this tax schedule as optimal (Blundell et al. (2009), Bourguignon and Spadaro (2010), Bargain et al. (2013), Jacobs, Jongen and Zoutman (2017)). While some papers have been intrinsically interested in inverse welfare functions because they encode the implicit interpersonal welfare comparisons that justify the observed tax schedule, other papers have focused on using inverse welfare functions to assess Pareto efficiency of tax schedules and identify Pareto improvements (e.g., Lorenz and Sachs (2016), Hendren (2020), Spiritus et al. (2022), and Bierbrauer, Boyer and Hansen (2023)). Importantly, almost all of these applications have been solely for income tax schedules and require a number of assumptions: individuals respond smoothly to tax reforms (e.g., there is no bunching or people with multiple optima), individuals do not face optimization frictions, there are no general equilibrium effects of taxation, and there are no externalities.

The goal of the present paper is to develop a general theory of inverse welfare functions that can be applied in much more general taxation settings. This paper has three main contributions. First and foremost, we prove two theorems (Theorems 1 and 2) establishing the existence of inverse welfare functions; importantly, these theorems show how to explicitly construct an inverse welfare function and can be applied to settings that feature important realisms: multidimensional tax schedules, bunching/jumping behavior, optimization frictions, general equilibrium effects, and externalities. Second, we contribute to the literature on Pareto efficiency of tax schedules (e.g., Werning (2007), Spiritus et al. (2022), Bierbrauer, Boyer and Hansen (2023), and Sturm and Sztutman (2023)) by showing that we can use the inverse welfare function to assess whether a tax schedule is Pareto efficient and construct Pareto improving reforms in more complex settings for which existing results cannot be applied. Third, we show that inverse welfare functions (in conjunction with society's actual welfare function) can be used to explicitly characterize the set of welfare-improving (marginal) non-linear tax reforms as well as determine the optimal local tax reform. Hence, we contribute to the literature on

the desirability of tax reforms (e.g., Saez (2001) and Saez and Stantcheva (2016)), and optimal tax reforms (e.g., Diewert (1978) and Dixit (1979)) by illustrating how inverse welfare functions can be used to answer these questions in more complicated settings. Finally, as a fourth minor contribution, we demonstrate numerically that allowing for various realisms (such as bunching/jumping, the presence of optimization frictions, multidimensional tax schedules, general equilibrium wage effects, or externalities) can have large and meaningful impacts on the inverse welfare function and, consequently, can have important policy implications. We now provide a brief outline of the paper.

Section 2 presents our first main theorem on existence and construction of inverse welfare functions: Theorem 1. Theorem 1 proves that in a partial equilibrium setting, we can compute an inverse welfare function if government revenue is Gateaux differentiable in the tax schedule (the Gateaux derivative is a generalization of the gradient). Theorem 1 is constructive: we show explicitly how to construct an inverse welfare function (which is a weighted sum of utilities) without restricting the dimension of the tax schedule, the dimension of the type space, the individual choice set, or the type of behavioral responses available to individuals. Intuitively, we construct the inverse welfare weight for individuals making particular choices by equating the revenue effect of an "instantaneous" tax change at that choice level with the welfare effect of such an "instantaneous" tax change.

Next, Section 3 provides a number of analytical constructions of Gateaux derivatives of government revenue along with associated inverse welfare functions, highlighting that the Gateaux derivative of revenue can be expressed in terms of behavioral responses to tax reforms. The next result of the paper, Proposition 1, provides a set of general sufficient conditions for Gateaux differentiability of government revenue, establishing that Gateaux differentiability of government revenue is a relatively mild restriction: revenue can be Gateaux differentiable even if the (potentially multidimensional) tax schedule is non-differentiable (generating bunching), individuals have multiple optima, individuals respond on the extensive margin, and/or individuals face limited choice sets. Hence, Proposition 1 combined with Theorem 1 establishes that inverse welfare functions typically exist even in complex settings.

Next, Section 4 illustrates how inverse welfare functions can be used to inform tax policy. First and foremost, the inverse welfare function can be used to assess whether the observed tax schedule is Pareto efficient (and if not, construct Pareto improvements).

We are not the first paper to recognize this: in fact, much of the existing literature on inverse welfare weights has been motivated with the goal of assessing Pareto efficiency (e.g., Lorenz and Sachs (2016), Hendren (2020), Spiritus et al. (2022) Bierbrauer, Boyer and Hansen (2023), Jacquet and Lehmann (2025)). We build on this literature by providing a simple characterization of Pareto efficient tax schedules that allows for complex behavioral responses (e.g., tax schedules can be non-differentiable generating bunching, individuals can face optimization frictions, individuals can "jump" between multiple optima) and allows for multidimensional heterogeneity and multidimensional tax schedules. We illustrate that even in these more complex environments, a tax schedule is Pareto efficient if the inverse welfare function constructed in Theorem 1 is positive (e.g., welfare weights are all positive); conversely, a tax schedule is Pareto inefficient if the inverse welfare function constructed in Theorem 1 is non-positive (e.g., some welfare weights are negative). Second, we show that the inverse welfare function and the actual welfare function are sufficient to characterize the set of welfare-improving tax perturbations. Loosely, if society's actual welfare weights are larger (smaller) than the inverse welfare weights at a particular choice level, then decreasing (increasing) taxes at that choice level (and closing the budget via changing the lump-sum transfer) is welfare improving. Section 4 also illustrates how to compute the optimal reform direction using only the inverse welfare function and the actual welfare function. We believe our results on the construction of welfare-improving marginal tax reforms are useful in complex taxation settings given that inverse welfare functions are relatively easy to compute (e.g., inverse welfare functions can be calculated in seconds even for the most complex models we consider) whereas solving for the optimal tax schedule computationally is often entirely intractable with existing methods once we move outside of smooth, unidimensional settings (other than some special cases as in Spiritus et al. (2022), Dodds (2023), or Krasikov and Golosov (2024)).

Next, Section 5 discusses how our theory of inverse welfare functions can be extended to settings with general equilibrium (GE) effects. To build intuition, we first illustrate how to construct inverse welfare weights in a model with endogenous wages similar to Sachs, Tsyvinski and Werquin (2020). We then prove Theorem 2, which shows that, more generally, an inverse welfare function can be constructed from the Gateaux derivative of government revenue *and* the Gateaux derivatives of equilibrium objects (e.g., endogenous

wages or prices) with respect to the tax schedule. In the GE case, the inverse welfare function is the fixed point of an integral equation. Next, we prove that inverse welfare functions can still be used to assess Pareto efficiency, identify Pareto improving reforms, and construct welfare improving tax reforms even with general equilibrium effects. Finally, Section 5 shows that our definition of "general equilibrium effects" is broad enough so that Theorem 2 can also allow for externalities; we provide an example application with the presence of inequality aversion (wherein individual utility depends directly on the level of inequality and therefore other individuals' incomes generate an externality by contributing to inequality).

Section 6 then presents a high-level discussion of several stylized numerical simulations to illustrate how various realistic complexities impact the inverse welfare function. Relative to a baseline calibration in which we infer inverse weights for a smoothed approximation of the U.S. income tax schedule (similar to Hendren (2020)), we numerically compute inverse weights in five alternative models that incorporate various realisms: (1) the non-smooth (piece-wise linear) U.S. income tax schedule which induces bunching and jumping behavior (under standard preferences), (2) sparsity-based frictions, (3) an additional tax instrument (property taxes), (4) general equilibrium wage effects, and (5) inequality aversion. The high-level conclusion from these simulations is that accounting for these realisms can substantially change the inverse welfare function and therefore can also substantially change policy conclusions about Pareto efficiency and the set of welfare improving tax reforms.

Finally, Section 7 illustrates that inverse welfare functions typically *do not* exist when the type space is smaller than the choice space. Section 7 shows that this allows for a strengthening of the Atkinson-Stiglitz Theorem: in the Atkinson-Stiglitz environment, it is often impossible to rationalize indirect taxes (e.g., savings or commodity taxes) even if the government wants to make some individuals as miserable as possible (via negative welfare weights). Section 8 concludes.

2 Construction of Inverse Welfare Functionals

2.1 Notation and Mathematical Preliminaries

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 distribution $F(\mathbf{n})$ with density $f(\mathbf{n})$. Individuals

choose $\mathbf{z} = (z_1, z_2, ..., z_J) \in A(\mathbf{n}) \subseteq \mathbb{R}^J$ to maximize a utility function subject to a budget constraint given a tax schedule, $T(\mathbf{z})$, which is a function of choice variables $\mathbf{z}$:

$$\begin{aligned} &\max_{\mathbf{z} \in A(\mathbf{n})} \; u(c, \mathbf{z}; \mathbf{n}) \\ &\text{s.t. } c = y(\mathbf{z}) - T(\mathbf{z}) \end{aligned} \tag{1}$$

where c is numeraire consumption and is a function of choices $\mathbf{z}$ as well as the tax schedule $T(\mathbf{z})$. For example, z_i might represent income from a particular source (e.g., labor or savings) or consumption of a particular good or the z_i's could represent incomes in various time periods. We will assume $u(c, \mathbf{z}; \mathbf{n})$ is twice differentiable in c and differentiable in $\mathbf{n}$ (we make no assumptions about differentiability in $\mathbf{z}$, thereby allowing for fixed costs of working). We assume that there is a societal budget constraint that total tax revenue, $R(T)$, is greater than or equal to some exogenous revenue requirement, E:

$$R(T) \equiv \int_{\mathbf{N}} T(\mathbf{z}(\mathbf{n}))dF(\mathbf{n}) \geq E \tag{2}$$

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 $\mathbf{z}(\mathbf{n})$ also depends on the tax schedule T. To save on notation, we will often write $u_c(\mathbf{n})$ as shorthand for $u_c(y(\mathbf{z}(\mathbf{n})) - T(\mathbf{z}(\mathbf{n})), \mathbf{z}(\mathbf{n}); \mathbf{n})$. Let $\mathbf{Z} = \{\mathbf{z}(\mathbf{n}) | \mathbf{n} \in \mathbf{N}\}$ denote the set of $\mathbf{z}$ that are chosen by some type $\mathbf{n}$ under the given tax schedule $T(\mathbf{z})$ and let $\mathbf{N}(\mathbf{z}) = \{\mathbf{n} | \mathbf{z}(\mathbf{n}) = \mathbf{z}\}$ denote the set of types $\mathbf{n}$ that choose $\mathbf{z}$ under the given tax schedule $T(\mathbf{z})$.

Next, we will frequently utilize the concept of a *Gateaux derivative* which we define as in the Encyclopedia of Mathematics:

Definition 1. *Let $M : \mathcal{G} \mapsto \mathbb{R}$ be a functional on a normed vector space $\mathcal{G}$. We say that M is Gateaux differentiable at a $g \in \mathcal{G}$ if $\exists$ a bounded linear functional, DM_g, which we call the Gateaux derivative, such that for any $\psi \in \mathcal{G}$:*[1]

$$\lim_{\epsilon \to 0} \frac{M(g + \epsilon\psi) - M(g)}{\epsilon} = DM_g(\psi)$$

We refer to $\lim_{\epsilon \to 0} \left(M(g + \epsilon\psi) - M(g)\right)/\epsilon$ as the Gateaux variation of M at g in the direction of ψ.

[1] Gateaux differentiability in Definition 1 is stronger than existence of the Gateaux variation (which does not require linearity and boundedness) but weaker than Frechet differentiability because we do not require uniform convergence in all directions ψ. We define these objects as in the Encyclopedia of Mathematics, but note that what we call the "Gateaux variation" is sometimes referred to by other authors as the Gateaux derivative.

Definition 2. *$B : \mathcal{G} \mapsto \mathbb{R}$ is a bounded linear functional if $B(a_1 g_1 + a_2 g_2) = a_1 B(g_1) + a_2 B(g_2)$ $\forall a_1, a_2 \in \mathbb{R}, g_1, g_2 \in \mathcal{G}$ and $\exists P$ s.t. for any $g \in \mathcal{G}$ then $|B(g)| < P \|g\|_{\mathcal{G}}$.*

To build intuition, when $\mathcal{G} = \mathbb{R}^k$, the Gateaux variation is the directional derivative and is always equal to the Gateaux derivative (the gradient) multiplied by the direction vector. However, in this paper $\mathcal{G}$ will typically be the set of continuous functions defined on a compact set $\mathbf{S}$, denoted $C(\mathbf{S})$. When $\mathcal{G} = C(\mathbf{S})$ endowed with the supnorm, there is a convenient representation of bounded linear functionals that we will utilize:

Remark 1. *By the Riesz-Markov-Kakutani representation theorem (Theorem 6.19 of Rudin (1974)), any bounded linear functional B on $C(\mathbf{S})$ (endowed with the supnorm) can be expressed as follows for some measure Γ:*

$$B(g) = \int_{\mathbf{S}} g(\mathbf{s}) d\Gamma(\mathbf{s}) \tag{3}$$

We are now ready to define welfare functionals.[2] First, let us denote $U(\mathbf{n}; T)$ as the indirect utility profile that arises when agents optimize under tax schedule $T(\mathbf{z})$ according to Equation 1. We define a welfare functional, W, as a bounded linear functional (on the space of continuous functions, equipped with the supnorm, with compact domain $\mathbf{N}$) that takes $U(\mathbf{n}; T)$ as its argument and returns a scalar which we call welfare.[3] By Remark 1, every such welfare functional $W(U(\mathbf{n}; T))$ can be represented as follows:

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

where Φ is a measure; hence, we restrict attention to welfare functions which are weighted sums of utilities.

Remark 2. *A measure Φ assigns a "weight" $\Phi(\tilde{\mathbf{N}})$ to each measurable subset $\tilde{\mathbf{N}}$ of its domain $\mathbf{N}$.[4] Throughout the paper, sets will be denoted by upper case letters (e.g., $\mathbf{N}$ or $\mathbf{Z}$) whereas elements of these sets will be denoted by lower case letters (e.g., $\mathbf{n}$ or $\mathbf{z}$). We will use the notation $\Phi(\tilde{\mathbf{n}})$ to denote the distribution function associated with measure Φ: $\Phi(\{\mathbf{n} \leq \tilde{\mathbf{n}}\})$ where the $\leq$ is understood as being component-wise. If $\Phi(\tilde{\mathbf{n}})$ is differentiable,*

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

[3] $U(\mathbf{n}; T)$ is (absolutely) continuous if the utility function $u(c, \mathbf{z}; \mathbf{n})$ is differentiable in $\mathbf{n}$ and the choice set $A(\mathbf{n})$ does not vary with $\mathbf{n}$ (Milgrom and Segal, 2002). Alternatively, $U(\mathbf{n}; T)$ is continuous if $u(c(\mathbf{z}), \mathbf{z}; \mathbf{n})$ is continuous in $\mathbf{z}$ and $\mathbf{n}$ and $A(\mathbf{n})$ is a continuous correspondence (by Berge's Maximum Theorem).

[4] The measure of a set, $\Phi(\tilde{\mathbf{N}})$, is allowed to be positive or negative; some authors refer to these as "signed measures", but we will drop the "signed" for brevity.

then Equation 4 can be expressed with more conventional "welfare weights" $\phi(\mathbf{n})$ so that:[5]

$$W(U(\mathbf{n};T)) = \int_{\mathbf{N}} \phi(\mathbf{n})U(\mathbf{n};T)f(\mathbf{n})d\mathbf{n} \tag{5}$$

Allowing for more general measures Φ in Equation 4 allows the welfare functional to contain mass points as in a Rawlsian welfare function (in which case Φ puts all weight on the lowest type $\mathbf{n}$ in society).

Finally, this brings us to the definition of a local inverse welfare functional. Consider the following Lagrangian (with Lagrange multiplier λ on the societal budget constraint) under a welfare functional W:

$$L(T;W) \equiv W(U(\mathbf{n};T)) + \lambda\left[R(T) - E\right] \tag{6}$$

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

Note that Definition 3 is inherently *local*; we will focus almost exclusively on *local* inverse welfare functionals because, as we will show in Section 4, local inverse welfare functionals encode all of the necessary information to assess Pareto efficiency, identify Pareto improvements, and construct welfare improving tax reforms.[6] Henceforth, when we refer to an "inverse welfare functional" this should be understood as "local inverse welfare functional"; we drop the "local" for brevity.[7]

Remark 3. *If the inverse welfare functional is expressed as $\int_{\mathbf{N}} U(\mathbf{n};T)d\Phi(\mathbf{n})$, then we will call Φ the inverse welfare measure. If the inverse welfare functional is expressed as $\int_{\mathbf{N}} \phi(\mathbf{n})U(\mathbf{n};T)f(\mathbf{n})d\mathbf{n}$, we will call ϕ inverse welfare weights.*

[5] Note that Φ incorporates the type distribution; for instance, a utilitarian welfare function sets Φ so that $\Phi(\mathbf{n}) = F(\mathbf{n})$ yielding $W(U(\mathbf{n};T)) = \int_{\mathbf{N}} U(\mathbf{n};T)dF(\mathbf{n})$. In contrast, for consistency with prior literature, $\phi(\mathbf{n})$ will *not* incorporate the type distribution: $W(U(\mathbf{n};T)) = \int_{\mathbf{N}} \phi(\mathbf{n})U(\mathbf{n};T)f(\mathbf{n})d\mathbf{n}$.

[6] In contrast, a *global* inverse welfare functional W for a given $T(\mathbf{z})$ is such that the indirect utility profile $U(\mathbf{n};T)$ maximizes W within the set $\mathcal{U}$, where $\mathcal{U}$ denotes the set of all utility profiles that are generated by maximization under some tax schedule that also satisfy the government's budget constraint. Establishing that a local inverse welfare functional is a global inverse welfare functional is difficult because $\mathcal{U}$ is not, in general, a convex set (i.e., if $U(\mathbf{n};T_1)$ and $U(\mathbf{n};T_2)$ are derived from individual optimization under $T_1(\mathbf{z})$ and $T_2(\mathbf{z})$, then the convex combination $U_3(\mathbf{n}) = \alpha U(\mathbf{n};T_1) + (1-\alpha)U(\mathbf{n};T_2)$ may not be consistent with individual optimization under *any* tax schedule). In Appendix B.1 we prove Proposition 5 establishing existence of a global inverse welfare functional under a concavity condition; however, this result is non-constructive and the concavity condition is not always easy to verify in practice.

[7] Note, Definition 3 defines a local inverse welfare functional so that the given tax schedule is a stationary point of the Lagrangian. One can check whether the tax schedule is a local maximum by checking that the second Gateaux variation of the Lagrangian is negative definite (although this can get quite complex); fortunately, all of the policy applications of inverse welfare functionals in Section 4 only require that the local inverse welfare functional be a stationary point of the Lagrangian.

2.2 A Constructive Existence Theorem

Our goal is to construct an inverse welfare functional $W(U(\mathbf{n};T))$ satisfying:

$$\lim_{\epsilon\to 0}\frac{W(U(\mathbf{n};T+\epsilon\tau))-W(U(\mathbf{n};T))}{\epsilon}+\lambda\lim_{\epsilon\to 0}\frac{R(T+\epsilon\tau)-R(T)}{\epsilon}=0\ \forall\tau\in C(\mathbf{Z})$$

By Remark 1, we know that the Gateaux derivative of revenue R can be expressed as:

$$\lim_{\epsilon\to 0}\frac{R(T+\epsilon\tau)-R(T)}{\epsilon}=\int_{\mathbf{Z}}\tau(\mathbf{z})d\Gamma(\mathbf{z}) \tag{7}$$

for some measure Γ. In words, the revenue effect of any given small tax perturbation in the direction of $\tau(\mathbf{z})$ is given by $\int_{\mathbf{Z}}\tau(\mathbf{z})d\Gamma(\mathbf{z})$. However, in a number of situations discussed throughout the paper it will turn out that the Gateaux derivative can be written as:[8]

$$\lim_{\epsilon\to 0}\frac{R(T+\epsilon\tau)-R(T)}{\epsilon}=\int_{\mathbf{Z}}\tau(\mathbf{z})\gamma(\mathbf{z})dH(\mathbf{z}) \tag{8}$$

where $H(\mathbf{z})$ is the distribution of $\mathbf{z}$. Loosely, $\gamma(\mathbf{z})$ represents the "instantaneous budgetary effect" of an infinitesimal tax perturbation that changes the tax schedule at a given choice level $\mathbf{z}$ (Figure 1a below illustrates such a perturbation for the unidimensional case in which agents choose an income z and consumption is given by $c=z-T(z)$).

This brings us to our first constructive existence result for inverse welfare functionals:

Theorem 1. *Consider continuous $T(\mathbf{z})$ such that $R(T)=E$, $\mathbf{Z}$ is compact, for every $\mathbf{z}$ almost all $\mathbf{n}$ that choose $\mathbf{z}$ have a unique optimum, and $R(T)$ is Gateaux differentiable.*

(1) If the Gateaux derivative of $R(T)$ can be expressed as in Equation 8, then we can construct an inverse welfare functional $W(U(\mathbf{n};T))=\int_{\mathbf{N}}\phi(\mathbf{n})U(\mathbf{n};T)f(\mathbf{n})d\mathbf{n}$ with:

$$\phi(\mathbf{n})=\frac{\gamma(\mathbf{z}(\mathbf{n}))}{\overline{u_c}(\mathbf{z}(\mathbf{n}))} \tag{9}$$

where $\overline{u_c}(\mathbf{z})$ represents average marginal utility of consumption at $\mathbf{z}$.

(2) More generally, we can construct an inverse welfare functional $W(U(\mathbf{n};T))=\int_{\mathbf{N}}U(\mathbf{n};T)d\Phi(\mathbf{n})$ with the inverse welfare measure Φ defined as:[9]

$$\Phi(\tilde{\mathbf{n}})=\int_{\mathbf{Z}}\int_{\{\mathbf{n}\leq\tilde{\mathbf{n}}\}\cap\mathbf{N}(\mathbf{z})}\frac{1}{\overline{u_c}(\mathbf{z})}dF(\mathbf{n}|\mathbf{z})d\Gamma(\mathbf{z}) \tag{10}$$

Proof. We provide a proof of Equation 9 as this allows us to avoid measure theory and skip over a number of technical details; see Appendix A.1 for the proof of Equation 10.

[8] Note, if the Gateaux derivative can be written as in Equation 8, we can easily turn this into the form of Equation 7 by defining the measure Γ for each $\mathbf{E}\subseteq\mathbf{Z}$ to satisfy $\Gamma(\mathbf{E})=\int_{\mathbf{E}}\gamma(\mathbf{z})dH(\mathbf{z})$.

[9] Equation 10 defines $\Phi(\tilde{\mathbf{n}})=\Phi(\{\mathbf{n}\leq\tilde{\mathbf{n}}\})$. Standard extension theorems imply that this uniquely defines $\Phi(\tilde{\mathbf{N}})$ for any measurable $\tilde{\mathbf{N}}\subseteq\mathbf{N}$ on a compact domain $\mathbf{N}$; see Appendix A.1.

First, the envelope theorem implies that for all $\mathbf{n}$ with a unique optimum, behavioral responses to tax changes have only second order impacts on indirect utility so that:

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

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

$$\frac{\partial W(U(\mathbf{n}; T + \epsilon\tau))}{\partial \epsilon} = -\int_{\mathbf{N}} \phi(\mathbf{n}) u_c(\mathbf{n}) \tau(\mathbf{z}(\mathbf{n})) dF(\mathbf{n}) = -\int_{\mathbf{Z}} \int_{\mathbf{N}(\mathbf{z})} \phi(\mathbf{n}) u_c(\mathbf{n}) \tau(\mathbf{z}) dF(\mathbf{n}|\mathbf{z}) dH(\mathbf{z}) \tag{11}$$

where we disintegrated the distribution $F(\mathbf{n})$ into $F(\mathbf{n}|\mathbf{z})$ and $H(\mathbf{z})$. Thus, the Gateaux derivative of the government's Lagrangian is given by:

$$-\int_{\mathbf{Z}} \int_{\mathbf{N}(\mathbf{z})} \phi(\mathbf{n}) u_c(\mathbf{n}) \tau(\mathbf{z}) dF(\mathbf{n}|\mathbf{z}) dH(\mathbf{z}) + \lambda \int_{\mathbf{Z}} \tau(\mathbf{z}) \gamma(\mathbf{z}) dH(\mathbf{z}) \tag{12}$$

where we used the simplifying assumption that the Gateaux derivative of revenue can be expressed as in Equation 8. To ensure that this Gateaux derivative equals zero, it suffices to ensure that for each $\mathbf{z}$:

$$\int_{\mathbf{N}(\mathbf{z})} \phi(\mathbf{n}) u_c(\mathbf{n}) dF(\mathbf{n}|\mathbf{z}) = \lambda \gamma(\mathbf{z}) \tag{13}$$

From here, let $\gamma(\mathbf{z}(\mathbf{n}))$ represent the value of $\gamma(\mathbf{z})$ at the $\mathbf{z}$ chosen by type $\mathbf{n}$ and $\overline{u_c}(\mathbf{z}(\mathbf{n})) \equiv \int_{\mathbf{N}(\mathbf{z}(\mathbf{n}))} u_c(\mathbf{n}') dF(\mathbf{n}'|\mathbf{z})$ represent average marginal utility of consumption at the $\mathbf{z}$ chosen by type $\mathbf{n}$. If we normalize $\lambda = 1$ (which simply rescales the inverse welfare functional multiplicatively everywhere), then the weights prescribed in Equation 9 imply that Equation 13 is satisfied.[10] □

There are three steps to identify the inverse welfare weights in the proof sketch to Theorem 1: (1) appeal to the envelope theorem to infer that the direct welfare impact of an infinitesimal tax perturbation on each person is equal to their welfare weighted marginal utility multiplied by the size of the tax change they face, (2) transform the welfare impact integrated over the type space $\mathbf{N}$ into a welfare impact integrated over the choice space $\mathbf{Z}$ using the disintegration theorem, and (3) choose the inverse weights so that the welfare effects exactly offset the budgetary impact of an infinitesimal tax perturbation at a given $\mathbf{z}$. The full proof in Appendix A.1 is conceptually the same, but

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

we construct a measure Φ to create an inverse welfare functional as in Equation 4.

There are several technical points to discuss. The assumption that $T(\mathbf{z})$ is continuous is WLOG as long as all individuals have indifference surfaces with bounded gradients: Lemma 2 in Appendix B.2 establishes that every utility profile derived from individual optimization under some tax schedule can also be derived from individual optimization under a continuous tax schedule as long as indifference surfaces have bounded gradients. Next, we can relax the requirement that at each $\mathbf{z}$ almost all $\mathbf{n}$ have a unique optima to instead only require that at least one $\mathbf{n}$ has a unique optima at each $\mathbf{z}$ (Appendix A.1 shows how to construct an inverse welfare functional in this case). This assumption cannot be further relaxed (i.e., an inverse welfare functional may not exist if there exists a $\mathbf{z}$ such that all $\mathbf{n}$ choosing that $\mathbf{z}$ have multiple optima). To see why, consider a unidimensional setting with an individual n_1 who has two optimal incomes, z^- and z^+. No other individual finds z^- or z^+ optimal other than the n_1 individual. Consider decreasing tax rates right around the z^- income level: there will be some welfare weight on type n_1 that ensures the Lagrangian remains unchanged. Similarly, consider decreasing tax rates right around the z^+ income level: there will be some welfare weight on type n_1 that ensures the Lagrangian remains unchanged. However, these two welfare weights need not coincide; hence, no matter which of these two welfare weight we choose, we can always improve welfare by changing taxes at either z^- or z^+ or both. Conversely, even if type n_1 has two optimal incomes, z^- and z^+, if there are types n_2 and n_3 with a single optimum at z^- and z^+, respectively, then we can set the welfare weight on type n_1 to zero and choose the weights on n_2 and n_3 to ensure that small tax perturbations at z^- and z^+ leave the government's Lagrangian unchanged.

Next, the inverse welfare functional constructed in Theorem 1 is typically not unique for two reasons. First, if many types locate at a given $\mathbf{z}$ so that $\mathbf{N}(\mathbf{z})$ is not a singleton, then there are typically many inverse welfare functionals of the form $\int_{\mathbf{N}} \phi(\mathbf{n})U(\mathbf{n};T)dF(\mathbf{n})$ that satisfy Equation 13 and therefore support the given tax schedule: the Gateaux derivative of revenue only pins down the *average* inverse welfare weight at each $\mathbf{z}$. Second, there are generically additional inverse welfare functionals that are non-linear functionals of $U(\mathbf{n};T)$. For example, consider constructing an inverse welfare functional of the form $W(U(\mathbf{n};T)) = \int_{\mathbf{N}} \Psi(U(\mathbf{n};T),\mathbf{n})dF(\mathbf{n})$ for some smooth Ψ. The same arguments in the proof sketch to Theorem 1 above illustrate that we can find such an alternative inverse

welfare functional as long as Ψ satisfies $\int_{\mathbf{N}(\mathbf{z})} \partial\Psi(U(\mathbf{n};T),\mathbf{n})/\partial U \, u_c(\mathbf{n}) dF(\mathbf{n}|\mathbf{z}) = \gamma(\mathbf{z})$. Hence, choosing any function Ψ with $\partial\Psi(U(\mathbf{n};T),\mathbf{n})/\partial U$ equal to the inverse welfare weight $\phi(\mathbf{n})$ from Equation 9 will also be a local inverse welfare functional.[11] Thus, there are generally an infinite number of functions Ψ satisfying the above restriction (in the same way that there are typically an infinite number of indifference curves that can rationalize an individual's choice over two goods). However, this non-uniqueness is not problematic because (as we will show in Section 4) we can test for Pareto efficiency and find Pareto improving reforms, determine the desirability of tax reforms, and construct optimal tax reforms just by using the *specific* linear inverse welfare functionals constructed in Theorem 1.[12] Finally, we *cannot* relax the requirement that government revenue is Gateaux differentiable in Theorem 1: we illustrate in Section 7 that inverse welfare functionals can fail to exist if we only have that the Gateaux variation $\lim_{\epsilon\to 0} \left(R(T+\epsilon\tau) - R(T)\right)/\epsilon$ exists in all directions τ (i.e., we drop the requirement that the Gateaux variations are related via a bounded linear functional). Without the structure of a bounded linear functional relating the Gateaux variations, Section 7 shows that inverse weights may need to satisfy an overdetermined infinite system of equations which has no solution.

3 Gateaux Differentiability of Government Revenue

The next natural question then is whether most tax schedules generate a government revenue function that is Gateaux differentiable and, if so, how do we compute this Gateaux derivative? In other words, how do we actually find the objects $\gamma(\mathbf{z})$ or $\Gamma(\mathbf{z})$ in Theorem 1? Previewing ahead, Propositions 1 and Corollary 1.1 will establish general sufficient conditions for Gateaux differentiability of government revenue: we will show that revenue can be Gateaux differentiable even when the tax schedule is a function of multiple

[11] Note, inverse welfare functionals of the form $\int_{\mathbf{N}} \Psi(U(\mathbf{n};T))dF(\mathbf{n})$ as in Mirrlees (1971) generally do not exist unless $\dim(\mathbf{N}) = \dim(\mathbf{Z}) = 1$. For instance, in a bidimensional tax system with bijective $\mathbf{n} \mapsto \mathbf{z}$, the arguments in the proof sketch to Theorem 1 imply that such a Ψ must satisfy $\partial\Psi(U(\mathbf{n};T))/\partial U = \gamma(\mathbf{z}(\mathbf{n}))/\overline{u_c}(\mathbf{z}(\mathbf{n}))$. Such a Ψ therefore only exists if $\gamma(\mathbf{z}(\mathbf{n}))/\overline{u_c}(\mathbf{z}(\mathbf{n}))$ is constant on every iso-utility curve, which is generically not true.

[12] One may be tempted to analyze large (i.e., non-local) policy reforms under the assumption that the inverse welfare functional is society's *true* welfare functional. While the inverse functionals we construct are weighted-utilitarian (and thereby avoid the critique of Sher (2024) that non-weighted-utilitarian local welfare weights generate global inconsistencies), they should not be used in this manner because they are not unique (and analysis of large policy reforms will depend on which inverse welfare functional one uses). In contrast, analysis of local reforms only depends on *average* inverse weights at each $\mathbf{z}$ (which are uniquely defined by Equation 13), see Section 4.

choices and is non-differentiable (generating bunching) and/or when individuals have multidimensional heterogeneity, multiple optima, and face optimization frictions. To build towards Propositions 1 and Corollary 1.1, we first provide a number of analytical examples illustrating how to calculate the Gateaux derivative of government revenue and apply Theorem 1 to calculate inverse welfare functionals. In doing so, we highlight how the Gateaux derivative of revenue can be expressed in terms of behavioral responses to tax perturbations and is thus, in principle, an empirically estimable object.

3.1 Smooth Unidimensional Example

First, let us consider an example with a unidimensional type $n \in N = [\underline{n}, \overline{n}]$ with utility function $u(c, z/n)$. Suppose that we want to find an inverse welfare functional for a smooth $T(z)$ under which all individuals have a unique optimum and the single crossing property holds so that $n \mapsto z$ is a strictly increasing bijection (Mirrlees, 1971). This setting has been analyzed previously (Bourguignon and Spadaro (2010); Bargain et al. (2013); Jacobs, Jongen and Zoutman (2017); Hendren (2020)), but it is useful to start here to build intuition and then move to more complex taxation settings. Let us first calculate the Gateaux variation of $R(T)$ in the direction of some $\tau(z)$ (recall that $z(n)$ is also a function of the tax schedule even though we omit T as an argument for brevity):

$$\lim_{\epsilon \to 0} \frac{R(T + \epsilon\tau) - R(T)}{\epsilon} = \int_N \frac{\partial}{\partial \epsilon} \left(T(z(n)) + \epsilon\tau(z(n)) \right) f(n) dn = \int_N \left(T'(z(n)) \frac{\partial z}{\partial \epsilon}(n) + \tau(z(n)) \right) f(n) dn \tag{14}$$

We have the individual first order condition:

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

For all individuals with a unique optimum and a strict second order condition, we can apply the implicit function theorem to determine the impacts of a tax perturbation:

$$\frac{\partial z}{\partial \epsilon}(n) = \frac{u_1 \tau'(z) + \left[u_{11}(1 - T'(z)) + \frac{1}{n} u_{12}\right] \tau(z)}{u_{11}(1 - T'(z))^2 + \frac{2}{n} u_{12}(1 - T'(z)) + \frac{1}{n^2} u_{22} - T''(z) u_1} \equiv \underbrace{\xi(n)}_{\substack{\text{Substitution} \\ \text{Effect}}} \times \tau'(z(n)) + \underbrace{\eta(n)}_{\substack{\text{Income} \\ \text{Effect}}} \times \tau(z(n)) \tag{15}$$

Plugging Equation 15 into Equation 14 and changing the variable of integration from n

to z (with $h(z)$ denoting the income density), we find that:[13]

$$\lim_{\epsilon \to 0} \frac{R(T + \epsilon\tau) - R(T)}{\epsilon} = \int_{\underline{z}}^{\overline{z}} \left(T'(z)\xi(z)\tau'(z) + \left[1 + T'(z)\eta(z)\right]\tau(z)\right) h(z)dz \tag{16}$$

where $\overline{z} \equiv z(\overline{n})$ and $\underline{z} \equiv z(\underline{n})$. However, Equation 16 is not linear in $\tau(z)$ and Theorem 1 requires that tax revenue be Gateaux differentiable (i.e., $\lim_{\epsilon \to 0} \left(R(T + \epsilon\tau) - R(T)\right)/\epsilon$ is a bounded *linear* functional of $\tau(z)$). Using integration by parts to get rid of the $\tau'(z)$ term, Equation 16 equals:

$$\int_{\underline{z}}^{\overline{z}} \left(-\frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right] + \left[1 + T'(z)\eta(z)\right]h(z)\right)\tau(z)dz + T'(z)\xi(z)h(z)\tau(z)\Big|_{\underline{z}}^{\overline{z}} \tag{17}$$

Note that all $\tau(z)$ terms enter Equation 17 linearly so that Equation 17 is a linear functional of $\tau(z)$. Assuming that behavioral responses are bounded, then the Gateaux variation of $R(T)$ is a bounded linear functional of $\tau(z)$; thus, $R(T)$ is Gateaux differentiable and we can express Equation 17 as $\int_Z \tau(z)d\Gamma(z)$ for some measure Γ by Remark 1.[14] Hence, for any function $g(z)$:

$$\int_Z g(z)d\Gamma(z) = \int_{\underline{z}}^{\overline{z}} \left(-\frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right] + \left[1 + T'(z)\eta(z)\right]h(z)\right)g(z)dz + T'(z)\xi(z)h(z)g(z)\Big|_{\underline{z}}^{\overline{z}} \tag{18}$$

Therefore, by Theorem 1, we can construct the following inverse welfare functional:

$$\begin{aligned} W(U) &= \int_N U(n;T)d\Phi(n) = \int_Z \int_{N(z)} \frac{U(n;T)}{\overline{u_c}(z)} dF(n|z)d\Gamma(z) \\ &= \int_{\underline{z}}^{\overline{z}} \left(-\frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right] + \left[1 + T'(z)\eta(z)\right]h(z)\right)\frac{U(n(z);T)}{u_c(n(z))}dz + T'(z)\xi(z)h(z)\frac{U(n(z);T)}{u_c(n(z))}\Big|_{\underline{z}}^{\overline{z}} \end{aligned} \tag{19}$$

where the second equality in Equation 19 follows from the definition of $\Phi(\tilde{n})$ in Equation 10 (see Appendix A.1 for proof). The third equality plugs in $g(z) = \int_{N(z)} U(n;T)/\overline{u_c}(z)dF(n|z)$ into Equation 18 and then uses the fact that $n \mapsto z$ is bijective so that $\int_{N(z)} U(n;T)/\overline{u_c}(z)dF(n|z) = U(n(z);T)/u_c(n(z))$ where $n(z)$ represents the type n that optimally chooses a given z. Changing variables back from z to n in the integral in Equation 19 after multiplying and dividing by $h(z)$, the inverse welfare functional therefore takes the form:

$$W(U) = \int_N \phi(n)U(n;T)f(n)dn + \overline{\phi}U(\overline{n};T)h(\overline{z}) + \underline{\phi}U(\underline{n};T)h(\underline{z}) \tag{20}$$

with

$$\phi(n)u_c(n)h(z) = -\frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right] + \left[1 + T'(z)\eta(z)\right]h(z) \tag{21}$$

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

[14] In this case, the distribution function of Γ is given by $\Gamma(z) = -T'(\underline{z})\xi(\underline{z})h(\underline{z})\mathbb{1}[z \geq \underline{z}] + \int_{\underline{z}}^{z}\left(-\frac{\partial}{\partial s}\left[T'(s)\xi(s)h(s)\right] + \left[1 + T'(s)\eta(s)\right]h(s)\right)ds + T'(\overline{z})\xi(\overline{z})h(\overline{z})\mathbb{1}[z \geq \overline{z}]$.

$$\overline{\phi} u_c(\overline{n}) h(\overline{z}) = T'(\overline{z}) \xi(\overline{z}) h(\overline{z}) \tag{22}$$

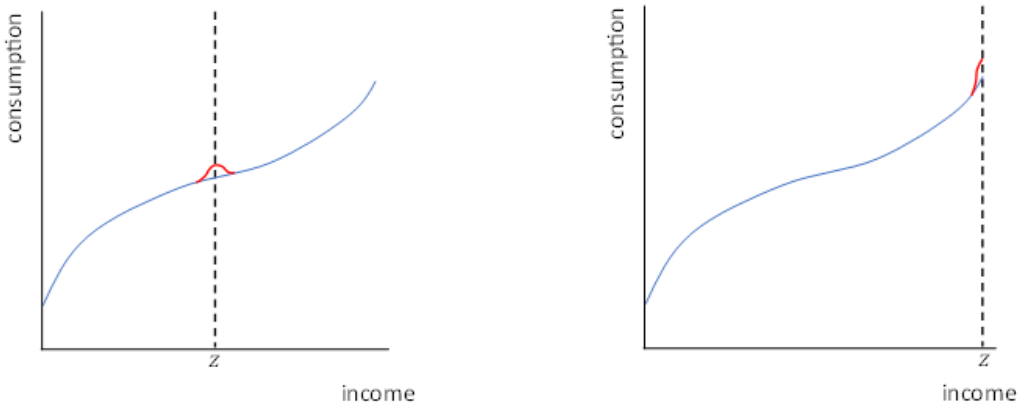
$$\underline{\phi} u_c(\underline{n}) h(\underline{z}) = -T'(\underline{z}) \xi(\underline{z}) h(\underline{z}) \tag{23}$$

where we have suppressed the argument of $z(n)$ in Equation 21 for readability. Equation 21 pins down $\phi(n)$ for each n that optimally chooses $z \in \text{Int}(Z)$ whereas Equations 22 and 23 pin down $\phi(n)$ choosing $z \notin \text{Int}(Z)$. Intuitively, Equations 21, 22, and 23 ensure that the total impact on the government's Lagrangian of every possible "bump" perturbation is zero.[15] Equation 21 ensures that at each interior z, adding a small bump to the tax schedule as in Figure 1a leaves the Lagrangian unchanged.[16] Conceptually, an interior bump perturbation leads to a mechanical welfare impact which, due to the envelope theorem, equals the left hand side of Equation 21. Moreover, an interior bump perturbation leads to a mechanical budgetary impact along with an income effect, $[1 + T'(z)\eta(z)]\, h(z)$, and also leads to a negative substitution effect to the right of z along with a positive substitution effect to the left of z; in the limit, this difference in substitution effects equals the (negative) derivative of the substitution effect, $-\partial/\partial z\, [T'(z)\xi(z)h(z)]$. Equation 22 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 22. Identical logic explains the intuition behind Equation 23. These "boundary weights" allow us to construct inverse welfare functionals that rationalize non-zero marginal tax rates at the top and bottom.[17] Note that if $h(\overline{z}) = h(\underline{z}) = 0$, then Equations 22 and 23 are vacuously satisfied; in this case these "boundary weights" can be set to zero.

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

[16] Note that Equation 21 is just a differentiated version of Equation (19) from Saez (2001).

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



(a) Interior Bump Function Perturbation (b) Boundary Bump Function Perturbation

Figure 1: Bump Function Perturbations

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, $c = z - T(z)$).

3.2 Non-Smooth Unidimensional Example

Next, let us consider an example with a unidimensional tax schedule $T(z)$ but with two dimensions of heterogeneity so that utility is given by $u(c, z/n; v)$ where the second dimension of heterogeneity is denoted by the parameter v with $(n, v) \in [\underline{n}, \overline{n}] \times [\underline{v}, \overline{v}]$. Suppose that we want to find an inverse welfare functional that rationalizes a piecewise linear tax schedule with three brackets; the marginal tax rates in the three brackets are denoted T_1, T_2, T_3 with $T_2 > T_1$ and $T_2 > T_3$ (the first kink point features increasing marginal rates and the second features decreasing marginal tax rates; generalizing to an arbitrary number of brackets will therefore be immediate).

Recall that there were two tricks used to calculate the Gateaux derivative of revenue in Section 3.1: (1) use the fact that if the tax schedule is differentiable and everyone has a unique optimum, then we can differentiate under the integral sign (Equation 14) and (2) use integration by parts to get rid of the dependence on $\tau'(z)$ that arises from individuals responding to marginal tax rate changes. However, the presence of kink points generates two problems that prevent us from using these two tricks: (1) the tax schedule is not differentiable at certain points leading to bunching and (2) the tax schedule has discontinuously decreasing marginal tax rates leading to individuals with multiple optima who "jump" between tax brackets in response to small tax perturbations.

In order to deal with these additional complexities, the high-level solution is to break

up the region of integration $\mathbf{N}$ at points where individuals bunch and/or have multiple optima. For each v, we split up the domain N into four regions: let $[\underline{n}, n_1]$ denote the set of individuals locating in the first tax bracket, $(n_1, n_2]$ denote the set of individuals bunching at the first kink, $(n_2, n_3]$ denote the set of individuals locating in the second tax bracket, and $(n_3, \overline{n}]$ denote the set of individuals locating in the third tax bracket. Hence, we write tax revenue as:

$$\int_V \Bigg\{ \underbrace{\int_{\underline{n}}^{n_1(v)} T(z(n,v)) f(n|v) dn}_{\text{First Bracket}} + \underbrace{\int_{n_1(v)}^{n_2(v)} T(z(n,v)) f(n|v) dn}_{\text{Bunch at First Kink}} + \underbrace{\int_{n_2(v)}^{n_3(v)} T(z(n,v)) f(n|v) dn}_{\text{Second Bracket}} + \underbrace{\int_{n_3(v)}^{\overline{n}} T(z(n,v)) f(n|v) dn}_{\text{Third Bracket}} \Bigg\} f(v) dv \tag{24}$$

Next, we will provide high-level intuition for how to calculate the Gateaux derivative of revenue. Starting from Equation 24, we use the Leibniz integral rule to evaluate the derivative w.r.t. ϵ on each region separately. For individuals with a unique optimum choosing z where $T(z)$ is smooth, we can differentiate under the integral sign and use integration by parts as in Section 3.1 to express the revenue impacts as a linear functional of $\tau(z)$. Next, those who bunch at the first kink point where $T(z)$ is non-differentiable do not respond to continuous tax perturbations because they are at a corner solution. As a result, the only revenue impacts of a tax perturbation for bunching individuals are mechanical effects, which are definitionally linear in $\tau(z)$. Next, marginal tax rates decrease at the second kink so this generates an individual $n_3(v)$ for each v with multiple optima (one in each of the second and third tax brackets) as in Bergstrom and Dodds (2021). These individuals have the following indifference condition where z^- and z^+ correspond to their two optima:

$$u(z^- - T(z^-) - \epsilon\tau(z^-), z^-/n_3(v); v) = u(z^+ - T(z^+) - \epsilon\tau(z^+), z^+/n_3(v); v) \tag{25}$$

Tax revenue changes as a result of $n_3(v)$ changing with the perturbation because $T(z^-) \neq T(z^+)$ (e.g., individuals moving from z^+ down to z^- reduces tax revenue).[18] Applying the implicit function theorem to Equation 25 and treating $n_3(v)$ as a function of ϵ (appealing to the envelope theorem to ignore derivatives of z^+ and z^- w.r.t. ϵ),

[18] In contrast, while the marginal bunchers $n_1(v)$ and $n_2(v)$ also change with ϵ, these Leibniz integral rule terms all cancel out because the left and right limits are equal; e.g., $\lim_{n \uparrow n_1(v)} T(z(n,v)) = \lim_{n \downarrow n_1(v)} T(z(n,v))$.

we see that $\partial n_3(v)/\partial\epsilon$ depends linearly on $\tau(z^+)$ and $\tau(z^-)$. Intuitively, the decision to move between optima depends solely on the tax levels (not marginal tax rates) at the two optima. As a result, all of the revenue impacts of a small tax perturbation are linear in $\tau(z)$ so that $R(T)$ is Gateaux differentiable. See Appendix B.3 for the explicit calculation of the Gateux derviative of government revenue and, in turn, the associated inverse welfare functional for this example.

3.3 Sparsity Based Frictions Example

There is a growing body of evidence suggesting that agents face a variety of labor supply frictions (Chetty, 2012). Fortunately, Theorem 1 does not require that individuals optimize in a frictionless environment: we made no assumptions about the choice set $A(\mathbf{n})$ for each agent $\mathbf{n}$ and we allowed the utility function to be non-differentiable in $\mathbf{z}$ (which allows there to be, for example, fixed costs of adjusting from some status quo). We will now discuss how to calculate the Gateaux derivative of revenue and apply Theorem 1 when agents face "sparsity-based frictions". In particular, suppose that rather than selecting among a continuum of choices, agents select from a sparse set of limited choices. As discussed in Anagol et al. (2022), this sparisty based model of frictions is quite general and can accommodate a variety of microfoundations, ranging from choices among full-time/part-time/no work, choices among professions, costly search, rational inattention, or imperfect targeting.

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

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

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

where type $n_1(a)$ is indifferent between earning 0 and $a/2$ and type $n_2(a)$ is indifferent between earning $a/2$ and a. Taking the Gateaux variation of revenue, we get the following, recognizing that the indifferent individuals $n_1(a)$ and $n_2(a)$ change with the tax schedule

(representing individuals "jumping" between multiple optima):

$$\int_A \Bigg\{ \underbrace{\int_{\underline{n}}^{n_1(a)} \tau(0) f(n|a) dn + \int_{n_1(a)}^{n_2(a)} \tau(a/2) f(n|a) dn + \int_{n_2(a)}^{\overline{n}} \tau(a) f(n|a) dn}_{\text{Mechanical Effect}} \Bigg\} f(a) da$$
$$+ \int_A \Bigg\{ \underbrace{[T(0) - T(a/2)] f(n_1(a)|a) \frac{\partial n_1(a)}{\partial \epsilon}}_{\text{Extensive Effect}} + \underbrace{[T(a/2) - T(a)] f(n_2(a)|a) \frac{\partial n_2(a)}{\partial \epsilon}}_{\text{Intensive Effect}} \Bigg\} f(a) da \tag{26}$$

where $\partial n_1(a)/\partial \epsilon$ and $\partial n_2(a)/\partial \epsilon$ come from applying the implicit function theorem to the following indifference conditions:

$$u(-T(0) - \epsilon\tau(0), 0; n_1(a)) = u(a/2 - T(a/2) - \epsilon\tau(a/2), a/2; n_1(a)) \tag{27}$$
$$u(a/2 - T(a/2) - \epsilon\tau(a/2), a/2; n_2(a)) = u(a - T(a) - \epsilon\tau(a), a; n_2(a)) \tag{28}$$

We show in Appendix B.4 that $\partial n_1(a)/\partial \epsilon$ and $\partial n_2(a)/\partial \epsilon$ are linear in $\tau(z)$ (for the same reason that $\partial n_3(v)/\partial \epsilon$ is linear in $\tau(z)$ in Section 3.2) so that revenue is Gateaux differentiable. Appendix B.4 also shows how to use this Gateaux derivative of revenue to construct an inverse welfare functional.

3.4 Smooth Multidimensional Example

Next, we will construct an inverse welfare functional for a smooth tax schedule $T(\mathbf{z})$ in a higher dimensional setting. At a high-level, the tricks used to calculate the Gateaux derivative of revenue are identical to Section 3.1: (1) differentiate under the integral sign and (2) use (multidimensional) integration by parts to get rid of the dependence on the vector of derivatives $\nabla_{\mathbf{z}} \tau(\mathbf{z})$.

Suppose all individuals have a unique optimum and that second order conditions hold strictly. Then, via the implicit function theorem, we show in Appendix B.5 that the derivative of $\mathbf{z}(\mathbf{n})$ with respect to ϵ can be written as:

$$\frac{\partial \mathbf{z}}{\partial \epsilon}(\mathbf{n}) = \vec{\eta}(\mathbf{n}) \tau(\mathbf{z}) + \mathbf{X}(\mathbf{n}) \cdot \nabla_{\mathbf{z}} \tau(\mathbf{z}) \tag{29}$$

where $\vec{\eta}(\mathbf{n})$ 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 Gateaux derivative of government

revenue equals:

$$\begin{aligned}
&\int_{\mathbf{N}} \left\{\tau(\mathbf{z}) + \nabla_{\mathbf{z}} T(\mathbf{z}(\mathbf{n}))\left[\vec{\eta}(\mathbf{n})\tau(\mathbf{z}) + \mathbf{X}(\mathbf{n}) \cdot \nabla_{\mathbf{z}}\tau(\mathbf{z})\right]\right\} dF(\mathbf{n}) \\
&= \int_{\mathbf{Z}} \int_{\mathbf{N}(\mathbf{z})} \left\{\tau(\mathbf{z}) + \nabla_{\mathbf{z}} T(\mathbf{z})\left[\vec{\eta}(\mathbf{n})\tau(\mathbf{z}) + \mathbf{X}(\mathbf{n}) \cdot \nabla_{\mathbf{z}}\tau(\mathbf{z})\right]\right\} dF(\mathbf{n}|\mathbf{z}) dH(\mathbf{z}) \\
&= \int_{\mathbf{Z}} \left\{\tau(\mathbf{z}) + \nabla_{\mathbf{z}} T(\mathbf{z})\left[\overline{\vec{\eta}}(\mathbf{z})\tau(\mathbf{z}) + \overline{\mathbf{X}}(\mathbf{z}) \cdot \nabla_{\mathbf{z}}\tau(\mathbf{z})\right]\right\} dH(\mathbf{z})
\end{aligned} \tag{30}$$

The second equality in Equation 30 evaluates the inner integral, representing the Gateaux variation of revenue 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 30 to get rid of the derivatives of $\tau(\mathbf{z})$ by appealing to multi-dimensional integration by parts:

Lemma 1 (Multidimensional Integration by Parts)**.** *For a continuously differentiable function $\tau(\mathbf{z})$ and a continuously differentiable vector field $\mathbf{v}(\mathbf{z})$, where $\mathbf{Z} \in \mathbb{R}^J$ is connected, bounded, and open with piecewise smooth boundary $\partial\mathbf{Z}$, we have the 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.[19]

Assuming that the average behavioral effects $\overline{\mathbf{X}}(\mathbf{z})$ are smooth and the distribution of incomes $H(\mathbf{z})$ admits a differentiable density function $h(\mathbf{z})$, we can appeal to Lemma 1 (recognizing that $\nabla_{\mathbf{z}} T(\mathbf{z})\overline{\mathbf{X}}(\mathbf{z})h(\mathbf{z})$ is a vector field on $\mathbf{Z}$) to rewrite Equation 30 as:

$$\int_{\mathbf{Z}} \left\{ \left[1 + \nabla_{\mathbf{z}} T(\mathbf{z})\overline{\vec{\eta}}(\mathbf{z})\right] h(\mathbf{z}) - \nabla_{\mathbf{z}} \cdot \left[\nabla_{\mathbf{z}} T(\mathbf{z})\overline{\mathbf{X}}(\mathbf{z})h(\mathbf{z})\right] \right\}\tau(\mathbf{z}) d\mathbf{z} + \int_{\partial\mathbf{Z}} \left\{\nabla_{\mathbf{z}} T(\mathbf{z})\overline{\mathbf{X}}(\mathbf{z})h(\mathbf{z}) \cdot \rho\right\}\tau(\mathbf{z}) dS \tag{31}$$

Importantly, note that Equation 31 is *linear* in $\tau(\mathbf{z})$ so that revenue is Gateaux differentiable in the tax schedule (assuming all terms in Equation 31 are bounded) and we can express Equation 31 as $\int_Z \tau(z) d\Gamma(z)$ for some measure Γ so that for any function $g(z)$:

$$\begin{aligned}
\int_{\mathbf{Z}} g(\mathbf{z}) d\Gamma(\mathbf{z}) = &\int_{\mathbf{Z}} \left\{ \left[1 + \nabla_{\mathbf{z}} T(\mathbf{z})\overline{\vec{\eta}}(\mathbf{z})\right] h(\mathbf{z}) - \nabla_{\mathbf{z}} \cdot \left[\nabla_{\mathbf{z}} T(\mathbf{z})\overline{\mathbf{X}}(\mathbf{z})h(\mathbf{z})\right] \right\} g(\mathbf{z}) d\mathbf{z} \\
&+ \int_{\partial\mathbf{Z}} \left\{\nabla_{\mathbf{z}} T(\mathbf{z})\overline{\mathbf{X}}(\mathbf{z})h(\mathbf{z}) \cdot \rho\right\} g(\mathbf{z}) dS
\end{aligned} \tag{32}$$

[19] 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 $\int_{\mathbf{Z}} [\nabla_{\mathbf{z}} \cdot \mathbf{v}(\mathbf{z})]\tau(\mathbf{z}) d\mathbf{z}$.

Therefore, by Theorem 1, we can construct the following inverse welfare functional:

$$\begin{aligned} W(U) &= \int_N U(n;T)d\Phi(n) = \int_{\mathbf{Z}}\int_{\mathbf{N(z)}} \frac{U(\mathbf{n};T)}{\overline{u_c}(\mathbf{z})} dF(\mathbf{n}|\mathbf{z})d\Gamma(\mathbf{z}) \\ &= \int_{\mathbf{Z}} \left(\left[1 + \nabla_{\mathbf{z}} T(\mathbf{z})\overline{\overline{\eta}}(\mathbf{z})\right] h(\mathbf{z}) - \nabla_{\mathbf{z}} \cdot \left[\nabla_{\mathbf{z}} T(\mathbf{z})\overline{\mathbf{X}}(\mathbf{z})h(\mathbf{z})\right]\right) \left[\int_{\mathbf{N(z)}} \frac{U(\mathbf{n};T)}{\overline{u_c}(\mathbf{z})} dF(\mathbf{n}|\mathbf{z})\right] d\mathbf{z} \\ &+ \int_{\partial\mathbf{Z}} \nabla_{\mathbf{z}} T(\mathbf{z})\overline{\mathbf{X}}(\mathbf{z}) \left[\int_{\mathbf{N(z)}} \frac{U(\mathbf{n};T)}{\overline{u_c}(\mathbf{z})} dF(\mathbf{n}|\mathbf{z})\right] h(\mathbf{z}) \cdot \rho dS \end{aligned} \tag{33}$$

where the second equality in Equation 33 follows from the definition of Φ in Equation 10 (see Appendix A.1 for proof) and the third equality plugs in $g(\mathbf{z}) = \int_{\mathbf{N(z)}} U(\mathbf{n};T)/\overline{u_c}(\mathbf{z})dF(\mathbf{n}|\mathbf{z})$ into Equation 32. Hence, this is an inverse welfare functional of the form:

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

where $\phi(\mathbf{n})$ for those locating at each $\mathbf{z} \in \text{Int}(\mathbf{Z})$ is given by:[20]

$$\phi(\mathbf{n}) = \frac{\left[1 + \nabla_{\mathbf{z}} T(\mathbf{z})\overline{\overline{\eta}}(\mathbf{z})\right] h(\mathbf{z}) - \nabla_{\mathbf{z}} \cdot \left[\nabla_{\mathbf{z}} T(\mathbf{z})\overline{\mathbf{X}}(\mathbf{z})h(\mathbf{z})\right]}{\overline{u_c}(\mathbf{z})h(\mathbf{z})} \tag{34}$$

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

$$\phi(\mathbf{n}) = \frac{\nabla_{\mathbf{z}} T(\mathbf{z})\overline{\mathbf{X}}(\mathbf{z}) \cdot \rho}{\overline{u_c}(\mathbf{z})} \tag{35}$$

There are two takeaways from Sections 3.1, 3.2, 3.3, and 3.4. First, revenue can be Gateaux differentiable even with multidimensional tax schedules, bunching (generated by non-differentiable tax schedules), jumping (generated by individuals with multiple optima), and limited choice sets. Second, the Gateaux derivative of government revenue is an empirical object that depends on substitution effects, income effects, jumping effects, bunching masses, and the density of choices $\mathbf{z}$. In principle, all of these objects can be estimated given sufficient tax variation. In practice, it is difficult to observe sufficient tax variation to estimate the requisite heterogeneous behavioral responses to tax reforms; hence, practitioners attempting to construct inverse welfare functionals will typically need to make assumptions about how behavioral responses vary across the choice distribution, such as making structural assumptions on utility and calibrating or assuming elasticities are constant across the choice distribution.

[20] Rearrangements of the first order conditions in Equations 34 and 35 have been derived previously in Golosov, Tsyvinski and Werquin (2014) and Spiritus et al. (2022).

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

Next, we provide general sufficient conditions for revenue 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 on a finite set of measure zero surfaces in* $\mathbf{N}$
3. *The set* $\mathbf{Z}$ *of chosen* $\mathbf{z} = (z_1, z_2, ..., z_J)$ *is the closure of an open set in* $\mathbb{R}^J$
4. *The set of individuals whose second order conditions hold weakly is measure zero*
5. *(5 technical regularity conditions discussed in the proof)*

Proof. See Appendix B.6. □

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/or 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. Note that all of the assumptions of Proposition 1 can be readily verified under the given tax schedule and are therefore not endogenous to an unknown optimal schedule.

While the proof to Proposition 1 is quite long, much of the intuition has been discussed in the above examples. To show revenue is Gatueaux differentiable, first split up the type space into regions where people respond smoothly to tax changes, regions where people may "jump" between multiple optima, and regions where the tax schedule is non-differentiable causing bunching. Our assumptions in Proposition 1 ensure that at most $\mathbf{z}$, the tax schedule is smooth and individuals have a unique optimum with their second order condition holding strictly; hence, individuals who choose these $\mathbf{z}$ respond to tax changes according to the implicit function theorem (i.e., via standard income and substitution effects). For these individuals, we can use multidimensional integration by parts as in Section 3.4 to express the revenue impact of a tax change as a linear functional of $\tau(\mathbf{z})$. Next, there may be surfaces where individuals have multiple optima and thereby react to a tax change by jumping to a different $\mathbf{z}$; however, under the stated assumptions, we can

show that these jumping effects can be expressed as a linear functional of $\tau(\mathbf{z})$ because the decision to jump depends only on the tax level at each $\mathbf{z}$. Finally, there are surfaces along which the tax schedule is non-differentiable. Consider the case with two choice variables. A non-differentiable surface for the tax schedule in this case is a ridge in three dimensional space (visually, imagine a creased piece of paper). Almost all individuals who choose $\mathbf{z}$ on this ridge strictly prefer their chosen $\mathbf{z}$ to any $\mathbf{z}$ that is off the ridge (consider an indifference surface that is tangent to a given $\mathbf{z}$ on the ridge). Hence, in response to any small tax perturbation, these individuals may move *along* the ridge but do not move off the ridge. For small tax perturbations, we can then recast the optimization problem for individuals on the ridge as a choice over some parameter t which parameterizes the ridge. Thus, for these individuals, we can reduce their problem to a unidimensional optimization problem wherein we can use integration by parts (integrating over the parameter t) to express the revenue impact of a tax change as a linear functional just as in Section 3.1.[21]

Finally, we note that in the case when agents face sparsity based frictions as in Section 3.3, many of the conditions in Proposition 1 are not needed:

Corollary 1.1. *The following are sufficient conditions for $R(T)$ to be Gateaux differentiable if agents face sparsity based frictions:*

1. *Individuals have multiple optima only on a finite set of measure zero surfaces in* $\mathbf{N}$
2. *Almost all individuals with multiple optima just have two optima.*

Proof. See Appendix B.7. □

In summary, Proposition 1 and Corollary 1.1 prove that Gateaux differentiability of government revenue is a relatively mild restriction. Taken together, Theorem 1, Proposition 1, and Corollary 1.1 establish that one can construct inverse welfare functionals in a wide variety of situations.

4 Policy Relevance

We now explore how the inverse welfare functional can be used to answer important policy relevant questions. First and foremost, the inverse welfare functional can be used to assess whether the observed tax schedule is Pareto efficient (and, if not, identify Pareto improvements). We are not the first paper to recognize this: for example, Bourguignon

[21] 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.

and Spadaro (2010) illustrate that positive welfare weights are a necessary condition for Pareto efficiency of (unidimensional) income tax schedules if individuals respond smoothly to tax reforms, Lorenz and Sachs (2016) and Hendren (2020) extend this result to allow for participation responses, and Bierbrauer, Boyer and Hansen (2023) further extend this result to show that positive welfare weights are necessary for Pareto efficiency and sufficient for (local) Pareto efficiency.[22] Finally, Spiritus et al. (2022) prove that positive welfare weights are a necessary condition for Pareto efficiency of multidimensional tax schedules if individuals respond smoothly to tax reforms. Our Proposition 2 below builds on this previous work by providing a simple characterization of Pareto efficient tax schedules that allows for complex behavioral responses (e.g., tax schedules can be non-differentiable generating bunching, individuals can face optimization frictions, individuals can "jump" between multiple optima) and allows for multidimensional heterogeneity and multidimensional decisions:

Proposition 2. *Suppose the conditions of Theorem 1 hold.*

1. *(Pareto Inefficiency) If the inverse welfare functional W defined in Theorem 1 is not positive so that $\exists g(\mathbf{n}) \in C(\mathbf{N})$ with $g(\mathbf{n}) \geq 0$ $\forall \mathbf{n}$ and $W(g) < 0$, then there exist Pareto improving tax reforms.*
2. *(Local Pareto Efficiency) If the inverse welfare functional W defined in Theorem 1 is strictly positive so that $W(g) > 0$ for all functions $g(\mathbf{n}) \in C(\mathbf{N})$ with $g(\mathbf{n}) \geq 0$ $\forall \mathbf{n}$ and $g(\mathbf{n}) > 0$ on a positive measure set, then there do not exist any (marginal) Pareto improving tax reforms.*

Proof. See Appendix A.2. □

The intuition for Proposition 2 is that negative inverse welfare weights imply that the tax rate at that choice level is beyond the local Laffer rate. Hence, we can construct a Pareto improvement by decreasing taxes at $\mathbf{z}$'s where average welfare weights are negative (see the proof for explicit construction of these tax reforms). Conversely, if the inverse welfare functional is strictly positive, then any small tax perturbation that (weakly) raises revenue must decrease utility for at least some individuals.[23] While the intuitive

[22] A local Pareto efficient tax schedule is one in which there are no small tax reforms that generate Pareto improvements.

[23] If the inverse welfare functional defined in Theorem 1 is a strictly positive *global* inverse welfare functional then it is straight-forward to strengthen part 2 of Proposition 2 to state that there are no Pareto improving tax reforms of any size (rather than there are no Pareto improving marginal tax

connection between inverse welfare weights and Pareto efficiency has been understood for some time in simpler taxation models, Proposition 2 extends these results to more complex settings and thereby illustrates that inverse welfare functionals in these more complex settings are still useful, policy-relevant objects.

While previous literature on inverse welfare weights has focused primarily on the link with Pareto efficiency (perhaps because the conclusions are independent of the normative choice of true welfare weights), we will now also illustrate how inverse welfare functionals can be used along with knowledge of the true welfare weights to explicitly construct welfare-improving marginal tax reforms and determine the optimal local reform direction. These results are useful in situations for which existing methods are not yet powerful enough to solve for the optimal tax schedule: unidimensional income taxation with multidimensional heterogeneity where a continuum of types have multiple optima as in Section 3.2; unidimensional income taxation with frictions as in Section 3.3; and multidimensional taxation other than special cases where $\dim(\mathbf{N}) = \dim(\mathbf{Z})$ and individuals respond smoothly (Spiritus et al., 2022), or utility is linear in type (Boerma, Tsyvinski and Zimin (2022), Krasikov and Golosov (2024)), or utility satisfies a generalized single-crossing property (Dodds, 2023). Proposition 3 characterizes the set of welfare improving tax reform directions in terms of the inverse and actual welfare functionals:

Proposition 3. *Suppose that the conditions of Theorem 1 part (1) hold and consider the inverse welfare functional* $\int_{\mathbf{N}} \phi(\mathbf{n})U(\mathbf{n};T)f(\mathbf{n})d\mathbf{n}$ *where* $\phi(\mathbf{n})$ *is defined by Equation 9. Suppose that the actual welfare functional can be expressed as* $\int_{\mathbf{N}} \phi^A(\mathbf{n})U(\mathbf{n};T)f(\mathbf{n})d\mathbf{n}$.[24] *A tax perturbation in the direction* $\tau(\mathbf{z})$ *along with a transfer* $\tau_0 = -\int_{\mathbf{N}} \phi(\mathbf{n})u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n}))dF(\mathbf{n})$ *that closes the budget constraint is welfare improving if and only if:*

$$\int_{\mathbf{Z}} \left[\int_{\mathbf{N}(\mathbf{z})} [\phi(\mathbf{n}) - \phi^A(\mathbf{n})]u_c(\mathbf{n})dF(\mathbf{n}|\mathbf{z}) \right] \tau(\mathbf{z})dH(\mathbf{z}) > 0 \tag{36}$$

as long as the actual weights $\phi^A(\mathbf{n})$ *are normalized (via multiplication by a constant) to*

reforms). Appendix B.1 illustrates conditions which guarantee existence of a global inverse welfare functional of the form in Equation 4. Hence, under these conditions, if the inverse welfare functional from Theorem 1 is the unique (local) inverse welfare functional of the form in Equation 4, then this will be a global inverse welfare functional as well. Unfortunately, this will typically only be the case when $\mathbf{n} \mapsto \mathbf{z}$ is bijective.

[24] The assumption that the inverse and actual welfare functionals can be expressed using welfare weights is for expositional simplicity; Appendix A.3 shows that Proposition 3 holds more generally when the inverse and actual welfare functionals are arbitrary continuous linear functionals $\int_{\mathbf{N}} U(\mathbf{n};T)d\Phi(\mathbf{n})$ and $\int_{\mathbf{N}} U(\mathbf{n};T)d\Phi^A(\mathbf{n})$.

satisfy:

$$\int_{\mathbf{N}} \phi^A(\mathbf{n}) u_c(\mathbf{n}) dF(\mathbf{n}) = \int_{\mathbf{N}} \phi(\mathbf{n}) u_c(\mathbf{n}) dF(\mathbf{n}) = 1 \tag{37}$$

Proof. See Appendix A.3. □

Proposition 3 builds on the results from Saez and Stantcheva (2016), who illustrate how to assess the desirability of small budget neutral tax reforms (using only the actual welfare weights) but do not show how to construct budget neutral tax reforms. In contrast, we can use Proposition 3 to *explicitly* construct welfare improving budget-neutral tax perturbations using the inverse welfare functional. Proposition 3 shows that *any* tax reform that increases the tax burden where average inverse welfare weights are greater than average actual welfare weights $\left(\left[\int_{\mathbf{N}(\mathbf{z}}[\phi(\mathbf{n}) - \phi^A(\mathbf{n})]u_c(\mathbf{n})dF(\mathbf{n}|\mathbf{z})\right] > 0\right)$ and decreases the tax burden where inverse welfare weights are less than actual welfare weights (and closes the budget by changing the lump-sum transfer by $\tau_0 = -\int_{\mathbf{N}} \phi(\mathbf{n}) u_c(\mathbf{n}) \tau(\mathbf{z}(\mathbf{n})) dF(\mathbf{n})$) is welfare improving. For example, suppose that society's actual welfare weights for households earning more than \$500,000 are lower than the corresponding inverse welfare weights and society's actual welfare weights for households earning less than \$30,000 are higher than the corresponding inverse welfare weights. Then *any* small reform that reduces taxes on those earning less than \$30,000, increases taxes on those earning more than \$500,000, and closes the budget via the lump sum transfer is welfare improving.[25]

Finally, we establish how to construct the optimal tax reform direction, building off the work of Diewert (1978) and Dixit (1979) who derive optimal tax reforms in finite dimensional settings of commodity taxation. Corollary 3.1 below uses the Cauchy-Schwarz inequality to prove that the optimal reform direction is given by the average difference between inverse welfare weights and actual welfare weights at each choice level $\mathbf{z}$:

Corollary 3.1. *Suppose that the conditions of Proposition 3 hold with the Gateaux derivative of revenue given by* $\int_{\mathbf{Z}} \tau(\mathbf{z})\gamma(\mathbf{z})dH(\mathbf{z})$. *Suppose that we seek to choose a budget neutral tax perturbation direction* $\tau(\mathbf{z})$ *to maximize the (first order) welfare impact subject to an*

[25] Note, Propositions 2 and 3 implicitly assume that the tax schedule can be reformed in the direction of any continuous function $\tau(\mathbf{z})$. If the government faces exogenous constraints on the form of the tax schedule (e.g., bounds on marginal tax rates or a convexity requirement that marginal tax rates be increasing), then some reform directions may not be feasible; in this case, one must verify ex-post whether any given proposed Pareto improving or welfare improving reform direction is feasible.

L^2 norm constraint:[26]

$$\begin{aligned} &\max_{\tau(\mathbf{z})} \int_{\mathbf{N}} -\phi^A(\mathbf{n}) u_c(\mathbf{n}) \tau(\mathbf{z}(\mathbf{n})) dF(\mathbf{n}) \\ &s.t. \int_{\mathbf{Z}} |\tau(\mathbf{z})|^2 dH(\mathbf{z}) = 1 \\ &\int_{\mathbf{Z}} \tau(\mathbf{z}) \gamma(\mathbf{z}) dH(\mathbf{z}) = 0 \end{aligned} \tag{38}$$

The solution to Problem 38 is given by:

$$\tau(\mathbf{z}) = \frac{\int_{\mathbf{N}(\mathbf{z})} [\phi(\mathbf{n}) - \phi^A(\mathbf{n})] u_c(\mathbf{n}) dF(\mathbf{n}|\mathbf{z})}{|| \int_{\mathbf{N}(\mathbf{z})} [\phi(\mathbf{n}) - \phi^A(\mathbf{n})] u_c(\mathbf{n}) dF(\mathbf{n}|\mathbf{z}) ||_{L^2}} \tag{39}$$

where $\phi(\mathbf{n})$ are defined in Equation 9 and the actual weights $\phi^A(\mathbf{n})$ are normalized (via multiplying by a constant) to satisfy:

$$\int_{\mathbf{Z}} \left[\int_{\mathbf{N}(\mathbf{z})} \phi^A(\mathbf{n}) u_c(\mathbf{n}) dF(\mathbf{n}|\mathbf{z}) \right] \left[\int_{\mathbf{N}(\mathbf{z})} \phi(\mathbf{n}) u_c(\mathbf{n}) dF(\mathbf{n}|\mathbf{z}) \right] dH(\mathbf{z}) = \int_{\mathbf{Z}} \left[\int_{\mathbf{N}(\mathbf{z})} \phi(\mathbf{n}) u_c(\mathbf{n}) dF(\mathbf{n}|\mathbf{z}) \right]^2 dH(\mathbf{z}) \tag{40}$$

Proof. See Appendix A.4. □

The conclusion from this section is that even in complex settings, inverse welfare functionals can still be used to assess Pareto efficiency, construct Pareto improving reforms, and construct welfare-improving budget neutral tax reforms (which is useful given that current methods typically cannot solve for the optimal tax schedule when we allow for the aforementioned complexities).

5 General Equilibrium

The theory developed so far has made very few assumptions on the utility function, choice variables $\mathbf{z}$, or primitives $\mathbf{n}$. However, one key restriction that we have made, consistent with most of the optimal taxation literature, is that we have only considered a "partial equilibrium" setting in which individuals' decisions $\mathbf{z}$ do not impact the economy more broadly. Next, we will show how to construct inverse welfare functionals when there are "general equilibrium" effects.

[26] One has to restrict the size of the tax reform in order to ensure that the reform remains local; using an L^2 constraint is standard (Diewert (1978) or Dixit (1979)), although we could also consider L^p constraints on $\tau(\mathbf{z})$ for $1 < p < \infty$ and apply the Hölder inequality to construct optimal reform directions in these cases.

5.1 Example: Endogenous Wages

To build intuition, we will first construct an inverse welfare functional in the context of a labor demand/labor supply model where, contrary to much of the optimal taxation literature, the labor demand side of the market is *not* assumed to be infinitely elastic implying that wages are endogenous to the tax schedule. Much of the analysis of tax perturbations in this section is similar to Sachs, Tsyvinski and Werquin (2020) who explore optimal income taxation with endogenous wages. We illustrate how to compute inverse welfare functionals whereas Sachs, Tsyvinski and Werquin (2020) show how to compute optimal tax schedules (our model also allows for firms to earn non-zero profits which further complicates the analysis). Consider a population of individuals indexed by a unidimensional type n who choose an income $z = wnl$, where l is labor supply and w is a wage paid on effective effort nl, to maximize a quasi-linear iso-elastic utility function:

$$\begin{aligned} U(n;T,w) = \max_{z} \; & c - \frac{[z/(nw)]^{1+k}}{1+k} \\ \text{s.t. } & c = z - T(z) + s(n)\pi(w) \end{aligned} \tag{41}$$

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. Thus, firm profits are given by $\pi = Y(L) - wL$ where $Y(L)$ is the firm's production function. Market clearing requires:

$$L = \int_N nl(n)dF(n) \tag{42}$$

The government's Lagrangian is given by:

$$W(U(n;T,w)) + \lambda \left[\int_N T(z(n))dF(n) - E \right] \tag{43}$$

We will find an inverse welfare functional $W(U(n;T,w)) = \int_N \phi(n)U(n;T,w)dF(n)$. Next, let us take the Gateaux variation of Equation 43 in the direction of $\tau(z)$, assuming that the tax schedule is smooth, $n \mapsto z$ is a smooth bijective function, individual second order conditions hold strictly, and $\partial w/\partial \epsilon$ exists (importantly, recall that $z(n)$ is

also a function of the tax schedule and the wage but we omit these arguments for clarity):

$$\begin{aligned}&\int_N \phi(n)\left[-\tau(z(n)) + \left(\frac{z(n)}{nw}\right)^{1+k}\frac{1}{w}\frac{\partial w}{\partial \epsilon} + s(n)\pi'(w)\frac{\partial w}{\partial \epsilon}\right] dF(n)\\ &+ \lambda\int_N \left(\tau(z) + T'(z(n))\frac{\partial z(n)}{\partial \epsilon}\bigg|_w + T'(z(n))\frac{\partial z(n)}{\partial w}\bigg|_\epsilon \frac{\partial w}{\partial \epsilon}\right) dF(n)\end{aligned} \tag{44}$$

Next, we need to express $\partial w/\partial \epsilon$ in terms of $\tau(z)$. We show in Appendix B.9 that $\partial w/\partial \epsilon$ is Gateaux differentiable in T and that if $h(z) \to 0$ at the top and bottom of the income distribution (this assumption is not necessary - it just simplifies the subsequent expressions), then $\partial w/\partial \epsilon = \int_Z p(z)\tau(z)dz$ for a function $p(z)$. Appendix B.9 shows that we can express Equation 44 as the following linear functional:

$$\begin{aligned}&-\int_Z \left[\underbrace{\phi(n(z))h(z)}_{\text{Direct Welfare Effect}} \underbrace{- p(z)\left(\int_Z \phi(n(\tilde{z}))\left[\left(\frac{\tilde{z}}{n(\tilde{z})w}\right)^{1+k}\frac{1}{w} + s(n(\tilde{z}))\pi'(w)\right]h(\tilde{z})d\tilde{z}\right)}_{\text{Indirect Welfare Effect}}\right]\tau(z)dz\\ &+\lambda\int_Z \left(\underbrace{h(z) - \frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right]}_{\text{Direct Budgetary Effect}} + \underbrace{p(z)\int_N \left(T'(z(n))\frac{\partial z(n)}{\partial w}\bigg|_\epsilon\right)dF(n)}_{\text{Indirect Budgetary Effect}}\right)\tau(z)dz\end{aligned} \tag{45}$$

Intuitively, Equation 45 captures two separate impacts: the direct impacts of the tax change and the indirect impacts of the wage change that results from changes in labor supply as a result of the tax change. A local inverse welfare functional is a set of weights $\phi(n)$ such that Equation 45 equals zero for all possible $\tau(z)$. Normalizing $\lambda = 1$, this will hold as long as $\forall z$:

$$\begin{aligned}&\phi(n(z))h(z) - p(z)\left(\int_Z \phi(n(\tilde{z}))\left[\left(\frac{\tilde{z}}{n(\tilde{z})w}\right)^{1+k}\frac{1}{w} + s(n(\tilde{z}))\pi'(w)\right]h(\tilde{z})d\tilde{z}\right)\\ &= h(z) - \frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right] + p(z)\int_N \left(T'(z(n))\frac{\partial z(n)}{\partial w}\bigg|_\epsilon\right)dF(n)\end{aligned} \tag{46}$$

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 46 can be expressed as:

$$\phi(n(z)) = \chi(z) + \int_Z K(z,\tilde{z})\phi(n(\tilde{z}))d\tilde{z} \tag{47}$$

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 47 (i.e., a fixed point of the contraction mapping) follows immediately from the contraction mapping theorem.[27] Economically, the condition that $\int_Z |K(z,\tilde{z})|d\tilde{z} < 1 \; \forall z$ ensures that the direct welfare effect of a change in taxes is larger than the indirect welfare effect of a change in taxes (defined in Equation 45). Thus, we have constructed a local inverse welfare functional that supports a given tax schedule (under some regularity conditions) in a labor supply/labor demand model with endogenous wages.

5.2 Existence Theorem with General Equilibrium Effects

Next, we will state and prove Theorem 2, which illustrates how to construct inverse welfare functionals with GE effects more generally. We augment the model from Section 2.1 to allow individual utility to also depend on a vector, $\mathbf{w}$, of "general equilibrium" parameters: $\mathbf{w}$ might consist of prices, wages, externalities, or other quantities that are impacted in some way by aggregate individual decisions (and hence depend on taxes). Thus, individuals maximize the following utility function:

$$\begin{aligned} U(n;T,\mathbf{w}) = \max_{\mathbf{z}} \; & u\left(c,\mathbf{z};\mathbf{n},\mathbf{w}\right) \\ \text{s.t. } & c = y(\mathbf{z},\mathbf{w}) - T(\mathbf{z}) \end{aligned} \tag{48}$$

Theorem 2. *Consider $T(\mathbf{z})$ with $R(T) = E$ such that $\mathbf{Z}$ is compact and for every $\mathbf{z}$ almost all $\mathbf{n}$ have a unique optimum.*[28] *Suppose that:*

1. *$R(T)$ is Gateaux differentiable with*
$$\lim_{\epsilon \to 0} \frac{R(T+\epsilon\tau) - R(T)}{\epsilon} = \int_{\mathbf{Z}} \tau(\mathbf{z}) d\Gamma(\mathbf{z})$$
2. *Each $w_i \in \mathbf{w}$ is Gateaux differentiable with*
$$\lim_{\epsilon \to 0} \frac{w_i(T+\epsilon\tau) - w_i(T)}{\epsilon} = \int_{\mathbf{Z}} \tau(\mathbf{z}) dP_i(\mathbf{z})$$

Then we can construct an inverse welfare functional $\int_{\mathbf{N}} U(\mathbf{n};T)d\Phi(\mathbf{n})$ with:

$$\Phi(\tilde{\mathbf{n}}) = \int_{\mathbf{Z}} \int_{\{\mathbf{n} \leq \tilde{\mathbf{n}}\} \cap \mathbf{N}(\mathbf{z})} \frac{1}{\overline{u_c}(\mathbf{z})} dF(\mathbf{n}|\mathbf{z}) d\Phi_2(\mathbf{z}) \tag{49}$$

[27] 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 47 until convergence.

[28] Appendix A.5 shows how to relax this assumption to only require that a single $\mathbf{n}$ has a unique optimum at each $\mathbf{z}$.

where Φ_2 *is a measure defined by the following integral equation:*

$$\Phi_2(\tilde{\mathbf{z}}) = \Gamma(\tilde{\mathbf{z}}) + \sum_i P_i(\tilde{\mathbf{z}}) \int_{\mathbf{Z}} \overline{\frac{u_{w_i}}{u_c}}(\mathbf{z}) d\Phi_2(\mathbf{z}) \tag{50}$$

and where $\overline{u_{w_i}/u_c}(\mathbf{z})$ *represents average willingness-to-pay for an increase in* w_i *at* $\mathbf{z}$*. Equation 50 has a solution if the direct welfare impacts of changing taxes are larger than the maximum possible indirect welfare impacts of changing* $\mathbf{w}$*:*

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

where the supnorm is taken over $\mathbf{z}$ *and* $||P_i||_{TV}$ *denotes the total variation norm:* $||P_i||_{TV} \equiv \sup_{||\tau||_\infty \leq 1} \int_{\mathbf{Z}} \tau(\mathbf{z}) dP_i(\mathbf{z})$*.*

Proof. See Appendix A.5. □

The 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 key idea is as follows: if the "direct" impacts of a tax perturbation on utility are larger than the "indirect" impacts of a tax perturbation on utility (via impacts on general equilibrium objects $\mathbf{w}$), then the equation that pins down a local inverse welfare functional is a contraction and hence has a solution. The proof is much more technical because the government's first order condition is an integral equation formulated in a measure space, but all of the intuition for Theorem 2 can be understood from the previous labor supply/labor demand example.

5.3 Policy Relevance in General Equilibrium

Next, we will illustrate the policy relevance of calculating inverse welfare functionals in general equilibrium by establishing that the results from Section 4 in partial equilibrium settings carry over with minor modifications to the general equilibrium case. First, the inverse welfare functional can still be used to assess whether the observed tax schedule is Pareto efficient (and if not, identify Pareto improvements) in general equilibrium settings. To the best of our knowledge, this is the first result illustrating how to assess Pareto efficiency for tax schedules with general equilibrium effects with the recent exception of Jacquet and Lehmann (2025), who illustrate how to assess Pareto efficiency in general equilibrium under perfect competition:

Proposition 2 GE. *Suppose the conditions of Theorem 2 hold.*

1. *(Pareto Inefficiency) Suppose* $\mathbf{n} \mapsto \mathbf{z}$ *is bijective. If the inverse welfare functional* W *defined in Theorem 2 is not positive so that* $\exists g(\mathbf{n}) \in C(\mathbf{N})$ *with* $g(\mathbf{n}) \geq 0$ $\forall \mathbf{n}$ *and* $W(g) < 0$*, then there exist Pareto improving tax reforms.*
2. *(Local Pareto Efficiency) If the inverse welfare functional* W *defined in Theorem 2 is strictly positive so that* $W(g) > 0$ *for all functions* $g(\mathbf{n}) \in C(\mathbf{N})$ *with* $g(\mathbf{n}) \geq 0$ $\forall \mathbf{n}$ *and* $g(\mathbf{n}) > 0$ *on a positive measure set, then there do not exist any (marginal) Pareto improving tax reforms.*

Proof. See Appendix A.6. □

Proposition 2 GE presents very similar necessary and sufficient conditions for Pareto efficiency as Proposition 2 in the partial equilibrium setting. The only substantive difference relative to Proposition 2 comes in the "Pareto Inefficiency" statement above: in the general equilibrium case, we require that the mapping between types and choices, $\mathbf{n} \mapsto \mathbf{z}$, is bijective in order to confirm that a tax schedule is Pareto inefficient. Intuitively, if the inverse welfare functional is not positive (e.g., features negative welfare weights), then the proof to Proposition 2 GE shows how to construct a tax reform that increases *average* utility at each $\mathbf{z}$; however, without additional structure on what the GE effects are, we cannot rule out the possibility that any such tax reform hurts some individuals at each $\mathbf{z}$ while still improving utility on average at each $\mathbf{z}$.[29] However, if $\mathbf{n} \mapsto \mathbf{z}$ is bijective, then increasing average utility at each $\mathbf{z}$ implies utility is increased for each $\mathbf{n}$ so that we have a Pareto improving tax reform.[30]

Next, we illustrate that we can also characterize the set of welfare improving (small) tax reforms with general equilibrium effects:

Proposition 3 GE. *Suppose that the conditions of Theorem 2 hold, the inverse welfare function constructed in Theorem 2 can be written as* $\int_{\mathbf{N}} \phi(\mathbf{n}) U(\mathbf{n}; T) dF(\mathbf{n})$*, and for each* $w_i \in \mathbf{w}$ *that* $\lim_{\epsilon \to 0} \left(w_i(T + \epsilon\tau) - w_i(T)\right)/\epsilon = \int_{\mathbf{Z}} \tau(\mathbf{z}) p_i(\mathbf{z}) dH(\mathbf{z})$.[31] *Then a tax perturba-*

[29] In the partial equilibrium case, if a tax reform increases average utility at a given $\mathbf{z}$, then by the envelope theorem it must decrease the tax burden at that $\mathbf{z}$, meaning it must increase utility for all types choosing $\mathbf{z}$ given $u_c > 0$; this is why this condition is not needed in the partial equilibrium case.

[30] Note, this Pareto improving tax reform is no longer as simple as "lower taxes where inverse weights are negative" as in the partial equilibrium case; instead the Pareto improvement is expressed as an integral equation (which we prove always has a solution under the conditions of Theorem 2).

[31] Appendix B.10 shows how to extend Proposition 3 GE to allow for general inverse and actual welfare functionals $\int_{\mathbf{N}} U(\mathbf{n}; T) d\Phi(\mathbf{n})$ and $\int_{\mathbf{N}} U(\mathbf{n}; T) d\Phi^A(\mathbf{n})$ and general Gateaux derivatives of each w_i given by $\int_{\mathbf{Z}} \tau(\mathbf{z}) dP_i(\mathbf{z})$.

tion in the direction of $\tau(\mathbf{z})+\tau_0$ *where* τ_0 *is a lump sum transfer that balances the budget (and is explicitly defined in Equation 144 in the proof) is welfare improving if and only if:*

$$\int_{\mathbf{Z}}\left\{\int_{\mathbf{N}(\mathbf{z})}[\phi(\mathbf{n})-\phi^{A}(\mathbf{n})]u_c(\mathbf{n})dF(\mathbf{n}|\mathbf{z})-\sum_i\int_{\mathbf{N}}[\phi(\mathbf{n})-\phi^{A}(\mathbf{n})]u_{w_i}(\mathbf{n})dF(\mathbf{n})p_i(\mathbf{z})\right\}\tau(\mathbf{z})dH(\mathbf{z})>0 \tag{51}$$

as long as the actual weights and inverse weights have the same normalization:

$$\begin{aligned}&\left(\int_{\mathbf{N}}\phi(\mathbf{n})u_c(\mathbf{n})dF(\mathbf{n})-\sum_i\int_{\mathbf{N}}\phi(\mathbf{n})u_{w_i}(\mathbf{n})dF(\mathbf{n})\left[\int_{\mathbf{Z}}p_i(\mathbf{z})dH(\mathbf{z})\right]\right)=\\&\left(\int_{\mathbf{N}}\phi^{A}(\mathbf{n})u_c(\mathbf{n})dF(\mathbf{n})-\sum_i\int_{\mathbf{N}}\phi^{A}(\mathbf{n})u_{w_i}(\mathbf{n})dF(\mathbf{n})\left[\int_{\mathbf{Z}}p_i(\mathbf{z})dH(\mathbf{z})\right]\right)=1\end{aligned} \tag{52}$$

Proof. See Appendix B.10. □

Proposition 3 GE shows that *any* tax reform that increases the tax burden where the expression in curly brackets, $\{\cdot\}$, in Equation 51 is positive and decreases the tax burden where $\{\cdot\}$ is negative (and closes the budget via the lump-sum transfer) is welfare improving. Finally, even with general equilibrium effects, we can still use the Cauchy-Schwarz inequality to establish that the optimal reform direction is proportional to the following at each $\mathbf{z}$ using identical logic as in Corollary 3.1 (see Appendix B.11 for details):

$$\int_{\mathbf{N}(\mathbf{z})}[\phi(\mathbf{n})-\phi^{A}(\mathbf{n})]u_c(\mathbf{n})dF(\mathbf{n}|\mathbf{z})-\sum_i\int_{\mathbf{N}}[\phi(\mathbf{n})-\phi^{A}(\mathbf{n})]u_{w_i}(\mathbf{n})dF(\mathbf{n})p_i(\mathbf{z})$$

The conclusion from this section is that inverse welfare functionals still encode information about Pareto efficiency and the set of welfare improving tax reforms even with general equilibrium effects.

5.4 Externalities Example

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

We will illustrate how externalities can impact inverse welfare functions using a more complex externality: inequality aversion. Suppose that in addition to having preferences

over consumption and labor supply, agents also have an intrinsic distaste for inequality (e.g., Alesina and Giuliano (2011) or Støstad and Cowell (2024)). We assume that individuals n choose an income z to maximize utility but that individuals are also negatively impacted by the overall level of inequality (i.e., other individuals' incomes generate an externality by contributing to inequality), which we operationalize using the Gini coefficient, $G = 100 \int_Z H(z)(1-H(z))dz \left(\int_Z zh(z)dz\right)^{-1}$ where $H(z)$ is the CDF of z and $h(z)$ is the PDF of z:

$$\begin{aligned} U(n;T,G) &= \max_z \; c - \frac{\left(\frac{z}{n}\right)^{1+k}}{1+k} - \alpha G \\ &\text{s.t. } \; c = z - T(z) \end{aligned} \tag{53}$$

α represents the willingness to pay to decrease the Gini coefficient by 1. For reference, the Gini coefficient is ≈ 40 in the U.S. and is ≈ 30 in Sweden (OECD, 2024), so $\alpha = 100$ implies that the individual willingness-to-pay to live in an economy with Swedish inequality relative to U.S. inequality is $\approx \$1,000$ per year. Changing variables and using the fact that by monotonicity of $n \mapsto z$, $H(z(n)) = F(n)$, so that the Jacobian equals $dz/dn = f(n)/h(z(n))$, we can express $G = 100 \int_N H(z(n))(1 - H(z(n)))f(n)/h(z(n))dn \left(\int_N z(n)f(n)dn\right)^{-1}$. Assuming that $h(\underline{z}) = h(\overline{z}) = 0$ for simplicity, we can use Equation 15 and integration by parts to calculate the Gateaux derivative of G:

$$\lim_{\epsilon \to 0} \frac{G(T+\epsilon\tau) - G(T)}{\epsilon} = \int_Z p(z)\tau(z)dz =$$
$$\frac{100 \int_Z \frac{\partial}{\partial z}\left[\frac{G(T)}{100}\xi(z)h(z) - \xi(z)(1-2H(z))h(z) + \xi(z)H(z)(1-H(z))h'(z)/h(z)\right]\tau(z)dz}{\int_Z zh(z)dz}$$

where $\xi(z)$ is the substitution effect defined in Equation 15. Next, we can use the analysis from Section 3.1 to calculate the Gateaux derivative of revenue as $\int_Z \{h(z) - \partial/\partial z \left[T'(z)\xi(z)h(z)\right]\}\tau(z)dz$. Applying Theorem 2, the inverse weights satisfy the following integral equation (which can be solved numerically via a standard fixed-point algorithm):[32]

$$\phi(n(\tilde{z}))h(\tilde{z}) = h(\tilde{z}) - \frac{\partial}{\partial z}\left[T'(z)\xi(z)h(z)\right]\Big|_{z=\tilde{z}} - \alpha p(\tilde{z}) \int_Z \phi(n(z))h(z)dz \tag{54}$$

[32]Equation 54 uses the fact that $\Phi(\tilde{n}) = \Phi_2(z(\tilde{n}))$ in Equation 49 (by monotonicity of $n \mapsto z$) and then differentiates Equation 50 (because everything is sufficiently smooth) to express the integral equations in terms of inverse welfare weights rather than the inverse welfare measure (using the convention in footnote 5 so that $d\Phi(n(\tilde{z}))/d\tilde{z} = \phi(n(\tilde{z}))f(n(\tilde{z}))dn/d\tilde{z} = \phi(n(\tilde{z}))h(\tilde{z})$).

6 Numerical Simulations

Thus far, this paper has argued that (1) we can compute inverse welfare functionals even when allowing for complicated behavioral responses to taxes, frictions, multidimensional heterogeneity and choice sets, as well as general equilibrium effects and (2) inverse welfare functionals allow us to assess Pareto efficiency, assess the desirability of tax reforms, and construct the optimal tax reform direction. However, a natural question then is whether allowing for these additional features meaningfully changes the policy conclusions relative to simpler models that are more standard in the literature?

To explore this question we conducted a variety of numerical simulations; we relegate detailed results and discussion of the calibrations to Appendix C while presenting the high-level findings here. We begin with a baseline case, calculating the inverse welfare functional for a smoothed approximation of the U.S. income tax schedule. This exercise is similar to those performed in Lockwood and Weinzierl (2016), Hendren (2020), and Heathcote and Tsujiyama (2021) in the U.S. and should be viewed as a point of comparison for our other simulations that introduce realistic complexities into the model. Consistent with prior work, we find that the U.S. tax schedule is (locally) Pareto efficient and that the implied value of giving \$1 to an individual (i.e., the inverse welfare weights multiplied by marginal utility of consumption) is decreasing with income.

Next, we consider five alternative models that incorporate various realisms, calculating inverse weights (1) for a piece-wise linear tax schedule accounting for bunching and individuals with multiple optima (which arise under standard preferences), (2) for a piece-wise linear tax schedule when agents face sparsity-based frictions as in Section 3.3, (3) for a tax system with two instruments (income taxes and property taxes), (4) for the baseline model but allowing for general equilibrium wage effects as in Section 5.1, and (5) for the baseline model but allowing for inequality aversion as in Section 5.4. The high-level conclusion from these simulations is that inverse welfare functionals can change substantially when one accounts for these realisms; for example, the tax schedule is no longer Pareto efficient if individuals bunch and have multiple optima as a result of kinks, the tax schedule is no longer Pareto efficient when accounting for a second linear tax instrument (property taxes), and the inverse welfare weights are much flatter as a function of income when accounting for sparsity-based frictions or general equilibrium wage effects (rendering progressive tax reforms more desirable). See Appendix C for a com-

plete discussion. Hence, policy conclusions about Pareto efficiency and the set of welfare improving tax reforms can change considerably once we account for these realisms.

7 Non-Existence of Inverse Welfare Functionals

We have shown that our framework allows us to construct inverse welfare functionals for a large class of tax schedules because Gateaux differentiability of government revenue is a relatively mild restriction. We now illustrate that we cannot relax the requirement of Gateaux differentiability of revenue in Theorem 1 to only require that Gateaux *variations* exist in all directions (recall Definition 1). In Appendix B.12 we prove the following result:

Proposition 4. *When* $dim(\mathbf{Z}) > dim(\mathbf{N})$*, there are smooth* $T(\mathbf{z})$ *such that* $R(T) = E$*,* $\mathbf{Z}$ *is compact, all* $\mathbf{n}$ *have a unique optimum, and Gateaux variations of revenue exist for all directions* $\tau(\mathbf{z})$*, yet an inverse welfare functional does not exist.*

We prove Proposition 4 by showing non-existence of an inverse welfare functional in a simple model of income and savings taxation as in Atkinson and Stiglitz (1976). The intuition is as follows. Under some smoothness assumptions, there is a unique set of inverse welfare weights that ensures every perturbation to the *income* tax schedule leaves the government's Lagrangian unchanged. But there is also a unique set of inverse welfare weights that ensures every perturbation to the *savings* tax schedule leaves the government's Lagrangian unchanged. However, these two sets of potential inverse welfare weights do not necessarily need to coincide; and if they do not coincide then the Gateaux derivative of revenue does not exist.[33] This situation yields an overdetermined (infinite) system of equations that the inverse welfare weights must satisfy, which then leads to non-existence of inverse welfare weights.

Proposition 4 therefore has broader implications for the Atkinson-Stiglitz Theorem. The Atkinson-Stiglitz Theorem is one of the most famous results in public economics, showing that when agents differ in terms of a unidimensional parameter $n \in N$, 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 la-

[33]Recall that both multidimensional integration by parts, Lemma 1, and Proposition 1 require that the set $\mathbf{Z}$ is open (or the closure of an open set) in the ambient space (i.e., $\mathbf{Z}$ has non-empty interior). This condition fails when $\dim(\mathbf{Z}) > \dim(\mathbf{N})$ so that government revenue may not be Gateaux differentiable even if the tax schedule is smooth and everyone has a unique optimum. 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 $\mathbb{R}^2$.

bor and all other goods. Proofs of the Atkinson-Stiglitz Theorem typically invoke the Pareto principle by showing that any multidimensional tax schedule is Pareto dominated by some non-linear income tax schedule (Kaplow, 2006). The non-existence result of Proposition 4 can therefore be viewed as a strengthening of the classic Atkinson-Stiglitz Theorem: in settings with unidimensional type heterogeneity and multidimensional choice spaces, many tax schedules are not just Pareto inefficient but are actually not supported by *any* inverse welfare functionals, even those that allow for negative weights.[34] The economic takeaway is as follows: in the Atkinson-Stiglitz environment, it is often impossible to rationalize indirect taxes (e.g., savings or commodity taxes) even if the government wants to make some individuals as miserable as possible (via negative welfare weights).

8 Conclusion

This paper has developed a general theory to recover the inverse welfare functional that rationalizes a given tax schedule as optimal. The key component required to construct such an inverse welfare functional is the Gateaux derivative of government revenue with respect to the tax schedule. Our theory allows for complex environments including the presence of multidimensional tax schedules, bunching/jumping behavior, optimization frictions, general equilibrium effects, and externalities. From a policy perspective, inverse welfare functionals are a simple tool that can be used to assess Pareto efficiency and create Pareto improving reforms, construct welfare-improving budget-neutral tax reforms, and determine the optimal reform direction. Finally, we have illustrated numerically how allowing for these arguably realistic complexities can have large and meaningful impacts on the inverse welfare functional. Hence, allowing for these realisms can significantly alter policy conclusions.

Moving forward, we believe there is still substantial scope for innovation in so-called "inverse optima" methods. First, all of the analysis in this paper has assumed that agents correctly perceive the tax schedule and their own utility function. While a general analysis of inverse welfare functionals in the presence of misperceptions, behavioral biases, and internalities seems challenging, we believe this is a useful area for future research. Second, while this paper focuses on inverse welfare functions for tax schedules, much

[34]Note, in Appendix B.12, we show that this non-existence does not rely on separability in any way: hence most tax schedules in this setting will not have associated inverse welfare functionals regardless of whether utility is weakly separable or not.

of the analysis can likely be extended to non-tax policy spaces, such as in-kind good provision, minimum wages, or social insurance.

References

Aistleitner, Christoph, and Josef Dick. 2015. "Functions of bounded variation, signed measures, and a general Koksma–Hlawka inequality." *Acta Arithmetica*, 167(2): 143–171.

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

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

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

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

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

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

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

Boerma, Job, Aleh Tsyvinski, and Alexander Zimin. 2022. "Bunching and taxing multidimensional skills." *SSRN Electronic Journal*.

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

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

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

Diewert, W.E. 1978. "Optimal tax perturbations." *Journal of Public Economics*, 10(2): 139–177.

Dixit, Avinash. 1979. "Price changes and optimum taxation in a many-consumer economy." *Journal of Public Economics*, 11(2): 143–157.

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

Folland, Gerald B. 1999. *Real Analysis: Modern Techniques and Their Applications*. . 2 ed., New York:John Wiley & Sons.

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

Heathcote, Jonathan, and Hitoshi Tsujiyama. 2021. "Optimal Income Taxation: Mirrlees Meets Ramsey." *Journal of Political Economy*, 129(11): 3141–3184.

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.

Jacquet, Laurence, and Etienne Lehmann. 2025. "Generalized Production Efficiency." THEMA (Théorie Économique, Modélisation et Applications), CY Cergy-Paris Université, ESSEC and CNRS 2025-08. RePEc working paper.

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.

Krasikov, Ilia, and Mikhail Golosov. 2024. "Multidimensional Screening in Public Finance: The Optimal Taxation of Couples."

Lockwood, Benjamin B., and Matthew C. Weinzierl. 2016. "Positive and Normative Judgments Implicit in U.S. Tax Policy, and the Costs of Unequal Growth and Recessions." *Journal of Monetary Economics*, 77: 30–47.

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

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

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

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

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

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

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

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

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

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

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

Sher, Itai. 2024. "Generalized Social Marginal Welfare Weights Imply Inconsistent Comparisons of Tax Policies." *American Economic Review*, 114(11): 3551–3577.

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

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

Støstad, Morten Nyborg, and Frank Cowell. 2024. "Inequality as an externality: Consequences for tax design." *Journal of Public Economics*, 235: 105139.

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

A Appendix: Proofs

A.1 Proof of Theorem 1

Proof. $T(\mathbf{z})$ is assumed continuous on a compact set $\mathbf{Z}$, so we will consider perturbations in the direction of arbitrary continuous functions $\tau(\mathbf{z})$. Thus, if $R(T)$ is Gateaux differentiable then by the Riesz-Markov-Kakutani representation theorem, $\exists$ a (signed) Borel measure Γ such that the Gateaux derivative can be written:

$$\lim_{\epsilon \to 0} \frac{R(T + \epsilon\tau) - R(T)}{\epsilon} = \int_{\mathbf{Z}} \tau(\mathbf{z}) d\Gamma(\mathbf{z}) \tag{55}$$

We will construct an inverse welfare functional of the following form:

$$W(U(\mathbf{n}; T)) = \int_{\mathbf{N}} U(\mathbf{n}; T) d\Phi(\mathbf{n}) = \int_{\mathbf{Z}} \int_{\mathbf{N}(\mathbf{z})} U(\mathbf{n}; T) d\Phi_1(\mathbf{n}|\mathbf{z}) d\Phi_2(\mathbf{z}) \tag{56}$$

The second equality above uses the disintegration theorem to split the measure Φ into conditional and marginal measures Φ_1 and Φ_2. To take the Gateaux derivative of $W(U(\mathbf{n}; T))$, we will appeal to the envelope theorem to take the derivative of $U(\mathbf{n}; T) \equiv u(y(\mathbf{z}(\mathbf{n})) - T(\mathbf{z}(\mathbf{n})), \mathbf{z}(\mathbf{n}); \mathbf{n})$. The envelope theorem implies that for all $\mathbf{n}$ with a unique optimum:

$$\lim_{\epsilon \to 0} \frac{U(\mathbf{n}; T + \epsilon\tau) - U(\mathbf{n}; T)}{\epsilon} = -u_c(y(\mathbf{z}(\mathbf{n})) - T(\mathbf{z}(\mathbf{n})), \mathbf{z}(\mathbf{n}); \mathbf{n})\tau(\mathbf{z}(\mathbf{n})) \equiv -u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n})) \tag{57}$$

While application of the envelope theorem is standard in public economics, we nonetheless rigorously justify its use. Consider optimal utility for a given $\mathbf{n}$ as a function of ϵ where we explicitly note that $\mathbf{z}(\mathbf{n})$ is also a function of ϵ: $u(y(\mathbf{z}(\mathbf{n}, \epsilon)) - T(\mathbf{z}(\mathbf{n}, \epsilon)) - \epsilon\tau(\mathbf{z}(\mathbf{n}, \epsilon)), \mathbf{z}(\mathbf{n}, \epsilon); \mathbf{n})$. Note that by standard arguments, any $\mathbf{n}$ with a unique optimum will move continuously in response to a given tax perturbation for sufficiently small ϵ. Theorem 3 of Milgrom and Segal (2002) then implies that Equation 57 holds for any such $\mathbf{n}$ if we can show that $-\tau(\mathbf{z})u_c(y(\mathbf{z}) - T(\mathbf{z}), \mathbf{z}; \mathbf{n})$ is bounded as a function of $\mathbf{z}$ over $\mathbf{Z}$ (the chosen set of $\mathbf{z}$'s) and that

$$\frac{u(y(\mathbf{z}) - T(\mathbf{z}) - \epsilon\tau(\mathbf{z}), \mathbf{z}; \mathbf{n}) - u(y(\mathbf{z}) - T(\mathbf{z}), \mathbf{z}; \mathbf{n})}{\epsilon} \tag{58}$$

converges uniformly in ϵ for all $\mathbf{z}$. Given that $T(\mathbf{z})$ and $\tau(\mathbf{z})$ are both continuous functions on compact sets, we will make the technical assumptions that $\forall \mathbf{n}$ we have $\sup_{\mathbf{z}} |u_c(y(\mathbf{z}) - T(\mathbf{z}), \mathbf{z}; \mathbf{n})| < \infty$ and that $\sup_{\mathbf{z}} |u_{cc}(y(\mathbf{z}) - T(\mathbf{z}), \mathbf{z}; \mathbf{n})| < \infty$ (this second condition ensures that Equation 58 converges uniformly in ϵ for all $\mathbf{z}$ in the compact set $\mathbf{Z}$ via a Taylor series argument).

In order to apply the envelope theorem, we will choose Φ_1 so that Φ_1-a.e. $\mathbf{n}$ have a unique optimum. By assumption, $F(\mathbf{n}|\mathbf{z})$-a.e. $\mathbf{n}$ have a unique optimum for every $\mathbf{z}$;

hence, we can set $\Phi_1 = F/\int_{\mathbf{N}(\mathbf{z})} u_c(\mathbf{n})dF(\mathbf{n}|\mathbf{z}) = F/\overline{u_c}(\mathbf{z})$ so that:[35]

$$\left.\frac{\partial W(U(\mathbf{n};T+\epsilon\tau))}{\partial\epsilon}\right|_{\epsilon=0} = \int_{\mathbf{Z}} \frac{\int_{\mathbf{N}(\mathbf{z})} -u_c(\mathbf{n})\tau(\mathbf{z})dF(\mathbf{n}|\mathbf{z})}{\int_{\mathbf{N}(\mathbf{z})} u_c(\mathbf{n})dF(\mathbf{n}|\mathbf{z})} d\Phi_2(\mathbf{z}) = -\int_{\mathbf{Z}} \tau(\mathbf{z})d\Phi_2(\mathbf{z})$$

To ensure the Gateaux derivative of the government's Lagrangian is zero $\forall\tau$ we must have:

$$\left.\frac{\partial L(T+\epsilon\tau;W)}{\partial\epsilon}\right|_{\epsilon=0} = -\int_{\mathbf{Z}} \tau(\mathbf{z})d\Phi_2(\mathbf{z}) + \lambda\int_{\mathbf{Z}} \tau(\mathbf{z})d\Gamma(\mathbf{z}) = 0 \tag{59}$$

From here we can find an inverse welfare functional by normalizing λ to 1 and choosing $\Phi_2 = \Gamma$. Thus, we have constructed the following inverse welfare functional:

$$W(U(\mathbf{n};T)) = \int_{\mathbf{Z}} \frac{1}{\overline{u_c}(\mathbf{z})} \int_{\mathbf{N}(\mathbf{z})} U(\mathbf{n};T)dF(\mathbf{n}|\mathbf{z})d\Gamma(\mathbf{z})$$

Next, we want to show that the Φ defined in Equation 10 generates this same inverse welfare functional. Note, Φ in Equation 10 is defined on sets of the form $\{\mathbf{n} \leq \tilde{\mathbf{n}}\}$ (so that Φ is easier to calculate in practice), but it is a standard result that the "pre-measure" defined on all sets of the form $\{\mathbf{n} \leq \tilde{\mathbf{n}}\}$ uniquely extends to a measure on all Borel sets $\tilde{\mathbf{N}} \subseteq \mathbf{N}$ for compact $\mathbf{N}$ satisfying:[36]

$$\Phi(\tilde{\mathbf{N}}) = \int_{\mathbf{Z}} \int_{\mathbf{n}\in\tilde{\mathbf{N}}\cap\mathbf{N}(\mathbf{z})} \frac{1}{\overline{u_c}(\mathbf{z})} dF(\mathbf{n}|\mathbf{z})d\Gamma(\mathbf{z}) \tag{60}$$

By Theorem 1.29 of Rudin (1974) if Φ satisfies Equation 60, then for every measurable $U(\mathbf{n};T)$:

$$\int_{\mathbf{N}} U(\mathbf{n};T)d\Phi(\mathbf{n}) = \int_{\mathbf{Z}} \int_{\mathbf{N}(\mathbf{z})} U(\mathbf{n};T)\frac{1}{\overline{u_c}(\mathbf{z})} dF(\mathbf{n}|\mathbf{z})d\Gamma(\mathbf{z}) \tag{61}$$

Thus, the measure Φ defined in Equation 10 satisfies Equation 61 which implies that Φ defines an inverse welfare functional for the given $T(\mathbf{z})$.

□

A.2 Proof of Proposition 2

Proof. We prove each statement below using the inverse measure defined in Equation 10. All results can be derived with identical logic using the inverse weights in Equation 9 under the stronger conditions on the form of the Gateaux derivative of revenue in

[35] If instead, we only have that for each $\mathbf{z}$ $\exists$ at least one $\hat{\mathbf{n}}(\mathbf{z})$ that has a unique optimum, then we can choose $\Phi_1 = \delta_{\hat{\mathbf{n}}(\mathbf{z})}/\int_{\mathbf{N}(\mathbf{z})} u_c(\mathbf{n})d\delta_{\hat{\mathbf{n}}(\mathbf{z})}(\mathbf{n})$ 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:

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

This welfare functional also yields that $\partial W(U(\mathbf{n};T+\epsilon\tau))/\partial\epsilon = -\int_{\mathbf{Z}} \tau(\mathbf{z})d\Phi_2(\mathbf{z})$ because at each $\mathbf{z}$ we only assign positive measure to the type $\hat{\mathbf{n}}(\mathbf{z})$ with a unique optimum.

[36] In other words, there is a one-to-one correspondence between a distribution function and the associated measure. See Theorem 1.16 of Folland (1999) for a simple proof of this result in the unidimensional case or Theorem 3 of Aistleitner and Dick (2015) for a proof in the multidimensional case.

Theorem 1 required for their existence.

1. Recall that the inverse welfare functional in Theorem 1 takes the following form:

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

If this inverse welfare functional is not positive then $\exists g(\mathbf{n}) \in C(\mathbf{N})$ with $g(\mathbf{n}) \geq 0$ and[37]

$$\int_{\mathbf{Z}} \frac{1}{\overline{u_c}(\mathbf{z})} \int_{\mathbf{N}(\mathbf{z})} g(\mathbf{n})dF(\mathbf{n}|\mathbf{z})d\Gamma(\mathbf{z}) < 0 \tag{62}$$

Next, consider $\tau(\mathbf{z}) = -\frac{1}{\overline{u_c}(\mathbf{z})} \int_{\mathbf{N}(\mathbf{z})} g(\mathbf{n})dF(\mathbf{n}|\mathbf{z})$, yielding a $\tau(\mathbf{z}) \leq 0$ (with strict inequality at some $\mathbf{z}$) such that:[38]

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

Thus, the given $\tau(\mathbf{z})$ perturbation lowers taxes everywhere (which improves welfare by the envelope theorem given $u_c > 0$) and is also feasible insofar as it raises government revenue.[39] Hence, if the inverse welfare functional constructed in Theorem 1 is not positive for the given $T(\mathbf{z})$, then the $\tau(\mathbf{z})$ defined above is a Pareto improving tax reform.

2. We seek to show that there is no budget feasible (marginal) reform (i.e., $\int_{\mathbf{Z}} \tau(\mathbf{z})d\Gamma(\mathbf{z}) \geq 0$) such that everyone is weakly better off and some positive measure of people are strictly better off. Given that the utility impacts of a small tax reform on type $\mathbf{n}$ are given by $-u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n}))$ by the envelope theorem (for almost all individuals), we seek to show there is no budget feasible reform satisfying:

$$\begin{aligned} &- u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n})) \geq 0 \ \forall \mathbf{n} \\ &- u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n})) > 0 \text{ for } \mathbf{n} \in \tilde{\mathbf{N}} \text{ with } F(\tilde{\mathbf{N}}) > 0 \end{aligned} \tag{63}$$

If the inverse welfare functional from Theorem 1 is strictly positive then $-u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n})) \geq 0 \ \forall \mathbf{n}$ (with strict inequality on a set $\tilde{\mathbf{N}}$ with $F(\tilde{\mathbf{N}}) > 0$) implies that:

$$-\int_{\mathbf{N}} u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n}))d\Phi(\mathbf{n}) = -\int_{\mathbf{Z}} \frac{1}{\overline{u_c}(\mathbf{z})} \int_{\mathbf{N}(\mathbf{z})} u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n}))dF(\mathbf{n}|\mathbf{z})d\Gamma(\mathbf{z}) > 0$$

[37] Practically, one can construct such a $g(\mathbf{n})$ by finding a connected set $\tilde{\mathbf{Z}}$ such that $\Gamma(\mathbf{E}) < 0 \ \forall \mathbf{E} \in \tilde{\mathbf{Z}}$ (e.g., if the inverse welfare functional can be represented with inverse weights as in Equation 9 then this corresponds to a connected set $\tilde{\mathbf{Z}}$ where welfare weights are all negative) and then setting $g(\mathbf{n}) = v(\mathbf{z}(\mathbf{n}))$ for any function $v(\mathbf{z}(\mathbf{n})) \geq 0$ that vanishes outside $\tilde{\mathbf{Z}}$.

[38] $\tau(\mathbf{z}) \leq 0$ because $g(\mathbf{n}) \geq 0$ and $u_c > 0$. Strict equality holds because if $g(\mathbf{n}) = 0$ then $\int_{\mathbf{Z}} (1/\overline{u_c}(\mathbf{z})) \int_{\mathbf{N}(\mathbf{z})} g(\mathbf{n})dF(\mathbf{n}|\mathbf{z})d\Gamma(\mathbf{z}) = 0$, violating Equation 62.

[39] Note, because $T(\mathbf{z}), \tau(\mathbf{z})$ are assumed to be continuous functions, we have implicitly assumed that $\tau(\mathbf{z}) = -(1/\overline{u_c}(\mathbf{z})) \int_{\mathbf{N}(\mathbf{z})} g(\mathbf{n})dF(\mathbf{n}|\mathbf{z})$ is a continuous function of $\mathbf{z}$; this is WLOG as the given $\tau(\mathbf{z})$ is in L^1 (as it is a negative integrable function) and continuous functions are dense in L^1 (under the measure Γ, which is regular by the Riesz-Markov-Kakutani representation theorem) so that we can always find a continuous $\hat{\tau}(\mathbf{z})$ close to $\tau(\mathbf{z})$ with $\int_{\mathbf{Z}} \hat{\tau}(\mathbf{z})d\Gamma(\mathbf{z}) > 0$.

Canceling terms and multiplying by -1:

$$\int_{\mathbf{Z}} \frac{1}{\overline{u_c}(\mathbf{z})} \int_{\mathbf{N}(\mathbf{z})} u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n}))dF(\mathbf{n}|\mathbf{z})d\Gamma(\mathbf{z}) = \int_{\mathbf{Z}} \tau(\mathbf{z})d\Gamma(\mathbf{z}) < 0$$

which means that this reform is not budget feasible; hence, there do not exist any marginal Pareto improving tax reforms.

□

A.3 Proof of Proposition 3

By the envelope theorem, the total welfare impact of a small tax perturbation $\tau(\mathbf{z}) + \tau_0$ where τ_0 is a lump sum transfer that makes the Gateaux derivative of revenue equal to zero is:

$$-\int_{\mathbf{N}} \phi^A(\mathbf{n})u_c(\mathbf{n})[\tau(\mathbf{z}(\mathbf{n})) + \tau_0]dF(\mathbf{n})$$

Hence, this tax perturbation is welfare improving if and only if:

$$-\int_{\mathbf{N}} \phi^A(\mathbf{n})u_c(\mathbf{n})[\tau(\mathbf{z}(\mathbf{n})) + \tau_0]dF(\mathbf{n}) > 0 \tag{64}$$

However, we know that the inverse welfare weights $\phi(\mathbf{n})$ satisfy the following given $[\tau(\mathbf{z}(\mathbf{n})) + \tau_0]$ is budget neutral:

$$-\int_{\mathbf{N}} \phi(\mathbf{n})u_c(\mathbf{n})[\tau(\mathbf{z}(\mathbf{n})) + \tau_0]dF(\mathbf{n}) = 0$$

Thus, by the normalization in Equation 37:

$$-\int_{\mathbf{N}} \phi(\mathbf{n})u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n}))dF(\mathbf{n}) = \tau_0 \int_{\mathbf{N}} \phi(\mathbf{n})u_c(\mathbf{n})dF(\mathbf{n}) = \tau_0 \int_{\mathbf{N}} \phi^A(\mathbf{n})u_c(\mathbf{n})dF(\mathbf{n}) = \tau_0 \tag{65}$$

Plugging $-\int_{\mathbf{N}} \phi(\mathbf{n})u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n}))dF(\mathbf{n}) = \tau_0 \int_{\mathbf{N}} \phi^A(\mathbf{n})u_c(\mathbf{n})dF(\mathbf{n})$ from Equation 65 into Equation 64 we have that a tax perturbation in the direction $\tau(\mathbf{z})+\tau_0$ is welfare improving if and only if:

$$\int_{\mathbf{N}} [\phi(\mathbf{n})-\phi^A(\mathbf{n})]u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n}))dF(\mathbf{n}) = \int_{\mathbf{Z}} \int_{\mathbf{N}(\mathbf{z})} [\phi(\mathbf{n})-\phi^A(\mathbf{n})]u_c(\mathbf{n})dF(\mathbf{n}|\mathbf{z})\tau(\mathbf{z})dH(\mathbf{z}) > 0$$

Next, suppose that the inverse and actual welfare functionals are more generally given by $\int_{\mathbf{N}} U(\mathbf{n};T)d\Phi(\mathbf{n})$ and $\int_{\mathbf{N}} U(\mathbf{n};T)d\Phi^A(\mathbf{n})$. We again consider a small tax perturbation $\tau(\mathbf{z}) + \tau_0$ where τ_0 is a lump sum transfer that makes the Gateaux derivative of revenue equal to zero. Assuming that $\Phi^A(\mathbf{n})$-a.e. $\mathbf{n}$ have a unique optimum, then by the envelope theorem a tax perturbation is welfare improving if and only if:

$$-\int_{\mathbf{N}} u_c(\mathbf{n})[\tau(\mathbf{z}(\mathbf{n})) + \tau_0]d\Phi^A(\mathbf{n}) > 0 \tag{66}$$

However, assuming that $\Phi(\mathbf{n})$-a.e. $\mathbf{n}$ have a unique optimum, then by the envelope

theorem we know that the inverse welfare functional satisfies:

$$-\int_{\mathbf{N}} u_c(\mathbf{n})[\tau(\mathbf{z}(\mathbf{n})) + \tau_0]d\Phi(\mathbf{n}) = 0$$

Thus, assuming Φ and Φ^A have the same normalization so that $\int_{\mathbf{N}} u_c(\mathbf{n})d\Phi(\mathbf{n}) = \int_{\mathbf{N}} u_c(\mathbf{n})d\Phi^A(\mathbf{n}) = 1$:

$$-\int_{\mathbf{N}} u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n}))d\Phi(\mathbf{n}) = \tau_0 \int_{\mathbf{N}} u_c(\mathbf{n})d\Phi(\mathbf{n}) = \tau_0 \int_{\mathbf{N}} u_c(\mathbf{n})d\Phi^A(\mathbf{n}) = \tau_0 \tag{67}$$

Plugging Equation 67 into Equation 66 we have that a tax perturbation $\tau(\mathbf{z}) + \tau_0$ is welfare improving if and only if:

$$\int_{\mathbf{N}} u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n}))d[\Phi(\mathbf{n}) - \Phi^A(\mathbf{n})] > 0$$

A.4 Proof of Corollary 3.1

Proof. By the definition of the local inverse welfare functional we have that:

$$\int_{\mathbf{Z}} \tau(\mathbf{z})\gamma(\mathbf{z})dH(\mathbf{z}) = \int_{\mathbf{N}} \phi(\mathbf{n})u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n}))dF(\mathbf{n}) = \int_{\mathbf{Z}} \left[\int_{\mathbf{N}(\mathbf{z})} \phi(\mathbf{n})u_c(\mathbf{n})dF(\mathbf{n}|\mathbf{z})\right] \tau(\mathbf{z})dH(\mathbf{z})$$

Forming a Lagrangian for Equation 38 with Lagrange multiplier ν and using the previous expression for $\int_{\mathbf{Z}} \tau(\mathbf{z})\gamma(\mathbf{z})dH(\mathbf{z})$ we seek to solve:

$$\begin{aligned} &\max_{\tau(\mathbf{z}),\nu} \int_{\mathbf{Z}} \int_{\mathbf{N}(\mathbf{z})} [\nu\phi(\mathbf{n}) - \phi^A(\mathbf{n})]u_c(\mathbf{n})dF(\mathbf{n}|\mathbf{z})\tau(\mathbf{z})dH(\mathbf{z}) \\ &\text{s.t.} \int_{\mathbf{Z}} |\tau(\mathbf{z})|^2 dH(\mathbf{z}) = 1 \end{aligned} \tag{68}$$

For a fixed value of ν, the solution (by the Cauchy-Schwarz inequality) is to set:

$$\tau(\mathbf{z}) = \frac{\int_{\mathbf{N}(\mathbf{z})}[\nu\phi(\mathbf{n}) - \phi^A(\mathbf{n})]u_c(\mathbf{n})dF(\mathbf{n}|\mathbf{z})}{||\int_{\mathbf{N}(\mathbf{z})}[\nu\phi(\mathbf{n}) - \phi^A(\mathbf{n})]u_c(\mathbf{n})dF(\mathbf{n}|\mathbf{z})||_{L^2}} \tag{69}$$

where

$$\left|\left|\int_{\mathbf{N}(\mathbf{z})}[\nu\phi(\mathbf{n}) - \phi^A(\mathbf{n})]u_c(\mathbf{n})dF(\mathbf{n}|\mathbf{z})\right|\right|_{L^2} \equiv \sqrt{\int_{\mathbf{Z}} \left|\int_{\mathbf{N}(\mathbf{z})}[\nu\phi(\mathbf{n}) - \phi^A(\mathbf{n})]u_c(\mathbf{n})dF(\mathbf{n}|\mathbf{z})\right|^2 dH(\mathbf{z})}$$

However, with the normalization given by Equation 40, the budget neutrality constraint is satisfied when $\nu = 1$ as then:

$$\int_{\mathbf{Z}} \left[\int_{\mathbf{N}(\mathbf{z})} \phi(\mathbf{n})u_c(\mathbf{n})dF(\mathbf{n}|\mathbf{z})\right] \frac{\int_{\mathbf{N}(\mathbf{z})}[\phi(\mathbf{n}) - \phi^A(\mathbf{n})]u_c(\mathbf{n})dF(\mathbf{n}|\mathbf{z})}{||\int_{\mathbf{N}(\mathbf{z})}[\phi(\mathbf{n}) - \phi^A(\mathbf{n})]u_c(\mathbf{n})dF(\mathbf{n}|\mathbf{z})||} dH(\mathbf{z}) = 0$$

Hence, setting $\tau(\mathbf{z})$ according to Equation 69 and $\nu = 1$ is a solution to Equation 68 so that setting $\tau(\mathbf{z})$ according to Equation 69 is a solution to Equation 38 by the Lagrange multiplier theorem.

□

A.5 Proof to Theorem 2

Proof. $T(\mathbf{z})$ is assumed continuous on a compact set $\mathbf{Z}$, so we will consider perturbations in the direction of arbitrary continuous functions $\tau(\mathbf{z})$. Thus, if $R(T)$ is Gateaux differentiable then by the Riesz-Markov-Kakutani representation theorem, $\exists$ a (signed) Borel measure Γ (that is unique, regular, and countably additive) such that the Gateaux derivative 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 (signed) 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:

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

We will construct an inverse welfare functional that takes the following form for some signed Borel measure Φ:

$$W(U(\mathbf{n}; T, \mathbf{w})) = \int_{\mathbf{N}} U(\mathbf{n}; T) d\Phi(\mathbf{n}) = \int_{\mathbf{Z}} \int_{\mathbf{N}(\mathbf{z})} U(\mathbf{n}; T) d\Phi_1(\mathbf{n}|\mathbf{z}) d\Phi_2(\mathbf{z}) \tag{70}$$

The second equality above uses the disintegration theorem to split the measure Φ into conditional and marginal measures Φ_1 and Φ_2. To take the Gateaux derivative of $W(U(\mathbf{n}; T, \mathbf{w}))$ we will use 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 optimum:[40]

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

By assumption, $F(\mathbf{n}|\mathbf{z})$-a.e. $\mathbf{n}$ have a unique optimum at every $\mathbf{z}$; hence, we can set $\Phi_1 = F / \int_{\mathbf{N}(\mathbf{z})} u_c(\mathbf{n}) dF(\mathbf{n}|\mathbf{z}) = F / \overline{u_c}(\mathbf{z})$. Applying the envelope theorem to compute the

[40] We can apply the envelope theorem for all individuals with a unique optimum as long as u_c, u_{cc}, u_{cw_i}, u_{w_i}, and $u_{w_i w_i}$ are all bounded by the same Taylor series argument discussed in the proof to Theorem 1 in Appendix A.1.

Gateaux derivative of $W(U(\mathbf{n};T,\mathbf{w}))$, we get the following expression for the Gateaux derivative of the Lagrangian:[41]

$$\int_{\mathbf{Z}}\int_{\mathbf{N}(\mathbf{z})} \frac{-u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n})) + \sum_i u_{w_i}(\mathbf{n})\frac{\partial w_i}{\partial \epsilon}}{\int_{\mathbf{N}(\mathbf{z})} u_c(\mathbf{n})dF(\mathbf{n}|\mathbf{z})} dF(\mathbf{n}|\mathbf{z})d\Phi_2(\mathbf{z}) + \lambda \int_{\mathbf{Z}} \tau(\mathbf{z})d\Gamma(\mathbf{z}) \tag{71}$$

In this case, define $\overline{u_{w_i}/u_c}(\mathbf{z}) \equiv \int_{\mathbf{N}(\mathbf{z})} u_{w_i}(\mathbf{n})dF(\mathbf{n}|\mathbf{z}) / \int_{\mathbf{N}(\mathbf{z})} u_c(\mathbf{n})dF(\mathbf{n}|\mathbf{z})$ so that Equation 71 can be rewritten (where we also plug in $\partial w_i/\partial\epsilon = \int_{\mathbf{Z}} \tau(\mathbf{z})dP_i(\mathbf{z})$):

$$-\int_{\mathbf{Z}} \tau(\mathbf{z})d\Phi_2(\mathbf{z}) + \sum_i \int_{\mathbf{Z}} \overline{\frac{u_{w_i}}{u_c}}(\mathbf{z})d\Phi_2(\mathbf{z}) \int_{\mathbf{Z}} \tau(\mathbf{z})dP_i(\mathbf{z}) + \lambda \int_{\mathbf{Z}} \tau(\mathbf{z})d\Gamma(\mathbf{z}) \tag{72}$$

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

$$\int_{\mathbf{Z}} \tau(\mathbf{z}) \left(-d\Phi_2(\mathbf{z}) + \sum_i dP_i(\mathbf{z}) \int_{\mathbf{Z}} \overline{\frac{u_{w_i}}{u_c}}(\mathbf{s})d\Phi_2(\mathbf{s}) + \lambda d\Gamma(\mathbf{z}) \right) \tag{73}$$

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

$$\int_{\mathbf{E}} \left(-d\Phi_2(\mathbf{z}) + \sum_i dP_i(\mathbf{z}) \int_{\mathbf{Z}} \overline{\frac{u_{w_i}}{u_c}}(\mathbf{s})d\Phi_2(\mathbf{s}) + \lambda d\Gamma(\mathbf{z}) \right) = 0 \tag{74}$$

Or, expressing Equation 74 in terms of measures with $\Phi_2(\mathbf{E}) \equiv \int_{\mathbf{E}} d\Phi_2(\mathbf{z})$, $\Gamma(\mathbf{E}) \equiv \int_{\mathbf{E}} d\Gamma(\mathbf{z})$, and $P_i(\mathbf{E}) \equiv \int_{\mathbf{E}} dP_i(\mathbf{z})$, we have (normalizing $\lambda = 1$):

$$\Phi_2(\mathbf{E}) = \Gamma(\mathbf{E}) + \sum_i P_i(\mathbf{E}) \int_{\mathbf{Z}} \overline{\frac{u_{w_i}}{u_c}}(\mathbf{s})d\Phi_2(\mathbf{s}) \tag{75}$$

which is an integral equation formulated in a measure space as in Das (1974) or Sharma (1975). Next, we consider the map Q which takes a measure Φ_2 as an argument and then outputs a new measure defined by $(Q\Phi_2)(\mathbf{E}) = \Gamma(\mathbf{E}) + \sum_i P_i(\mathbf{E}) \int_{\mathbf{Z}} \overline{u_{w_i}/u_c}(\mathbf{s})d\Phi_2(\mathbf{s})$. We will now show that $Q\Phi_2$ is a contraction mapping on the space of regular, countably additive Borel measures. The space of regular, countably additive Borel measures on $\mathbf{Z}$ is a Banach space when equipped with the "total variation" norm (hence, we can apply the contraction mapping theorem):

$$||\mu||_{\text{TV}} = \sup_{||f||_\infty \leq 1} \int f d\mu \tag{76}$$

[41] As in Theorem 1, if we relax our assumptions to only assume that (at least) a single type $\hat{\mathbf{n}}(\mathbf{z})$ has a unique optimum at each $\mathbf{z}$ then we can set $\Phi_1 = \delta_{\hat{\mathbf{n}}(\mathbf{z})} / \int_{\mathbf{N}(\mathbf{z})} u_c(\mathbf{n})d\delta_{\hat{\mathbf{n}}(\mathbf{z})}(\mathbf{n})$ where $\delta_{\hat{\mathbf{n}}(\mathbf{z})}$ is the Dirac measure centered at $\hat{\mathbf{n}}(\mathbf{z})$. In this case, Equation 72 still holds if we define $\overline{u_{w_i}/u_c}(\mathbf{z}) \equiv u_{w_i}(\hat{\mathbf{n}}(\mathbf{z}))/u_c(\hat{\mathbf{n}}(\mathbf{z}))$.

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

$$\begin{aligned}
||(Q\Phi_2') - (Q\Phi_2)||_{\text{TV}} &= \left|\left|\sum_i P_i \int_{\mathbf{Z}} \frac{\overline{u_{w_i}}}{u_c}(\mathbf{s}) d(\Phi_2'(\mathbf{s}) - \Phi_2(\mathbf{s}))\right|\right|_{\text{TV}} \\
&\leq \sum_i ||P_i||_{\text{TV}} \left| \int_{\mathbf{Z}} \frac{\overline{u_{w_i}}}{u_c}(\mathbf{s}) d(\Phi_2'(\mathbf{s}) - \Phi_2(\mathbf{s}))\right| \\
&= \sum_i ||P_i||_{\text{TV}} \left|\left|\frac{\overline{u_{w_i}}}{u_c}\right|\right|_\infty \left| \int_{\mathbf{Z}} \frac{\overline{u_{w_i}}}{u_c}(\mathbf{s}) \Big/ \left|\left|\frac{\overline{u_{w_i}}}{u_c}\right|\right|_\infty d(\Phi_2'(\mathbf{s}) - \Phi_2(\mathbf{s}))\right| \\
&\leq \sum_i ||P_i||_{\text{TV}} \left|\left|\frac{\overline{u_{w_i}}}{u_c}\right|\right|_\infty ||\Phi_2' - \Phi_2||_{\text{TV}} \\
&< ||\Phi_2' - \Phi_2||_{\text{TV}}
\end{aligned} \tag{77}$$

Let us explain the steps in Equation 77. The first line simply uses the definition of the measure $(Q\Phi_2') - (Q\Phi_2)$ from Equation 75. The second line uses the triangle inequality and the absolute homogeneity of the norm. The third line just multiplies and divides by $\left|\left|\overline{u_{w_i}/u_c}\right|\right|_\infty$ (recognize that $\overline{u_{w_i}/u_c}$ 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 76. The final line uses our assumption on the size of $\sum_i ||P_i||_{\text{TV}} \left|\left|\overline{u_{w_i}/u_c}\right|\right|_\infty$. Hence, $(Q\Phi_2)(E)$ is a contraction mapping, which implies the existience of a (unique) fixed point Φ_2 which solves Equation 75. Hence, we have proved the existence of an inverse welfare functional taking the form of Equation 70 where the conditional and marginal measures, Φ_1 and Φ_2, are defined as in the proof.

Next, by the same standard extension arguments at the end of Appendix A.1, the "pre-measures" Φ_2 and Φ defined on sets of the form $\{\mathbf{z} \in \mathbf{Z} : \mathbf{z} \leq \tilde{\mathbf{z}}\}$ and $\{\mathbf{n} \in \mathbf{n} : \mathbf{n} \leq \tilde{\mathbf{n}}\}$ (respectively) in Equations 50 and 49 extend uniquely (because $\mathbf{Z}$ and $\mathbf{N}$ are assumed compact) to signed measures on all Borel sets satisfying, respectively, Equation 75 and Equation 78:

$$\Phi(\tilde{\mathbf{N}}) = \int_{\mathbf{Z}} \int_{\mathbf{n} \in \tilde{\mathbf{N}} \cap \mathbf{N}(\mathbf{z})} \frac{1}{\overline{u_c}(\mathbf{z})} dF(\mathbf{n}|\mathbf{z}) d\Phi_2(\mathbf{z}) \tag{78}$$

By Theorem 1.29 of Rudin (1974), if Φ satisfies Equation 78, then for every measurable $U(\mathbf{n}; T)$:

$$\int_{\mathbf{N}} U(\mathbf{n}; T) d\Phi(\mathbf{n}) = \int_{\mathbf{Z}} \int_{\mathbf{N}(\mathbf{z})} U(\mathbf{n}; T) \frac{1}{\overline{u_c}(\mathbf{z})} dF(\mathbf{n}|\mathbf{z}) d\Phi_2(\mathbf{z}) \tag{79}$$

which implies that the measure Φ defined by Equation 49 yields an inverse welfare functional of the form $\int_{\mathbf{N}} U(\mathbf{n}; T) d\Phi(\mathbf{n})$.

□

A.6 Proof of Proposition 2 GE

Proof. We prove each statement below using the measure defined in Equation 49.

1. Recall that the inverse welfare functional in Theorem 2 takes the form:

$$W(U(\mathbf{n};T)) = \int_{\mathbf{N}} U(\mathbf{n};T)d\Phi(\mathbf{n}) = \int_{\mathbf{Z}} \frac{1}{\overline{u_c}(\mathbf{z})} \int_{\mathbf{N(z)}} U(\mathbf{n};T)dF(\mathbf{n}|\mathbf{z})d\Phi_2(\mathbf{z})$$

where Φ_2 is defined in Equation 50. If this inverse welfare functional is not positive then $\exists g(\mathbf{n}) \in C(\mathbf{N})$ with $g(\mathbf{n}) \geq 0 \ \forall \mathbf{n}$ and[42]

$$\int_{\mathbf{Z}} \frac{1}{\overline{u_c}(\mathbf{z})} \int_{\mathbf{N(z)}} g(\mathbf{n})dF(\mathbf{n}|\mathbf{z})d\Phi_2(\mathbf{z}) < 0 \tag{80}$$

Next, we will consider a perturbation in the direction $\tau(\mathbf{z})$ implicitly defined by the following integral equation:

$$\int_{\mathbf{N(z)}} \frac{\left[-u_c(\mathbf{n})\tau(\mathbf{z}) + \sum_i u_{w_i}(\mathbf{n}) \int_{\mathbf{Z}} \tau(\tilde{\mathbf{z}})dP_i(\tilde{\mathbf{z}})\right]}{\overline{u_c}(\mathbf{z})} dF(\mathbf{n}|\mathbf{z}) = \int_{\mathbf{N(z)}} \frac{g(\mathbf{n})}{\overline{u_c}(\mathbf{z})} dF(\mathbf{n}|\mathbf{z}) \geq 0 \tag{81}$$

where the final ≥ 0 follows because $g(\mathbf{n}) \geq 0$ and $u_c > 0$ (noting that the inequality is strict for some types otherwise $\int_{\mathbf{Z}} \int_{\mathbf{N(z)}} g(\mathbf{n})/\overline{u_c}(\mathbf{z})dF(\mathbf{n}|\mathbf{z})d\Phi_2(\mathbf{z})$ would equal 0, violating Equation 80). The LHS of Equation 81 is the average utility impact of a small tax reform in the direction $\tau(\mathbf{z})$ (this follows from the envelope theorem, see proof to Theorem 2). Hence, any $\tau(\mathbf{z})$ satisfying Equation 81 increases average utility for types choosing each $\mathbf{z}$. Next, under the assumptions in Theorem 2, there is a function $\tau(\mathbf{z})$ that satisfies Equation 81. To see this, rewrite Equation 81 as:

$$\tau(\mathbf{z}) = -\int_{\mathbf{N(z)}} \frac{g(\mathbf{n})}{\overline{u_c}(\mathbf{z})} dF(\mathbf{n}|\mathbf{z}) + \int_{\mathbf{N(z)}} \frac{\sum_i u_{w_i}(\mathbf{n}) \int_{\mathbf{Z}} \tau(\tilde{\mathbf{z}})dP_i(\tilde{\mathbf{z}})}{\overline{u_c}(\mathbf{z})} dF(\mathbf{n}|\mathbf{z}) \tag{82}$$

By the exact same logic as in the proof to Theorem 2, Equation 82 is a contraction mapping under the assumption that:

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

Finally, the inverse welfare function satisfies the following equation for any $\tau(\mathbf{z})$ (see Equation 71 in the proof to Theorem 2, noting we normalize $\lambda = 1$):

$$\int_{\mathbf{Z}} \int_{\mathbf{N(z)}} \frac{-u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n})) + \sum_i u_{w_i}(\mathbf{n}) \int_{\mathbf{Z}} \tau(\tilde{\mathbf{z}})dP_i(\tilde{\mathbf{z}})}{\overline{u_c}(\mathbf{z})} dF(\mathbf{n}|\mathbf{z})d\Phi_2(\mathbf{z}) = -\int_{\mathbf{Z}} \tau(\mathbf{z})d\Gamma(\mathbf{z}) \tag{83}$$

The proposed perturbation $\tau(\mathbf{z})$ makes the LHS of Equation 83 negative (by Equations 80 and 81), meaning that $\int_{\mathbf{Z}} \tau(\mathbf{z})d\Gamma(\mathbf{z}) > 0$. In other words, we have found a perturbation that is feasible insofar as it *raises* government revenue yet also (weakly) increases *average* utility at each choice level $\mathbf{z}$ (with some choice levels $\mathbf{z}$

[42] Practically, one can construct such a $g(\mathbf{n})$ by finding a connected set $\tilde{\mathbf{Z}}$ such that $\Phi_2(\mathbf{E}) < 0 \ \forall \mathbf{E} \in \tilde{\mathbf{Z}}$ and then setting $g(\mathbf{n}) = v(\mathbf{z}(\mathbf{n}))$ for any function $v(\mathbf{z}(\mathbf{n})) \geq 0$ that vanishes outside $\tilde{\mathbf{Z}}$.

strictly having strictly higher average utility). Finally, given we assume that $\mathbf{n} \mapsto \mathbf{z}$ is bijective, if average utility is weakly increased at each $\mathbf{z}$ (with some strictly increased), we know that all individuals must be made weakly better off (with some strictly better off); hence, we have constructed a Pareto improvement.

2. We seek to show that there is no feasible reform (i.e., $\int_{\mathbf{Z}} \tau(\mathbf{z}) d\Gamma(\mathbf{z}) \geq 0$) such that everyone is weakly better off and some positive measure of people are strictly better off. By the envelope theorem, for almost all individuals $\mathbf{n}$ the utility impact of a tax reform is given by the following (see proof to Theorem 2):

$$-u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n})) + \sum_i u_{w_i}(\mathbf{n}) \int_{\mathbf{Z}} \tau(\mathbf{z}) dP_i(\mathbf{z})$$

Hence, we seek to show that there is no budget feasible reform $\tau(\mathbf{z})$ such that:

$$\begin{aligned} &- u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n})) + \sum_i u_{w_i}(\mathbf{n}) \int_{\mathbf{Z}} \tau(\mathbf{z}) dP_i(\mathbf{z}) \geq 0 \ \forall \mathbf{n} \\ &- u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n})) + \sum_i u_{w_i}(\mathbf{n}) \int_{\mathbf{Z}} \tau(\mathbf{z}) dP_i(\mathbf{z}) > 0 \text{ for } \mathbf{n} \in \tilde{\mathbf{N}} \text{ with } F(\tilde{\mathbf{N}}) > 0 \end{aligned} \tag{84}$$

If the inverse welfare functional is strictly positive then we know that $\forall g(\mathbf{n}) \in C(\mathbf{N})$ with $g(\mathbf{n}) \geq 0 \ \forall \mathbf{n}$ and $g(\mathbf{n}) > 0$ for $\mathbf{n} \in \tilde{\mathbf{N}}$ with $F(\tilde{\mathbf{N}}) > 0$:

$$\int_{\mathbf{N}} g(\mathbf{n}) d\Phi(\mathbf{n}) = \int_{\mathbf{Z}} \frac{1}{\overline{u_c}(\mathbf{z})} \int_{\mathbf{N}(\mathbf{z})} g(\mathbf{n}) dF(\mathbf{n}|\mathbf{z}) d\Phi_2(\mathbf{z}) > 0 \tag{85}$$

Hence, because $-u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n})) + \sum_i u_{w_i}(\mathbf{n}) \int_{\mathbf{Z}} \tau(\mathbf{z}) dP_i(\mathbf{z}) \geq 0 \ \forall \mathbf{n}$ (with strict inequality on a set $\tilde{\mathbf{N}}$ with $F(\tilde{\mathbf{N}}) > 0$) then:

$$\int_{\mathbf{Z}} \int_{\mathbf{N}(\mathbf{z})} \frac{-u_c(\mathbf{n})\tau(\mathbf{z}(\mathbf{n})) + \sum_i u_{w_i}(\mathbf{n}) \int_{\mathbf{Z}} \tau(\tilde{\mathbf{z}}) dP_i(\tilde{\mathbf{z}})}{\overline{u_c}(\mathbf{z})} dF(\mathbf{n}|\mathbf{z}) d\Phi_2(\mathbf{z}) > 0$$

Given that the inverse welfare functional satisfies Equation 83 (again, see Equation 71 in the proof to Theorem 2, noting we normalize $\lambda = 1$), any reform satisfying Equation 84 must yield $\int_{\mathbf{Z}} \tau(\mathbf{z}) d\Gamma(\mathbf{z}) < 0$. Thus, any small reform that weakly increases utility for all types is not budget feasible; hence, there do not exist any marginal Pareto improving tax reforms.

□