Diffuse Bunching with Frictions: Theory and Estimation*

Santosh Anagol Allan Davids
Benjamin B. Lockwood Tarun Ramadorai†

October 9, 2025

Abstract

We incorporate a general model of frictions into the bunching-based elasticity estimator. The new estimator replaces bunching-window bounds with a "lumpiness parameter" and uses fewer parameters than the conventional approach while delivering additional economically-interpretable quantities such as the size of frictions and unobserved adjustment costs. The model matches rich observed bunching patterns such as sharp-peaked diffuse bunching around tax kinks and depressed density in notch-adjacent dominated regions. We apply the estimator to canonical settings including the U.S. EITC kink, and to administrative tax data from South Africa, where we uncover novel insights on unobserved VAT compliance costs and kink misperceptions. *JEL* codes: H30, J20, O12

*We wish to acknowledge the National Treasury of South Africa for providing us with access to anonymized tax administrative data. We thank Analytics at Wharton and the Penn Wharton Budget Model for funding support. The views expressed in this paper are our own and do not necessarily reflect the views of the National Treasury of South Africa. We are grateful to Wian Boonzaaier, Ana Gamarra Rondinel, Henrik Kleven, Dylan Moore, Jacob Mortenson, Alex Rees-Jones, Juhana Siljander, Joel Slemrod, Jakob Søgaard, David Thesmar, Andrew Whitten, Eric Zwick, and seminar participants at the University of Pretoria, Economic Research South Africa (ERSA), the European Bank for Reconstruction and Development, Imperial College, IIPF 2023, LAGV 2021, LMU Munich, CREST, the NBER Public Economics Meetings, the Toulouse School of Economics, University of Cape Town, University of Michigan, and the South African Revenue Services for helpful comments and to Michael Partridge, Afras Sial, Zamir Ticknor, and Laila Voss for excellent research assistance. All errors are our own. This paper subsumes and replaces the working paper titled "Do Firms Have a Preference for Paying Exactly Zero Tax?" Computational code for the uniform sparsity model is available at `https://github.com/bblockwood/DiffuseBunching`.

†Anagol: Wharton School, University of Pennsylvania. Email: `anagol@wharton.upenn.edu`. Davids: School of Economics, University of Cape Town. Email: `allan.davids@uct.ac.za`. Lockwood: Wharton School, University of Pennsylvania and NBER. Email: `ben.lockwood@wharton.upenn.edu`. Ramadorai: Imperial College London and CEPR. Email: `t.ramadorai@imperial.ac.uk`

1 Introduction

The elasticity of taxable income is among the most central parameters in public economics, appearing as an input in many economic forecasts. It is a key statistic in models of optimal taxation, governing the optimal asymptotic top tax rate on high earners as well as the revenue-maximizing tax rate. One influential estimation strategy, proposed by Saez (2010), seeks to quantify this elasticity by measuring the amount of excess bunching mass in the income distribution around tax bracket thresholds where there is a change in the marginal tax rate (a "kink") or tax level (a "notch").[1]

The conventional bunching estimator computes the excess mass in an income density around a tax kink or notch relative to a counterfactual density that would arise if the tax were linear. This excess mass is converted into an elasticity using a formula which Saez (2010) derives using a frictionless model that predicts an atom of excess mass at the kink. In contrast, bunching seen in empirical distributions is typically diffuse, blurring the distinction between bunching induced by tax incentives and fluctuations in the underlying productivity distribution that would appear even under a linear tax. A now-conventional solution originally proposed by Chetty et al. (2011), is to measure diffuse excess bunching mass as the difference between the observed density and a smooth function fitted to the observed density outside of a visually specified "bunching window" around the kink; Kleven and Waseem (2013) extend this approach to notches.[2] These approaches abstract from the frictions that produce diffusion, leaving open the possibility that they do not successfully recover the (potentially friction-affected) structural elasticity of taxable income. Moreover, by compressing density distortions around the tax bracket threshold into a single excess mass statistic, they risk discarding useful information about taxpayers' behavioral responses to kinks and notches.

In this paper, we present a flexible yet parsimonious model of income adjustment frictions, which we label the sparsity-based approach. Rather than choosing their income frictionlessly from a continuum, taxpayers choose between a discrete set of options representing constraints on their action space. Depending on the process from which these income opportunities are drawn, this approach can nest search or adjustment costs (Chetty, 2012; Gelber, Jones and Sacks, 2020; Mavrokonstantis and Seibold, 2022), lumpy adjustment (Rees-Jones, 2018), unpredictable bargaining outcomes (Andersen et al., 2022), and inattention (Sims,

[1] For examples, see Chetty et al. (2011), Kleven and Waseem (2013), Mortenson and Whitten (2020), Rees-Jones (2018) and others reviewed in Kleven (2016). Bunching estimation has also been applied to domains of retirement incentives (Manoli and Weber, 2016), mobile phone services (Grubb and Osborne, 2015), and educational test scores for both students (Diamond and Persson, 2016; Dee et al., 2019) and teachers (Brehm, Imberman and Lovenheim, 2017), marathon times (Allen et al., 2017), and home sales (Andersen et al., 2022).

[2] To account for non-zero density at "dominated incomes" above the notch, Kleven and Waseem (2013) assume a fraction of taxpayers are unresponsive, leaving unmodeled the source of diffusion in (one-sided) excess mass.

2003; Gabaix, 2014).[3] We then prove a surprising result: a wide variety of such income adjustment frictions share a common underlying structure and are well approximated by a limiting case—formalized in our Proposition 3—of *uniform sparsity*, in which income opportunities are drawn from a Poisson distribution with a single money-metric "lumpiness" parameter quantifying the average distance between discrete income opportunities. Varying this lumpiness parameter allows for these income opportunities sets to be very dense—approaching the continuous model—or very disperse, with the extent of empirical diffusion in the bunching mass identifying the parameter.

The uniform sparsity model predicts many features of empirical bunching behavior that are commonly observed around tax kinks (cf. Saez, 2010; Mortenson and Whitten, 2020) and notches (cf. Kopczuk and Munroe, 2015; Best and Kleven, 2018). Tax kinks induce tent-shaped (i.e., sharp-peaked and fat-tailed) symmetric bunching, as the many taxpayers who would exactly bunch under the frictionless model select their income opportunity nearest to the kink. Tax notches, in contrast, produce asymmetric bunching in the model, with diffuse excess mass below the notch and depressed, yet still positive, density at dominated incomes just above the notch. Intuitively, although such incomes are dominated by earning at the threshold—where effort is lower and post-tax income is higher—some taxpayers' next-best opportunities are sufficiently far away that an opportunity in the dominated region is preferred.

To demonstrate the usefulness of this new approach, we apply it to four different settings. Two of these settings reconsider results using data from prior literature, and two generate new insights from newly-utilized administrative data. In all four cases, our model is able to accurately replicate observed diffuse bunching distributions.

First, we re-analyze data on bunching at the canonical Earned Income Tax Credit (EITC) kink studied in Saez (2010) and more recently in Mortenson and Whitten (2020). Our model predicts a peaked, tent-shaped pattern of bunching around this threshold that closely matches empirical data. This contrasts with the conventional approach, which produces only a smoothed counterfactual density and an excess mass estimate, but no positive prediction about the pattern of that excess mass around the threshold. We show that accounting for income adjustment frictions leads to estimated elasticities that are nearly twice as large as in previous work. These results are consistent with simulations reported in the Appendix, where we show that under a simulated model of sparsity-based frictions the standard estimators tend to substantially underestimate the true elasticities. This finding is consistent with recent work demonstrating that conventional bunching-based estimators produce lower elasticity estimates than other methods even when applied to the same data (Einav, Finkelstein and

[3] See Søgaard (2019) for a review of different models of frictions in the context of labor supply adjustments and bunching patterns.

Schrimpf, 2017; He, Peng and Wang, 2021).

Second, we re-analyze data from a setting with a pure notch of known size: a fiscal stimulus coupon program in China that gave consumers a fixed subsidy amount for spending in excess of a specified amount, first studied by Ding et al. (2025). They estimate the marginal propensity to consume (MPC) in this setting using the notch estimator in Kleven and Waseem (2013). Our estimator is able to fit the observed bunching distributions well; in this setting the source of lumpiness has a particularly clear interpretation, as consumers must add or remove discrete items in order to adjust their spending around the subsidy threshold. This allows us to extract an additional economically meaningful parameter, namely, the average distance between consumption size opportunities.

Third, we analyze data on bunching around a compliance-cost notch of *unknown* size: value-added tax (VAT) registration thresholds. This application illustrates a strength of our method: the ability to recover an unobserved money-metric notch value from observed bunching behavior.For this application we use novel administrative tax data from South Africa, which (like many developing countries) relies more heavily on VAT and corporate income relative to income taxes. In addition to estimating the elasticity of taxable turnover due to VAT tax rate changes around the threshold, this allows us to estimate the compliance cost of registering for VAT. This compliance cost is an important parameter in the optimal setting of taxable turnover thresholds (see, e.g., Keen and Mintz (2004)).

Finally we study a setting in which taxpayers appear treat a statutory tax kink *as if* it were a notch, suggesting potential unobserved real or psychological costs corresponding the tax bracket threshold, for example due to confusion between average and marginal tax rates. Here too we employ novel South African administrative tax data where we observe parallel "notch-like" bunching behavior at tax kinks among both individuals and small businesses. Our method allows us to estimate the extent to which behavior departs from what would be expected under a pure kink. Exploring heterogeneity, we find that small businesses with paid tax practitioners exhibit less bunching diffusion than other firms, and also less notch-like bunching behavior, consistent with practitioners helping these businesses to understand and take advantage of tax incentives by fine-tuning income.

The main contribution of our work is to the theoretical and empirical literatures on bunching estimators and frictions. Our empirical findings also connect to two other areas of the public finance literature. First, our finding of notch-like behavior at statutory kink points contributes to the literature on behavioral frictions and misperceptions about the tax code. Rees-Jones (2018) uses bunching behavior around the threshold at which taxpayers face a net refund or balance due in order to quantify their degree of loss aversion. Rees-Jones and Taubinsky (2020) experimentally study misperceptions of the income tax code, finding that

a large number of respondents "iron," misinterpreting an average tax rate as the relevant marginal rate. Using exogenous variation in worker knowledge about a notch in the Norwegian income tax system, Kostøl and Myhre (2021) estimate that at least 30 percent of estimated optimization frictions are due to workers' imperfect knowledge about the tax system. Outside the domain of taxes, Ito (2014) present evidence that consumers respond (at the margin) to average rather than marginal electricity prices.

Second, some of the empirical applications that we study add to the literature quantifying behavioral responses to taxation in developing economies. Particularly relevant is the subset of papers estimating the elasticity of corporate taxable income (e.g., Devereux, Liu and Loretz, 2014), which is an important parameter in developing economies, given their greater relative reliance on the corporate income tax base (Gordon and Li, 2009). For examples, see Best et al. (2015) in Pakistan, Bachas and Soto (2021) in Costa Rica, and Pillay (2021), Kemp (2019), Boonzaaier et al. (2019) and Lediga, Riedel and Strohmaier (2019) in our setting of South Africa.

The rest of the paper proceeds as follows. In Section 2, we show how sparsity-based frictions modify the frictionless model of bunching at tax kinks and notches. We then prove that a wide range of income adjustment processes representing sparsity-based frictions can be approximated by a remarkably parsimonious limiting case, where income opportunities are drawn from a Poisson distribution with a single statistic quantifying the magnitude of the frictions. Section 3 presents our four empirical applications. Section 4 concludes. Computational code for the the uniform sparsity model is available at `https://github.com/bblockwood/DiffuseBunching`.

2 Model

2.1 Baseline bunching model with frictionless choice

Our starting point is the canonical bunching estimator presented in Saez (2010) and illustrated in Figure 1. Taxpayers with heterogeneous productivities—for example, individuals earning labor income or firms earning profits—choose incomes z under an income tax $T(z)$. The top panel of Figure 1 plots after-tax income as a function of pre-tax income; for consistency with the bunching literature, we will refer to these as "consumption" (c) and "earnings" (z), respectively.

Different tax schedules cause taxpayers to select different incomes, giving rise to different income distributions. In the example plotted in Figure 1, the linear tax $T_0(z)$ results in the cumulative distribution function (CDF) of incomes labeled $H_0(z)$, while the linear tax $T_1(z)$ results in the CDF $H_1(z)$. Corresponding income densities $h_0(z)$ and $h_1(z)$ are shown in the bottom panel.

The horizontal distance between H_0 and H_1 quantifies the income response Δz to a reform from T_0 to T_1 at each quantile n of the distribution. Formally,

$$\Delta z(n) = H_1^{-1}(n) - H_0^{-1}(n). \tag{1}$$

This remains well defined even if taxpayers reorder in response to the tax reform, for example due to heterogeneous elasticities or income adjustment frictions of the kind considered in the next section. Indeed, in the presence of such heterogeneity or adjustment frictions, the response Δz is generally the statistic of primary interest for policy makers, because it quantifies the fiscal effects from tax-induced behavioral distortions. $\Delta z(n)$ can be expressed in elasticity form as:[4]

$$e(n) = \frac{d\ln z}{d\ln(1-T')} = \frac{\ln\left(\frac{z+\Delta z(n)}{z}\right)}{\ln\left(\frac{1-T_1'}{1-T_0'}\right)}. \tag{3}$$

This being a static model, the CDFs H_0 and H_1 should be interpreted as steady-state distributions under different tax schedules, so that $e(n)$ represents the long-run response to a reform from T_0 to T_1.[5]

The insight in Saez (2010) is that the observed income distribution under the piecewise-linear income tax plotted as the kinked solid line in the top panel of Figure 1,

$$T(z) := \begin{cases} T_0(z) & \text{if } z \le k \\ T_1(z) & \text{if } z > k, \end{cases} \tag{4}$$

can provide information about the income response Δz to a reform from T_0 to T_1. This logic is illustrated in Figure 2. Suppose that each income choice is the solution to a taxpayer's frictionless optimization problem, so that their selected income represents a point of tangency between their budget constraint and an upward-sloping indifference curve arising from utility function $u(c,z)$ which trades off the utility of consumption against the disutility of exerting effort to earn income, as illustrated in the case of a linear tax in panel (a). Then taxpayers who

[4]When Δz is small relative to z, $e(n)$ is approximately the compensated elasticity. Formally, it is a weighted average of the compensated and uncompensated elasticities of taxable income $e_c(n)$ and $e_u(n)$:

$$e(n) = \left(\frac{k}{z_0(n)}\right) e_c(n) + \left(1 - \frac{k}{z_0(n)}\right) e_u(n). \tag{2}$$

where $z_0(n) := H_0^{-1}(n)$ denotes the nth quantile of the income distribution under T_0 and k is the income at which T_0 and T_1 intersect. See Kleven (2016), footnote 5, where this is derivedusing the Slutsky equation.

[5]In adjustment cost models, such as Gelber, Jones and Sacks (2020), agents either pay an adjustment cost and fully reoptimize or else do not respond at all. Such models produce predictions about bunching dynamics, i.e., bunching increases over time as more agents adjust. Our static model complements such models, predicting diffusion even in the static (steady state) equilibrium.

choose incomes below k under the kinked tax schedule T (see panel (b)) face the same local budget constraint as they would under the linear tax T_0, while taxpayers earning above k under T face the same local budget constraint as they would under the linear tax T_1. As a result, theory predicts that the observed income distribution under the kinked tax T will coincide with $H_0(z)$ at incomes below k and with $H_1(z)$ at incomes above k, with a discontinuous vertical jump at k of magnitude

$$B := H_1(k) - H_0(k), \tag{5}$$

corresponding to an atom of mass in the observed income density at k. Appendix A illustrates this logic in the presence of a tax notch, which also produces an atom of mass at the threshold k as well as a "hole" (density equal to zero) at incomes just above the bracket threshold which are utility-dominated by income k.

The vertical distance B between these CDFs is related to the horizontal distance Δz by the following equation:

$$H_0(k) + B = H_0(k + \Delta z). \tag{6}$$

Strictly speaking, this condition identifies the income response only at a specific quantile of the income distribution—the quantile that earns k under the linear tax T_1—but applications of the bunching estimator approach often assume that the elasticity is constant across incomes, either globally or in a region around the kink k.

Employing equation (6), we can use a Taylor expansion of $H_0(z)$ around $z = k$ to write B in terms of Δz and the local density $h_0(k)$ and its derivatives:

$$B = h_0(k)\Delta z + \frac{h_0'(k)}{2}\Delta z^2 + \frac{h_0''(k)}{3!}\Delta z^3 + \ldots \tag{7}$$

If we regard terms of order exceeding $h_0'(k)$ as negligible (i.e., if we assume the density $h_0(z)$ is locally linear near k), then we can write the income response Δz as an explicit function of B,

$$\Delta z = \frac{B}{h_0(k) + \frac{h_0'(k)}{2}\Delta z} = \frac{B}{\left(\frac{h_0(k) + h_0(k+\Delta z)}{2}\right)}, \tag{8}$$

where the second equality uses the Taylor approximation $h_0(k+\Delta z) \approx h_0(k) + h_0'(k)\Delta z$ for small Δz. Substituting equation (8) into equation (3) yields the bunching estimator derived in Saez (2010).[6]

[6]To produce equation (5) in Saez (2010), rearrange equation (8) above to be $\Delta z = k\left[\left(\frac{1-T_1'}{1-T_0'}\right)^e - 1\right]$ and substitute it into equation (3), noting that $h_1(k) = h_0(k+\Delta z)\left(\frac{1-T_0'}{1-T_1'}\right)^e$.

This frictionless model predicts an atom of excess bunching mass in the observed income density. In practice, excess bunching mass—when it is observed—is typically diffuse, like the green density line $h(z)$ in Figure 1 (see, for example, Saez (2010)). Equivalently, observed income CDFs generally do not jump discontinuously at tax kinks; instead, they transition gradually from $H_0(z)$ to $H_1(z)$ across a range of incomes around k, as shown by the green CDF $H(z)$. Such diffuse bunching is usually interpreted as evidence of income-adjustment frictions.

In the presence of adjustment frictions, the vertical distance B between the latent CDFs $H_0(k)$ and $H_1(k)$ still quantifies the size of the key empirical parameter of interest—the steady-state income response to a marginal tax rate increase, Δz—according to equation (7). But in contrast to the frictionless setting, B cannot be estimated by measuring a vertical discontinuity in the income CDF at k.

To estimate B in the presence of adjustment frictions, Saez (2010) proposes quantifying the excess mass in the observed density $h(z)$ around k. To formalize this logic, note that the vertical distance B is identical to the integral of the excess mass in the observed density $h(z)$ around k, relative to the counterfactual densities $h_0(z)$ and $h_1(z)$:[7]

$$B \equiv \int_{-\infty}^{k} (h(z) - h_0(z))\, dz + \int_{k}^{\infty} (h(z) - h_1(z))\, dz. \tag{9}$$

This relationship is illustrated in Figure 1, where the shaded region of excess mass in the bottom panel is equal to B.

Although this equation provides a strategy for estimating B in the presence of adjustment frictions, it requires knowledge of the counterfactual densities h_0 and h_1, one or both of which is often unknown. Thus much of the empirical bunching literature proposes strategies for estimating excess bunching mass B from the observed density $h(z)$ without full knowledge of the counterfactual densities h_0 and h_1. Chetty et al. (2011) influentially proposes estimating B by integrating over the observed density $h(z)$ relative to an estimated counterfactual density obtained by fitting a flexible polynomial to $h(z)$ outside of a visually specified bunching window. Kleven and Waseem (2013) extends this approach to handle notches, wherein the tax *level* (and not just the marginal tax rate) jumps discontinuously at a threshold k. While these approaches implicitly accommodate adjustment frictions by allowing for bunching to be diffuse, they leave the underlying optimization process that produces diffuse bunching unmodeled.

As we show in the next section, a more explicit treatment of the diffusion process can improve the estimation of the bunching mass B in the presence of adjustment frictions by better distinguishing between the diffuse bunching that is "expected" around a kink and

[7] This identity follows from the fact that $B = H_1(k) - H_0(k) = [H(k) - H_0(k)] + [H_1(k) - H(k)]$, where the bracketed terms are equal to the two integrals in equation (9), respectively.

the underlying counterfactual densities. Moreover, by explicitly modeling these adjustment frictions and then fitting the observed shape of the diffuse mass around a kink, we can exploit additional information about income-adjustment frictions, costs, and tax misperceptions that is discarded when the integrated bunching mass is reduced to a scalar measure of excess mass.

2.2 A model of sparsity-based frictions

We introduce a simple modification to the static frictionless model of Saez (2010): rather than selecting incomes from a continuum, taxpayers choose their preferred income from a sparse set of opportunities.

This notion of sparsity-based frictions spans a diverse set of microfoundations. In a setting where incomes are produced by performing discrete jobs or gigs that are discovered via a search process, the income opportunity set represents the incomes available from the set of jobs (or combinations of jobs) that a taxpayer faces after searching with a given intensity. If an employee works for a single employer, they may face discrete choices ("lumpiness") over work shifts or overtime opportunities, rather than being able to adjust their labor hours continuously. The model can also be interpreted to allow for rational inattention, where a taxpayer learns about the precise income that arises from each potential combination of actions—such as claiming a specific set of tax deductions—only by paying information-gathering costs.[8] Taxpayers' action spaces could span both real responses—e.g., deciding which income-earning opportunities to pursue—or reporting responses—e.g., a business owner deciding which of their (lumpy) payments to realize in the current tax year, or which potentially tax-deductible expenses to claim.[9]

Formally, we assume that each taxpayer faces an *income opportunity set*,

$$\left\{z^* + \varepsilon_1,\, z^* + \varepsilon_2,\, \ldots,\, z^* + \varepsilon_M\right\}, \tag{10}$$

consisting of M income opportunities, which are offset from the taxpayer's preferred ("target") income z^* by a random error term ε_i. We assume that the error terms are independent and identically distributed (iid) within taxpayer type, with distribution $F_\varepsilon(x|n)$ and density $f_\varepsilon(x|n)$. We denote the error distribution F_ε and the number of income draws M. The pair (F_ε, M) characterizes the *income opportunity process*.

[8] This interpretation is in line with Jung et al. (2019), who microfound the compression of an underlying continuous distribution of actions into a lower-dimensional discrete set when information processing is costly. Such information-gathering costs have also been used in the literatures on firm price-setting and household trading in financial markets (Alvarez, Lippi and Paciello, 2011; Alvarez, Guiso and Lippi, 2012; Abel, Eberly and Panageas, 2013; Andersen et al., 2020).

[9] For the case of lumpy tax deductions, see Rees-Jones (2018) and also discussions in the accounting literature, e.g., Kothari, Leone and Wasley (2005).

Taxpayers of an identical type (productivity, elasticity, etc.) will nevertheless choose different incomes due to their different opportunity sets. The *type-conditional density* of incomes among taxpayers of a given type is characterized by the following proposition.

Proposition 1. *The type-conditional density of incomes at $\tilde{z}$ among taxpayers of type n with utility $u(c,z|n)$ under income tax $T(z)$ is given by*

$$g(\tilde{z}|n) = \underbrace{M \cdot f_\varepsilon\left(\tilde{z} - z^*(n)|n\right)}_{\text{probability of drawing } \tilde{z}} \times \underbrace{\left[1 - \int_{z \in \Theta(\tilde{z}|n)} f_\varepsilon\left(z - z^*(n)|n\right) dz\right]^{M-1}}_{\text{probability of choosing } \tilde{z} \text{ conditional on drawing it}}, \tag{11}$$

where

$$\Theta(\tilde{z}|n) := \left\{ z \middle| u(z - T(z), z|n) \geq u(\tilde{z} - T(\tilde{z}), \tilde{z}|n) \right\}, \tag{12}$$

denoting the set of incomes that utility-dominate $\tilde{z}$ for a taxpayer of type n.

The logic behind this result is illustrated in Figure 3, which plots the indirect utility function $v(z|a)$ over income for a taxpayer of type a (see Figure 2) who faces a locally linear income tax. The type-conditional income density at income $\tilde{z}$ is equal to the probability that $\tilde{z}$ is drawn as an element of taxpayer's opportunity set multiplied by the probability that none of the taxpayer's $M-1$ other opportunities fall in the pink region of incomes that provide higher utility than $\tilde{z}$.

The first term in equation (11) is the probability that $\tilde{z}$ is drawn in the taxpayer's income opportunity set—it is the probability of drawing the error value $\varepsilon = \tilde{z} - z^*(n)$ that would produce income $\tilde{z}$ multiplied by the M chances to draw that error value. The second term in equation (11) represents the probability that $\tilde{z}$ is *chosen* from the opportunity set conditional its being drawn, which is equal to the probability that none of the other $M-1$ income opportunities are in the region of dominating incomes, $\Theta(\tilde{z}|n)$, shaded in pink in Figure 3. Formally, when the indirect utility function is convex, as is the case for convex preferences under a linear income tax, then $\Theta(\tilde{z}|n)$ is the interval $\left[\underline{Z}(\tilde{z}|n), \overline{Z}(\tilde{z}|n)\right]$, where $\underline{Z}(\tilde{z}|n)$ and $\overline{Z}(\tilde{z}|n)$ are the minimal and maximal values of z satisfying the equation $v(z|n) = v(\tilde{z}|n)$. In the figure, the type-conditional density is maximized at the taxpayer's target income $z^*(n)$, at which point the set of dominating incomes $\Theta(z|n)$ is a singleton containing only $z^*(n)$.

Figure 4 illustrates how to extend this logic to a tax kink. Figures 4(a) and 4(b) plot the budget constraints and indifference curves for types b and c (from Figure 2) under the kinked tax schedule $T(z)$. Figures 4(c) and 4(d) plot the resulting indirect utility functions for each type, which are constructed by retaining the relevant segments of the indirect utility functions that would arise under each of the linear tax schedules $T_0(z)$ and $T_1(z)$.

Figure 5 illustrates the implications for the type-conditional income densities $g(z|b)$ and $g(z|c)$. The formula for the type-conditional density in Proposition 1 carries through: the

distortionary effect of the tax kink operates entirely through its impact on the set of dominating incomes $\Theta(z|n)$, which are again shaded as pink regions in Figure 5. Because the kink point k maximizes indirect utility for both types b and c (and all types in between), the type-conditional income density is also maximized at k for those types. Appendix A extends this logic to notches, where adjustment frictions lead to diffuse excess mass in the type-conditional density below the notch threshold and a density above the threshold that, while discontinuously lower, is still positive. This illustrates an important contrast between the predictions of this model, in which some taxpayers choose an income opportunity in the "dominated region" just above the notch—because all of their competing opportunities are more distant and yield lower utility—and the frictionless model, in which the density is zero above the notch, because it is strictly dominated by the threshold income k.

By aggregating these type-conditional income densities across the continuum of types $f_n(n)$ we obtain the *observed* income density:

$$h(z) = \int_n g(z|n) f(n) dn. \tag{13}$$

If the type density is smooth and the tax schedule is linear, then the type-conditional density effectively acts as a smoothing filter through which the type density is passed, producing an observed density that is also smooth. But in the presence of a kink, the diffuse bunching in type-conditional densities means that the excess mass in the observed income distribution is not concentrated at the kink. Rather, it is spread out as illustrated in the left panels of Figure 6. The right panels illustrate the effects of a notch, which produces diffuse excess mass below the notch and a depression in the income density above it.

2.3 A tractable case: uniform sparsity

For a given income opportunity process—combined with the usual parameters of a frictionless bunching model—Proposition 1 allows us to numerically compute the observed income distribution. In general, this computation may be a demanding one, as it will depend on several structural parameters, such as the number of income opportunity draws and the parameters of the error distribution F_ε.

In this section, we consider a parsimonious case, *uniform sparsity*, which is characterized by only one parameter and turns out to be a good approximation for a broad range of other income opportunity processes.

To build intuition, consider the income process simulated in Saez (1999)—the working

paper that preceded Saez (2010)—in which taxpayers have isoelastic utility

$$u(c, z|n) = c - \frac{n}{1+1/e}\left(\frac{z}{n}\right)^{1+1/e}, \tag{14}$$

and each taxpayer draws M income opportunities from a uniform distribution of width W centered around their target income $z^*(n)$.[10] This is a model of sparsity-based frictions with two additional parameters relative to the frictionless model: the number of income opportunities M, and the width of the uniform distribution W from which they are drawn.

Figure 7 displays several simulated income densities that arise from this model, and variations on it, plotted around a tax kink.[11] Figure 7(a) displays four simulated income densities in which each taxpayer draws $M = 1$, 2, 3, or 5 income opportunities from a uniform distribution of width $W = 50{,}000$ around their target income. When $M = 1$, the bunching mass is a rectangular plateau centered around k, which is produced by the mass of bunchers who target the income k but then draw an income opportunity that is offset from that target by a uniformly distributed error. When $M = 2$, the plateau disappears and the bunching mass approximates an inverted "V," reflecting that when taxpayers targeting income k face two opportunities, they choose the one that is closer to their target. As the number of income opportunities increases, this pattern becomes more pronounced, with a higher peak at k.

The limit of the series of densities in Figure 7(a) as $M \to \infty$ is the frictionless model with no diffusion in the bunching mass. However, there is an alternative notion of a limiting case in which diffusion remains non-degenerate. Consider the simulation in which $M = 5$. In this case, taxpayers' income choices are effectively determined by the distribution of just two income opportunities, the lowest opportunity above their target and the highest opportunity below their target, which dominate all other more distant opportunities. As a result, an income opportunity process with $M = 5$ uniform draws from an income window of $W = 50{,}000$ produces very similar behavior to an income opportunity process with $M = 10$ draws from a window of $W = 100{,}000$. Both specifications produce distributions of the two nearest-to-target opportunities that are uniformly drawn in the vicinity of the target with the same density $M/W = 5/50{,}000 = 10/100{,}000 = 0.0001$.

[10] Formally, the error distribution for this income opportunity process is

$$F_\varepsilon(x|n) = \begin{cases} 0 & \text{if } x < -W/2, \\ \frac{x - W/2}{W} & \text{if } -W/2 \le x \le W/2, \\ 1 & \text{if } x > W/2. \end{cases}$$

[11] These simulations use tax parameters with similar nominal magnitudes to our empirical setting: the marginal tax rate rises from $t_0 = 0.1$ to $t_1 = 0.2$ at the bracket threshold of $k = 300{,}000$, and we assume a locally linear density of productivity n and an elasticity of taxable income of $e = 0.3$; see Appendix C for further simulation details.

Motivated by this observation, we consider the behavior of the series of income densities that arises when each agent draws M opportunities from a window of width $M \times 10{,}000$ around their target. Figure 7(b) plots this series with the same values of M as in Figure 7(a). The income density with $M = 1$ exhibits a rectangular plateau centered around the kink, though this time with a width of 10,000. But as M increases—and the width W increases proportionally—the bunching density appears to converge toward a distinctive "tent shape" with a peak at k.

The apparent convergence exhibited in Figure 7(b) motivates a natural question: does this series converge to well-defined limiting density? The answer turns out to be yes, and it is this limiting density that we call the "uniform sparsity model," formally defined as follows.

Definition 1 (Uniform sparsity)**.** *Under* ***uniform sparsity****, the income opportunity set for each taxpayer is a Poisson process with arrival rate* λ*.*

Intuitively, each taxpayer draws an infinite set of income opportunities spanning the number line. The probability of drawing any particular income opportunity is the same, with an average of λ opportunities drawn from any \$1 interval. The limiting density of the series in Figure 7(b) is the uniform sparsity model with $\lambda = 0.0001$. (For a proof of this claim, see Proposition 3 and its proof in the Appendix.)

Under uniform sparsity, the type-conditional density has a tractable form, as shown in the following proposition.

Proposition 2. *Under the uniform sparsity model, the type-conditional density of incomes at* $\tilde{z}$ *among taxpayers of type* n *with utility* $u(c, z|n)$ *is*

$$g(\tilde{z}|n) = \lambda \exp\left[-\lambda \left|\Theta(\tilde{z}|n)\right|\right] \tag{15}$$

where $|\Theta(\tilde{z}|n)|$ *denotes the Lebesgue measure of the set of dominating incomes* $\Theta(\tilde{z}|n)$ *as defined in Proposition 1.*

Proof. Due to the Poisson income process, the probability that any particular income is in the income opportunity set is a constant equal to λ. The type-conditional density $g(\tilde{z}|n)$ is equal to the probability that $\tilde{z}$ is in the taxpayer's income opportunity set, which is λ, multiplied by the probability that no income opportunity is drawn from the dominating income region $\Theta(\tilde{z}|n)$, which is $\exp\left[-\int_{z\in\Theta(\tilde{z}|n)} \lambda dz\right] = \exp\left[-\lambda \left|\Theta(\tilde{z}|n)\right|\right]$. Multiplying these two terms produces the result. □

The parameter λ has a natural economic interpretation: the expected distance between adjacent income opportunities is $1/\lambda$. We call this distance the "lumpiness" of the income opportunity process. Greater lumpiness implies that taxpayers face a sparser set of income

opportunities, and thus bunching around tax kinks is more diffuse. Because it is sometimes more natural to think of these adjustment frictions as parameterized by their lumpiness rather than its inverse, we define the "lumpiness parameter" $\mu := 1/\lambda$, which has a lower bound of 0—corresponding to the frictionless model—and is unbounded above.

The convergence exhibited in Figure 7(b) is not unique to uniformly distributed error terms. Figure 7(c) displays an analogous series of income densities in the case where income opportunities are drawn from a normal (rather than uniform) distribution centered around the target income. As in Figure 7(b), the spread of the distribution from which income opportunities are drawn is adjusted to preserve the density of draws in the neighborhood of the target income—this time by rescaling the standard deviation of F_ε in proportion to M.[12] As in Figure 7(b), the series appears to converge quickly toward a distinctive tent shape as M increases.

Our next proposition demonstrates that this series of income densities also converges to the uniform sparsity model as $M \to \infty$. More generally, it shows that the uniform sparsity model is the limit of such a series for *any* distribution of income opportunities with positive continuous density around the income target, suggesting that this single-parameter model is a parsimonious approximation for a broad class of adjustment frictions.

To formalize this statement, we begin with an arbitrary distribution of income opportunity errors $F_\varepsilon(x)$ for which $f_\varepsilon(0) > 0$. We then define a transformation that controls the "spread" of this distribution around the target income, $F_\varepsilon^M(x) := F_\varepsilon(x/M)$, so that as M increases, the density of opportunities around the target income, $M \cdot f_\varepsilon^M(0)$ remains constant.[13] We can then show the following proposition.

Proposition 3. *For any error distribution $F_\varepsilon(x)$ with positive continuous density at $x = 0$, the income density arising from a model in which each agent draws M income opportunities offset from their target by iid disturbances $\varepsilon \sim F_\varepsilon^M$ converges pointwise to the density produced by the uniform sparsity model with $\lambda = f_\varepsilon(0)$.*

The proof is presented in Appendix B.

A striking feature of Figures 7(b) and 7(c) is that these series in M converge to the uniform

[12] Formally, the probability density function of this error term distribution is $f_\varepsilon(x|n) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{x^2}{2\sigma^2}\right]$, where σ is the standard deviation of the normal distribution from which errors are drawn. The probability of a specific income opportunity draw being equal to $\tilde{z}$ is $f_\varepsilon(x|n)$, and thus the probability of *any* of the M draws being equal to $\tilde{z}$ is $M \cdot f_\varepsilon(\tilde{z}|n) = \frac{M}{\sigma\sqrt{2\pi}} \exp\left[-\frac{x^2}{2\sigma^2}\right]$. Therefore the density of income opportunities in the vicinity of the target is equal to $M \cdot f_\varepsilon(0|n) = \frac{M}{\sigma\sqrt{2\pi}}$. As in the case with uniform distributions above, we jointly adjust M and the spread of the distribution—this time by adjusting the standard deviation σ—to hold constant the density of opportunities around the target income. This amounts to scaling M and σ proportionally.

[13] Note that the transformations used to construct the series of densities in Figures 7(b) and 7(c) are special cases of this general construction.

sparsity model quite quickly. The simulation with just two income opportunities looks similar to uniform sparsity, and the simulations with $M = 3$ and $M = 5$ are nearly indiscernible.

Panels (d)–(f) of Figure 7 reproduce the simulations in panels (a)–(c) in the case of a tax notch, where the tax liability increases by 1000 at the bracket threshold. In panel (d)—as in panel (a)—the case with $M = 1$ has distinctive features produced by the specifics of the uniform distribution from which income opportunities are drawn, with the bunching mass again having a plateau-like shape around the bracket threshold.[14] As the number of opportunities M increases, the mass develops a distinctive shape with diffuse mass to the left of the threshold and a depression to the right. Although convergence is slightly less rapid in the case of a notch than the case of a kink, both densities appear quite similar to the uniform sparsity model for $M = 5$.

Taken together, Proposition 3 and the simulations in Figure 7 suggest that the uniform sparsity model is a parsimonious approximation for a broad class of adjustment frictions in which taxpayers choose their final income from a sparse set containing multiple income opportunities. This model has a number of attractive features. First, the patterns of bunching it produces are strikingly similar to those observed empirically in settings with tax kinks and notches. The tent shape (high kurtosis) of the uniform sparsity specification in Figure 7(b) resembles both the shape of diffuse bunching observed in canonical "bunching at the kink" papers (e.g., Saez, 2010; Chetty et al., 2011; Mortenson and Whitten, 2020). Similarly, the bunching pattern around a notch in Figure 7(c) matches key empirical features observed in Kleven (2016), with diffuse mass to the left of the notch and a positive (but depressed) density in the "dominated region" to the right of the notch.[15] Both types of patterns appear in our empirical setting, which we discuss in Section 3.

Figure 7 also provides some reassurance that although the uniform sparsity model is not entirely general—in particular, it does a poor job of approximating the income distributions produced when $M = 1$—those cases are not the ones that appear to exhibit the key patterns of empirical bunching in many settings of interest. This is particularly evident in the case of

[14] In panel (d), for $M = 1$, the downward slope at the left end of the plateau comes from the interaction between the mass of bunching taxpayers who target their income at the bracket threshold k and the absence of taxpayers targeting income in the dominated region to the right of k. When $M = 1$, the observed density at any particular income is simply the average of the density of target incomes—which resembles Figure 7(d)—in a \$50,000 window centered at that point. At \$275,000 (the left end of the plateau), this averaging window spans from \$250,000 to \$300,000, across which the target income density is positive. As income increases from \$275,000, the average falls due to the absence of target incomes to the right of k. The plateau levels out again at \$300,000 when the upper end of the averaging window rises above the upper bound of the dominated income region.

[15] Kleven and Waseem (2013) propose an alternative model that predicts positive mass in the dominated income range, in which a subset of taxpayers are insensitive to the presence of the notch due to adjustment or informational frictions. Such a model explains positive density in the dominated region, but in contrast to the uniform sparsity model, it does not predict leftward diffusion in the excess density at the bracket threshold.

notches, where specifications with $M = 1$ predict an income density that is *continuous* across the notch threshold at k, in contrast to specifications with $M > 1$, which exhibit a sharp drop in density at the notch due to the taxpayers' endogenous selection of their preferred draw. Empirical densities like those in Kleven (2016) that clearly exhibit such a discontinuity suggest they are better represented by a model with $M > 1$, for which the uniform sparsity model performs well.[16]

A second strength of the uniform sparsity model is its parsimony. It distills the many details underlying particular sparsity-based models of adjustment frictions—such as the parametric distribution of income opportunities, the number of income opportunity draws, and the spread of the distribution (controlled, e.g., by the width of the uniform distribution or variance of the normal distribution)—into a single lumpiness parameter with a clear economic interpretation.

A third feature of the uniform sparsity model is that it has no "center," and as a result it can be conceptually separated from the taxpayer's choice of target income. This is particularly attractive in the case of notches, where a taxpayer may be indifferent between two different incomes in the frictionless model. This case is illustrated in Figure A3, where type c is exactly indifferent between earning two different incomes. Under a model of targeting or directed search, this would raise the question of which of the two equally desirable incomes should be modeled as the "target" around which opportunities are drawn. Under uniform sparsity, this question is irrelevant, because the choice of target is inconsequential.[17]

2.4 Quadratic Approximation for Numerical Implementation

The simulations above assume the isoelastic utility function in equation (14) as in Saez (2010). In general, the model can be solved numerically for any particular specification of utility. However, if the indirect utility function is approximated by its second-order Taylor expansion around the target income, the model turns out to be particularly tractable. We impose this approximation in the following assumption.

Assumption 1. *Each taxpayer's indirect utility over incomes under a linear tax is approximated*

[16] As a counterexample,Allen et al. (2017) evidence of bunching in marathon times, consistent with a psychological payoff for recording a time under 4 hours, for example. Notably, the bunching patterns in that paper resemble the shape of the $M = 1$ simulation in Figure 7(f), suggesting that marathon times are better modeled by an imperfect targeting model with a single draw—representing one's finish time—rather than a uniform sparsity model with multiple options from which to choose.

[17] This feature also allows us to sidestep the question of whether taxpayers *anticipate* frictions when selecting their choice of target. In the case illustrated by Figure A2, for example, if type c is sophisticated, they would do better to target an income slightly below the threshold k, rather than k, in order to reduce the probability of drawing income opportunities on the "wrong side" of the notch; a naive taxpayer might instead target k. The uniform sparsity model does not require specifying a target, so this issue is irrelevant.

by its second-order Taylor approximation around their target income:

$$v(z|n) \approx v(z^*(n)|n) + v'(z^*(n)|n)(z - z^*(n)) + \frac{1}{2}v''(z^*(n)|n)(z - z^*(n))^2. \tag{16}$$

Assumption 1 is a useful simplification because it implies that under a linear income tax, the distance between a given income z and the taxpayer's optimal income $z^*(n)$ is the same as the distance between $z^*(n)$ and the utility-equivalent income on the "far side" of the optimum.[18]

In the context of Figure 3, this assumption implies that the dominating income region shaded in pink is centered around the taxpayer's preferred income $z^*(n)$. As a result, the size (Lebesgue measure) of the dominating income set $\Theta(\tilde{z}|n)$ under a linear tax takes a particularly simple form:

$$|\Theta(\tilde{z}|n)| = 2\left|z - z^*(n)\right|. \tag{17}$$

An implication of this result is that under Assumption 1 the size of the dominating region Θ does not depend on the taxpayer's elasticity of taxable income, which governs the curvature of the indirect utility function, because under quadratic indirect utility the incomes $z^*(n) - \delta$ and $z^*(n) + \delta$ both generate the same utility for every δ, regardless of curvature. By Proposition 2, the type-conditional density $g(z|n)$ depends on a taxpayer's type n only through the dominating income set $\Theta(z|n)$, which in turn depends on type only through target income $z^*(n)$, so we can refine our notation to write the type-conditional density as $g(z|z^*)$, where z^* is the taxpayer's preferred target income under a specified linear tax schedule. We will often use this notation in the remainder of the paper.

The following proposition characterizes the type-conditional density of incomes under the uniform sparsity model and Assumption 1.

Proposition 4. *Under Assumption 1, the type-conditional income density under the uniform sparsity model among taxpayers of type n for any linear income tax is given by*

$$g\left(z|z^*(n)\right) = \lambda \exp\Big[-\lambda \cdot 2\left|z - z^*(n)\right|\Big], \tag{18}$$

where $z^*(n)$ is the taxpayer's target income.

In words, because income opportunities are drawn uniformly from the number line, the probability that no opportunity falls in the dominating region $\Theta(z|n)$ depends only on the size of the interval, which under a linear tax is $2|z - z^*(n)|$. This probability declines exponentially at rate λ with the size of this interval, producing a tent-shaped density function centered at

[18] We describe the numerical implementation of the model in detail in our computational appendix document. There, Figure B1 demonstrates that this approximation also produces results very similar to the exact solution for the isoelastic utility function used in the simulations above.

the taxpayer's preferred income declining exponentially in both directions. This distribution is also known as the Laplace distribution with location parameter $z^*(n)$ and scale parameter $\mu(n)$. This tractable analytic representation of the type-conditional density function significantly aids the computational estimation described in Section 2.6 below.

2.5 Identification of Model Parameters

To explain the identification of the elasticity e and lumpiness parameter μ in our model, we follow the approach suggested in Andrews, Gentzkow and Shapiro (2020). That is, (1) we show simulations of how different parameter values generate different income densities, and (2) we provide an analytical proof that the parameters of the uniform sparsity model are separately identified in the sense of Matzkin (2013): for a given underlying distribution of types (equivalently, target incomes), no two combinations of elasticity e and lumpiness μ lead to the same observed income distribution.

Intuition for identification is provided by Appendix Figure A4, which plots simulated income densities under various combinations of values for the elasticity parameter e and the lumpiness parameter μ in the presence of a kink and a notch. Panel (a) plots simulated income densities under the baseline parameter values, as well as with lower and higher values of the elasticity e. A higher elasticity raises the overall amount of diffuse bunching mass around the kink. Panel (b) holds fixed the elasticity but varies the lumpiness parameter μ, altering the spread of the bunching mass around the kink while preserving the total amount of excess mass. Although at first glance the bunching masses in panels (a) and (b) might appear similar, on inspection the source of identification is clear. A higher elasticity e and a lower μ both lead to a higher peak at the kink, but the former achieves that higher peak by increasing the total amount of bunching mass, while the latter holds fixed the bunching mass, and thus has a much narrower spread of excess mass around the kink. Panels (c) and (d) plot such simulations in the presence of a notch.

In Appendix D we provide a proof of conditions under which separate identification of e and μ can be shown analytically in the case of a kink. The proof proceeds in three steps. First, we demonstrate that if $h_0(z)$—the counterfactual income density that would be observed under the linear tax $T_0(z)$—is locally linear, then the corresponding density of *target incomes*, denoted $h_0^*(z)$, is given by the same locally linear function.[19] Second, we show that in a region of constant elasticity, the leftward shift in the CDF of target incomes $H_1^*(z)$ is a monotonic function

[19] More generally, in Appendix Lemma 4 we show that if the observed density around income z under a linear tax is equal to a polynomial function $h(\tilde{z}) = \alpha_0 + \alpha_1(\tilde{z}-z) + \alpha_2(\tilde{z}-z)^2 + \ldots$, then the density of target incomes around z is also equal to a polynomial with coefficients $\alpha_n^* = \alpha_n - \frac{\alpha_{n+2}}{(2\lambda)^2}$. By implication, if the observed density is linear (i.e., $\alpha_n = 0$ for $n > 1$), then the target income density is also linear.

of the elasticity e, and at any income above the kink, $k+\varepsilon$, the observed CDF $H(z)$ converges arbitrarily closely to the CDF $H_1(z)$ when the tax change at the kink is sufficiently small. The integrated difference between $h_0(z)$ and the observed density $h(z)$ up to $k+\varepsilon$ identifies the bunching mass B, and thus the elasticity e. Finally, we show that elasticity e and the observed density at the kink $h(k)$ identify the lumpiness parameter $\mu = 1/\lambda$.

2.6 Estimation

We now describe how the parameters of this model can be estimated from data. The empirical strategy is to select the model parameters that maximize the likelihood of observing a given empirical density. To do so, we search over the parameter values for the elasticity e and the lumpiness parameter μ. If desired, we can also allow the tax notch size dT to be an estimated parameter, treating it as a revealed feature of taxpayer behavior.

In settings where the counterfactual income density is unobserved, we must infer the underlying ability density $f(n)$ by imposing some parametric structure. As in much of the bunching literature, (see Chetty et al., 2011) the key identifying assumption is that the ability distribution, and thus the counterfactual income density $h_0(z)$, is smooth in the vicinity of the bracket threshold k, in a sense that will be made precise. Intuitively, this amounts to assuming that the bracket threshold is not located at a point in the income distribution that happens to coincide with a distortion in the underlying ability distribution.[20]

We operationalize this identification strategy by assuming that the ability density follows a polynomial of order Q, i.e.,

$$f(n|\theta) = \theta_0 + \theta_1 n + \theta_2 n^2 + \ldots = \sum_{q=0}^{Q} \theta_q n^q \tag{19}$$

for a vector $\theta = \{\theta_0, \theta_1, \ldots, \theta_Q\}$.

We then estimate the parameters of the model—e, μ, θ, and (if desired) dT—using maximum likelihood. Letting i index the observations in the data, with X_i denoting each observation's income, our starting point for the likelihood function is

$$L(e, \mu, dT, \theta) = \Pi_i h(X_i = z|e, \mu, dT, \theta). \tag{20}$$

The numerical details of this estimation method are described in Appendix B.5.

[20] Blomquist et al. (2021) explore this identification strategy and its limitations at length. See Moore (2022) for a discussion of what can be identified by bunching estimators without estimating the elasticity directly, and Bertanha, McCallum and Seegert (2023) for the potential to exploit notches for elasticity identification with weaker assumptions on the density.

While this method may appear to resemble the widely-used Chetty et al. (2011) approach of fitting a flexible polynomial to the observed income distribution outside of a selected bunching window, it is substantively different in two important ways. First, by structurally accounting for the distortion pattern produced by the bracket threshold, there is no need to select (visually or otherwise) an excluded bunching window around the threshold when computing the best-fit values of θ. Instead, even data near the bracket threshold helps to identify θ. This logic points to the greater robustness of this estimation method to choices about the polynomial degree Q relative to standard bunching estimators, where additional flexibility may attempt to fit excess mass that spills outside the excluded bunching window. We confirm this reasoning in simulations in Appendix C.3.

Second, this approach assumes that the smooth polynomial structure is a feature of the underlying ability distribution, $f(n)$, rather than of the observed income distribution outside the bunching window. As illustrated in Figure 1(c), the frictionless model predicts a discontinuity in the income density around the bracket threshold due to the jump in types and the condensed mapping from types to income under higher marginal tax rates. By estimating the polynomial coefficients on the type distribution directly, this approach does not impose smoothness across the bracket threshold.

In Appendix C we compare our proposed estimation method to the conventional bunching estimator. This comparison suggests that when sparsity-based frictions are present, the conventional estimator can be substantially biased downward—by up to 50% when there is substantial bunching diffusion—in a way that our method is not. To show this, we simulate data containing sparsity-based frictions with known parameters. We then apply both our method and the conventional estimation method to the simulated data, to compare the procedures' ability to recover the (known) elasticity of taxable income. We perform this estimation repeatedly for different random samples from the data-generating process in order to assess the precision of the estimates. We compute standard errors for our maximum likelihood point estimates using the standard maximum likelihood estimator.

Results are shown in Appendix Figure A7. Panel (a) displays a histogram of the estimates from each method, and panel (b) shows the mean and 95 percent confidence interval of the estimates at each value of the lumpiness parameter μ. As μ increases, the conventional estimator exhibits substantial bias, underestimating the true elasticity by about 50 percent at larger values of μ. The conventional estimator also gives a misleading impression of precision, with 95% confidence intervals that lie entirely below the true elasticity in the data-generating process in all the plotted simulations.

Bias in the conventional estimator arises primarily from mispecifying the counterfactual density. As discussed, the conventional method requires specifying a bunching window

outside of which no bunching is assumed to be present, allowing the counterfactual density to be estimated outside this specified window. However, under a data generating process where there are sparse income choices, some bunching mass can spill beyond any finite bunching window, and in particular, beyond the bunching windows typically selected by visual inspection or current algorithms (see Appendix C.2). This leaves a substantial amount of bunching mass outside the window, which the conventional approach confuses with the counterfactual density. The counterfactual density in this approach is therefore pulled into the bunching region, thus leading to an under-estimate of the bunching mass and elasticity.[21]

Although we focus our formal comparison on the conventional kink-based bunching estimator, our estimator also differs from the methods normally used to estimate an elasticity around a notches. Kleven and Waseem (2013) (abbreviated KW) pioneered the application of bunching estimators to notches, and unlike most kink-based estimators, they do explicitly incorporate a specification of frictions into their model. They note that such frictions are necessary to explain the presence of mass in the dominated income range above the tax bracket threshold. In the KW model, a subset of agents are unresponsive to the presence of the notch, explaining the presence of mass in the dominated income range above the tax bracket threshold. To estimate the structural elasticity, the KW method therefore scales up the observed excess mass to estimate the bunching that would arise if all taxpayers were responsive. While this rescaling is appropriate if frictions take the form assumed in the KW—with a set of unresponsive agents—our method may be more appropriate if the frictions take the form of agents optimizing under sparsity-based constraints. Our procedure therefore complements KW by providing an alternative notch-based estimator based on an alternative model of frictions. Because the models predict somewhat different patterns of bunching around a notch, the choice between them can be informed by the data.[22]

[21] A second source of bias relates to the imposition of an integration constraint, which requires that the observed density to the right of the kink within the bunching window equal the density to the right of the kink in the counterfactual. This constraint can again distort the counterfactual density upward because it reallocates the bunching mass to be smoothed over the counterfactual density, rather than allowing some to have appeared from beyond the upper bound of the region of analysis.

[22] There are three differing predictions about the bunching mass which might be used to choose between these models of frictions. The KW model predicts (1) tight bunching at the bracket threshold among the subset of responsive tax payers, (2) upward-sloping density above the notch in the dominated income range, and (3) empirical density in the dominated range that is strictly below the estimated counterfactual density. (See KW Figure II.) In contrast, the sparsity-based frictions model predicts (1) leftward diffusion in bunching at the bracket threshold, (2) U-shaped density above the notch, and (3) empirical density in the dominated range that may be above the counterfactual density. (See our Figure A4(d).)

3 Empirical Applications

We apply our estimation method to four empirical bunching contexts. First, we consider the canonical example of bunching at a tax kink: the earned-income tax credit (EITC) threshold first studied by Saez (2010). We then consider three additional examples meeting two criteria: 1) the tax policy design (i.e. kink or notch, presence of compliance costs, etc.) is widely utilized around the world; and 2) our method is able to glean economically-important insights from moments of the data discarded by conventional approaches.

We first study the canonical case of diffuse bunching at a kink, originally studied in Saez (2010) and later in Mortenson and Whitten (2020). We then apply our method to the case of a known notch value, namely, the consumption coupon fiscal stimulus program studied in Ding et al. (2025). Our method extracts useful new insights here, because the estimated lumpiness parameter has a clear interpretation in the data, corresponding to the lumpiness in adjusting purchase sizes with discrete additions or deletions from the consumption basket. The estimated elasticity in this application then provides an estimate of the size of the consumption response to fiscal stimulus coupons. Third, we study a notch of unknown size, focusing on the frequently-studied case of a VAT revenue threshold for small-business VAT compliance; our method is useful in this setting because it allows us to exploit bunching behavior at the threshold to estimate the unobserved compliance cost of being subject to VAT. Finally, we study the case in which the statutory tax policy is a kink, but the observed pattern of bunching exhibits patterns more consistent with a notch. This could arise because taxpayers treat the kink as if it were a notch due to a misperception or an unobserved cost.[23] Our method is useful here because we can exploit the "notch-like" asymmetry in the bunching mass to estimate the size of the "as-if" notch consistent with observed taxpayer behavior.

3.1 Application 1: Diffuse Bunching at a Canonical Kink

Here we apply our bunching estimator to the canonical example of bunching at a kink: the EITC threshold in the United States explored in Saez (2010) and more recently in Mortenson and Whitten (2020). We focus on the first kink, where the EITC amount "plateaus" (i.e. stops increasing). At this point in the schedule, we see substantial diffuse bunching in the distribution, illustrated in the income histograms in Figure 8.[24]

We use two datasets: the first, comprising replication data from Mortenson and Whitten (2020), has finer granularity as it covers a larger number of years. The second, comprising

[23] To employ a term coined by Joel Slemrod, such misperceptions are sometimes called "botches".

[24] See Saez (1999) Table 1 for details on the kink-sizes and income brackets. Here we follow Saez (2010) in centering the data around the first kink to make all years comparable.

replication data from Saez (2010), allows us to observe the self-employed workers separately. In both datasets we focus on the sub-sample of workers with one child.

Figure 8(a) plots the histogram of incomes among all workers from Mortenson and Whitten (2020).[25] It also displays the best fit of the uniform sparsity model bunching estimator, which estimates an elasticity of 0.21, substantially higher than 0.12, the elasticity estimated using the conventional Chetty et al. (2011) method. Our method also produces a μ value (the average distance between earning opportunities) of \$500 relative to the kink-point of approximately \$8,600.

Figure 8b presents the original Saez (2010) bunching data at the first kink, along with our model's best fit. We estimate an elasticity of 0.54, again substantially higher than the original Saez (2010) estimate of 0.21, and also higher than the estimate of 0.28 using the Chetty et al. (2011) method with a polynomial approximation for the counterfactual density.[26] The estimated μ in this sample is \$1,700.

Figure 8(c) applies the uniform-sparsity estimator to the income histogram for self-employed workers from Saez (2010). Our estimated elasticity is 1.60, again higher than the original Saez (2010) elasticity of 1.10, and also higher than the Chetty et al. (2011) estimate of 1.11. Estimated μ is again \$1,700, with a tighter confidence interval than in Figure 8(b).

Together, these estimates show that the uniform sparsity model consistently estimates EITC elasticities that are substantially higher than conventional estimation approaches. These differences are consistent with the differences in elasticity estimates delivered by the simulation results that compare the two methods in Appendix C.

3.2 Application 2: Bunching at a Statutory Notch of Known Size

This application demonstrates our method's ability to estimate an elasticity in the presence of a notch of known size. Here we focus on a rather different setting: Ding et al. (2025) study a subsidy program in China that paid consumers bonus coupons for making purchases over a specified order size.[27] This policy structure, which was designed as a fiscal stimulus program, produces a notch in the mapping between costs and the value of goods purchased; the cost falls discontinuously when this value exceeds a specified threshold. Ding et al. (2025) focus their estimation on the marginal propensity to consume out of the subsidy amount, but we note that this setting is also well-suited to estimating other parameters of behavior, including the spending elasticity (the analog of the elasticity of taxable income in this setting) and

[25] We thank the authors for access to the data displayed in the histograms in their paper.

[26] The Saez (2010) elasticity results come from Table 2, panel (A) in that paper.

[27] We thank the authors for their openness to our use of the histograms data in their paper; no identifiable data is necessary for the production of these estimates.

the lumpiness parameter. The latter has a particularly clear interpretation in this setting, as consumers adjust their order size by adding or removing discrete items from their cart.

This setting also illustrates the generality of the estimation method in three ways. First, the running variable is consumer expenditures rather than earnings (our focus in the discussions thus far). Second, the beneficial side of the notch is to the right of the threshold, rather than to the left (more commonly the case with earnings tax kinks and notches). And third, because the policy was a temporary treatment, the data also contain an observable counterfactual distribution of purchases that obtains in the absence of the notch subsidy, so there is no need to estimate the counterfactual in this setting.

Figure 9 displays the histogram of purchase sizes among consumers treated with the notch subsidy. It also displays (with smaller markers) the counterfactual distribution of purchases observed in the absence of the subsidy. Consistent with the model of sparsity-based frictions, the bunching mass at the notch exhibits pronounced asymmetry, with diffuse mass on the desirable side of the notch. That diffusion, combined with the lumpiness inherent in consumers' choices over discrete menu items, suggest to us that this is a setting in which sparsity-based frictions are likely to be a more realistic model of underlying frictions than the Kleven and Waseem (2013) notch-bunching model of heterogeneous responsiveness.

The solid orange line in Figure 9 displays the best fit of the uniform sparsity bunching estimator. When we apply the method, we use the counterfactual density provided by the data, rather than an estimated polynomial. Like the observed density, the best-fit model-predicted density exhibits diffuse asymmetric bunching to the right of the notch. The notch value (dT) in this setting is known (18 yuan), and we impose it as a parameter of the tax function. We estimate a lumpiness parameter equalling $\mu = 22.8$, which represents the cost in yuan of the average item that is marginal to being added or dropped from a consumer's basket, and a spending elasticity of 0.48. Because these parameters are distinct from the marginal propensity to consume, they can be used to inform how consumers would respond to other types of subsidies, such as a flat subsidy on all purchases rather than a discontinuous notch.

Heterogeneity in lumpiness. Although the model-predicted (solid-line) density in Figure 9 captures key qualitative features of the observed income histogram—most notably the diffuse bunching to the right of the notch—we also note that it exhibits a smaller "peak" at the notch, and a slower rate of decay (moving rightward) from that peak, than is evident in the data. We suspect that this mismatch is driven by our imposition of a homogeneous lumpiness parameter μ. If that parameter is instead heterogeneous across consumers—for example, if some consumers are choosing between small marginal items in their order baskets while others are choosing between more expensive items—then the observed density would be

a mixture of densities, some with a narrow sharp peak above the notch, and others with a lower peak and more diffusion. The mixture would produce a taller peak with a faster rate of decay. To test this idea, we re-estimate the model allowing for heterogeneous lumpiness μ. For illustrative purposes, we constrain this heterogeneity to add just one degree of freedom, finding the best-fit combination of two μ values while retaining homogeneity in the spending elasticity. As shown by the dashed line in Figure 9, allowing for this heterogeneous lumpiness indeed produces a better fit to the high peak and fast rate of decay observed in the bunching mass.

3.3 Application 3: Bunching at Regulatory Notch of Unknown Size

Our third application makes use of the uniform sparsity model's ability to estimate a notch value of unknown size from observed bunching behavior. We focus on a widely-prevalent feature of tax policies, namely, value-added tax (VAT) registration thresholds which require that firms with revenues above a specified threshold must register for, collect, and remit VAT, imposing costs on them. Keen and Mintz (2004) note that these VAT compliance costs are an important determinant of optimal VAT thresholds, and Liu and Lockwood (2015), among others, show bunching at VAT registration thresholds—consistent with VAT registration imposing compliance costs on affected firms. By estimating the notch value at the VAT registration threshold, our method allows us to infer the money-metric size of compliance costs distinct from the change in statutory tax incentives at that threshold.

We study the VAT registration threshold in a novel setting, South Africa, using extensive administrative tax data spanning 2018-2022. Firms are required to register for the Value-Added Tax (VAT) with the South African Revenue Service (SARS) if their total taxable turnover exceeds R1 million in any consecutive 12-month period. This registration is mandatory, and once registered, the business must charge VAT—currently at a standard rate of 15%—on most goods and services it supplies. Firms with taxable turnover below this threshold may also register voluntarily if their turnover exceeds ZAR 50,000 over the past 12 months. Registered businesses can claim input tax credits on VAT paid for business-related purchases, thereby offsetting their VAT liability. Registration is done online through the SARS eFiling system or at a SARS branch.

To approximate the change in the tax level associated with crossing the VAT registration threshold—and therefore distinguish that change from compliance costs—we compute the marginal income tax rate for firms below and above the notch. We approximate the marginal income tax rate by taking the slope of a smoothed binscatter of average tax liability across revenues for forms below and above the notch. We compute this marginal tax rate below the notch only for the subset of firms that do not *voluntarily* register for VAT, in order to identify

the change in tax level that arises due to the change in VAT registration status.[28] Voluntary registration is allowed in South Africa, and we find that 49% of firms below the notch choose to register for VAT voluntarily.

Figure 10 shows the distribution of firms in 10,000 ZAR bins of taxable turnover around the VAT registration thresholds. There is a clear pattern of bunching in firm turnover to the left of the VAT registration threshold. Our model is able to fit the raw bunching data well, including the diffuse bunching to the left and the positive density to the right of the notch—both patterns that are apparent in previous VAT registration applications. We estimate an elasticity of taxable turnover to VAT tax rates of $e = 0.54$. Our estimate of the lumpiness parameter μ equals R69900, representing to the average distance between turnover opportunities being 8.6% of taxable turnover (for firms with taxable turnover of 1 million rand). Finally, we estimate an additional discrete notch value beyond the change in tax rates of dT= R6470 in taxable turnover. To map this back to taxable income, we use the fact that between 2018-2022, the average taxable profit margin for all firms in South Africa was 6.18%, suggesting a relatively small compliance cost of dT= R399.846. This estimate is potentially helpful for evaluating optimal VAT turnover thresholds.

3.4 Application 4: Notch-like Bunching Around a Kink

For our final application, we consider a variation on standard bunching-at-a-kink behavior, in which the observed pattern of bunching around a statutory kink resembles observed behavior around a tax notch. This application is motivated by our observation, also in South African administrative data, that (both personal and small business) taxpayers appear to exhibit strong notch-like patterns of asymmetric bunching around tax thresholds that are statutory kinks.[29]

Figure 11 exhibits this pattern of asymmetric bunching for two distinct populations. In panel (a), we plot the histogram of taxable incomes among small businesses around their first corporate income tax kink, at which the marginal tax rate rises from 0 to 7 percent.[30] The

[28]This issue is studied extensively in Liu et al. (2021), who find that approximately half of British firms just below a VAT notch voluntarily register. They find that firms are more likely to voluntarily register for the VAT when they benefit more from input credits than they are hurt by the VAT tax on output (i.e. firms that do more business-to-business sales relative to business-to-consumer sales, and when they face lower product market competition, because more of the VAT tax can be passed on to their customers.

[29]Like many developing countries, South Africa relies more heavily on corporate income taxes than most developed economies. In 2017, corporate taxes accounted for 16.2 percent of total tax revenue in South Africa, considerably higher than the OECD average of 9.7 percent, but in line with the average shares for Africa (18.5 percent) and Latin America (15.4 percent). Data from the OECD accessed at: `https://stats.oecd.org/Index.aspx?DataSetCode=CTS_REV`. The South African corporate tax system is tiered, with a progressive, kinked tax schedule for "Small Business Corporations" (SBCs), and a flat 28 percent tax applying to other (larger) resident corporations. Businesses can optionally register as an SBC if they meet a set of requirements, the most pertinent being that their annual revenue must be below R20 million (about $1.4 million US).

[30]SBCs account for 38 percent of all formally registered companies, although due to their smaller size, they

bunching pattern exhibits pronounced asymmetric diffuse bunching to the left of the threshold, with no excess mass—and even a small apparent depression—to the right of the threshold. Panel (b) displays a similar pattern for self-employed individuals around the first kink in the personal income tax schedule, where the income tax rate rises from 0% to 18%.[31] Although the asymmetry remains pronounced in panel (b), the density to the right of the threshold does not exhibit a depression, but rather a muted excess mass relative to smooth interpolation from the far left to the far right. The sparsity-based frictions model can fit either of these patterns around a notch, depending on the combination of the elasticity and lumpiness parameters, as illustrated by Appendix Figure A4(d).

Applying the sparsity-based frictions estimator to the density in Figure 11(a), we estimate a notch value of R340, or about $24 US. This estimate is statistically significant, indicating that the model strongly rejects pure kink behavior. That said, the as-if notch value is not large, consistent with a modest perceived cost of earning above the threshold. Figure 11(a) focuses on the lowest of three prominent tax kinks in the small business income tax schedule. Appendix Figure A10 shows bunching patterns in the taxable incomes of small businesses around the middle kink (at 360K rand) and the upper kink (at 550K rand). For reasons explored below, we partition these populations by whether they use a tax practitioner. Although much more muted than at the lower kink, asymmetry remains present at these higher tax kinks as well, with estimated notch values significantly above zero in all cases except Figure A10(d), suggesting that notch-like behavior at a kink is not only observed in the very smallest businesses.

A potential mechanism underlying notch-like bunching around a tax kink is taxpayer confusion about average vs. marginal tax rates. A large literature suggests agents may confuse marginal versus average changes in incentives (Ito (2014), Rees-Jones and Taubinsky (2020)). If taxpayers earning just below the threshold mistakenly believe that a small increase in earnings would cause their average (rather than their marginal) tax rate to rise, that would produce a notch-like pattern of bunching of the sort observed in Figure 11.

Heterogeneity in "Botching"

Here we explore differences in notch-like bunching at a kink across a dimension of taxpayer heterogeneity that is observable in our tax data and may be related to taxpayers' ability to

account for less than 20 percent of total corporate tax revenues. For the purposes of this paper, we focus on the population of SBCs from 2014 to 2018 because they face a piecewise-linear kinked tax schedule ideal for bunching estimation.

[31] The lowest kink point in the schedule is a result of the standard deduction—all South Africans receive a tax deduction equal to the amount of taxable income at the lowest threshold (R70,000 or $4,860 per annum in 2017). This means that any income below this threshold is taxed at a 0% marginal tax rate, whereafter the marginal tax rate jumps to 18%, in a similar fashion to that of the first SBC kink. We find little evidence of bunching at kinks for wage earning South African workers.

understand and take advantage of tax incentives, namely, the use of a hired tax preparer.

Figures 11(c) and 11(d) partition small businesses by their tax practitioner usage. At the lowest kink, there is a clear hole in the distribution for firms that do not use a tax practitioner (panel (c))—consistent with firms interpreting the lowest kink as a notch; this hole is not clearly evident at the lowest kink for firms that use a tax practitioner (panel (d)). Our method allows us to quantify the fact that firms using tax practitioners are less prone to "botching" (i.e., they have less "notch-like" behavior around a kink); at the lowest kink firms without tax practitioners act as if the average tax liability increases by approximately 44,000 rand at the first kink, while firms with tax practitioners act as if the average tax liability increases by 30,000 rand (a 32% reduction).

Figure A10 shows plots like those in Figure 11(c) and (d) but for the middle and upper tax kinks. Panels partition businesses on their tax practitioner usage, and each reports estimated elasticities. The raw histograms exhibit more pronounced bunching among firms that use tax practitioners. We do not observe a consistent pattern in the relative income elasticity between businesses that do and do not use tax practitioners: although the elasticity is higher among firms that use tax practitioners at the lowest kink, the reverse is true at the middle kink. At the upper kink, the elasticities in each group are statistically indistinguishable. Yet, there is a clear and consistent difference between the lumpiness parameters of the different groups of firms: adjustment frictions appear to be smaller for firms who use tax practitioners at every kink. This is consistent with such firms fine-tuning their incomes more precisely in response to tax incentives, or paying closer attention to the set of possible actions—real economic activity or reporting behavior—which can be used to target incomes more precisely to a desired level.

4 Conclusion

This paper develops a new bunching-based elasticity estimator using a novel approach to modelling income adjustment frictions. We adopt a general model of frictions in which taxpayers choose between discrete income opportunities, capturing a range of microfoundations including directed search, limited attention, and lumpy adjustment. We show that many different income adjustment processes of this type are well approximated by a parsimonious limiting case, in which opportunities are drawn from a Poisson process governed by a single "lumpiness" parameter. This new parameter quantifies the expected distance between adjacent opportunities. The model predicts key patterns observed in empirical bunching settings, such as diffuse bunching around kinks and positive mass above notches.

We demonstrate the versatility and applicability of this new approach by using it to study

a set of four canonical and novel settings. The new method—consistent with simulations that we conduct—uncovers estimated elasticities in the classic US EITC kink that are nearly twice as large as in previous work. The method is also well able to capture the bunching patterns observed in a recently studied fiscal stimulus coupon programme in China featuring a pure notch of known size, additionally extracting an ecoonomically-meaningful parameter, namely, the average distance between consumption opportunities.

In two other applications on recently-harnessed administrative data from South Africa, the method yields useful new economic insights. When we apply the method to observed bunching around VAT registration thresholds, we provide a new money-metric estimate of the compliance and hassle factors of VAT registration for small businesses. And when we apply this method to administrative tax data on small firms and individuals around kinks in both the corporate and personal income tax schedule in South Africa, we find evidence that both firms and individuals treat the lowest tax kink like a notch. We estimate substantially lower income adjustment frictions among firms with paid tax preparers, consistent with finer income targeting among that group. By quantifying the extent of adjustment frictions, our approach allows us to recover the "as-if" notch value when taxpayers treat a kink like a notch.

Our proposed bunching estimator can be applied to estimate behavioral responses in other settings with kinked budget sets, including with non-income tax instruments, nonlinear pricing schedules, or non-monetary payoffs. More generally, the model of uniform sparsity, as an approximation of sparsity-based frictions, can be extended to a wide range of settings, including multidimensional choices.

References

Abel, Andrew B, Janice C Eberly and Stavros Panageas. 2013. "Optimal inattention to the stock market with information costs and transactions costs." *Econometrica* 81(4):1455–1481.

Allen, Eric J, Patricia M Dechow, Devin G Pope and George Wu. 2017. "Reference-Dependent Preferences: Evidence from Marathon Runners." *Management Science* 63(6):1657–1672.

Alvarez, Fernando E, Francesco Lippi and Luigi Paciello. 2011. "Optimal price setting with observation and menu costs." *The Quarterly Journal of Economics* 126(4):1909–1960.

Alvarez, Fernando, Luigi Guiso and Francesco Lippi. 2012. "Durable consumption and asset management with transaction and observation costs." *American Economic Review* 102(5):2272–2300.

Andersen, Steffen, Cristian Badarinza, Lu Liu, Julie Marx and Tarun Ramadorai. 2022. "Reference Dependence in the Housing Market." *American Economic Review* 173(10):3398–3440.

Andersen, Steffen, John Y Campbell, Kasper Meisner Nielsen and Tarun Ramadorai. 2020. "Sources of inaction in household finance: Evidence from the danish mortgage market." *American Economic Review* 110(10):3184–3230.

Andrews, Isaiah, Matthew Gentzkow and Jesse M Shapiro. 2020. "Transparency in structural research." *Journal of Business & Economic Statistics* 38(4):711–722.

Bachas, Pierre and Mauricio Soto. 2021. "Corporate Taxation under Weak Enforcement." *American Economic Journal: Economic Policy* 13(4):36–71.

Bertanha, Marinho, Andrew H McCallum and Nathan Seegert. 2023. "Better bunching, nicer notching." *Journal of Econometrics* 237(2):105512.

Best, Michael Carlos, Anne Brockmeyer, Henrik Jacobsen Kleven, Johannes Spinnewijn and Mazhar Waseem. 2015. "Production versus Revenue Efficiency with Limited Tax Capacity: Theory and Evidence from Pakistan." *Journal of Political Economy* 123(6):1311–1355.

Best, Michael Carlos and Henrik Jacobsen Kleven. 2018. "Housing market responses to transaction taxes: Evidence from notches and stimulus in the UK." *The Review of Economic Studies* 85(1):157–193.

Blomquist, Sören, Whitney K. Newey, Anil Kumar and Che-Yuan Liang. 2021. "On Bunching and Identification of the Taxable Income Elasticity." *Journal of Political Economy* 129(8):2320–2343.

Boonzaaier, Wian, Jarkko Harju, Tuomas Matikka and Jukka Pirttilä. 2019. "How Do Small Firms Respond to Tax Schedule Discontinuities? Evidence from South African Tax Registers." *International Tax and Public Finance* 26(5):1104–1136.

Bosch, Nicole, Vincent Dekker and Kristina Strohmaier. 2020. "A Data-Driven Procedure to Determine the Bunching Window: An Application to the Netherlands." *International Tax and Public Finance* 27(4):951–979.

Brehm, Margaret, Scott A Imberman and Michael F Lovenheim. 2017. "Achievement Effects of Individual Performance Incentives in a Teacher Merit Pay Tournament." *Labour Economics* 44:133–150.

Chetty, Raj. 2012. "Bounds on Elasticities with Optimization Frictions: A Synthesis of Micro and Macro Evidence on Labor Supply." *Econometrica* 80(3):969–1018.

Chetty, Raj, John N. Friedman, Tore Olsen and Luigi Pistaferri. 2011. "Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records." *The Quarterly Journal of Economics* 126(2):749–804.

Dee, Thomas S, Will Dobbie, Brian A Jacob and Jonah Rockoff. 2019. "The Causes and Consequences of Test Score Manipulation: Evidence from the New York Regents Examinations." *American Economic Journal: Applied Economics* 11(3):382–423.

Devereux, Michael P., Li Liu and Simon Loretz. 2014. "The Elasticity of Corporate Taxable Income: New Evidence from UK Tax Records." *American Economic Journal: Economic Policy* 6(2):19–53.

Diamond, Rebecca and Petra Persson. 2016. "The Long-Term Consequences of Teacher Discretion in Grading of High-Stakes Tests." *Working Paper no. 22207, National Bureau of Economic Research* .

Ding, Jing, Lei Jiang, Lucy Msall and Matthew J. Notowidigdo. 2025. "Consumer-Financed Fiscal Stimulus: Evidence from Digital Coupons in China." *American Economic Review: Insights* 7(3):411–27.

Einav, Liran, Amy Finkelstein and Paul Schrimpf. 2017. "Bunching at the kink: implications for spending responses to health insurance contracts." *Journal of Public Economics* 146:27–40.

Gabaix, Xavier. 2014. "A Sparsity-Based Model of Bounded Rationality." *The Quarterly Journal of Economics* 129(4):1661–1710.

Gelber, Alexander M., Damon Jones and Daniel W. Sacks. 2020. "Estimating Adjustment Frictions Using Nonlinear Budget Sets: Method and Evidence from the Earnings Test." *American Economic Journal: Applied Economics* 12(1):1–31.

Gordon, Roger and Wei Li. 2009. "Tax Structures in Developing Countries: Many Puzzles and a Possible Explanation." *Journal of Public Economics* 93(7-8):855–866.

Grubb, Michael D and Matthew Osborne. 2015. "Cellular Service Demand: Biased Beliefs, Learning, and Bill Shock." *American Economic Review* 105(1):234–271.

He, Daixin, Langchuan Peng and Xiaxin Wang. 2021. "Understanding the elasticity of taxable income: A tale of two approaches." *Journal of Public Economics* 197:104375.

Ito, Koichiro. 2014. "Do Consumers Respond to Marginal or Average Price? Evidence from Nonlinear Electricity Pricing." *American Economic Review* 104(2):537–563.

Jung, Junehyuk, Jeong Ho Kim, Filip Matějka and Christopher A Sims. 2019. "Discrete actions in information-constrained decision problems." *The Review of Economic Studies* 86(6):2643–2667.

Keen, Michael and Jack Mintz. 2004. "The optimal threshold for a value-added tax." *Journal of Public Economics* 88(3-4):559–576.

Kemp, Johannes Hermanus. 2019. "The elasticity of taxable income: The case of South Africa." *South African Journal of Economics* 87(4):417–449.

Kleven, Henrik J. and Mazhar Waseem. 2013. "Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan." *The Quarterly Journal of Economics* 128(2):669–723.

Kleven, Henrik Jacobsen. 2016. "Bunching." *Annual Review of Economics* 8:435–464.

Kopczuk, Wojciech and David Munroe. 2015. "Mansion tax: The effect of transfer taxes on the residential real estate market." *American economic Journal: economic policy* 7(2):214–257.

Kostøl, Andreas R. and Andreas S. Myhre. 2021. "Labor Supply Responses to Learning the Tax and Benefit Schedule." *American Economic Review* 111(11):3733–3766.

Kothari, S.P., Andrew J. Leone and Charles E. Wasley. 2005. "Performance Matched Discretionary Accrual Measures." *Journal of Accounting and Economics* 39(1):163–197.

Lediga, Collen, Nadine Riedel and Kristina Strohmaier. 2019. "The elasticity of corporate taxable income—Evidence from South Africa." *Economics Letters* 175:43–46.

Liu, Li and Ben Lockwood. 2015. VAT notches. In *Proceedings. Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association*. Vol. 108 JSTOR pp. 1–51.

Liu, Li, Ben Lockwood, Miguel Almunia and Eddy HF Tam. 2021. "VAT notches, voluntary registration, and bunching: Theory and UK evidence." *Review of Economics and Statistics* 103(1):151–164.

Manoli, Day and Andrea Weber. 2016. "Nonparametric Evidence on the Effects of Financial Incentives on Retirement Decisions." *American Economic Journal: Economic Policy* 8(4):160–182.

Matzkin, Rosa L. 2013. "Nonparametric identification in structural economic models." *Annu. Rev. Econ.* 5(1):457–486.

Mavrokonstantis, Panos and Arthur Seibold. 2022. "Bunching and Adjustment Costs: Evidence from Cypriot Tax Reforms." *Journal of Public Economics* 214:104727.

Moore, Dylan. 2022. "Evaluating Tax Reforms without Elasticities: What Bunching Can Identify." *Working paper* .

Mortenson, Jacob A. and Andrew Whitten. 2016. "How Sensitive Are Taxpayers to Marginal Tax Rates? Evidence from Income Bunching in the United States." *Working Paper* .

Mortenson, Jacob A. and Andrew Whitten. 2020. "Bunching to Maximize Tax Credits: Evidence from Kinks in the US Tax Schedule." *American Economic Journal: Economic Policy* 12(3):402–432.

Pillay, Neryvia. 2021. "Taxpayer responsiveness to taxation: Evidence from bunching at kink points of the South African income tax schedule." *Working Paper* .

Rees-Jones, Alex. 2018. "Quantifying Loss-Averse Tax Manipulation." *The Review of Economic Studies* 85(2):1251–1278.

Rees-Jones, Alex and Dmitry Taubinsky. 2020. "Measuring "Schmeduling"." *The Review of Economic Studies* 87(5):2399–2438.

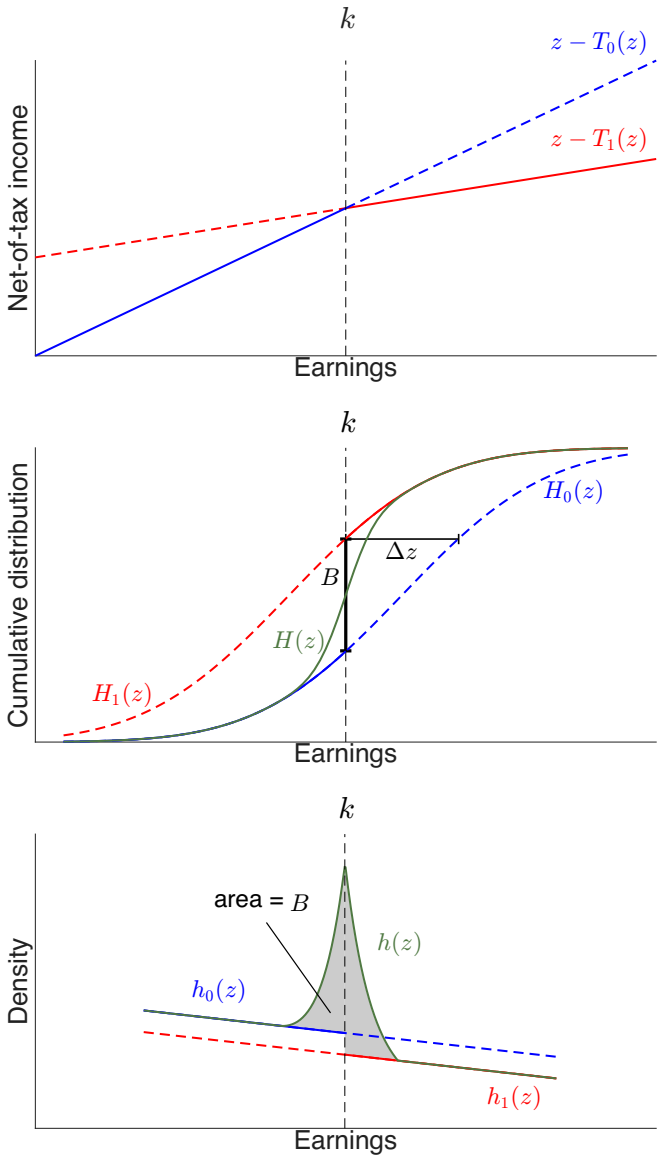
Saez, Emmanuel. 1999. "Do Taxpayers Bunch at Kink Points?" *Working Paper no. 7366, National Bureau of Economic Research* .

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

Sims, Christopher A. 2003. "Implications of Rational Inattention." *Journal of Monetary Economics* 50(3):665–690.

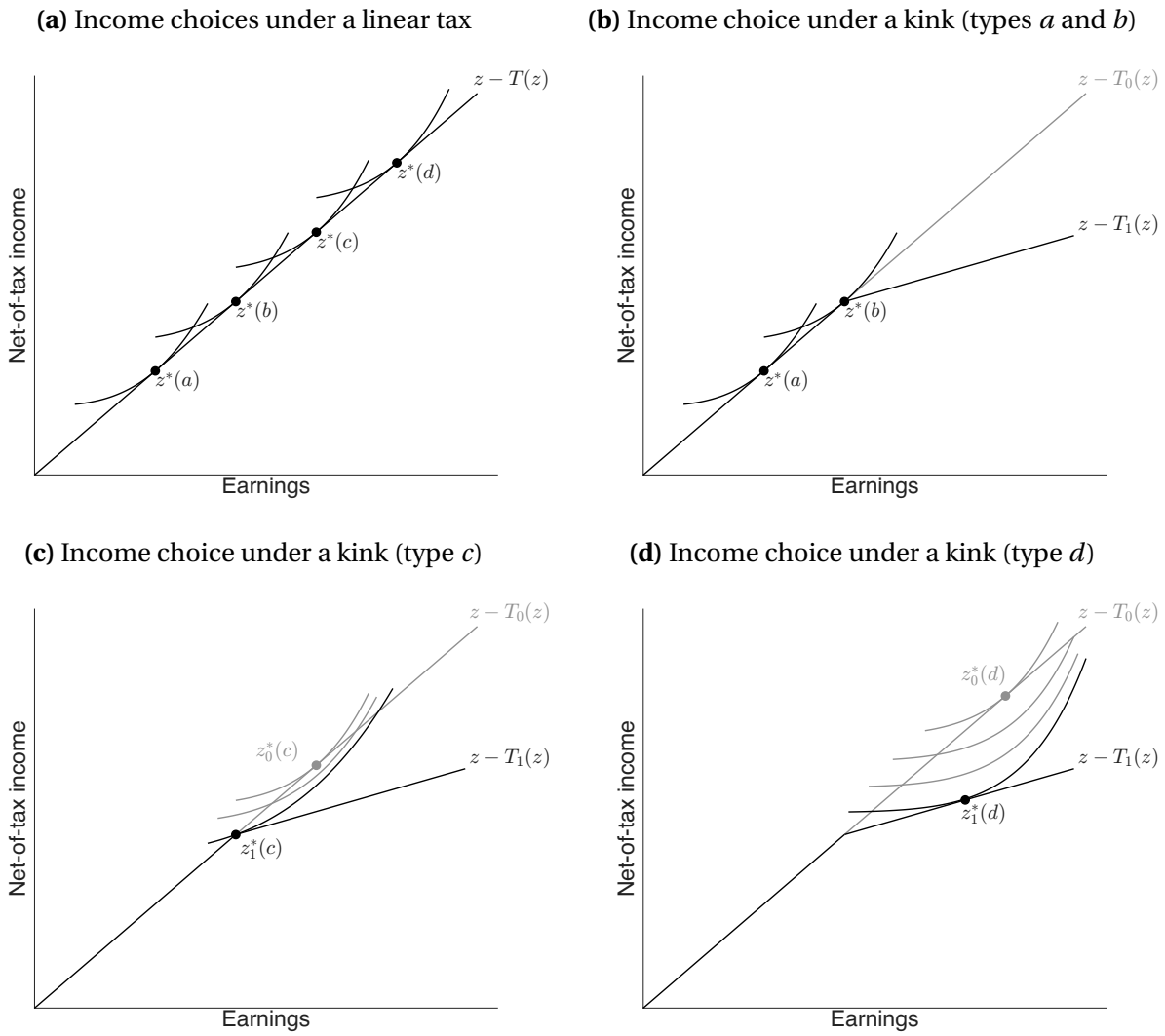
Søgaard, Jakob Egholt. 2019. "Labor Supply and Optimization Frictions: Evidence from the Danish Student Labor Market." *Journal of Public Economics* 173:125–138.

Figure 1: Income Distributions and Densities Under a Kinked Tax Schedule



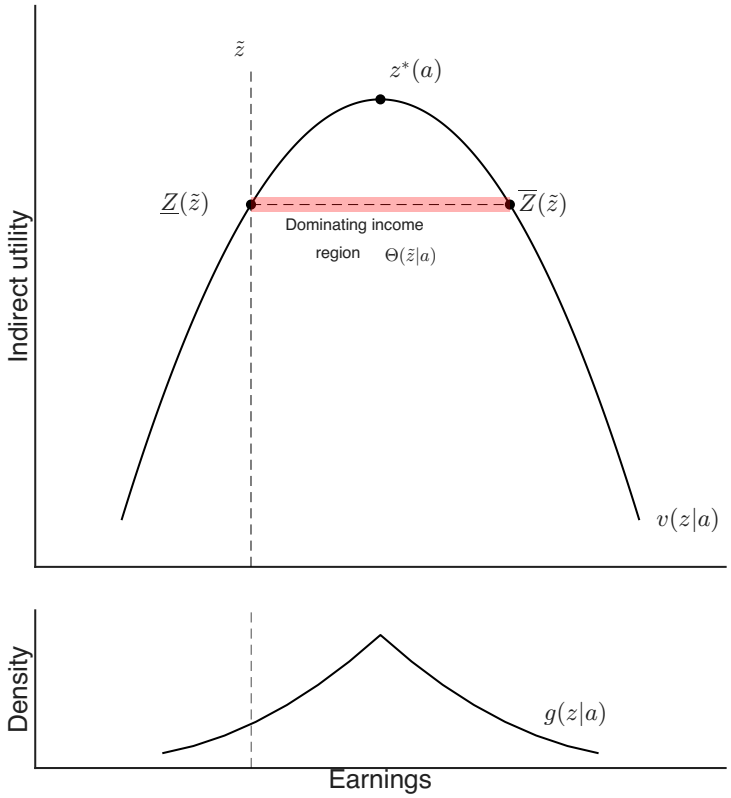
The top panel shows a taxpayer's budget constraints under two linear income tax functions, $T_0(z)$ and $T_1(z)$. These can construct a "kinked" tax schedule (with resulting budget constraint plotted by the solid line) consisting of $T_0(z)$ and $T_1(z)$ below and above income k, respectively. The middle panel plots the income CDFs $H_0(z)$ and $H_1(z)$ that would arise under each linear tax function. Absent frictions, the CDF under the kinked tax schedule coincides with $H_0(z)$ below k and with $H_1(z)$ above k, with a discontinuous jump B at the threshold k. With frictions, the transition from $H_0(z)$ to $H_1(z)$ around the kink at k is gradual, as plotted by the CDF labeled $H(z)$. The bottom panel plots the income densities $h_0(z)$, $h_1(z)$, and $h(z)$, corresponding to the CDFs $H_0(z)$, $H_1(z)$, and $H(z)$, respectively.

Figure 2: Frictionless bunching model with a progressive tax kink



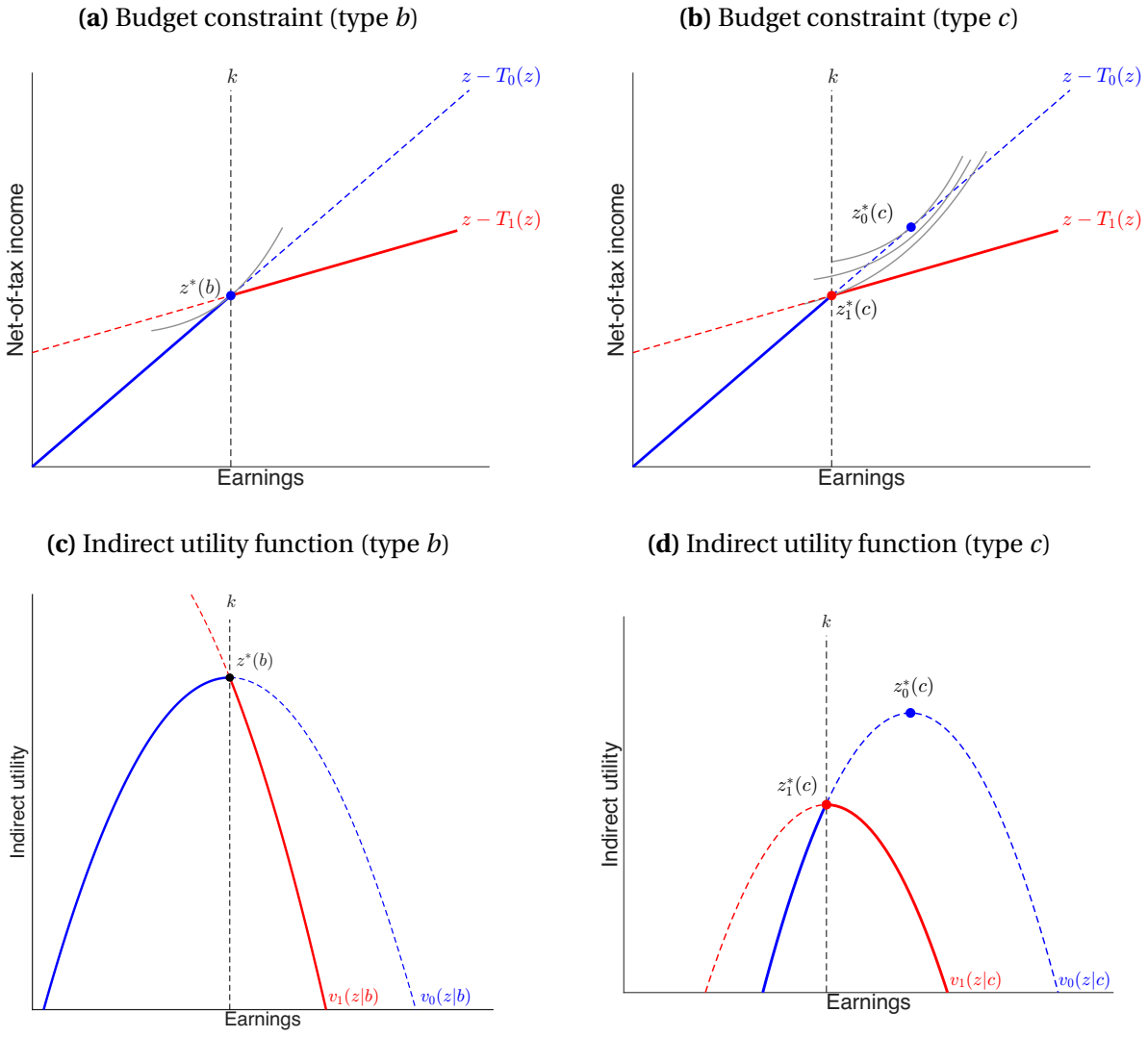
This figure illustrates income choices around a tax kink under the conventional frictionless model. Panel (a) illustrates the optimal choice of income, z^*, for four selected types of taxpayers under a linear income tax. Panels (b), (c), and (d) illustrate the optimal choice for each type under a kinked income tax, where the tax changes from the $T_0(z)$ to $T_1(z)$ at the threshold k. Incomes z_0^* and z_1^* denote the optimal choice under the linear tax $T_0(z)$ and $T_1(z)$, respectively.

Figure 3: Type-conditional income density under a linear tax (type a)



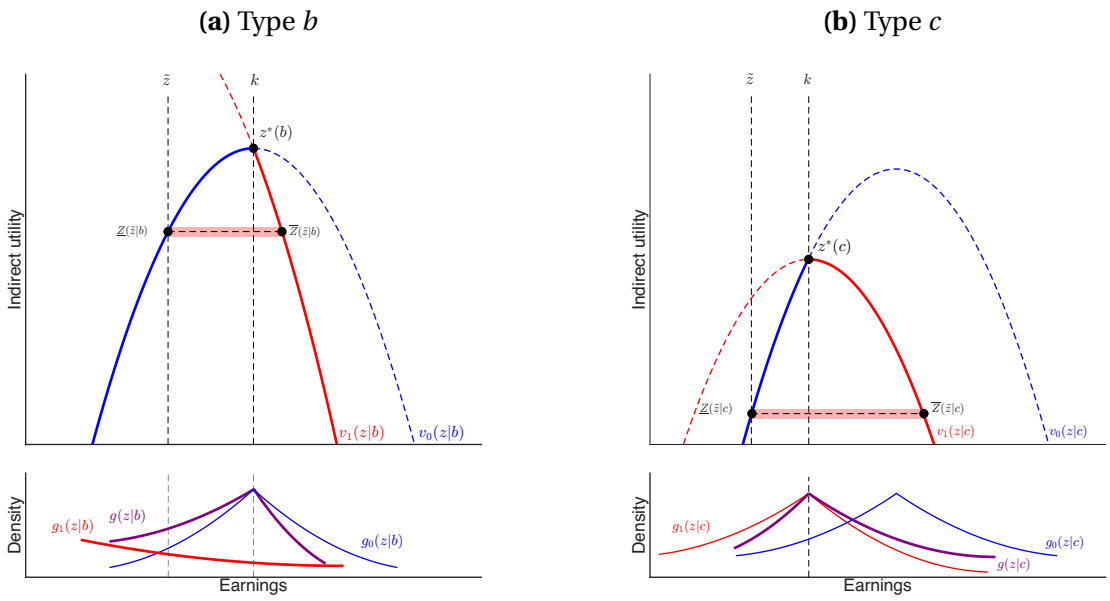
This figure illustrates the calculation of the type-conditional income density in the uniform sparsity model among a-type agents at a particular income level $\tilde{z}$, under a locally linear income tax. The top plot displays the taxpayer's indirect utility function over incomes. An agent who has $\tilde{z}$ in their income opportunity set will select this income iff they do not have some other income opportunity in the shaded "dominating region." The type-conditional density $g(\tilde{z}|a)$ is equal to this conditional probability multiplied by the probability of drawing $\tilde{z}$.

Figure 4: Utility from income choices around a tax kink



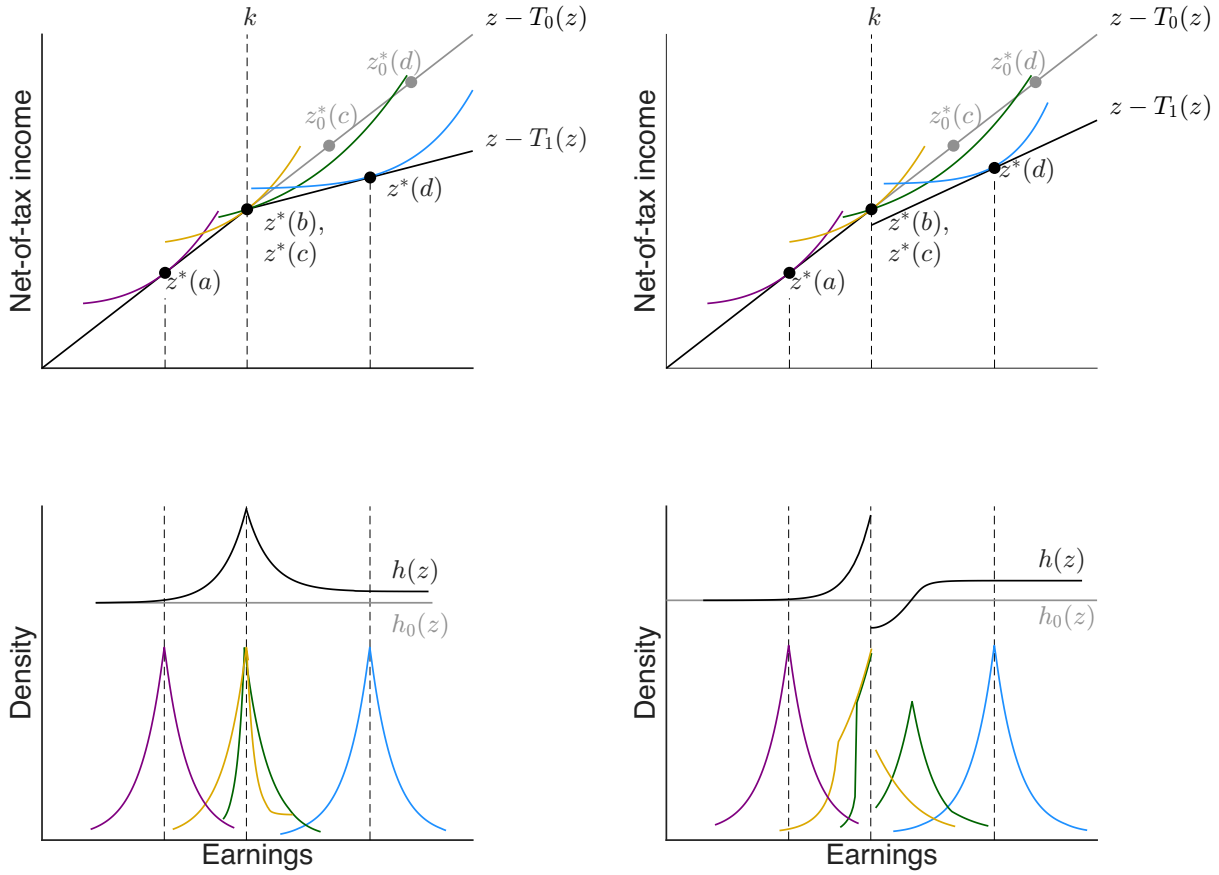
Panels (a) and (c) illustrate the construction of the indirect utility function around a progressive tax kink for the marginal non-buncher (type b). Panel (a) shows the taxpayer's budget constraint, plotted as a solid line, where $T_0(z)$ and $T_1(z)$ are the linear income taxes below and above the bracket threshold k, respectively. Panel (c) plots the indirect utility functions $v_0(z|b)$ and $v_1(z|b)$, which would be obtained if the linear tax functions $T_0(z)$ or $T_1(z)$ applied across all incomes. Type b's indirect utility function under the kinked tax schedule, plotted as a solid line, is given by $v_0(z|b)$ below k and $v_1(z|b)$ above k. Panels (b) and (d) show analogous illustrations for the marginal buncher (type c). This taxpayer's optimal frictionless income choices under the linear taxes $T_0(z)$ and $T_1(z)$ are denoted $z_0^*(c)$ and $z_1^*(c)$.

Figure 5: Type-conditional income density around a kink



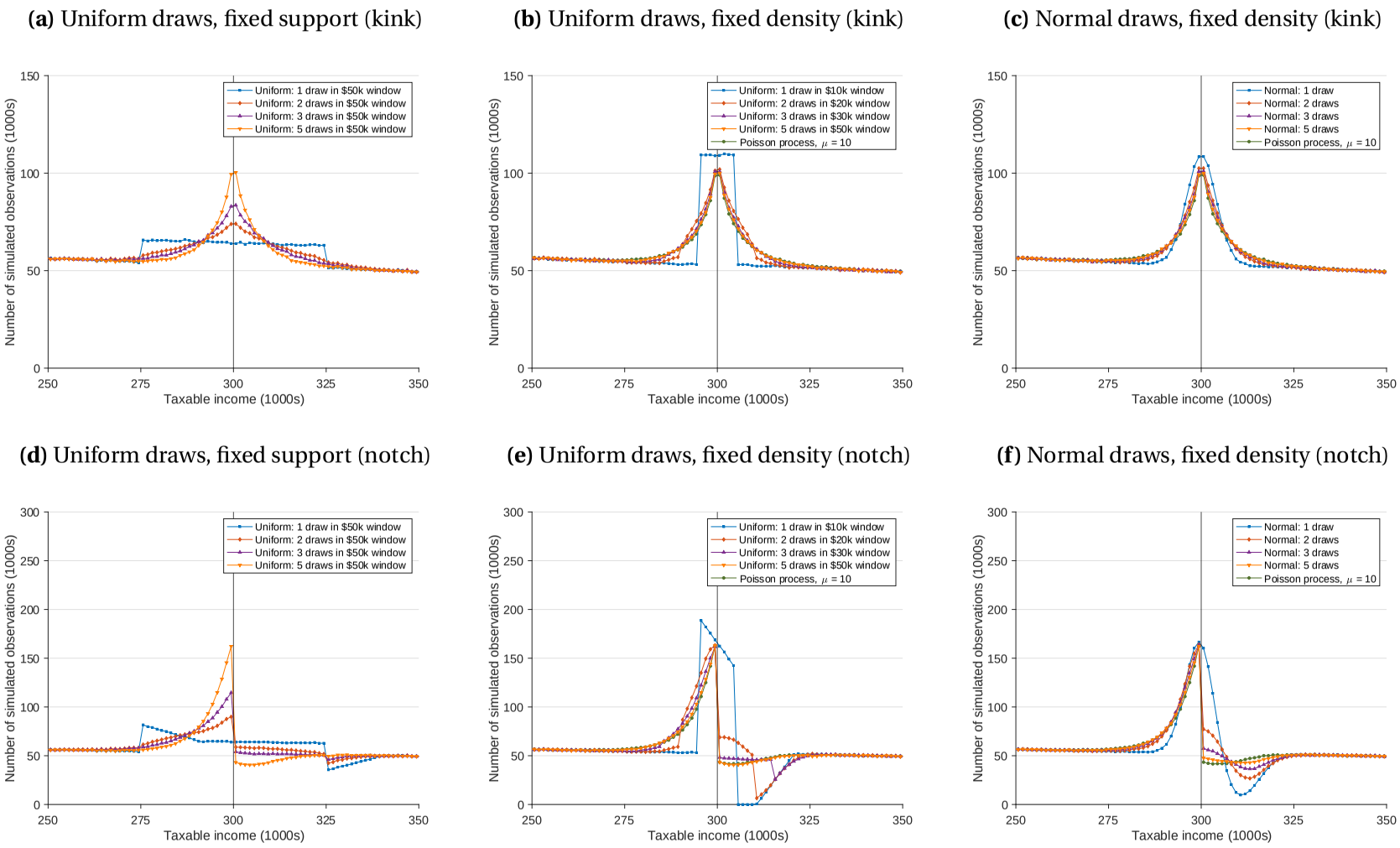
This figure illustrates how the indirect utility functions from Figure 4 are used to compute the type-conditional income densities. The panels show the calculation for the marginal non-buncher (panel (a)) and the marginal buncher (panel (b)). Each panel illustrates the calculation of the type-conditional income density $g(z|n)$ at a (different) income $\tilde{z}$. We first identify the range of incomes that utility-dominate $\tilde{z}$ for each taxpayer, corresponding to the horizontal dashed line, and we proceed as in Figure 3. The type-conditional densities are plotted in purple. For reference, the type-conditional density under the counterfactual linear taxes $T_0(z)$ and $T_1(z)$ are plotted in blue and in red, respectively.

Figure 6: Aggregating type-conditional densities into an observable income density



The top left panel of this figure shows the optimal frictionless income choice for agents of types a, b, c, and d in the presence of a kink at k, with each type's maximal indifference curve plotted in a different color. The panel below it illustrates the type-conditional income densities in corresponding colors, plotted on the same horizontal axis. Summing across the type-conditional densities of these and the continuum of intervening types produces the observed income density, $h(z)$, which exhibits diffuse bunching around the bracket threshold. The counterfactual income density $h_0(z)$, which would be observed under the linear tax function $T_0(z)$, is plotted in gray for reference. The top right panel shows the analogous construction for a tax notch, with the resulting observed income density plotted below.

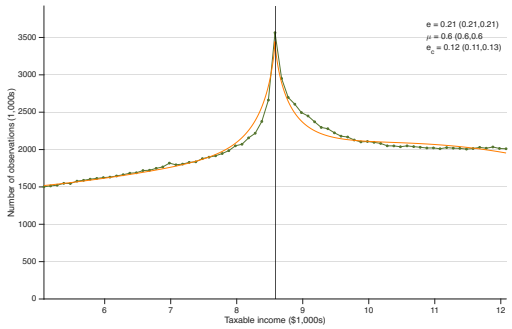
Figure 7: Simulated bunching patterns with sparsity-based frictions



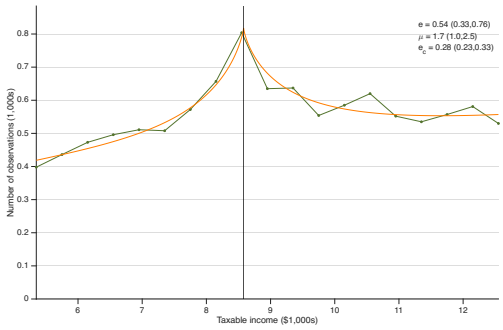
This figure plots simulated income densities around a bracket threshold under a model of frictions in which each taxpayer faces a sparse set of income opportunities drawn from around their preferred frictionless ("target") income. In all simulations, the marginal tax rate rises from $t_0 = 0.1$ to $t_1 = 0.2$ at the bracket threshold of $300,000. In panels (d)–(f), the tax *level* also increases by $1000 at the threshold, creating a notch. In panels (a) and (d), each taxpayer chooses from M income opportunities drawn from a uniform distribution of width $50,000 around their target income, for $M = 1$, 2, 3, and 5. In panels (b) and (e), each taxpayer chooses from M income opportunities drawn from a uniform distribution whose width is adjusted to hold fixed the density of opportunities around the target income. These panels also plot the "uniform sparsity model"—the limiting case as $M \to \infty$—in which income opportunities are a Poisson process with the same density of income opportunity draws at the target income. In panels (c) and (f), income opportunities are drawn from a *normal* distribution centered at taxpayers' target incomes, with variance adjusted so that the density of opportunities is the same as in panels (b) and (e).

Figure 8: Bunching at the EITC kink

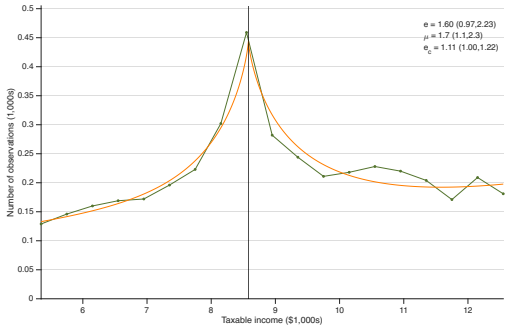
(a) Mortenson-Whitten All Taxpayers



(b) Saez (2010) All Taxpayers

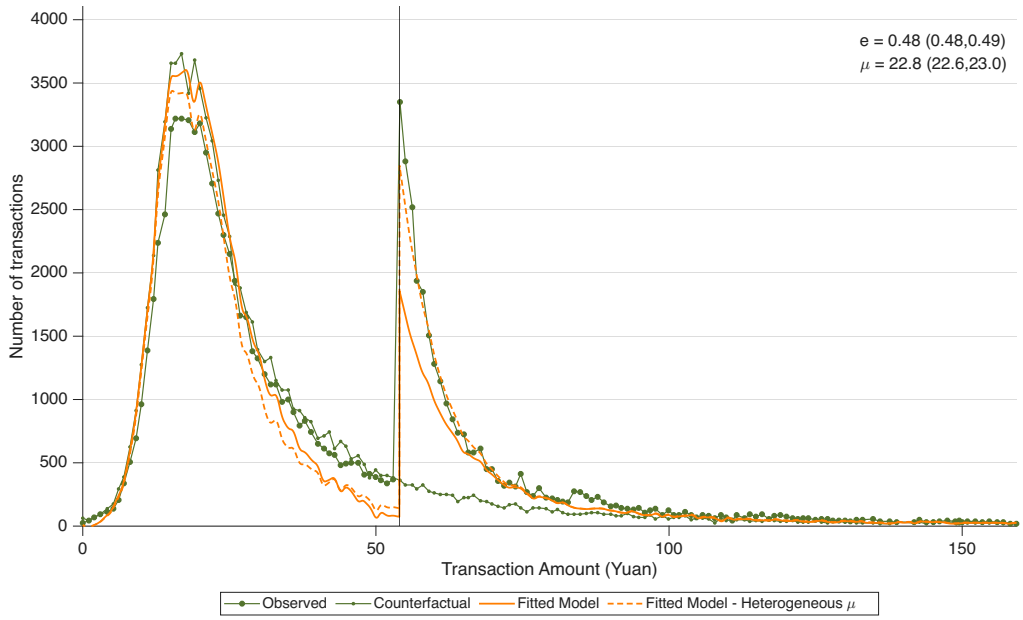


(c) Saez (2010) Self-Employed



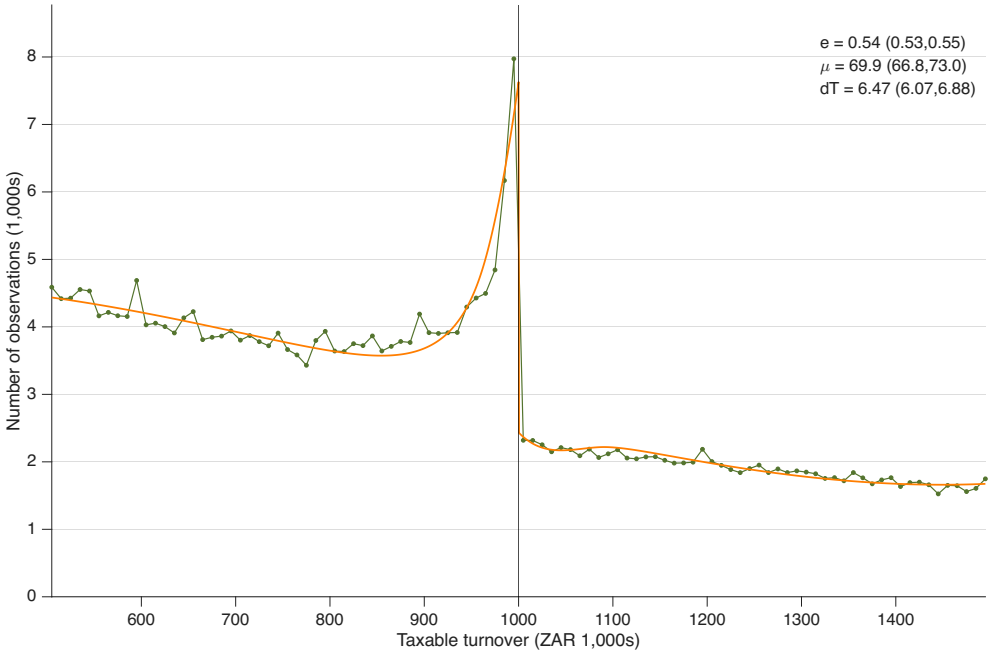
Green points plot the income histograms of single parents with one child in the United States around the first EITC kink. Panel (a) uses the microdata from Mortenson and Whitten (2016). Panel (b) uses the replication data from all taxpayers in Saez (2010). Panel (c) uses the self-employed sample from Saez (2010). Orange lines plot the best-fit income density under our uniform sparsity model. Each panel reports our parameter estimates e (elasticity of taxable income) and μ (average distance between income opportunities in $ 1000s), as well as the elasticity e_c estimated using the conventional method. Numbers in parentheses indicate the 95 percent confidence interval on parameter estimates generated by the MLE method.

Figure 9: Bunching at a Known Notch Value: Threshold Coupons



This figure plots bunching in the distribution of online order amounts around the 54 yuan coupon threshold studied in Ding et al. (2025). The coupon gives 18 yuan off a purchase of 54. Series are labeled as follows. "Observed": the observed distribution of order amounts when consumers had access to the coupons (period t). "Counterfactual": the distribution of order amounts among the same sample of consumers during the pre-period before coupon access. "Fitted model": the best-fit model-generated distribution from our uniform sparsity model. "Fitted Model - Heterogeneous μ": the best-fit model-generated distribution from the uniform-sparsity model when we allow for two values of the μ parameter.

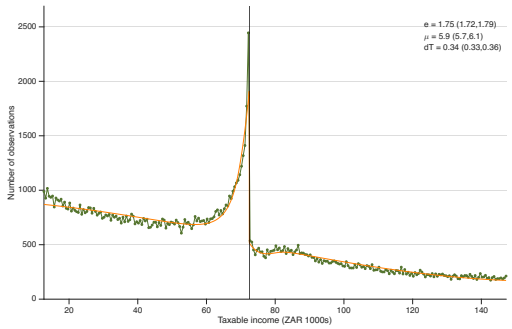
Figure 10: Bunching at a VAT registration threshold



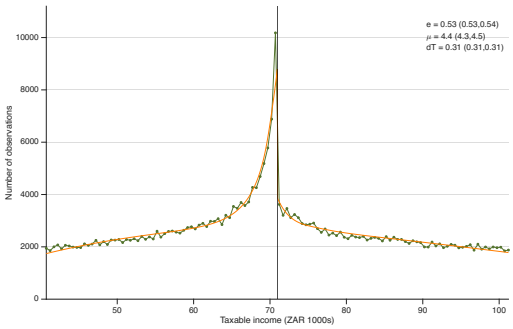
Green points plot the histograms of taxable income of all businesses in South Africa around a VAT registration threshold, after which VAT registration becomes compulsory. Orange lines plot the best-fit income density under the uniform sparsity model. The figure reports parameter estimates e (elasticity of taxable turnover), μ (average distance between taxable turnover opportunities in ZAR 1000s), and dT (the estimated "as-if" discrete change in tax liability at the bracket threshold, in ZAR 1000s). Numbers in parentheses indicate the 95 percent confidence interval on parameter estimates generated by the MLE method.

Figure 11: Asymmetric bunching at a kink: South African Evidence

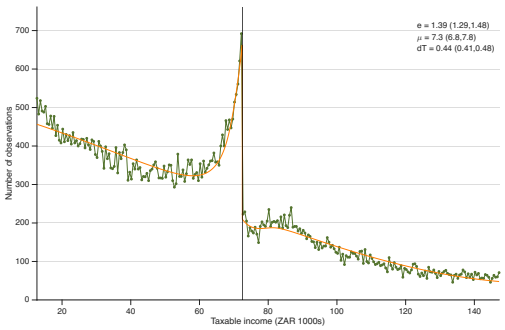
(a) CIT: Full Sample



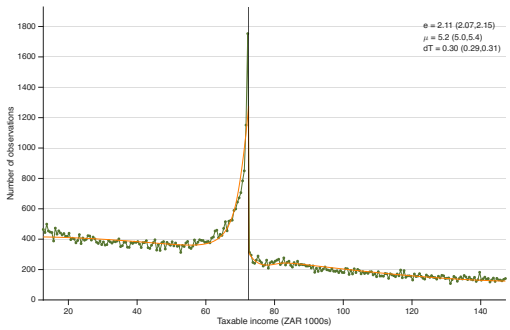
(b) PIT: Self-employed individuals



(c) CIT: No tax practitioner



(d) CIT: Uses tax practitioner



Panel (a) displays the distribution of corporate taxable incomes around the first kink in the corporate income tax (CIT) Small Business tax schedule, where the marginal tax rate rises from 0% to 7%. Green points represent the empirical histograms of taxpayer incomes; orange lines plot the best-fit income density under the uniform sparsity model. Panel (b) displays the corresponding distributions of personal incomes for self-employed taxpayers at the first kink in the personal income tax schedule, where marginal tax rates change from 0% to 18%. Panels (c) and (d) display the CIT distributions for at the first kink among firms without and with tax practitioners, respectively. Each figure also reports parameter estimates e (elasticity of taxable income), μ (average distance between income opportunities in ZAR 1000s), and dT (the estimated "as-if" discrete change in tax liability at the bracket threshold, in ZAR 1000s). Numbers in parentheses indicate the 95 percent confidence interval on parameter estimates generated by the MLE method.

Online Appendix

Diffuse Bunching with Frictions: Theory and Estimation

Santosh Anagol, Allan Davids, Benjamin B. Lockwood, Tarun Ramadorai

A Details of Bunching in the Presence of a Tax Notch

Figures A1–A3 illustrate details of the bunching model around kinks and notches in the frictionless model and with sparsity-based frictions.

Figure A1 illustrates bunching patterns that arise from a kink (panels (a) and (b)) or a notch (panels (c) and (d)) in a model without income frictions. The left panels (a) and (c) illustrate the choices of individual types of taxpayers, each of whom is depicted by a discrete dot. The right panels translate this discrete choice behavior into income densities with a continuum of types. In the presence of a kink, the frictionless model predicts an atom of mass at the bracket threshold with smooth densities on either side, generally with a discontinuity at the threshold due to the leftward shift and income compression of taxpayers who face the higher marginal tax rate above the threshold. In the presence of a notch, the model again predicts an atom of mass at the threshold and a smooth density to the left, but with an absence of mass (density equal to zero) in a dominated region of incomes just above the bracket threshold.

Figures A2 and A3 are analogous to Figures 4 and 5 in the paper, but in the presence of a notch rather than a kink. The notch in the budget constraint produces a discontinuity in the indirect utility functions plotted in Figures A2(c) and A2(d). For illustrative purposes, type c is selected to be the type that is just indifferent between two levels of income, k and their optimal income choice under $T_1(z)$. (This is distinct from Figure 6 in the paper, where type c strictly prefers k.) As illustrated in Figure A3, the logic of Proposition 1 again carries through—the type-conditional density scales with the probability of drawing an income opportunity inside the shaded region of dominating incomes—although the notch sometimes produces a set of dominating income $\Theta(z|n)$ consisting of disjoint intervals, as shown in A3b. Under sparsity-based frictions, type-conditional densities exhibit positive density even in the so-called "dominated region" to the right of the notch because for some taxpayers an income opportunity in that region may be preferable to all of their other opportunities.

B Proof of Proposition 3

Setup: We consider a sequence of income opportunity processes indexed by M, where each agent draws M income opportunities

$$\{z_1, z_2, \ldots, z_M\} = \{z^* + \varepsilon_1, z^* + \varepsilon_2, \ldots, z^* + \varepsilon_M\}$$

from a distribution $F_\varepsilon^M(x) = F_\varepsilon(x/M)$ with density $f_\varepsilon^M(x) = \frac{1}{M} f_\varepsilon(x/M)$, where F_ε is any distribution with $f_\varepsilon(0) > 0$. This transformation preserves the local density of income opportunities around the target: $M \cdot f_\varepsilon^M(0) = f_\varepsilon(0) =: \lambda$.

Our goal is to show that as $M \to \infty$, the type-conditional density $g^M(\tilde{z}|n)$ converges to $\lambda \exp[-\lambda|\Theta(\tilde{z}|n)|]$.

Choice behavior: An agent chooses income $z^* + x$ if: (i) this income is drawn, and (ii) no better income is drawn. The probability of drawing $z^* + x$ among M opportunities is $M \cdot f_\varepsilon^M(x) = f_\varepsilon(x/M)$. The probability that no better income is drawn equals the probability that all other $M-1$ draws fall outside the dominating region $\Theta(z^* + x|n)$.

Let $\Theta(z^* + x|n) = \{z|u(z - T(z), z|n) \geq u(z^* + x - T(z^* + x), z^* + x|n)\}$ be the set of incomes that utility-dominate $z^* + x$. Under the convexity assumption, this set forms an interval $[\underline{Z}(z^* + x|n), \overline{Z}(z^* + x|n)]$. In error space, this corresponds to the interval $[\underline{\phi}(x), \overline{\phi}(x)]$, where $\underline{\phi}(x) = \underline{Z}(z^* + x|n) - z^*$ and $\overline{\phi}(x) = \overline{Z}(z^* + x|n) - z^*$.

By Proposition 1, the type-conditional density for a given M is

$$g^M(z^* + x|n) = M \cdot f_\varepsilon^M(x) \left[1 - \left(F_\varepsilon^M(\overline{\phi}(x)) - F_\varepsilon^M(\underline{\phi}(x))\right)\right]^{M-1}$$

Taking the limit: Substituting the transformation $F_\varepsilon^M(x) = F_\varepsilon(x/M)$:

$$g^M(z^* + x|n) = f_\varepsilon(x/M) \left[1 - \left(F_\varepsilon\left(\frac{\overline{\phi}(x)}{M}\right) - F_\varepsilon\left(\frac{\underline{\phi}(x)}{M}\right)\right)\right]^{M-1}$$

As $M \to \infty$, we have $f_\varepsilon(x/M) \to f_\varepsilon(0) = \lambda$.

For the second term, since F_ε is differentiable at 0 with $F_\varepsilon'(0) = f_\varepsilon(0) = \lambda$:

$$F_\varepsilon\left(\frac{\overline{\phi}(x)}{M}\right) - F_\varepsilon\left(\frac{\underline{\phi}(x)}{M}\right) = \lambda \cdot \frac{\overline{\phi}(x) - \underline{\phi}(x)}{M} + o(1/M)$$

Therefore:

$$\lim_{M\to\infty} g^M(z^*+x|n) = \lambda \lim_{M\to\infty}\left[1-\lambda\cdot\frac{\overline{\phi}(x)-\underline{\phi}(x)}{M}\right]^{M-1} \tag{21}$$

$$= \lambda \lim_{M\to\infty}\left[1-\lambda\cdot\frac{\overline{\phi}(x)-\underline{\phi}(x)}{M}\right]^{M}\cdot\left[1-\lambda\cdot\frac{\overline{\phi}(x)-\underline{\phi}(x)}{M}\right]^{-1} \tag{22}$$

$$= \lambda\exp\left[-\lambda(\overline{\phi}(x)-\underline{\phi}(x))\right]\cdot 1 \tag{23}$$

$$= \lambda\exp\left[-\lambda(\overline{\phi}(x)-\underline{\phi}(x))\right] \tag{24}$$

where the third line uses the standard limit $\lim_{M\to\infty}(1-a/M)^M = e^{-a}$ and the fact that $\lim_{M\to\infty}(1-a/M)^{-1} = 1$.

Connection to uniform sparsity: Note that $\overline{\phi}(x)-\underline{\phi}(x)$ represents the total length of the error interval that yields utility at least as high as error x. In income space, this corresponds to $|\Theta(\tilde{z}|n)|$, the measure of the set of incomes that dominate income $\tilde{z} = z^* + x$.

Therefore, the limiting density is:

$$g^\infty(\tilde{z}|n) = \lambda\exp[-\lambda|\Theta(\tilde{z}|n)|]$$

This demonstrates that any income opportunity process with continouous positive density around the target income converges to the uniform sparsity model with $\lambda = f_\varepsilon(0)$ as $M\to\infty$, proving Proposition 3.

C Simulations

Using simulated data with known underlying parameters, we can assess the performance of our proposed estimation method, as compared to the conventional approach, in the presence of sparsity-based frictions. Given the large literature estimating elasticities from kinks, as opposed to notches, we focus primarily on the conventional "kink-based" bunching estimators as in Saez (2010) and Chetty et al. (2011).

We specify a simulated tax kink using the same parameters as in Figures 7 and A4: the marginal tax rate rises from $t_0 = 0.1$ to $t_1 = 0.2$ at the threshold $k = 300{,}000$. We simulate income densities assuming a baseline elasticity of $e_0 = 0.3$ and a lumpiness parameter of $\mu_0 = 10{,}000$, where the "0" subscript denotes the true parameters of the data-generating process, as distinct from model estimates of the parameters, which are denoted $\hat{e}$ and $\hat{\mu}$. We construct these simulations using linear underlying ability density, $f(n|\theta) = \theta_0 + \theta_1 n$, with $\theta_0 = 1000$ and

$\theta_1 = -50$. Each simulation uses a taxpayer population of 100,000, which produces an amount of sampling noise similar to our empirical distributions in Figure A10. (The simulations in Figures 7 and A4 used a much higher population size of 2 million to illustrate the shape of the bunching mass with less sampling noise.)

We simulate these income distributions in two steps. First, we draw ability values (n_i) from the known ability density $f(n|\theta)$ in the vicinity of the tax bracket threshold.[32] For each ability draw, we then simulate a set of income opportunities drawn from a Poisson process, from which we choose, for each agent, the highest-utility option.[33]

C.1 Performance of our estimator and the conventional bunching estimator

To assess the performance of our estimation method, we simulate many rounds of data from the same data-generating process with sparsity-based frictions, and in each case, we apply our estimation procedure to jointly estimate $\hat{e}$ and $\hat{\mu}$. We are interested in whether the distribution of these estimates is centered around the parameters of the data-generating process e_0 and μ_0, and how often the estimated confidence intervals contain the true value.

One example round of simulated data is displayed in Figure A5(a). The green dots plot the simulated income histogram. The estimated parameters $\hat{e}$ and $\hat{\mu}$ resulting from our maximum likelihood estimation are reported in the upper corner, along with the 95 percent confidence interval for each estimate. The orange line plots the model-predicted income density under these estimated parameter values.

Before comparing to the conventional estimate, we analyze multiple simulation draws from the uniform-sparsity model to understand (1) standard error performance and (2) simulation based evidence for separate identification of the elasticity and friction parameters.

Figure A6(a) plots the joint distribution of estimates $(\hat{e}, \hat{\mu})$ across 1000 simulation rounds, using the same data generating process underlying Figure A5(a). The histogram of each marginal distribution is displayed outside of each axis. A number of notable features emerge.

[32] Specifically, we draw 100,000 values of n_i between a set of bounds $\underline{n}$ and $\overline{n}$, with the probability of drawing any value n proportional to $f(n|\theta)$. To choose the lower bound $\underline{n}$, we note that due to frictions, the set of agents who earn a given z will include types whose target incomes are well below and well above z. Therefore, to simulate the income density near the bounds of an income range $[\underline{z}, \overline{z}]$, we must draw from an ability density with target incomes well outside that range. We choose $\underline{n}$ and $\overline{n}$ such that $z^*(\underline{n}) = \underline{z} - 100{,}000$ and $z^*(\overline{n}) = \overline{z} + 100{,}000$.

[33] To simulate income opportunity sets, we exploit the fact that differences between adjacent elements in a Poisson process are iid draws from an exponential distribution with mean μ. Thus, we can construct a random income opportunity set spanning an arbitrarily wide range around a type's preferred income $z^*(n)$ by joining a random set of above-target opportunities, $\{z^*(n) + \varepsilon_a, z^*(n) + \varepsilon_a + \varepsilon_b, z^*(n) + \varepsilon_a + \varepsilon_b + \varepsilon_c, \ldots\}$, with a random set of below-target opportunities $\{z^*(n) - \varepsilon_i, z^*(n) - \varepsilon_i - \varepsilon_j, z^*(n) - \varepsilon_i - \varepsilon_j - \varepsilon_k, \ldots\}$, where the ε values are iid draws from an exponential distribution with mean μ. In the context of a kink, where indirect utility functions are concave, only a single element must be drawn in each set, since more distant draws are guaranteed to yield lower utility. For a notch, with non-concave indirect utility functions, a larger number of opportunities is drawn, such that each agent's range of income opportunities spans across the local maxima in their indirect utility functions.

First, for both $\hat{e}$ and $\hat{\mu}$, the distribution of estimates is centered around the true parameter value. Averaging across simulation rounds, and the average value of $\hat{\mu}$ is 10.1 (measured in 1000s), close to the true values of $e_0 = 0.3$ and $\mu_0 = 10$.

Second, the spread of both distributions provides an indication of sampling error. In each round of simulated data, the maximum likelihood estimation procedure also provides a standard error estimate, and so a key question is whether this estimate gives an accurate picture of the degree of precision in the estimate. To explore this, we can compare the standard deviation of the distribution of $\hat{e}$ estimates, which is 0.026, to the average *estimate* of the standard error, which is 0.026, indicating that the maximum likelihood estimate of the standard error provides a good sense of the true degree of sampling uncertainty. In the case of μ, the standard deviation of the distribution of $\hat{\mu}$ is 1.121, and the average value of the estimated standard error is 1.099.

A third notable feature of Figure A6(a) is the upward slope in the cluster of joint estimates. This indicates that when $\hat{e}$ is overestimated due to sampling bias, it is likely that $\hat{\mu}$ is overestimated as well. To explore this phenomenon, Figure A6(b) plots model-generated income densities for five combinations of (e, μ). The thick solid line plots the baseline density with $e = 0.3$ and $\mu = 10$. The other four lines correspond to the (e, μ) pairs corresponding to the four square-shaped points in Figure A6(a).

In Figure A6(b), the densities corresponding to the points to the northwest and southeast of the baseline are easy to visually distinguish from the baseline, exhibiting substantially lower and higher densities at the kink, respectively. A higher elasticity e increases the density at the kink point by raising the total amount of bunching mass. A *lower* value of the lumpiness parameter also increases the density at the kink point, by concentrating the excess mass more tightly around the kink (Figure A6(b)). Thus, the parameter combinations to the southeast of the baseline in Figure A6(a) correspond to densities with substantially higher density around the kink point, like the tallest density displayed in Figure A6(b). The reverse is true for parameter combinations to the northwest of the baseline values, where the levels of both parameters (low e and high μ) reinforce each other to push down the density at the kink. In contrast, parameter combinations to the northeast and southwest of the baseline have opposing effects on the density at the kink. They are still distinct, indicating that the model is identified, but their difference is more subtle, involving the density at intermediate points in between the kink point and the bounds of the income window. The pattern of points in Figure A6(a) corresponds to this visual impression: in the presence of sampling error, it is easier to distinguish—in a statistical sense—between data-generating processes with parameter pairs on the northwest-southeast axis than those on the northeast-southwest axis in Figure A6(b).[34]

[34] Put differently, the estimator that we propose would find it easier to distinguish between data generated from

In sum, these points paint a clear picture of the performance of the maximum likelihood estimator when the model is correctly specified. Estimates of the elasticity and the lumpiness parameter appear consistent in that they are distributed around the true parameters of the data-generating process, and standard errors estimated by maximum likelihood are very close to the standard deviation of the distribution of estimates. They also highlight an important aspect of this model: estimation error in e and μ are likely to have the same sign. This result has important implications for the comparison of this model to the conventional elasticity estimator based on bunching mass.

C.2 Specifying the Conventional Bunching Estimator

We now apply the conventional bunching estimator to the same simulation data. We apply the conventional bunching estimator based on Saez (2010) to estimate the income elasticity of the simulated data sets underlying Figure A5. We use as our baseline the implementation described in Chetty et al. (2011), which builds on Saez (2010) by estimating a counterfactual using a smoothed polynomial regression. We also later discuss comparisons to the methods used in Saez (2010) and Mortenson and Whitten (2016), the working paper that preceded Mortenson and Whitten (2020).

This estimation procedure involves two steps, first estimating a counterfactual income density based on the income density excluding data points near the kink, and then using the counterfactual density to estimate the excess mass from which the elasticity is recovered. To estimate the counterfactual density, we fit a polynomial of a specified degree to the observed income density, excluding the data in a specified window around the kink, using the following specification:

$$C_j = \sum_{i=0}^{q} \beta_i^0 \cdot (Z_j)^i + \sum_{i=R_l}^{R_u} \gamma_i^0 \cdot \mathbf{1}[Z_j = i] + \epsilon_j^0. \tag{25}$$

Here, q denotes the order of the polynomial, and R_l and R_u denote the lower and upper bounds of the "bunching window" near the kink, which is excluded from the polynomial estimation.[35] When estimating the polynomial regression, we follow Chetty et al. (2011) and impose an "integration constraint" such that the total count of observations across the empirical distribution equals the integral of observations under the counterfactual density across the plotted region.[36]

"low e, high μ" and "low μ, high e" combinations than between "low μ, low e" and "high μ, high e" combinations.

[35] The convention in Chetty et al. (2011) is to set a symmetric bunching window, such that $R_l = -R_u$. We allow for the possibility of an asymmetric bunching window, following the approach in Bosch, Dekker and Strohmaier (2020), which we detail below.

[36] Kleven (2016) notes that imposing an integration constraint may bias the elasticity estimate: "This approach may introduce bias, especially in relatively flat distributions in which interior responses do not affect bin counts

The second step is to compute the excess mass of incomes around the kink relative to this counterfactual density. Using equation (25), we compute the counterfactual mass in each bin within the bunching window, $\hat{C_j^0}$. Subtracting this predicted mass from the observed density yields the estimated excess number of individuals who report incomes near the kink relative to this counterfactual distribution:

$$\hat{B} = \sum_{i=R_l}^{R_u} C_j - \hat{C_j^0} = \sum_{i=R_l}^{R_u} \hat{\gamma}^0. \tag{26}$$

We then map this excess mass estimate to an estimated elasticity using the approximation from Chetty et al. (2011):

$$\hat{e} \approx \frac{\hat{B}}{z^* \cdot h_0(z^*) \cdot \log(\frac{1-t_0}{1-t_1})}. \tag{27}$$

Standard errors for $\hat{e}$ are estimated using a bootstrap procedure. We resample with replacement from the underlying distribution of firms 1000 times, re-estimating the elasticity each time, and defining the standard error as the standard deviation of the distribution of $\hat{e}$ estimates.

This conventional estimation method relies on three parameter inputs: the lower and upper bounds of the bunching window (R_l and R_u) and the order of the polynomial (q). These are often left to the discretion of the researcher to be chosen via "visual inspection." We instead follow the algorithmic approach proposed in Bosch, Dekker and Strohmaier (2020), which allows the polynomial order and the bunching region to be informed by the data itself.[37]

Figure A5(b) presents the results. The bunching window is bounded by dashed lines, and the orange line displays the fitted counterfactual density outside that window.[38] The estimated elasticity and bootstrap-based 95 percent confidence interval is reported in the corner.

(except at the very top of the distribution away from the threshold being analyzed). It would be feasible to implement a conceptually more satisfying approach that does not have this potential bias, but for the reasons stated above, it will matter very little in most applications." As we discuss below, our results confirm that the integration constraint introduces bias, and that the introduced bias may be substantial.

[37]This approach proceeds in five steps: (1) Estimate equation (25) with no bunching window—so that the polynomial estimation excludes only the bins adjacent to the kink—for a range of polynomial orders, retaining the specification that minimizes the Bayesian Information Criterion (BIC). (2) Define the lower bound of the bunching window as the leftmost set of two adjacent bins below the threshold where the actual count in each bin exceeds the 95 percent confidence interval of the predicted bin counts from equation (25), and define the upper bound using an analogous procedure to the right of the kink. (3) Repeat steps (1) and (2), widening the bunching window by one bin above and below the kink each time. Each such iteration produces a candidate set of bounds for a bunching window. (4) From the resulting distributions of candidate bounds, choose the modal lower bound and upper bound to define the preferred bunching window. (5) Using this preferred bunching window, re-estimate the final counterfactual regression with the preferred polynomial order as in Step (1).

[38]Strictly speaking, this line represents the counterfactual *frequency*, equal to the counterfactual density scaled up by the bin width of the empirical histogram in order to render the plots visually comparable.

C.3 Comparison of Uniform Sparsity to Conventional Method Elasticities

Comparing the elasticity estimates from the two methods, we note that the conventional bunching estimator in Figure A5(b) underestimates the true elasticity of the data-generating process by 25 percent. It also provides a misleading sense of precision: the 95 percent confidence interval does not contain the true elasticity. In contrast, the sparsity-based friction estimator in panel (a) is close to the true value of $e_0 = 0.3$, which is spanned by the 95 percent confidence interval.

To compare the relative performance of these estimators more generally, we apply them to 1000 different rounds of simulated data. Figure A7(a) plots the histogram of elasticity estimates from the conventional bunching estimator and from our proposed estimation method. Consistent with the results from the single simulation round, the distribution of elasticity estimates from the conventional bunching estimator lies substantially below the true elasticity e_0. The average of elasticity estimates under the conventional approach is 0.243, and the bootstrap-based 95 percent confidence intervals contain the true e_0 in less than 10 percent of the cases. In contrast, the distribution of elasticity estimates from our proposed estimation method is centered around e_0, with an average value of $\hat{e}$ across these simulation rounds of 0.307. The estimated confidence intervals from our approach also provide an accurate sense of precision: across the 1000 estimation rounds, the 95 percent confidence intervals contained the true e_0 in 95.3 percent of cases.

The downward bias in the conventional bunching estimator appears to be driven by frictions, as illustrated by Figure A7(b). To construct this figure, we reproduce distributions like those in Figure A5(a) using several different values of the lumpiness parameter μ_0. Figure A7(b) plots the mean and the 95 percent quantile interval of each distribution at each value of μ. When the lumpiness parameter is small—approaching the continuous-income-choice model—the mean estimate of $\hat{e}$ under the conventional approach is close to the true value of $e_0 = 0.3$. However, as μ_0 rises, the conventional estimator exhibits substantial bias, underestimating the true parameter by more than 50 percent at the highest plotted value of μ_0. These estimates also provide a misleading sense of precision: the 95 percent quantile intervals remain about the same size as μ_0 rises, and their upper bound falls far below e_0. In contrast, under our method, the distribution of $\hat{e}$ remains centered around e_0 as frictions increase. The 95 percent quantile interval grows with μ_0, reflecting the increasing imprecision in the elasticity estimate as lumpiness increases. This imprecision accurately reflects the greater difficulty of discerning diffuse bunching mass from underlying features of the smooth ability density when frictions are substantial.

Why do frictions cause the conventional bunching estimator to be biased downward? We highlight two contributing factors. The first arises because diffusion in the bunching

mass makes it difficult to distinguish excess mass from patterns in the counterfactual income density. In the uniform sparsity model of frictions—and in many of the other sparsity-based friction models it approximates, such as when income opportunities are drawn from a normal distribution around the target income—there is no window outside of which the bunching mass falls to zero. As a result, some excess bunching mass will spill over outside of any particular bunching window—including the window chosen visually or algorithmically when implementing the conventional approach. This spillover mass tends to "pull up" the estimated polynomial fit in the vicinity of the kink, causing the procedure to underestimate the difference between the observed density and the counterfactual, and hence the bunching mass. In our model, in contrast, the distortions due to frictions are endogenously modeled throughout the income distribution, including at points far from the threshold, and so they should not exert an upward pull on the ability density around the threshold.

To explore the role of this factor in producing the bias evident in Figure A7, we note that this source of bias should become more severe when the polynomial fit is allowed to be more flexible. In Figure A8, we reproduce the estimates in Figure A5 with different polynomial degrees of 1 (linear), 3, 5, and 10. Consistent with this story, Figure A8(b) shows that when the polynomial degree is higher, the counterfactual density bends farther up into the bunching mass, and the elasticity estimate is more severely biased downward. In contrast, Figure A8(a) demonstrates that our proposed method continues to estimate $\hat{e}$ close to e_0 across all polynomial degrees, suggesting that this method is robust to misspecification in the shape of the ability density in a way that the conventional approach is not.

Although this first factor appears to play an important role in the downward bias of the conventional bunching estimator, it does not appear to be the sole explanation, because even the linear polynomial specification in Figure A5(a) produces a substantial underestimate of the true elasticity.

The second factor contributing to downward bias in the conventional method relates to the integration constraint imposed when estimating the counterfactual polynomial fit. The logic for such a constraint comes from the observation that any taxpayers bunching around a threshold must come from points to the right of the threshold under the counterfactual, and so the total population under the actual and counterfactual income densities must be the same.[39] However, as illustrated by the hollow blue points in Figure A1(a), the presence of a kink may

[39] Describing the rationale for imposing the integration constraint, Chetty et al. (2011) remarks that an unadjusted polynomial fit "... overestimates [the bunching mass] because it does not account for the fact that the additional individuals at the kink come from points to the right of the kink. That is, it does not satisfy the constraint that the area under the counterfactual must equal the area under the empirical distribution. To account for this problem, we shift the counterfactual distribution to the right of the kink upward until it satisfies the integration constraint."

induce taxpayers to appear inside the plotted region who were previously outside of it. In other words, although such an integration constraint does apply to the global income density, it need not apply within the particular region over which the bunching estimator is applied.[40]

We now compare our estimator to other conventional bunching estimators that differ in how they construct counterfactuals, namely Saez (2010) and Mortenson and Whitten (2016), the working paper that preceded Mortenson and Whitten (2020), with an ultimate focus on the importance of integration constraint. We illustrate the differences in counterfactuals between these approaches in Figure A9(a). The approach developed in Saez (2010) constructs two linear counterfactuals on either side of the kink with the assumption that the densities are uniformly distributed on either side of the threshold. In order to construct the counterfactual, the approach takes the average value of the densities that occur outside of the bunching window and extrapolates that density through to the kink threshold. This is done on either side of the kink resulting in two counterfactuals. An alternative approach is developed in Mortenson and Whitten (2016) who construct a piecewise linear counterfactual on either side of the kink, in a similar vein to Saez (2010), but to allow for that counterfactual to take into account the slope of the observed densities on either side of the kink. Finally, we also consider an implementation of the approach in Chetty et al. (2011) where we do not impose the "integration constraint." This allows for the possibility that the bunching mass may be reallocated to income bins beyond the region depicted in the histogram, which would cause the total integral under the counterfactual distribution to be smaller than the total integral of population across the empirical distribution. The practical implication of this is that the counterfactual distribution is shifted downward relative to the approach which imposes the integration constraint, as is depicted in Figure A9(a).

Next, we compare the elasticities produced under these four approaches to our estimates for varying values of the lumpiness parameter. We report these results in Figure A9(b). Imposing the integration constraint in the Chetty et al. (2011) approach produces lower elasticities than when the constraint is not imposed. Intuitively, by imposing the constraint, the counterfactual density is shifted upward, which causes the estimate of bunching to fall, leading to a lower elasticity. The Mortenson and Whitten (2016) elasticities are very similar to the Chetty et al. (2011) elasticities without the integration constraint. In that sense, the specification is nearly identical to the counterfactual specification in Mortenson and Whitten (2016), apart from the fact that the latter approach estimates two counterfactuals on either side of the threshold,

[40] Indeed, Figure A1(b), which illustrates the observed income density in the frictionless model with a continuous uniform type density, demonstrates that the kink may induce both extra mass at the kink *and* higher density at incomes above the kink, in which case the true counterfactual density under T_0—which extends the uniform density below k to points above it—clearly has a lower integral over the plotted region than the observed density does.

thereby allowing for a different slope on the counterfactual on either side of the kink threshold. The visual similarity between these counterfactuals is evident in Figure A9(a). Out of all of the conventional approaches, the Saez (2010) approach produces the largest elasticities. The reason for this becomes evident when considering the counterfactuals produced in Figure A9(a). Given the empirical distribution slopes downwards, by assuming uniformly distributed densities, the Saez (2010) approach produces a counterfactual that is significantly lower than the other counterfactuals in the bunching region above the kink, leading to a higher measure of bunching, and a higher estimated elasticity.

For smaller values of the lumpiness parameter, only Mortenson and Whitten (2016) and the Chetty et al. (2011) approach without the integration constraint can recover the true elasticity. However, for large values of the lumpiness parameter, not even these approaches are able to recover the true elasticity and exhibit a significant downward bias, whereas our approach can consistently recover the elasticity, irrespective of the extent of lumpiness in the observed empirical distribution.

Under the uniform sparsity model, for a given observed income density $h(z)$ and counterfactual density $h_0(z)$, if the target income response Δz (which identifies the elasticity e) and sparsity parameter λ are constant in a sufficiently wide region around the tax kink, then Δz and λ are separately identified.[41]

D Formal Identification Proof

In this Section, we present a formal proof that the elasticity ϵ and the diffusion parameter λ are separetely identified in our model. We begin by showing ϵ is identified - for a given counterfactual income distribution $H_0(z)$ an observed income distribution $H(z)$ can be consistent with only one target income response Δz (this target income response Δz is isomorphic to our elasticity ϵ.

We begin with the fact that the total bunching mass around the kink k is defined as the excess mass at income $k+\epsilon$ relative to the counterfactual density that would arise under the local linear tax function:

$$b(\epsilon) = \begin{cases} h(k+\epsilon) - h_0(k+\epsilon) & \text{if } \epsilon < 0, \\ h(k+\epsilon) - h_1(k+\epsilon) & \text{if } \epsilon > 0. \end{cases} \tag{28}$$

In our numerical implementation companion document we show:

[41] Note that our simulations in the paper assume a constant elasticity e, whereas for analytic simplicity this proof assumes a constant target income response Δz in levels. The two assumptions are quantitatively similar provided the region of bunching is small relative to the level of income at the kink k.

$$\int_{-\infty}^{\infty} b(\epsilon)\,d\epsilon = H_1(k) - H_0(k) = B. \tag{29}$$

By construction of this $b(\epsilon)$ (and equation 29), the integral $\int_{-\infty}^{\infty} b(\epsilon)\,d\epsilon$ is finite, and thus so is the portion in the positive domain, $\int_{0}^{\infty} b(\epsilon)\,d\epsilon$. Further, in our numerical companion document we show that the bunching mass function at income $k+\epsilon > k$ under the uniform sparsity model is:

$$b(\epsilon) = \int_{-\infty}^{k+\epsilon/2} \lambda \exp\left[-\lambda 2\left|k+\epsilon-z_1^*\right|\right]\left(\exp[\lambda\delta(\epsilon|z_1^*)]-1\right) h_1^*(z_1^*)\,dz_1^*. \tag{30}$$

Intuitively, this calculates the excess mass at $k+\epsilon$ given the income opportunities agents face combined with the size of their dominating regions. Equation 30 shows that the bunching mass function is strictly positive ($b(\epsilon) > 0$ for all ϵ).

Together, these facts imply that the integral $\int_{\bar{\epsilon}}^{\infty} b(\epsilon)\,d\epsilon$ converges to zero as $\bar{\epsilon} \to \infty$. That is, for any desired precision $\varepsilon^* > 0$, there is a $\bar{\epsilon}$ such that $\int_{\bar{\epsilon}}^{\infty} b(\epsilon)\,d\epsilon < \varepsilon^*$. And from the definition of $b(\epsilon)$ we have

$$\begin{aligned}\int_{\bar{\epsilon}}^{\infty} b(\epsilon)\,d\epsilon &= \int_{\bar{\epsilon}}^{\infty} \left[h(k+\epsilon) - h_1(k+\epsilon)\right] d\epsilon \\ &= 1 - H\left(k+\bar{\epsilon}\right) - \left(1 - H_1\left(k+\bar{\epsilon}\right)\right) \\ &= H_1\left(k+\bar{\epsilon}\right) - H\left(k+\bar{\epsilon}\right).\end{aligned} \tag{31}$$

Thus by choosing $\bar{\epsilon}$ sufficiently large, we can ensure that $H_1(k+\bar{\epsilon}) - H(k+\bar{\epsilon}) < \varepsilon^*$, implying that the observed income distribution $H(z)$ approximates the counterfactual income distribution $H_1(z)$ to within ε^*. By assumption of our proposition, the target income response Δz is constant in a sufficiently wide region around the tax kink; we now exploit that assumption to assert that it is constant in a region that includes $k+\bar{\epsilon}$. By the definition in Equation 29, we can thus write

$$H_1(k+\bar{\epsilon}-\Delta z) - H_0(k+\bar{\epsilon}) = 0 \implies \left|H\left(k+\bar{\epsilon}-\Delta z\right) - H_0\left(k+\bar{\epsilon}\right)\right| < \varepsilon^*. \tag{32}$$

For known $H(z)$ and $H_0(z)$, this equation can be used to identify Δz within an arbitrary degree of precision controlled by ε^*. This method depends on λ only through the speed at which $H(z)$ converges to $H_1(z)$ above k, and thus provided one chooses a sufficiently high $\bar{\epsilon}$ to achieve the desired degree of precision, variation in λ has an arbitrarily small effect on the estimated B, and thus on the estimated Δz. It is in this sense that Δz is separately identified from λ under the uniform sparsity model. The intuition for this result stems from the fact that the elasticity is identified entirely by the integral of the bunching mass function—B—which is not affected by

the value of λ.

Having identified Δz, we have also identified $H_1(z)$ and $h_1(z)$. Our next lemma demonstrates how the sparsity parameter λ can be identified from the observed income density $h(z)$ and the target income response Δz.

Lemma 1. *Under the uniform sparsity model, the density at the kink satisfies*

$$h(k) = \lambda B + \int_{-\infty}^{k} h_0^*(z_0^*)\lambda \exp\left[-\lambda 2|k - z_0^*|\right] dz_0^* + \int_{k}^{\infty} h_1^*(z_1^*)\lambda \exp\left[-\lambda 2|k - z_1^*|\right] dz_1^*. \quad (33)$$

Proof. Any agent with a target income $z_0^* < k$ prefers income k to any $z > k$, regardless of whether a kink is present, and therefore the contribution of such agents to the observed income density $h(z)$ is equal to their contribution to the counterfactual density $h_0(z)$. Similarly, any agent with a target income $z_1^* > k$ prefers income k to any $z < k$, and therefore their contribution to $h(z)$ is equal to their contribution to $h_1(z)$.

Finally, for any agent with $z_1^* < k < z_0^*$, or equivalently, $k < z_0^* < k + \Delta z$, their preferred income under the kink is k, meaning the dominating income region is just a single value, k, implying that their type-conditional density at k is $\lambda \exp[-\lambda \cdot 0] = \lambda$. Combining these facts, the observed density at k can be written

$$h(k) = \int_{-\infty}^{k} h_0^*(z_0^*)\lambda \exp\left[-\lambda 2|k - z_0^*|\right] dz_0^* + \int_{k}^{\infty} h_1^*(z_1^*)\lambda \exp\left[-\lambda 2|k - z_1^*|\right] dz_1^* + \lambda \int_{k}^{k+\Delta z} h_0^*(z_0^*) dz_0^*. \quad (34)$$

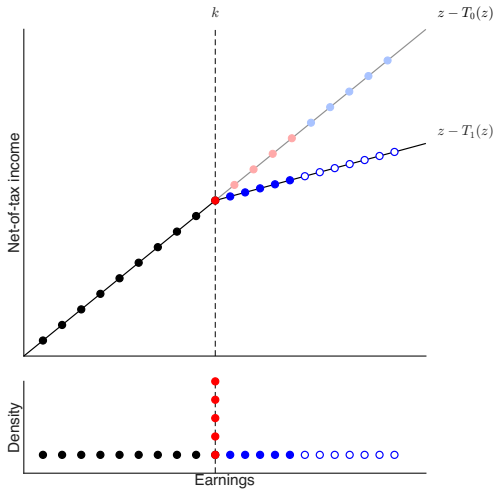
Note that the integral in the final term is equal to $\lambda\left(H_0(k + \Delta z) - H_0(k)\right)$, and by Lemma 3, this is equal to $\lambda\left(H_1(k) - H_0(k)\right) = \lambda B$. This proves the lemma. □

We can then use Lemma 1 to compute λ.

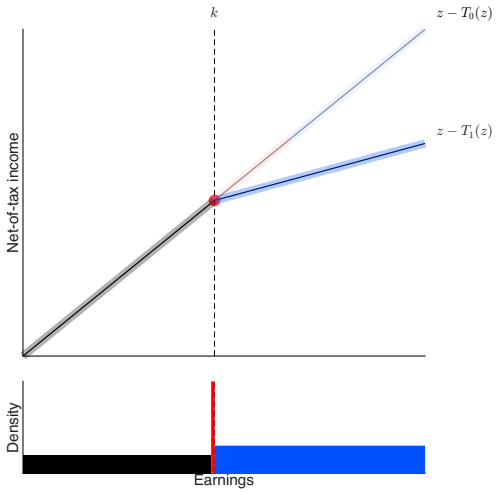
This completes the proof that Δz and λ are separately identified under the uniform sparsity model.

Figure A1: Bunching patterns under a frictionless model

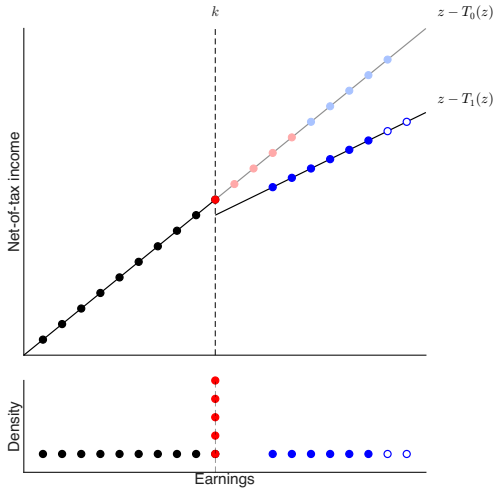
(a) Income choice around kink



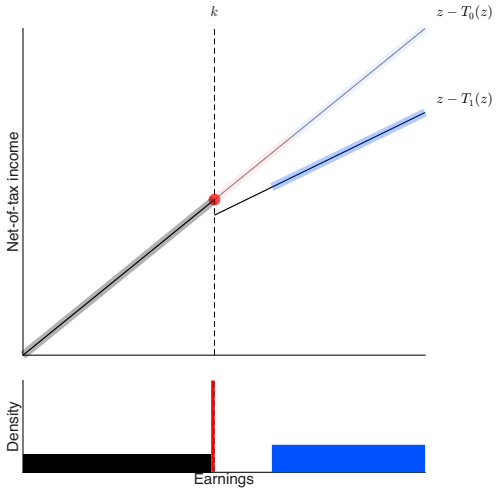
(b) Income density around kink



(c) Income choice around notch

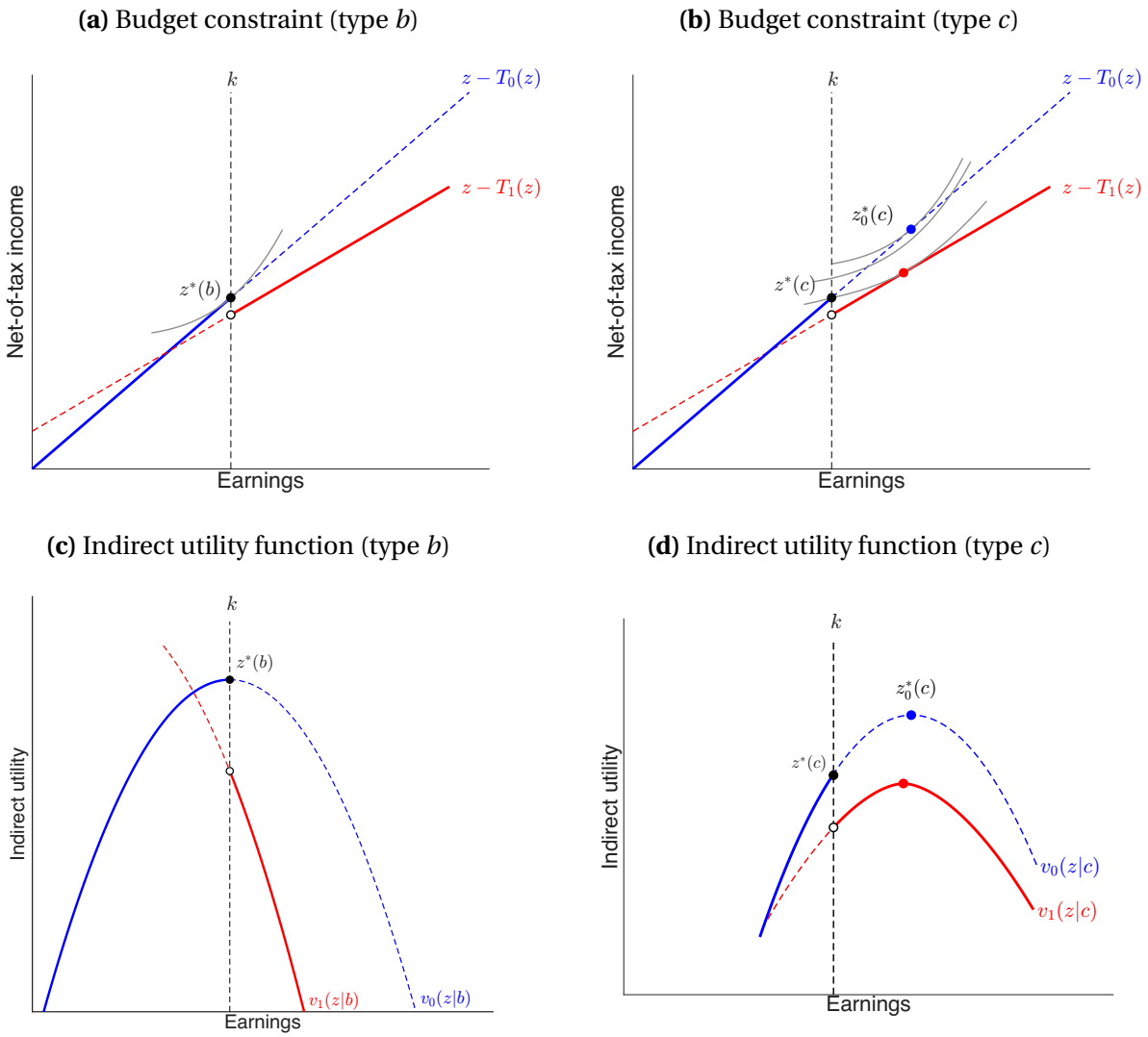


(d) Income density around notch



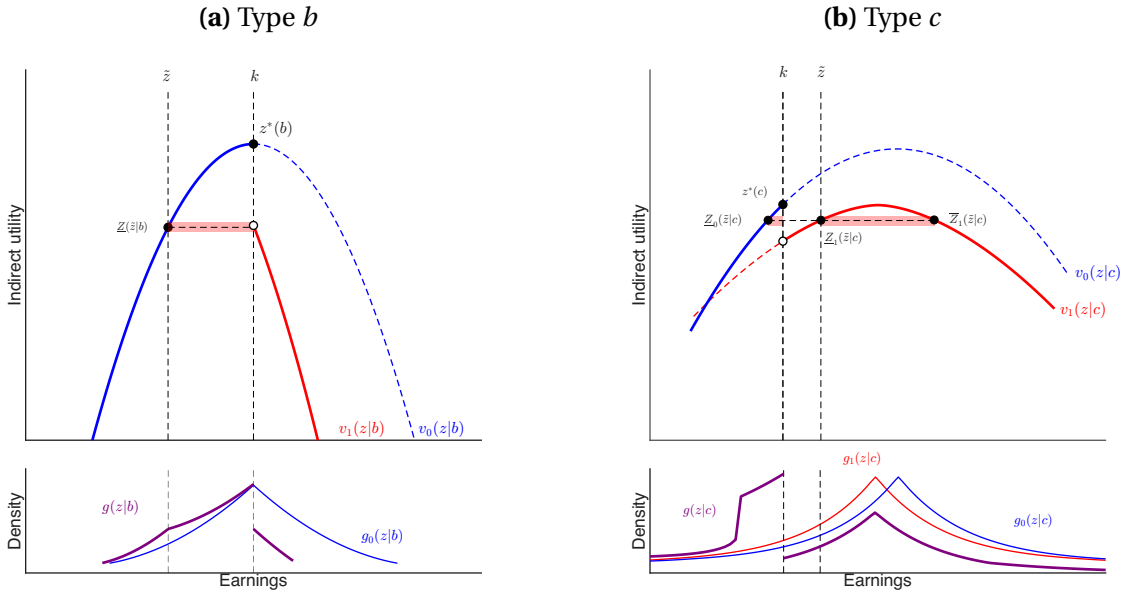
Panel (a) plots income choices for discrete types in the presence of a kink, for a uniform type distribution. Black points denote types who choose the same income under the linear tax $T_0(z)$ and the kinked tax schedule. Red points denote types who bunch at the bracket threshold k under the kinked tax schedule; their counterfactual income choices under $T_0(z)$ are plotted in light red for reference. Blue points denote types who choose incomes above the threshold under the kink. Hollow blue points denote agents whose counterfactual incomes under $T_0(z)$ lie outside the displayed range of incomes. The lower portion of panel (a) displays the observed probability density function from these choices. Panel (b) translates to the case of continuous types, which exhibits an atom of mass at the threshold k and a jump in the density around that threshold, due to the compression of incomes in response to the higher marginal tax rate above the kink. Panels (c) and (d) are the same as panels (a) and (b), but in the presence of a tax notch.

Figure A2: Utility from income choices around a tax notch



This figure is analogous to Figure 4, but in the presence of a notch, which produces a discontinuity in the indirect utility function (panels (c) and (d)). In the case of type c, the notch produces a non-monotonic indirect utility function with two local maxima (panel (d)).

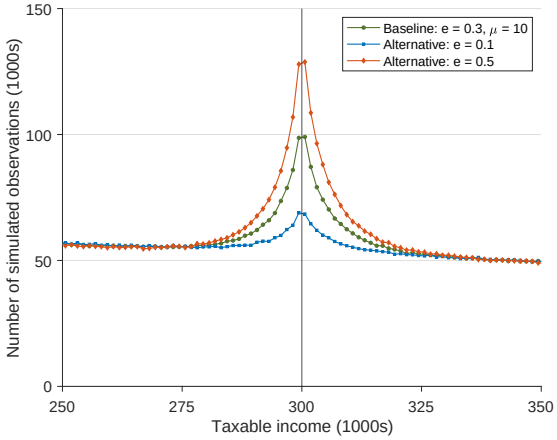
Figure A3: Type-conditional income density around a notch



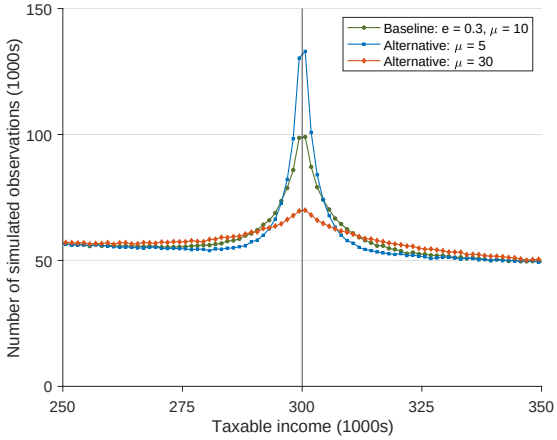
This figure is analogous to Figure 5, but in the presence of a notch. As shown in panel (b), when the indirect utility function has multiple local maxima, the dominating income range may be a disjoint set, in which case the type-conditional density is multimodal.

Figure A4: Simulated effects of parameter variations on income densities

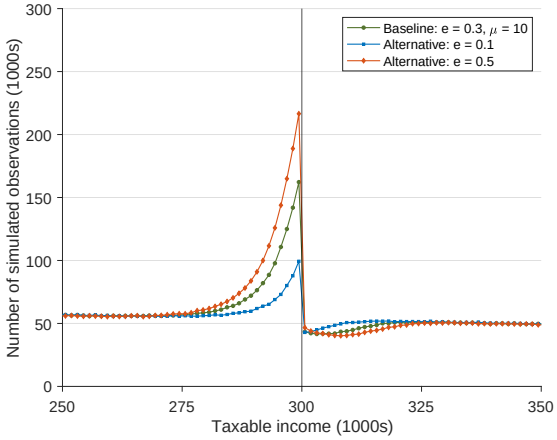
(a) Different elasticities (kink)



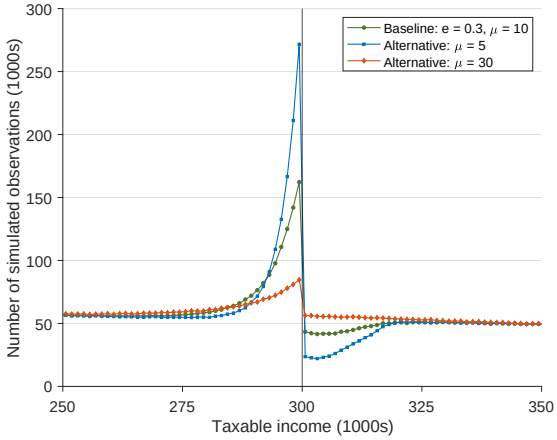
(b) Different lumpiness parameters (kink)



(c) Different elasticities (notch)



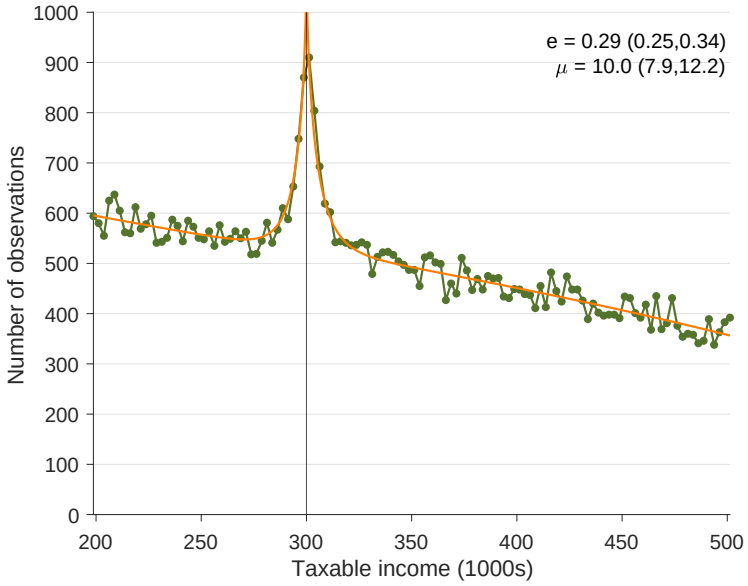
(d) Different lumpiness parameters (notch)



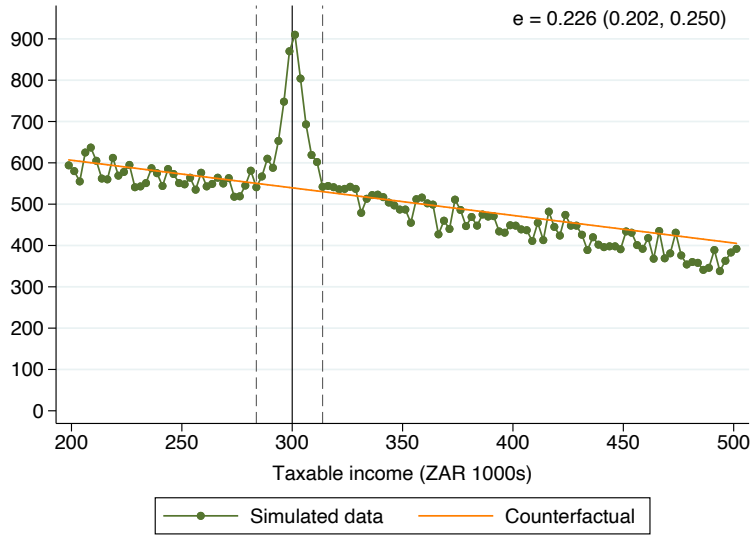
This figure plots income histograms from simulated data sets under the uniform sparsity model of frictions. For each simulation, we draw agents from an ability distribution with a linear density. We assume agents have a homogeneous income elasticity, e_0, and for each agent we then draw a sparse set of income opportunities from a Poisson process with a specified lumpiness parameter, μ_0. Each agent chooses the income opportunity that delivers the highest utility. We bin the resulting incomes to construct the income histograms displayed above. Panels (a) and (b) display simulated income histograms around a progressive tax kink for different values of e_0 and μ_0, respectively. Panels (c) and (d) display histograms around a tax notch. In each case, the marginal tax rate rises from 0.1 to 0.2 at \$300,000, and for the notch simulations in panels (c) and (d), the level of tax liability increases by \$1000.

Figure A5: Parameter estimates from simulated data

(a) Sparsity-based frictions estimator for a single simulation round



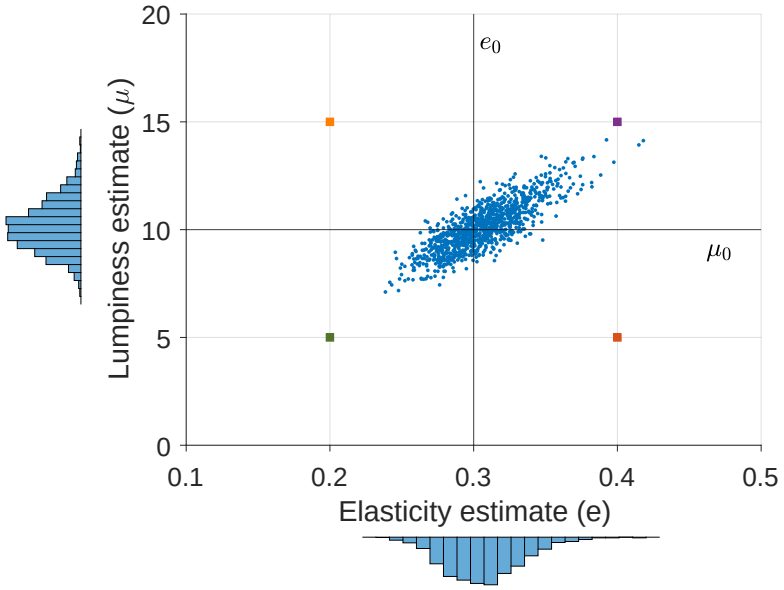
(b) Conventional bunching estimator for a single simulation round



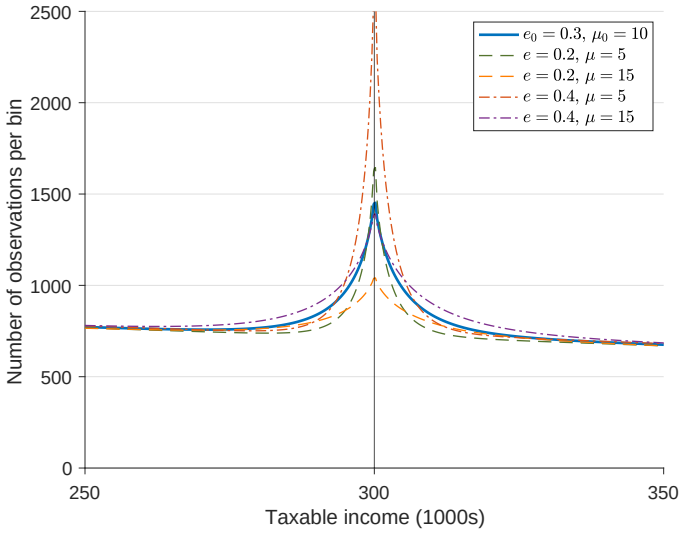
This figure displays the estimation of the maximum likelihood model and the conventional bunching estimator for one round of simulated data. The simulation is constructed as in Figure A4, with a true elasticity of $e_0 = 0.3$ and a lumpiness parameter of $\mu_0 = 10$, but using a smaller number of drawn observations ($M = 100,000$) to produce a level of sampling noise similar to that in our empirical application in Section 3. Panel (a) displays the results of applying the our maximum likelihood estimation method described in Section 2.6. Estimates of $\hat{e}$ and $\hat{\mu}$ and their 95 percent confidence intervals are reported in the upper corner. Panel (b) illustrates the conventional bunching estimator, applied to the same round of simulated data, resulting in an elasticity estimate well below the true value $e_0 = 0.3$. The vertical dashed lines display the algorithmically selected bunching window, and the orange line plots the best-fit polynomial to the data points outside the bunching window.

Figure A6: Joint identification of elasticity and lumpiness estimates

(a) Joint distribution of $\hat{e}$ and $\hat{\mu}$ estimates



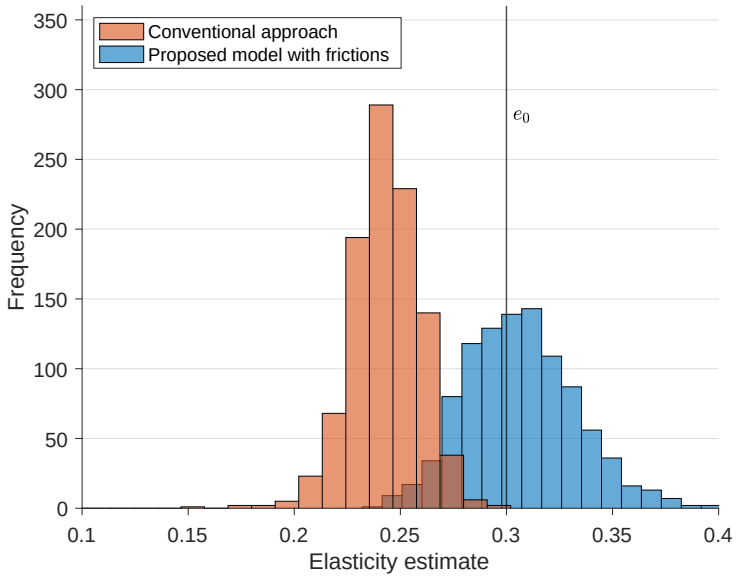
(b) Income densities for different combinations of e and μ



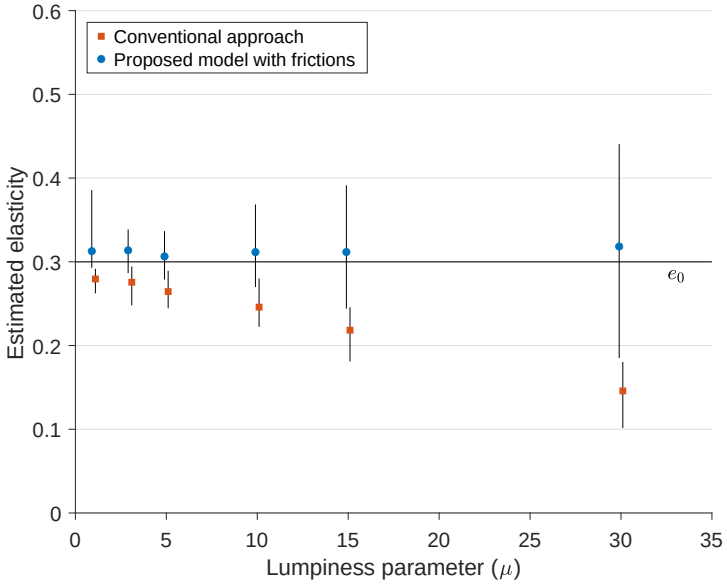
In Panel (a), each blue point plots the combination of parameter estimates $(\hat{e}, \hat{\mu})$ from one round of simulated data like that in Figure A5(a). Marginal histograms of the estimates are plotted for each axis. Panel (b) plots the model-generated income density under the true parameters of the data-generating process, $e_0 = 0.3$ and $\mu_0 = \$10,000$, as well as under the four different combinations corresponding to the colored square points in panel (a).

Figure A7: Elasticity estimates using the conventional approach

(a) Distribution of elasticity estimates under each approach



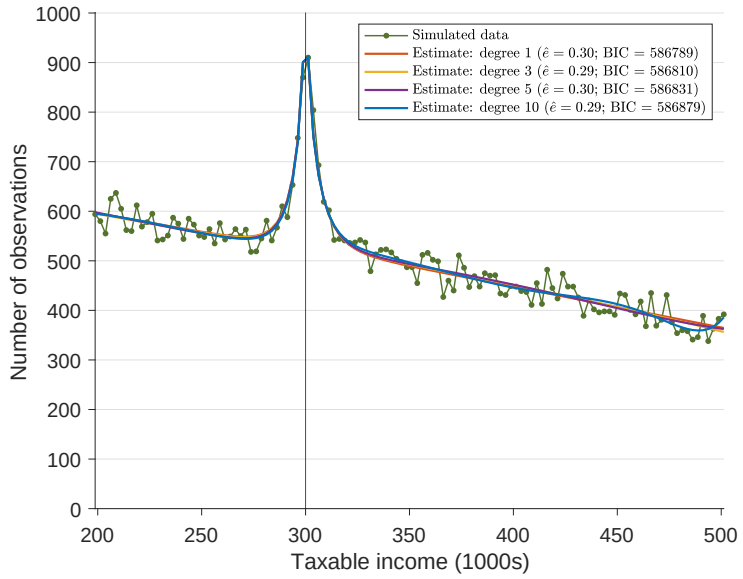
(b) Elasticity estimates under each approach for different lumpiness parameters



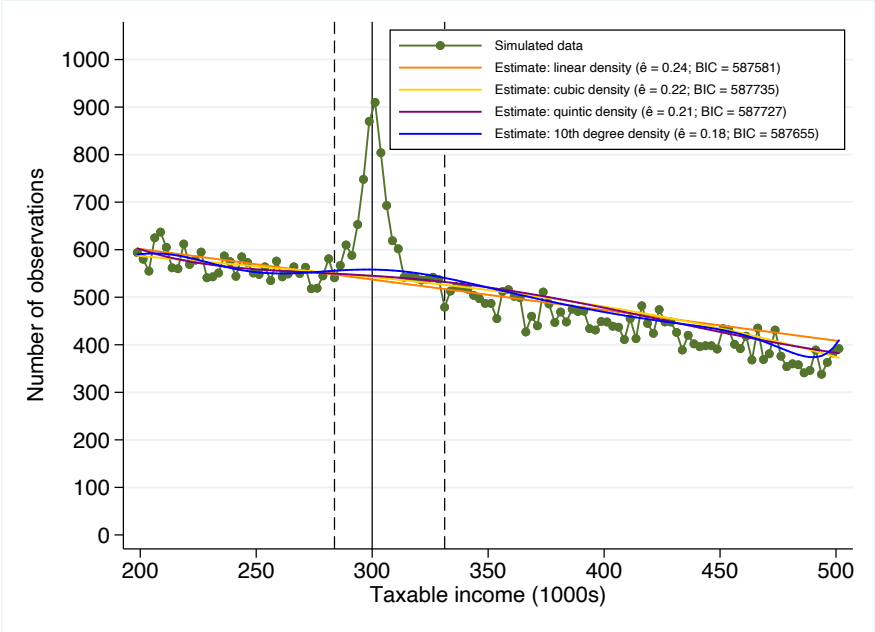
Panel (a) plots the histogram of elasticity estimates under the conventional approach (orange) and the maximum likelihood method allowing for frictions (blue). The vertical line at e_0 locates the true elasticity of the data generating process used to construct the simulated data sets. To construct panel (b), we produce histograms like those in panel (a) using simulated data with several different lumpiness parameters, holding fixed the true elasticity. Panel (b) displays the mean and 95 percent confidence intervals for the distribution of elasticity estimates in each case.

Figure A8: Estimated elasticities assuming different polynomial degrees

(a) Estimator with frictions



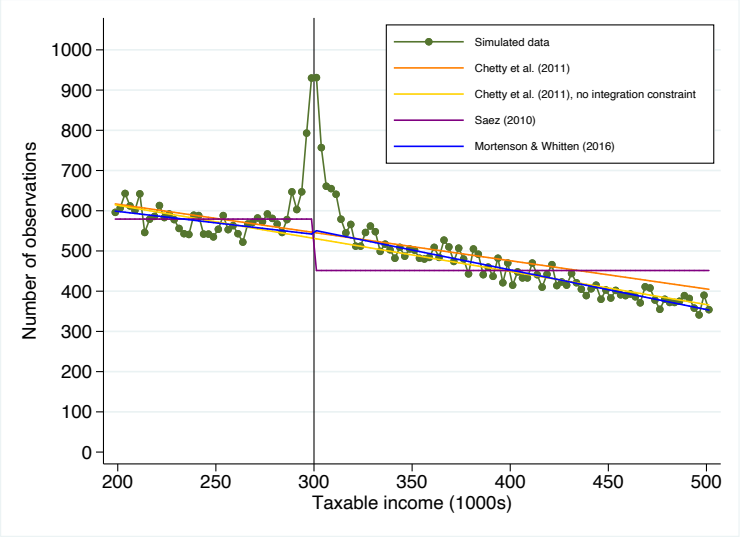
(b) Conventional estimator



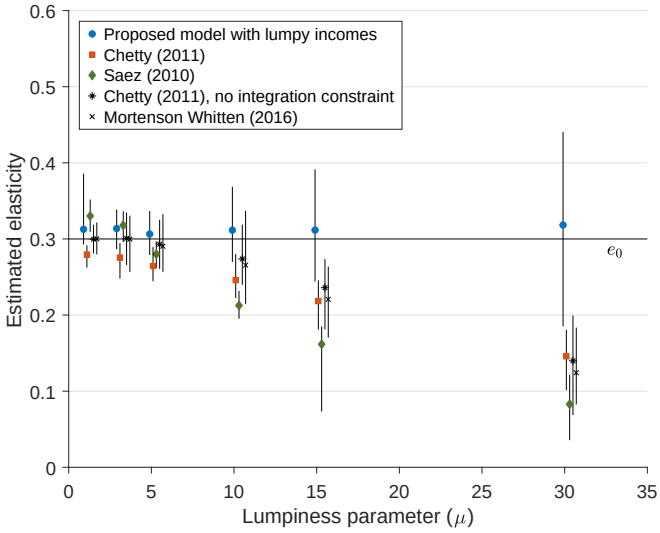
This figure reports the estimated elasticity for one round of simulated data, plotted in green, using both the conventional approach with frictionless income choice (panel a) and our estimation method with lumpy income choice (panel b), assuming different polynomial degrees for the counterfactual (or ability) density. The true ability density of the data-generating process is linear, with a true elasticity value of $e_0 = 0.3$.

Figure A9: Counterfactuals and elasticity estimates using various conventional approaches and our approach, for varying lumpiness parameters

(a) Alternative approaches to constructing counterfactuals



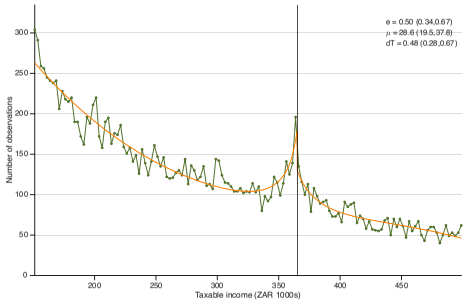
(b) Comparing the elasticity estimates



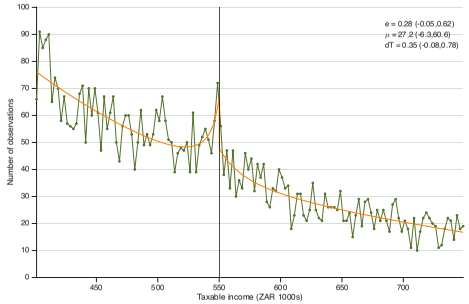
In panel (a), we illustrate the counterfactuals produced under four different conventional bunching approaches to estimating elasticities for a simulated dataset where $\mu = 10$. In panel (b), we simulate 100 rounds of data using a constant elasticity $e_0 = 0.3$ at each value of the lumpiness parameter μ_0 shown in the plots. We then estimate the elasticity $\hat{e}$ using our estimation approach and four conventional bunching estimators. The vertical lines indicate the 95 percent confidence intervals for the $\hat{e}$ estimates. For the conventional methods, we adapt the automated bunching window approach in Bosch, Dekker and Strohmaier (2020) in order to account for each method's approach to constructing a counterfactual distribution.

Figure A10: Differences in lumpiness by tax practitioner usage, CIT middle/upper kinks

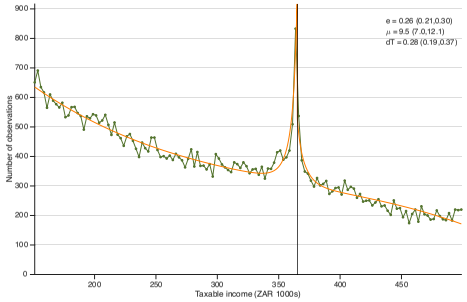
(a) No tax practitioner, middle kink



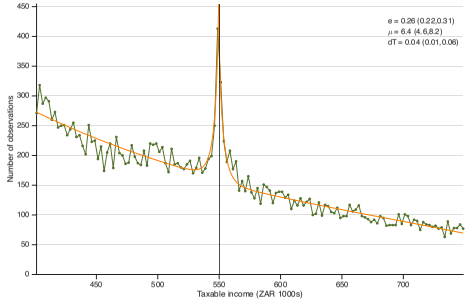
(b) No tax practitioner, upper kink



(c) Uses tax practitioner, middle kink



(d) Uses tax practitioner, upper kink



Green points plot the income histograms of South African Small Businesses. In all panels, orange lines plot the best-fit income density under the uniform sparsity model. Figures report parameter estimates e (elasticity of taxable income), μ (average distance between income opportunities in ZAR 1000s), and dT (the estimated "as-if" discrete change in tax liability at the bracket threshold, in ZAR 1000s). Numbers in parentheses indicate the 95 percent confidence interval on parameter estimates generated by the MLE method. The sample is split into firms that do not use professional tax practitioners to prepare their tax returns in Panels (a)–(c) and firms that use professional tax practitioners in Panels (d)–(e).

Numerical Implementation

Diffuse Bunching with Frictions: Theory and Estimation

Santosh Anagol, Allan Davids, Benjamin B. Lockwood, Tarun Ramadorai

In this document we describe our numerical implementation of the uniform sparsity model. This computation is implemented in the Matlab function `compute_income_density_quad.m`, which computes the model-implied income density for a given tax function, elasticity e, lumpiness μ. In this section we sometimes describe results in terms of the rate parameter $\lambda = 1/\mu$, which can be useful because λ can then be interpreted as a rate of decay in the exponential functions where it repeatedly appears.

We begin with some preliminaries. Section A defines notation and characterizes the bunching distortion in a way that facilitates the numerical computation. Section B describes how we use the second-order Taylor expansions of indirect utility to compute the bunching distortion analytically at high speed using matrix operations. These calculations require relating polynomial approximations of the counterfactual income densities to the implied density of underlying types. Section C describes how this can be done analytically with arbitrary precision.

A Characterizing the Bunching Distortion

In the derivations to follow, we use the notation of the income tax $T(z)$, defined as in equation (4) in the paper:

$$T(z) := \begin{cases} T_0(z) & \text{if } z \leq k, \\ T_1(z) & \text{if } z > k, \end{cases} \tag{35}$$

where $T_0(z)$ and $T_1(z)$ are linear tax functions with a kink or a notch at income threshold k. As in the paper, we define the "notch value" dT as the discrete change in tax liability that occurs at the threshold k:

$$dT := T_1(k) - T_0(k).$$

The specification in equation (35) assumes that earnings of $z = k$ are subject to T_0. If there is a discontinuity (notch) at k, this assumption is not innocuous, and if k is instead subject to T_1, then the inequalities to follow should be adjusted accordingly.

We let $h(z)$ denote the observed density of incomes under the status quo income tax $T(z)$. We define $h_0(z)$ and $h_1(z)$ as the counterfactual income densities that would be observed under the linear tax functions $T_0(z)$ and $T_1(z)$, as in Figure 1. We also define the *composite counterfactual density* $h_{cf}(z)$ as the counterfactual density that would obtain under the linear tax at a given income z, i.e.,

$$h_{cf}(z) = \begin{cases} h_0(z) & \text{if } z \leq k, \\ h_1(z) & \text{if } z > k. \end{cases} \tag{36}$$

We use "H" to denote the cumulative distribution function corresponding to any density h.

When numerically computing the income density $h(z)$, it turns out to be useful to compute the *distortion* to the observed density relative to the composite counterfactual density in (36). We define this distortion as the *bunching mass function* $b(z)$:

$$b(z) = h(z) - h_{cf}(z) = \begin{cases} h(z) - h_0(z) & \text{if } z \leq k, \\ h(z) - h_1(z) & \text{if } z > k. \end{cases} \tag{37}$$

As incomes get very far above or below k, the bunching mass function decays to zero at a predictable rate controlled by $\lambda = 1/\mu$. Thus for given model parameters we can expediently compute the computationally intensive $b(z)$ across only the range of incomes over which $b(z)$ is nonnegligible, and add that to our composite counterfactual $h_{cf}(z)$—which is simple to compute at any income for given polynomial parameters using Lemma 3—to obtain the observed income density $h(z)$.

To compute the bunching mass function $b(z)$, we first define some additional notation. Recall from Proposition 1 that $\Theta(\tilde{z}|n)$ denotes the set of incomes that utility-dominate $\tilde{z}$ for a taxpayer of type n:

$$\Theta(\tilde{z}|n) := \left\{ z \middle| u\left(z - T(z), z \middle| n\right) \geq u\left(\tilde{z} - T(\tilde{z}), \tilde{z} \middle| n\right) \right\}.$$

It is also useful to define the set of incomes that would dominate $\tilde{z}$ under the counterfactual linear income tax $T_i(z)$, where $i = 0$ or 1:

$$\Theta_i(\tilde{z}|n) := \left\{ z \middle| u\left(z - T_i(z), z \middle| n\right) \geq u\left(\tilde{z} - T_i(\tilde{z}), \tilde{z} \middle| n\right) \right\}. \tag{38}$$

We then define the counterfactual utility-dominating income set at z as

$$\Theta_{cf}(\tilde{z}|n) := \begin{cases} \Theta_0(\tilde{z}|n) & \text{if } \tilde{z} \leq k, \\ \Theta_1(\tilde{z}|n) & \text{if } \tilde{z} > k. \end{cases} \tag{39}$$

The bunching mass function depends critically on the *difference* between the (Lebesgue) measure of utility-dominating incomes under the local linear tax $\Theta_i(\tilde{z}|n)$ vs. under actual tax, $\Theta(\tilde{z}|n)$, at each income. We define that difference as

$$\delta(\tilde{z}|n) := \begin{cases} \left|\Theta_0(\tilde{z}|n)\right| - \left|\Theta(\tilde{z}|n)\right| & \text{if } \tilde{z} \leq k, \\ \left|\Theta_1(\tilde{z}|n)\right| - \left|\Theta(\tilde{z}|n)\right| & \text{if } \tilde{z} > k. \end{cases} \tag{40}$$

An example illustration of Θ, Θ_i, and δ is shown in Figure B2. We can now characterize the bunching mass function $b(z)$ using the following lemma.

Lemma 2. *The bunching mass function at income $\tilde{z}$ under the uniform sparsity model is*

$$b(\tilde{z}) = \int_{-\infty}^{\infty} \lambda \exp\left[-\lambda\left|\Theta_{cf}(\tilde{z}|n)\right|\right]\left(\exp[\lambda\delta(\tilde{z}|n)] - 1\right) f(n)dn. \tag{41}$$

Proof. Using equation (13) from the text and the definitions above, we can write the income density at $\tilde{z}$ as

$$\begin{aligned} h(\tilde{z}) &= \int_{-\infty}^{\infty} g(\tilde{z}|n) f(n)dn & (42)\\ &= \int_{-\infty}^{\infty} \lambda \exp\left[-\lambda|\Theta(\tilde{z}|n)|\right] f(n)dn & (43)\\ &= \int_{-\infty}^{\infty} \lambda \exp\left[-\lambda\left(\left|\Theta_{cf}(\tilde{z}|n)\right| - \delta(\tilde{z},n)\right)\right] f(n)dn & (44)\\ &= \int_{-\infty}^{\infty} \lambda \exp\left[-\lambda\left|\Theta_{cf}(\tilde{z}|n)\right|\right] \exp\left[\lambda\delta(\tilde{z},n)\right] f(n)dn & (45) \end{aligned}$$

Similarly, we can express the counterfactual income density at $\tilde{z}$ under the local linear tax $T_i(z)$ as

$$h_{cf}(\tilde{z}) = \int_{-\infty}^{\infty} \lambda \exp\left[-\lambda\left|\Theta_{cf}(\tilde{z}|n)\right|\right] f(n)dn. \tag{46}$$

Using the definition in (37), we can find $b(\tilde{z})$ by subtracting equation (46) from (45), which yields equation (41) in the lemma. □

B Numerically Computing the Income Density

We now describe the numerical calculation of the bunching mass function $b(z)$ defined in equation (37). This requires computing $\delta(\tilde{z}|n)$ as defined in equation (40) across a wide range of productivity types n at each income $\tilde{z}$ and then integrating across those n using equation (41) to find the value of the bunching mass function at that income, $b(\tilde{z})$. It is computationally expedient to use matrix operations to perform these calculations, defining a column vector $\tilde{\boldsymbol{z}}$ of

incomes at which to compute the bunching mass function, and a matrix of types $\boldsymbol{n}$ with each row corresponding to a different income in $\tilde{\boldsymbol{z}}$ and containing the vector of types n across which we will numerically integrate equation (41). For reasons that will become clear below, it is useful to perform these calculations separately for incomes that lie below vs. above the threshold k, and we therefore define two vectors of incomes, $\tilde{\boldsymbol{z}}_{\boldsymbol{0}}$ and $\tilde{\boldsymbol{z}}_{\boldsymbol{1}}$, containing the incomes below and above the threshold k, with corresponding type matrices $\boldsymbol{n}_{\boldsymbol{0}}$ and $\boldsymbol{n}_{\boldsymbol{1}}$.

We now describe how we can use quadratic approximations of indirect utility functions and the quadratic formula to efficiently compute $\delta(\tilde{\boldsymbol{z}}|\boldsymbol{n})$ using element-wise matrix operations.

B.1 Local approximations of indirect utility functions

We employ 2nd-order (quadratic) Taylor approximations of indirect utility function around their target income under a given linear tax $T_i(z)$, which we denote

$$v(z) = az^2 + bz + c, \tag{47}$$

suppressing the dependence on productivity n and tax schedule index i for the moment.

By construction, the agent's target income is the z^* that maximizes this indirect utility function. Setting $v'(z^*) = 0$, we find

$$z^* = -\frac{b}{2a}.$$

Noting that a reform to the marginal income tax rate $T_i'(z)$ affects the linear term b in equation (47), the income response to a change in the marginal tax rate can be found by differentiating z^* with respect to b, i.e., $\frac{\partial z^*}{\partial b} = -\frac{1}{2a}$. We therefore define

$$\zeta := -\frac{1}{2a}, \tag{48}$$

where ζ is the behavioral response of interest for our estimation purposes. Note that ζ is locally proportional to the elasticity of taxable income with respect to the net-of-tax rate, e, through the formula $e = \zeta \cdot \frac{1-T'(z^*)}{z^*}$.

Using equation (47) we now define the quadratic approximations of utility under each of the linear taxes $T_0(z)$ and $T_1(z)$:

$$v_0(z) := a_0 z^2 + b_0 z + c_0$$

and

$$v_1(z) := a_1 z^2 + b_1 z + c_1.$$

The structure of the problem imposes restrictions on the relationship between these

coefficients, and we can use these to rewrite the coefficients a_1, b_1, and c_1 in terms of the coefficients a_0, b_0, and c_0, and the parameters of the tax function $T(z)$.

First, our estimation assumption is that individuals earning in the vicinity of the threshold k exhibit a common behavioral response ζ, and thus from equation (48) we have

$$a_1 = a_0,$$

both of which are equal to $-\frac{1}{2\zeta}$. Therefore we hereafter write a without a subscript 0 or 1.

Second, the marginal tax rate (and thus the linear earning incentive) changes by $(1-t_1) - (1-t_0) = t_0 - t_1$ at the threshold k, and thus

$$b_1 = b_0 - (t_1 - t_0).$$

Note that by the first-order conditions for optimization, we have

$$z_1^* = -\frac{b_1}{2a} = -\frac{b_0 - (t_1 - t_0)}{2a} = z_0^* + \frac{1}{2a}(t_1 - t_0) = z_0^* + \zeta \times [(1-t_1) - (1-t_0)],$$

which is consistent with our definition of ζ quantifying the behavioral response of earnings to changes in the net-of-tax rate.

Third, under a pure kink (when $dT = 0$) utility under T_0 and T_1 is the same under each linear schedule at the threshold: $v_0(k) = v_1(k)$. More generally in the presence of a notch, because indirect utility is quasilinear in consumption, utility changes discontinuously by $-dT$ at the threshold k, implying that $v_1(k) = v_0(k) - dT$. Because the problem is insensitive to level shifts in both v_0 and v_1, we can choose the normalization $v_0(k) = 0$, implying

$$c_0 = -ak^2 - b_0 k. \tag{49}$$

Then imposing $v_1(k) = -dT$, we have

$$\begin{aligned} c_1 &= -dT - ak^2 - b_1 k \\ &= -dT - ak^2 - (b_0 - (t_1 - t_0))k \\ &= c_0 - dT + (t_1 - t_0)k. \end{aligned}$$

Combining these restrictions yields

$$v_1(z) := az^2 + \underbrace{(b_0 - (t_1 - t_0))}_{b_1} z + \underbrace{c_0 - dT + (t_1 - t_0)k}_{c_1}. \tag{50}$$

Up to this point, all derivations have been individual specific and we have notationally ignored any dependence on type. To notationally account for heterogeneity in individuals (and thus heterogeneity in earnings under a given tax) we assume that individuals differ in their productivity or ability type n, and thus we now denote indirect utility as $v_0(z|n)$ with individual-specific coefficients $a(n)$, $b(n)$, and $c(n)$, representing the quadratic approximations of indirect utility around the optimal frictionless choice under each linearized tax schedule for an agent of a given type n. We maintain the assumption that the behavioral response ζ is common across individuals in this vicinity, implying that the coefficient a, which is equal to $-\frac{1}{2\zeta}$, does not actually depend on n. Using this fact, and equation (49), we can write n's indirect utility as follows:

$$v_0(z|n) = az^2 + b_0(n)z \underbrace{-ak^2 - b_0(n)k}_{c_0(n)}$$

Thus the heterogeneity in earnings capacity can be written in terms of differences in the coefficient $b_0(n)$. Moreover, noting that the individual's first-order condition implies $z_0^* = -\frac{b_0(n)}{2a}$, we can adopt z_0^* as a convenient parameterization of individual type, i.e., we define $n = z_0^*$, and thus $b_0(n) = -2az_0^*$. In sum, we have a quadratic specification of indirect utility functions under $T_0(z)$ which we can write entirely in terms of the behavioral response ζ, and each individual's target income z_0^*:

$$v_0(z|z_0^*) = \underbrace{\left(-\frac{1}{2\zeta}\right)}_{a} z^2 + \underbrace{\left(\frac{z_0^*}{\zeta}\right)}_{b_0(n)} z + \underbrace{\frac{1}{2\zeta}k^2 - \frac{z_0^*}{\zeta}k}_{c_0(n)}. \tag{51}$$

Using equation (50), we can also write the quadratic specification of indirect utility under $T_1(z)$ in terms of ζ, z_0^*, and the tax parameters dT, t_0, and t_1: :

$$v_1(z|z_0^*) = \underbrace{\left(-\frac{1}{2\zeta}\right)}_{a} z^2 + \underbrace{\left(\frac{z_0^*}{\zeta} - (t_1 - t_0)\right)}_{b_1(n)} z + \underbrace{\frac{1}{2\zeta}k^2 - \left(\frac{z_0^*}{\zeta} - (t_1 - t_0)\right)k - dT}_{c_1(n)}. \tag{52}$$

These coefficients will allow us to compute $\delta(\tilde{\boldsymbol{z}}|\boldsymbol{n})$ using element-wise matrix operations using the quadratic formula, as described below.

B.2 Baseline case: pure kink

We begin by considering the case of a pure kink, and we first show how we numerically compute $b(\tilde{\boldsymbol{z}}_{\boldsymbol{0}})$, the bunching mass function across our vector of specified incomes below the bracket threshold k. Two examples of a pure kink with $\tilde{z} < k$ are illustrated in Figures 5(a) and 5(b).

The computation is conceptually symmetric for computing $b(\tilde{\boldsymbol{z}}_{\mathbf{1}})$, the bunching mass at incomes above the threshold k, but performing the calculations separately allows us to use $z_0^*(n)$ and $z_1^*(n)$ as our index of types when computing $b(\tilde{\boldsymbol{z}}_{\mathbf{0}})$ vs. $b(\tilde{\boldsymbol{z}}_{\mathbf{1}})$, respectively.

Recall that we use $\underline{Z}(\tilde{z}|n)$ and $\overline{Z}(\tilde{z}|n)$ to denote the lower and upper incomes that give type n the same utility as $\tilde{z}$ under the tax function $T(z)$. We extend this notation to define $\underline{Z}_0(\tilde{z}|n)$ and $\overline{Z}_0(\tilde{z}|n)$ as the lower and upper incomes that give type n the same utility as $\tilde{z}$ under the linear tax function $T_0(z)$. Similarly, $\underline{Z}_1(\tilde{z}|n)$ and $\overline{Z}_1(\tilde{z}|n)$ denote the lower and upper incomes that give type n the same utility as $\tilde{z}$ under the linear tax function $T_1(z)$.

Constructing the type grid $\boldsymbol{n_0}$

Note that in Figures 5(a) and 5(b), the agent's target income $z_0^*(n)$ is above $\tilde{z}$ under the local linear tax $T_0(z)$. That is by design: in cases where z_0^* is below $\tilde{z}$, the entire dominating set of incomes is below the threshold k, and thus the dominating income region under the kinked tax is the same as it would be under the linear tax $T_0(z)$, so $\delta(\tilde{z}|n)$ is trivially zero and these type-income combinations do not contribute to the bunching mass function. In fact, because z_0^* is our index of types, under quadratic indirect utility any type with z_0^* below the *midpoint* between $\tilde{z}$ and k will have a dominating set of incomes that is entirely below the threshold k. This insight allows us to judiciously choose the bounds of the type vector $\boldsymbol{n}$ at each $\tilde{z}$ to restrict computations to combinations of n and $\tilde{z}$ that contribute positively to the bunching mass function. Namely, we set the minimal type for each $\tilde{z}$ to be $n_{min}(\tilde{z}) = k - \frac{k-\tilde{z}}{2} = \frac{k+\tilde{z}}{2}$, so that the leftmost column of the type matrix $\boldsymbol{n_0}$ is $\frac{k+\tilde{\boldsymbol{z}}_{\mathbf{0}}}{2}$.

In principle there is no upper bound to the set of types that contribute to the bunching mass function at a given $\tilde{z} < k$. However, for types very far above $\tilde{z}$ this contribution becomes negligible. From equation (15) in Proposition 2, under a linear tax the contribution from type with target income $z_0^* > \tilde{z}$ is $\lambda \exp\left[-\lambda \cdot 2(z_0^* - \tilde{z})\right]$. We therefore set $n_{max}(\tilde{z})$ to satisfy

$$\varepsilon = \lambda \exp\left[-\lambda \cdot 2(n_{max}(\tilde{z}) - \tilde{z})\right] \tag{53}$$

$$\Rightarrow n_{max}(\tilde{z}) = \tilde{z} + \frac{\log(\varepsilon/\lambda)}{-2\lambda} \tag{54}$$

for some small $\varepsilon > 0$, such as $\varepsilon = 10^{-7}$. Thus the rightmost column of the type matrix $\boldsymbol{n_0}$ is $\tilde{\boldsymbol{z}}_{\mathbf{0}} + \frac{\log(\varepsilon/\lambda)}{-2\lambda}$. We complete the type matrix by linearly interpolating between the leftmost and rightmost columns using a large number of intermediate columns, such as 10,000. A higher number of intermediate gridpoints produces smoother variation in the eventual income density in response to variations in model parameters, facilitating convergence of a maximum likelihood solver.

Computing $\boldsymbol{\delta(\tilde{z}_0|n_0)}$

With quadratic indirect utility, $\underline{Z}_0(\tilde{z}|n)$ and $\overline{Z}_0(\tilde{z}|n)$ are equidistant from the target income $z_0^*(n)$. This implies that (for $\tilde{z} < k$)

$$\Theta_{cf}(\tilde{z}|n) = 2(z_0^*(n) - \tilde{z}),$$

which will be employed when using equation (41) to compute the density distortion.

The *change* in the set of utility-dominating incomes created by the kink, relative to this counterfactual $\Theta_{cf}(\tilde{z}|n)$, is

$$\delta(\tilde{z}|n) = \overline{Z}_0(\tilde{z}|n) - \overline{Z}_1(\tilde{z}|n). \tag{55}$$

Note that in the case of a pure kink $\delta(\tilde{z}|n) \leq 0$ because the dominating set of incomes shrinks due to the kink. (This need not be the case in the presence of a notch, as we will see below.) We can find explicit expressions for both $\overline{Z}_0(\tilde{z}|n)$ and $\overline{Z}_1(\tilde{z}|n)$ in terms of model parameters.

The equidistance of $\underline{Z}_0(\tilde{z}|n)$ and $\overline{Z}_0(\tilde{z}|n)$ from $z_0^*(n)$ implies

$$\overline{Z}_0(\tilde{z}|n) = z_0^*(n) + \left[z_0^*(n) - \tilde{z}\right]. \tag{56}$$

By construction, $\overline{Z}_1(\tilde{z}|n)$ is the (upper) income that gives the same utility under v_1 as $\tilde{z}$ gives under v_0, so we can write:

$$v_0(\tilde{z}|n) = v_1\left(\overline{Z}_1(\tilde{z}|n)\middle|n\right) \tag{57}$$

$$0 = a\overline{Z}_1(\tilde{z}|n)^2 + b_1(n)\overline{Z}_1(\tilde{z}|n) + c_1(n) - v_0(\tilde{z}|n). \tag{58}$$

We can rearrange to solve for the unknown $\overline{Z}_1(\tilde{z}|n)$ using the quadratic formula:

$$\overline{Z}_1(\tilde{z}|n) = \frac{-b_1(n) - \sqrt{b_1(n)^2 - 4a(c_1(n) - v_0(\tilde{z}|n))}}{2a}. \tag{59}$$

The choice of the negative root ensures that this calculation returns $\overline{Z}_1(\tilde{z}|n)$ rather than $\underline{Z}_1(\tilde{z}|n)$, which also gives the same utility as $v_0(\tilde{z}|n)$. Using equations (51) and (52), this provides an explicit formula for $\overline{Z}_1(\tilde{z}|n)$ in terms of model parameters.

Computing the bunching mass function $\boldsymbol{b(\tilde{z}_0)}$

Using equations (55), (56), and (59), we can compute the value inside the integral in equation (41) using element-wise matrix operations corresponding to each cell of the matrix $\boldsymbol{n_0}$. This

calculation requires weighting the integrand by the type density $f(n)$, which in our case corresponds to the density of target incomes $h_0^*(z)$. Given a polynomial specification of the counterfactual income density $h_0(z)$, we can compute this type density using the relationship derived in Section C below. In practice, the difference between the polynomial coefficients of $h_0^*(z)$ and $h_0(z)$ is negligible when the high-order coefficients are sufficiently small, suggesting that it is often a good approximation to use the polynomial coefficients of $h_0(z)$ without transformation—an approximation that we employ in our implementation. Our code also allows for the specification of a provided, non-polynomial counterfactual density $h_0(z)$, although in this case it may not be a good approximation to use the same density for $h_0^*(z)$ and $h_0(z)$.[42]

We then integrate horizontally (across columns) to obtain the bunching mass function $b(\tilde{\boldsymbol{z}}_{\boldsymbol{0}})$. We perform this integration numerically using the trapezoidal rule.

Computing the income density $h(\tilde{\boldsymbol{z}}_{\boldsymbol{0}})$

We can now compute the income density $h(\tilde{\boldsymbol{z}}_{\boldsymbol{0}})$ by adding the bunching mass function $b(\tilde{\boldsymbol{z}}_{\boldsymbol{0}})$ to the counterfactual income density $h_{cf}(\tilde{\boldsymbol{z}}_{\boldsymbol{0}})$. This computation is trivial when the the polynomial coefficients of the counterfactual density $h_0(z)$ are provided, because we can simply add $b(z)$ to the specified $h_0(z)$ to obtain $h(z)$. This completes the computation of the income density $h(\tilde{\boldsymbol{z}}_{\boldsymbol{0}})$ for the case of a pure kink across all incomes below the threshold k.

Computing the income density $h(\tilde{\boldsymbol{z}}_{\boldsymbol{1}})$ across incomes above k

To compute the income density $h(\tilde{\boldsymbol{z}}_{\boldsymbol{1}})$ for the case of a pure kink across all incomes above the threshold k, we use the same method as above with the following modifications:

- We use $z_1^*(n)$ as our index of types, rather than $z_0^*(n)$.
- For our type grid $\boldsymbol{n}_{\boldsymbol{1}}$, we set the $\tilde{z}$-dependent lower and upper bounds (the leftmost and rightmost columns) to be $\tilde{\boldsymbol{z}}_{\boldsymbol{1}} - \frac{\log(\varepsilon/\lambda)}{-2\lambda}$ and $\frac{k+\tilde{\boldsymbol{z}}_{\boldsymbol{1}}}{2}$, respectively.
- We compute $\delta(\tilde{z}|n) = \underline{Z}_1(\tilde{z}|n) - \underline{Z}_0(\tilde{z}|n)$, with $\underline{Z}_1(\tilde{z}|n) = z_1^*(n) + \left[\tilde{z} - z_1^*(n)\right]$ and

$$\underline{Z}_0(\tilde{z}|n) = \frac{-b_0(n) + \sqrt{b_0(n)^2 - 4a(c_0(n) - v_1(\tilde{z}|n))}}{2a}. \tag{60}$$

[42] From equation (13) in the paper, we have $h_0(z) = \int_n g(z|n) f(n) dn$, so the counterfactual density $h_0(z)$ can be viewed as the convolution of the type-conditional density $g(z|n)$ and the type density $h_0^*(z)$. Therefore when specifying a non-polynomial counterfactual density $h_0(z)$, the type density $h_0^*(z)$ can be obtained by deconvolving $h_0(z)$ using the Fourier transform.

B.3 Case with a positive notch ($dT > 0$)

With a positive notch, the indirect utility function jumps down discontinuously at k, as illustrated in Figure A2. This complicates the calculation of $\delta(\tilde{z}|n)$ because the dominating set of incomes may be dominated by k (as in Figure A2(a)) or it may consist of two separate intervals (as in Figure A2(b)).

Constructing the type grids n_0 and n_1

For incomes $\tilde{z}$ below k, the minimal type at which the dominating set of incomes begins to be distorted by the notch is the same as in the case of a pure kink: the midpoint between $\tilde{z}$ and k. Thus in this case we use the same bounds for the type grid $\boldsymbol{n_0}$ as in the case of a pure kink described above.

For incomes $\tilde{z}$ above k, we use the same lower bound on the type matrix $\boldsymbol{n_1}$ as in the case of a pure kink. However, the maximal type at which the dominating set of incomes stops being distorted by the notch isn't the midpoint between $\tilde{z}$ and k, as with a pure kink. Instead it is the highest type that may prefer an income opportunity at k (on the "attractive" side of the notch) over an income in the set $\Theta_1(\tilde{z}|n)$ that would be relevant under the counterfactual linear tax $T_1(z)$. That is, $n_{max}(\tilde{z})$ satisfies

$$\begin{aligned} v_0(k|n_{max}(\tilde{z})) &= v_1(\tilde{z}|n_{max}(\tilde{z})) \\ ak^2 + b_0(n_{max}(\tilde{z}))k + c_0(n_{max}(\tilde{z})) &= a\tilde{z}^2 + b_1(n_{max}(\tilde{z}))\tilde{z} + c_1(n_{max}(\tilde{z})) \\ b_0(n_{max}(\tilde{z}))(k-\tilde{z}) &= a(\tilde{z}^2 - k^2) - dT \\ n_{max}(\tilde{z}) &= \frac{\zeta}{k-\tilde{z}}\left[a(\tilde{z}^2 - k^2) - dT\right], \end{aligned}$$

where we have used the coefficient derivations in equations (51) and (52).

As with the pure kink case, we then linearly interpolate between the leftmost and rightmost columns of the type grids $\boldsymbol{n_0}$ and $\boldsymbol{n_1}$ across a large number of intermediate columns.

Finally, for the purposes of numerically integrating across types to obtain $b(z)$, it is useful to add a few particular "threshold" types to the type grids $\boldsymbol{n_0}$ and $\boldsymbol{n_1}$ at which dominating income range changes qualitatively. One such threshold type is illustrated in Figure A2(a), where the upper bound of the dominating income range switches from being k (for lower types) to $\overline{Z}_1(\tilde{z}|n)$

(for higher types). Given $\tilde{z} < k$, this threshold is the type $\hat{n}$ such that

$$\begin{aligned} v_0(\tilde{z}|\hat{n}) &= v_1(k|\hat{n}) \\ a\tilde{z}^2 + b_0(\hat{n})\tilde{z} + c_0(\hat{n}) &= ak^2 + b_1(\hat{n})k + c_1(\hat{n}) \\ b_0(\hat{n})(\tilde{z} - k) &= a(k^2 - \tilde{z}^2) - dT \\ \hat{n} &= \frac{\zeta}{\tilde{z} - k}\left[a(k^2 - \tilde{z}^2) - dT\right]. \end{aligned}$$

The two other threshold types are the ones at which the dominating income range switches from being a single interval to being two separate intervals. These cases are relevant only for incomes $\tilde{z}$ on the "attractive" side of the notch, where the dominating income range is a single interval. Consider the case illustrated in Figure A2(b), with $\tilde{z} > k$. Fixing $\tilde{z}$ and k, if the type is reduced from the illustrated case, the red and blue indirect utility curves will shift down until the portion of the dominating income range below k shrinks to a singleton, and then disappears. This threshold type $\hat{n}$ satisfies

$$\begin{aligned} v_0(k|\hat{n}) &= v_1(\tilde{z}|\hat{n}) \\ ak^2 + b_0(\hat{n})k + c_0(\hat{n}) &= a\tilde{z}^2 + b_1(\hat{n})\tilde{z} + c_1(\hat{n}) \\ b_0(\hat{n})(k - \tilde{z}) &= a(\tilde{z}^2 - k^2) - dT \\ \hat{n} &= \frac{\zeta}{k - \tilde{z}}\left[a(\tilde{z}^2 - k^2) - dT\right]. \end{aligned}$$

Finally, there is also a threshold type at which the portion of the dominating region under the counterfactual indirect utility function disappears on the "far" side of the threshold. For example, in Figure A2(b), suppose we were interested in computing the density at a $\tilde{z} < k$ that happens to equal the value $\underline{Z}_0(\tilde{z}|c)$ plotted in that figure. Then there would be a type that gives rise to v_0 and v_1 indirect utility functions like the ones plotted. Notice that if the type is decreased while holding fixed $\tilde{z}$ and k, the portion of the dominating income range above k will shrink to a singleton, and then disappear. This threshold type $\hat{n}$ satisfies $v_0(\tilde{z}, \hat{n}) = v_1(z_1^*(\hat{n}), \hat{n})$, implying

$$a\tilde{z}^2 + b_0(\hat{n})\tilde{z} + c_0(\hat{n}) = az_1^*(\hat{n})^2 + b_1(\hat{n})z_1^*(\hat{n}) + c_1(\hat{n})$$

Substituting the expressions for the coefficients previously derived we have

$$\left(-\frac{1}{2\zeta}\right)\tilde{z}^2 + \left(\frac{z_0^*}{\zeta}\right)\tilde{z} = \left(-\frac{1}{2\zeta}\right)z_1^*(\hat{n})^2 + \left(\frac{z_0^*}{\zeta} - (t_1 - t_0)\right)z_1^*(\hat{n}) + \underbrace{(t_1 - t_0)k - dT}_{c_1(\hat{n}) - c_0(\hat{n})}$$

By assumption of constant local income shifts we can substitute $z_1^*(\hat{n}) = z_0^*(\hat{n}) + \Delta z$, giving

$$\left(-\frac{1}{2\zeta}\right)\tilde{z}^2 + \left(\frac{z_0^*}{\zeta}\right)\tilde{z} = \left(-\frac{1}{2\zeta}\right)(z_0^* + \Delta z)^2 + \left(\frac{z_0^*}{\zeta} - (t_1 - t_0)\right)(z_0^* + \Delta z) + (t_1 - t_0)k - dT$$

Recalling that we use z_0^* as our index of types over this range of incomes with $\tilde{z} < k$, we substitute $\hat{n} = z_0^*$ and rearrange to obtain a quadratic equation in $\hat{n}$:

$$0 = \frac{1}{2\zeta}\hat{n}^2 - \left(\frac{\tilde{z}}{\zeta} + (t_1 - t_0)\right)\hat{n} + \frac{1}{2\zeta}(\tilde{z}^2 - (\Delta z)^2) + (t_1 - t_0)(k - \Delta z) - dT$$

We can then solve for $\hat{n}$ using the quadratic formula.

Finally, we add these threshold types to our type grids $\boldsymbol{n_0}$ and $\boldsymbol{n_1}$ and sort them horizontally so that they appear at their appropriate locations.

Computing $\boldsymbol{\delta(\tilde{z}_0|n_0)}$

Having constructed the type grids $\boldsymbol{n_0}$ and $\boldsymbol{n_1}$, we compute $\delta(\tilde{z}|n)$ for each $\tilde{z}$ and n using the same method as in the case of a pure kink, with two modifications:

- As illustrated in Figure A2(a), the bound of the dominating interval on the "far" side of the threshold may now be k itself, rather than $\overline{Z}_1(\tilde{z}|n)$ (for incomes $\tilde{z}$ below k) or $\underline{Z}_0(\tilde{z}|n)$ (for incomes $\tilde{z}$ above k).

- As illustrated in Figure A2(b), there may be a "hole" in the dominating interval between $\overline{Z}_0(\tilde{z}|n)$ and k (for incomes $\tilde{z}$ below k) or between k and $\underline{Z}_1(\tilde{z}|n)$ (for incomes $\tilde{z}$ above k). In this case, we still compute the critical values $\overline{Z}_0(\tilde{z}|n)$ and $\underline{Z}_1(\tilde{z}|n)$ as before, using the quadratic formula when necessary.

B.4 Case with a negative notch ($dT < 0$)

To compute the income density with a negative notch, we use the same procedure described above, except that the indirect utility now jumps *up* discontinuously at k. We therefore modify the above procedure accordingly.

B.5 Estimating Polynomial Coefficients

The preceding description assumed that polynomial coefficients θ are given. Our estimation code also provides an alternative behavior, in which the coefficients are estimated from the data for given parameters and a given tax schedule. This amounts to the maximum likelihood estimation problem presented in Section 2.6:

$$\max_{e,\mu,dT,\theta} L(e,\mu,dT,\theta). \tag{61}$$

Performing maximum likelihood estimation with this likelihood function will not result in an interior maximum, however, because we have imposed no constraint on the integral of the income density function $h(z|e,\mu,dT,\theta)$. For example, the solver can make equation (20) arbitrarily high by letting the polynomial intercept θ_0 become large. To address this, we can normalize the population density within a desired range $[z_{min}, z_{max}]$ around the bracket threshold (e.g., the income range reflected in the empirical support of the taxable income distribution). In principle, we could then perform maximum likelihood estimation by computationally searching for the vector (e,μ,θ,dT) that solves the following constrained maximization problem:

$$\max_{e,\mu,\theta,dT} \sum_i \log h(X_i = z|e,\mu,dT,\theta) \quad \text{subject to} \quad \int_{z_{min}}^{z_{max}} h(z|e,\mu,dT,\theta)dz = 1. \tag{62}$$

This estimation can be implemented directly with raw microdata on incomes reported to the tax authority.[43] In many settings, however, privacy or logistical constraints restrict the analyst to operate with a binned histogram of incomes; that is the usual data input in the bunching literature. The approach in equation (62) can be modified for use with binned data using interval censoring by letting i index bins (rather than observations) and replacing the maximand in equation (62) with $\sum_i H_i \log h(Z_i|e,\mu,\theta,dT)$, where (Z_i, H_i) denotes the income and frequency values for each bin i, and letting $h(Z_i)$ denote the model-predicted probability of drawing an observation from bin Z_i. We adopt this modification for our estimations in the simulations and empirical exercises that follow.

Computationally solving the constrained maximization problem in equation (62) presents a challenge. The likelihood function is

$$h(z|e,\mu,dT,\theta) = \int_{-\infty}^{\infty} g(z|n,e,\mu,dT) f(n|\theta) dn. \tag{63}$$

This is difficult because numerically integrating over a large grid of types n is time consuming, and the parameter space is very large when allowing for even a cubic polynomial, which we adopt as our baseline specification.

The problem can be converted into one that is numerically tractable by viewing the selection of the polynomial coefficients θ as an inner problem that is computed conditional on

[43] With such microdata, our method does not require the researcher to specify a bin-size before estimation, effectively removing a researcher degree of freedom relative to the conventional approach.

the other parameters, so that we can write the maximum likelihood problem as

$$\max_{e,\mu,dT} \sum_i H_i \log h(Z_i = z|e,\mu,\theta(e,\mu,dT)), \tag{64}$$

with the integration constraint (62) enforced by appropriate selection of polynomial coefficients $\theta(e,\mu,dT)$. If the inner function $\theta(e,\mu,dT)$ were selected to solve the constrained maximization in equation (62), then this approach would amount to concentrating out the parameter vector θ. For numerical expediency, we adopt a close approximation to that, where we exploit the structure of the problem in a way that allows us to compute $\theta(e,\mu,dT)$ very quickly using polynomial regression. In effect, we select θ to minimize the sum of squared differences between the observed histogram (normalized to sum to one) and the predicted income density:

$$\theta(e,\mu,dT) = \min_{\theta} \sum_i \left(\frac{H_i}{\sum_j H_j} - h(Z_i|e,\mu,dT) \right)^2. \tag{65}$$

To illustrate, this problem can be written in regression form as follows for the case in which $f(n|\theta)$ is cubic, where the θ coefficients are selected to minimize the sum of squared residuals $\sum_i \varepsilon_i^2$:

$$\begin{aligned} \frac{H_i}{\sum_j H_j} &= h(Z_i|e,\mu,dT) + \varepsilon_i \\ &= \int_{-\infty}^{\infty} g(Z_i|n|e,\mu,dT) f(n|\theta) dn + \varepsilon_i \\ &= \int_{-\infty}^{\infty} g(Z_i|n,e,\mu,dT) \left(\theta_0 + \theta_1 n + \theta_2 n^2 + \theta_3 n^3\right) dn + \varepsilon_i \\ &= \left[\int_{-\infty}^{\infty} g(Z_i|n,e,\mu,dT) dn\right] \theta_0 + \left[\int_{-\infty}^{\infty} g(Z_i|n,e,\mu,dT) n dn\right] \theta_1 \\ &\quad + \left[\int_{-\infty}^{\infty} g(Z_i|n,e,\mu,dT) n^2 dn\right] \theta_2 + \left[\int_{-\infty}^{\infty} g(Z_i|n,e,\mu,dT) n^3 dn\right] \theta_3 + \varepsilon_i. \end{aligned} \tag{66}$$

The terms in brackets require only a single numerical computation of $g(z|n,e,\mu,dT)$, after which the θ polynomial coefficients can be calculated efficiently using standard matrix inversion. This facilitates rapidly computing equation (64) searching over only the three parameters e, μ, and dT. The integration constraint in equation (62) can be enforced by using a two-step procedure in which, after selecting a provisional θ vector to solve equation (65), we adjust the intercept θ_0 so that the constraint holds exactly.

C Relating the Type Density and Income Density

In this appendix, we show how the counterfactual income density $h_0(z)$ can be related to $h_0^*(z)$, the density of types when target income z_0^* is used as the type index.

Invoking Assumption 1 (quadratic indirect utility) the measure of Θ_{cf} in Lemma 2 takes a simple form:

$$\left|\Theta_i\left(\tilde{z}|n\right)\right| = 2\left|\tilde{z} - z_i^*(n)\right|. \tag{67}$$

Motivated by this fact, we note that we can use the target income that each type would choose, $z_i^*(n)$, as an index for type itself. Noting that types map to incomes under a linear tax, we let types be indexed by z_0^* or z_1^* when computing the bunching distortion at incomes below or above k, respectively. We will then use $h_0^*(z)$ to denote the density of target-income-indexed types, so that $\int_{n_1}^{n_2} f(n)dn = \int_{z_0^*(n_1)}^{z_0^*(n_2)} h_0^*(z)dz$ for any n_1, n_2, and similarly for $h_1^*(z)$.

Throughout this discussion we maintain the assumption from the paper that the change in marginal tax rates at k induces a locally constant shift Δz in target incomes, i.e., $z_1^*(n) = z_0^*(n) + \Delta z$ for all n earning in the vicinity of k. An immediate implication is that $H_0^*(\tilde{z}) = H_1^*(\tilde{z} - \Delta z)$ for all $\tilde{z}$ in the vicinity of z. A slightly less immediate implication is that the counterfactual distributions bear the same relationship:

Lemma 3. *With constant target income response Δz in the vicinity of k, the counterfactual income CDFs and densities satisfy*

$$H_0(z) = H_1\left(z - \Delta z\right) \tag{68}$$

and

$$h_0(z) = h_1\left(z - \Delta z\right). \tag{69}$$

Proof. From equation (13)—though written using target incomes to index types—and Proposition 4, we can write the observed income CDF under $T_0(z)$ as

$$H_0(z) = \int_{\tilde{z}=-\infty}^{z} h_0(\tilde{z})d\tilde{z} = \int_{\tilde{z}=-\infty}^{z}\int_{z_0^*=-\infty}^{\infty} \lambda \exp\left[-\lambda \cdot 2\left|\tilde{z} - z_0^*\right|\right] h_0^*(z_0^*)dz_0^*d\tilde{z}. \tag{70}$$

Similarly we can write

$$H_1(z) = \int_{\tilde{z}=-\infty}^{z} h_1(\tilde{z})d\tilde{z} = \int_{\tilde{z}=-\infty}^{z}\int_{z_1^*=-\infty}^{\infty} \lambda \exp\left[-\lambda \cdot 2\left|\tilde{z} - z_1^*\right|\right] h_1^*(z_1^*)dz_1^*d\tilde{z}, \tag{71}$$

and therefore

$$H_1(z - \Delta z) = \int_{\tilde{z}=-\infty}^{z-\Delta z}\int_{z_1^*=-\infty}^{\infty} \lambda \exp\left[-\lambda \cdot 2\left|\tilde{z} - z_1^*\right|\right] h_1^*(z_1^*)dz_1^*d\tilde{z}, \tag{72}$$

Using a change of variables with $\check{z} = \tilde{z} + \Delta z$ we have

$$H_1(z - \Delta z) = \int_{\check{z}=-\infty}^{z} \int_{z_1^*=-\infty}^{\infty} \lambda \exp\left[-\lambda \cdot 2\left|\check{z} - \Delta z - z_1^*\right|\right] h_1^*(z_1^*) dz_1^* d\check{z}. \tag{73}$$

By the assumption of constant target income response, in the vicinity of the tax kink, *target incomes* under the linear tax $T_1(z)$ are lower than under $T_0(z)$ by Δz. That is, $z_1^* = z_0^* - \Delta z$. Substituting this into (73) gives

$$H_1(z - \Delta z) = \int_{\check{z}=-\infty}^{z} \int_{z_0^*=-\infty}^{\infty} \lambda \exp\left[-\lambda \cdot 2\left|\check{z} - \Delta z - z_0^* + \Delta z\right|\right] h_1^*(z_0^* - \Delta z) dz_0^* d\check{z} \tag{74}$$

$$= \int_{\check{z}=-\infty}^{z} \int_{z_0^*=-\infty}^{\infty} \lambda \exp\left[-\lambda \cdot 2\left|\check{z} - z_0^*\right|\right] h_1^*(z_0^* - \Delta z) dz_0^* d\check{z}. \tag{75}$$

Because target income rankings are preserved under a constant shift, $z_1^* = z_0^* - \Delta z$ implies

$$H_0^*(z_0^*) = H_1^*(z_1^*) = H_1^*(z_0^* - \Delta z). \tag{76}$$

and differentiating with respect to z_0^* gives

$$h_0^*(z_0^*) = h_1^*(z_0^* - \Delta z). \tag{77}$$

Substituting (77) into (75) gives the first equation in the lemma, (68), and differentiating (68) with respect to z gives (69), proving the lemma. □

This result is useful because if we believe that these counterfactual distributions are approximated by a polynomial function of some order around k, then the lemma implies that the polynomial function approximating h_0 is the same as the polynomial function approximating h_1 but shifted by Δz.

Next, we produce the following lemma, which demonstrates that when the observed income density under a linear tax can be written as a polynomial, then the density of *target incomes* under that tax can also be written as a polynomial, with a simple relationship between the coefficients of the two polynomials. This lemma is useful to estimate the counterfactual θ distribution in the absence of the kink, because in many settings the analyst does not observe the density of counterfactual incomes. (We note that in some applications the analyst does observe this distribution, for example, in our application to the data in Ding et al. (2025)).

Lemma 4. *Under the uniform sparsity model, if the Taylor approximation around income $\tilde{z}$ of*

the observed income density under a given linear tax $T_i(z)$ has coefficients α_j, i.e.,

$$h_i(z) \approx \alpha_0 + \alpha_1(z-\tilde{z}) + \frac{\alpha_2}{2}(z-\tilde{z})^2 + \ldots, \tag{78}$$

then the Taylor approximation of the (unobservable) density of target incomes, $h_i^*(z)$, around $\tilde{z}$ has coefficients α_j^*, where

$$\alpha_j^* = \alpha_j - \frac{\alpha_{j+2}}{(2\lambda)^2}. \tag{79}$$

Proof. Under the uniform sparsity model, the density of observed and target incomes are related by the following equation (the subscript i indexing the linear tax is suppressed for readability, as the result turns out not to depend on the tax, provided that it is linear):

$$h(z) = \int_{z^*=-\infty}^{z} \lambda \exp\left[-2\lambda\left(z-z^*\right)\right] h^*(z^*)dz^* + \int_{z^*=z}^{\infty} \lambda \exp\left[-2\lambda\left(z^*-z\right)\right] h^*(z^*)dz^*. \tag{80}$$

Employing a change of variables in each integral, with $x = z - z^*$ (so $\frac{dx}{dz^*} = -1$) and $y = z^* - z$ (so $\frac{dy}{dz^*} = 1$), we can write

$$h(z) = \frac{1}{2}\left\{\int_{x=0}^{\infty} 2\lambda \exp\left[-2\lambda x\right] h^*(z-x)dx + \int_{y=0}^{\infty} 2\lambda \exp\left[-2\lambda y\right] h^*(z+y)dy\right\}. \tag{81}$$

Intuitively, this equation states that the observed density at z is a weighted average of the target income density around z, where the weights decline exponentially (with parameter 2λ) as the target income gets farther from z in either direction.

This insight allows us to approximate the relationship between $h(z)$ and $h^*(z)$ around $\tilde{z}$ arbitrarily well using Taylor expansions. Let $\hat{h}^*(z)$ denote the Taylor expansion of the target income density $h^*(z)$ around income $\tilde{z}$:

$$\hat{h}^*(z) = \alpha_0^* + \alpha_1^*(z-\tilde{z}) + \frac{\alpha_2^*}{2}(z-\tilde{z})^2 + \ldots \tag{82}$$

where $\alpha_0^* = h^*(\tilde{z})$, $\alpha_1^* = h^{*\prime}(\tilde{z})$, $\alpha_2^* = h^{*\prime\prime}(\tilde{z})$, etc., with "$*$" superscripts indicating that these are the Taylor expansion coefficients of the *target* income (type) density. Substituting this approximation into equation (81) evaluated at $z = \tilde{z}$, we have

$$\begin{aligned} h(\tilde{z}) = \frac{1}{2}\Bigg\{&\int_{x=0}^{\infty} 2\lambda \exp\left[-2\lambda x\right]\left(\alpha_0^* - \alpha_1^* x + \frac{\alpha_2^*}{2}x^2 - \frac{\alpha_3^*}{3!}x^3 + \frac{\alpha_4^*}{4!}x^4 + \ldots\right)dx + \\ &\int_{y=0}^{\infty} 2\lambda \exp\left[-2\lambda y\right]\left(\alpha_0^* + \alpha_1^* y + \frac{\alpha_2^*}{2}y^2 + \frac{\alpha_3^*}{3!}y^3 + \frac{\alpha_4^*}{4!}y^4 + \ldots\right)dy\Bigg\}. \end{aligned} \tag{83}$$

Collecting terms by coefficients, we can rewrite this as

$$\begin{aligned}
h(\tilde{z}) = \; & \alpha_0^* \cdot \frac{1}{2}\left\{\int_{x=0}^{\infty} 2\lambda \exp\left[-2\lambda x\right] dx + \int_{y=0}^{\infty} 2\lambda \exp\left[-2\lambda y\right] dy\right\} \\
& + \alpha_1^* \cdot \frac{1}{2}\left\{-\int_{x=0}^{\infty} 2\lambda \exp\left[-2\lambda x\right] x dx + \int_{y=0}^{\infty} 2\lambda \exp\left[-2\lambda y\right] y dy\right\} \\
& + \frac{\alpha_2^*}{2} \cdot \frac{1}{2}\left\{\int_{x=0}^{\infty} 2\lambda \exp\left[-2\lambda x\right] x^2 dx + \int_{y=0}^{\infty} 2\lambda \exp\left[-2\lambda y\right] y^2 dy\right\} \\
& + \frac{\alpha_3^*}{3!} \cdot \frac{1}{2}\left\{-\int_{x=0}^{\infty} 2\lambda \exp\left[-2\lambda x\right] x^3 dx + \int_{y=0}^{\infty} 2\lambda \exp\left[-2\lambda y\right] y^3 dy\right\} \\
& + \frac{\alpha_4^*}{4!} \cdot \frac{1}{2}\left\{\int_{x=0}^{\infty} 2\lambda \exp\left[-2\lambda x\right] x^4 dx + \int_{y=0}^{\infty} 2\lambda \exp\left[-2\lambda y\right] y^4 dy\right\} \\
& + \ldots
\end{aligned}$$

All terms multiplying odd-numbered Taylor coefficients (α_1^*, α_3^*, etc.) consist of two equal integrals with opposite signs, and they therefore cancel. The term multiplying α_0^* contains two identical integrals, each of which is an integral over an exponential density function, and therefore they each integrate to 1. More generally, terms multiplying even-numbered Taylor coefficients (α_0^*, α_2^*, etc.) consist of two equal integrals, and therefore their average is equal to either of them. Putting this together, we can simplify the above expression to

$$h(\tilde{z}) = \alpha_0^* + \frac{\alpha_2^*}{2!} \cdot \int_{x=0}^{\infty} 2\lambda \exp\left[-2\lambda x\right] x^2 dx + \frac{\alpha_4^*}{4!} \cdot \int_{x=0}^{\infty} 2\lambda \exp\left[-2\lambda x\right] x^4 dx + \ldots \tag{84}$$

We can now make use of a convenient property of exponential distributions: when x is exponentially distributed with rate parameter β, the expectation of x^m is $m!/\beta^m$. Each of the above integrals represents such an expectation, with x exponentially distributed with rate parameter 2λ. Therefore we can rewrite equation (84) as

$$\begin{aligned}
h(\tilde{z}) &= \alpha_0^* + \frac{\alpha_2^*}{2!} \cdot \frac{2!}{(2\lambda)^2} + \frac{\alpha_4^*}{4!} \cdot \frac{4!}{(2\lambda)^4} + \ldots \\
&= \alpha_0^* + \frac{\alpha_2^*}{(2\lambda)^2} + \frac{\alpha_4^*}{(2\lambda)^4} + \ldots \\
&= \alpha_0^* + \sum_{j=1}^{\infty} \frac{\alpha_{2j}^*}{(2\lambda)^{2j}}.
\end{aligned} \tag{85}$$

We do not observe the coefficients α_j^* because they are features of the unobserved distribution of target incomes. But we do observe features of $h(z)$. And the two sets of coefficients can be related using equation (85).

Letting $\hat{h}(z)$ denote the Taylor expansion around $\tilde{z}$ of the *observed* income density with

coefficients $\alpha_0 = h(\tilde{z})$, $\alpha_1 = h'(\tilde{z})$, $\alpha_2 = h''(\tilde{z})$, etc., we have, from equation (85),

$$\alpha_0 = h(\tilde{z}) = \alpha_0^* + \sum_{j=1}^{\infty} \frac{\alpha_{2j}^*}{(2\lambda)^{2j}}. \tag{86}$$

To find α_1, we differentiate equation (85) with respect to $\tilde{z}$, noting that the coefficients α_j^* are functions of $\tilde{z}$:

$$\alpha_1 = h'(\tilde{z}) = \alpha_0^{*\prime}(\tilde{z}) + \sum_{j=1}^{\infty} \frac{\alpha_{2j}^{*\prime}(\tilde{z})}{(2\lambda)^{2j}}. \tag{87}$$

By definition of the Taylor expansion, $\alpha_j^{*\prime}(\tilde{z}) = \alpha_{j+1}^*(\tilde{z})$, so we have

$$\alpha_1 = \alpha_1^* + \sum_{j=1}^{\infty} \frac{\alpha_{2j+1}^*}{(2\lambda)^{2j}}. \tag{88}$$

Similarly,

$$\begin{aligned} \alpha_2 &= \alpha_2^* + \sum_{j=1}^{\infty} \frac{\alpha_{2j+2}^*}{(2\lambda)^{2j}} \\ &= (2\lambda)^2 \left(\sum_{j=1}^{\infty} \frac{\alpha_{2j}^*}{(2\lambda)^{2j}} \right). \end{aligned} \tag{89}$$

Combining equations (86) and (89), we have

$$\alpha_0 - \frac{\alpha_2}{(2\lambda)^2} = \alpha_0^*. \tag{90}$$

By a similar process, it can be shown that

$$\alpha_1 - \frac{\alpha_3}{(2\lambda)^2} = \alpha_1^*, \tag{91}$$

and more generally,

$$\alpha_j - \frac{\alpha_{j+2}}{(2\lambda)^2} = \alpha_j^*. \tag{92}$$

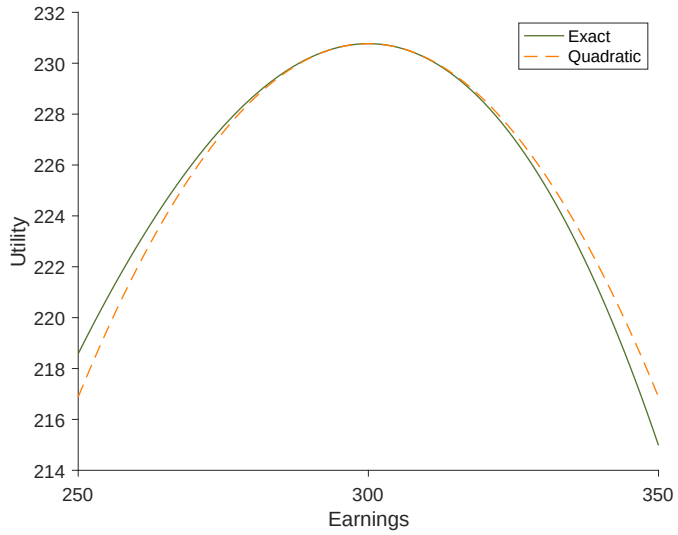
This proves the lemma. □

This lemma is useful for two reasons. First, it provides a method for converting between target income (type) densities and observed income densities for polynomial specifications of any order, with arbitrary precision. Second, it demonstrates that the polynomial coefficients differ from each other only in proportion to the coefficient two orders higher, e.g., the difference between coefficients α_2 and α_2^* is proportional to α_4^*. As a result, if the polynomial

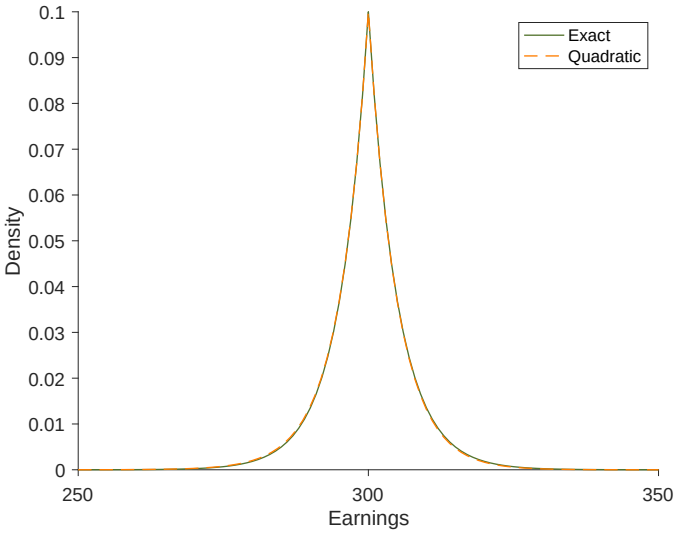
approximations are low-order, or if their higher order terms are small, then the difference between the target and observed income densities is also small, and in many settings it may be reasonable to approximate the target income density with the observed income density's polynomial approximation. We adopt this simplifying approximation in our numerical implementation.

Figure B1: Quadratic approximations of utility

(a) Indirect utility vs. quadratic approximation

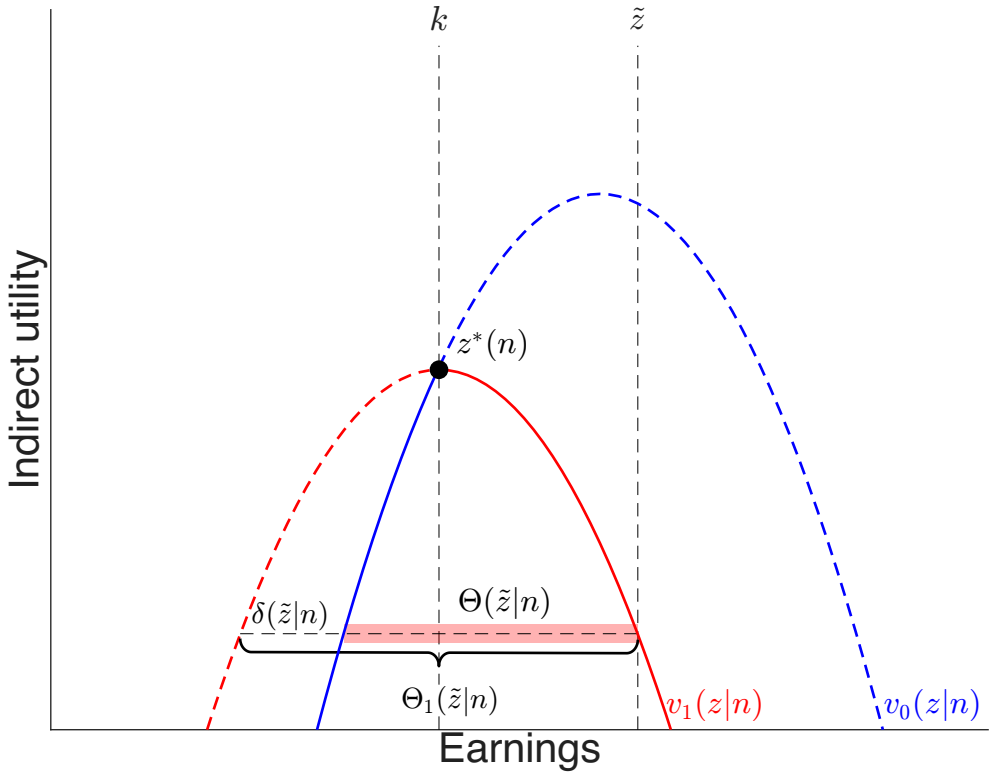


(b) Type-conditional income density vs. quadratic-utility approximation



This figure illustrates the quantitative effect of imposing Assumption 1. Panel (a) plots exact indirect utility under a linear tax for the specification used in the simulations from Section C, wherein taxpayers have constant elasticity of taxable income, and compares it to the quadratic approximation arising from the second-order Taylor approximation of that indirect utility around the taxpayer's target income. The approximation can be seen to be quite accurate across incomes near the target. Panel (b) plots the type-conditional income density under the linear tax and compares it to the density produced by the quadratic approximation. This approximation is extremely accurate, because the quadratic approximation in panel (a) only diverges meaningfully from true utility at incomes far from the target, which are very unlikely to be chosen from the opportunity set.

Figure B2: Illustration of dominated income regions with and without kink



This figure illustrates the notation used in the numerical computation of the bunching distortion. $\Theta(\tilde{z}|n)$ represents the set of incomes that utility-dominate income opportunity $\tilde{z}$ under the kinked tax function $T(z)$. $\Theta_1(\tilde{z}|n)$ represents the set of incomes that utility-dominate $\tilde{z}$ under the linear tax $T_1(z)$ which applies above the kink. $\delta(\tilde{z}|n)$ represents the difference between these sets.