Taxes and Heat

Stefan Steinerberger University of Washington

Aleh Tsyvinski Yale University*

December 18, 2021

Abstract

We show that the static optimal income tax problem is tightly connected to the heat equation. The optimal tax satisfies a particular version of tax smoothing – averaging welfare gains and behavioral losses at all income scales with the Gaussian averaging function. As heat in a steady state is spread evenly across all possible areas of the medium, the optimal tax spreads the benefits and the costs of taxes across all ranges of income. Tax smoothing also holds along the trajectory of the gradient flow that reforms existing taxes in the direction of the increase in welfare. As heat smoothes the most uneven parts of the medium and quickly spreads through the most sensitive parts, the tax reform is most important in the income regions where the marginal taxes are the most suboptimal and where the behavioral effect is large.

*Steinerberger is supported by the NSF (DMS-1763179) and the Alfred P. Sloan Foundation. We thank Arpad Abraham, Manuel Amador, Andy Atkeson, Jess Benhabib, Felix Bierbrauer, Job Boerma, Patrick Bolton, Jaroslav Borovicka, Pierre Boyer, Hector Chade, Alfred Galichon, Jonathan Heathcote, Ricardo Lagos, Chris Moser, Narayana Kocherlakota, Pricila Maziero, Abdoulaye Ndiaye, Yena Park, Alessandro Pavan, Chris Phelan, Georgii Riabov, Florian Scheuer, Michael Sockin, Stefanie Stantcheva, Kjetil Storesletten, and, especially, Nicolas Werquin. The paper was previously circulated under the title "Tax Mechanisms and Gradient Flows".

Introduction

One of the classic principles of optimal taxation is smoothing benefits and distortions that taxes create. We show that a particular version of this principle is also an important characteristic of the static optimal nonlinear tax (Mirrlees 1971). The optimal solution of this problem is tightly connected to the heat equation, one of the most well-studied and well-behaved partial differential equations, that possesses strong smoothing properties.

Our first main result shows that the optimal tax satisfies a certain invariance relationship – tax smoothing across different income scales. We show that the equity-efficiency tradeoff is represented by the social planner averaging the social welfare gains of redistribution and the behavioral responses of optimal taxes at all income scales. Just as heat in a steady state is spread evenly across all possible areas of the underlying medium, the optimal tax spreads the underlying costs and benefits of the taxes at all income ranges.

Specifically, the results for the optimal tax are as follows. It is well known that the optimal tax is characterized by a second-order ordinary differential equation. We associate with that equation a heat kernel which is a solution to the heat equation, the second-order partial differential equation that shows how heat diffuses in a medium. The steady state of this heat equation is the optimal tax. We show that the optimal tax at any given income satisfies the invariance relationship – it is determined as a weighted average of the welfare gains and the weighted average of the optimal taxes at other incomes. This proposition states that the central tradeoff of the optimal Mirrlesean taxation, the equity versus efficiency tradeoff, is represented in a particular way. Similarly to heat spreading evenly across all areas of a medium, the planner wants to spread benefits and the costs of taxes across all possible income ranges. We

show that the averaging function behaves like a Gaussian and, hence, possesses strong smoothing properties. For any income range, we provide a Gaussian upper bound that shows that the heat kernel decays as the Gaussian so that the more distant incomes are exponentially downweighted. For the small scale asymptotics, when averaging is done over the small neighborhood of incomes, the heat kernel is essentially the Gaussian. Different scales are represented by the wider Gaussians that are connected to each other by gradually encompassing larger ranges of incomes. In other words, the equity-efficiency tradeoff is smoothed at every income range similarly to heat spreading evenly across all possible ranges in a medium.

Our second set of results show that smoothing holds also for the tax reform of any initial (optimal or suboptimal) tax function and identifies the most important areas of such reform. We construct a gradient flow tax reform that changes the underlying tax function in the direction of the steepest descent of the change in the social welfare and show that it is a heat equation. The optimal tax is the steady state of this gradient flow. We prove that the short-term reform, that is the small-time behavior of the gradient flow, is such that the evolved tax is equal to the Gaussian average of the welfare gains and the Gaussian average of the taxes at other incomes. Intuitively, a heat equation diffuses initial temperature such that at each point of the medium the new temperature distribution is a Gaussian average of the nearby points of the initial distribution and, hence, is smoothed over. In our taxation context, the short-term tax reform similarly smoothes the equity and efficiency considerations by taking a Gaussian average of the welfare gains and the behavioral effects at the nearby incomes.

We then show that the short-term tax reform is most important in the income regions where either the marginal taxes are far from the short-term

optimum or where the behavioral effects determined by the elasticity and the density of incomes are large. The intuition for this result is as follows. The heat equation diffuses heat starting from its initial distribution and smoothes the most uneven and the most conductive points. The most uneven points are given by those areas where the differences in the initial temperature are the largest, that is, where the derivative of the initial temperature distribution is the largest. In our taxation context, these uneven points are given by the areas of income where the marginal taxes under the current systems differ significantly from the short-term optimum. The points of the high conductivity are the most sensitive points of the medium for heat diffusion. In the taxation context, those are the areas where the behavioral responses of taxes are the largest.

We now briefly discuss the related literature. The smoothing results for optimal taxes that we derive provides a different perspective on the representation of the classic equity-efficiency tradeoff. Similarly to the ABC formulas of Diamond (1998) and Saez (2001), the key determinants of the optimal tax are the elasticities, social marginal utility of income, and the density of incomes but our results highlight the underlying new smoothing properties. While tax smoothing is present in various forms in optimal taxation literature, the result that it happens at every income scale in a unified way and is done with the Gaussian is new.

Our paper is also a generalization of Tirole and Guesnerie (1981) who construct a process for linear taxes based on gradient projections, leading to an ordinary differential equation. The environment with nonlinear taxes is significantly more challenging as now the whole tax function is evolved as opposed to just one linear tax. We therefore have the steepest descent path in a space of functions that corresponds to a partial differential equation.

The construction of the gradient flow of the tax reforms follows the variational approach to taxation that considers a potentially suboptimal tax and proceeds with varying it locally to derive the formulas for the effects of the tax reforms.[1] The gradient flow tax reform that we analyze is a dynamical process that corresponds to the variational approach. We show that the optimal tax is the stationary point of the gradient flow and is invariant under averaging.

Sonnenschein (1981, 1982) and Artzner, Simon, and Sonnenschein (1986) derive a heat equation as a gradient process of the firms adjusting the commodity they produce by maximizing the rate of change in profit subject to a quadratic cost of adjustment. McCann (2014) argues that this result is a precursor to some of the results on the gradient flows in the optimal transport literature. Some of the techniques that we use have parallels in the optimal transport literature (see, e.g., Villani (2003)) in which there is a renewal of interest in economics (see, e.g., early work of Chiappori, McCann, and Nesheim (2010), a comprehensive book by Galichon (2016), or a review in the context of matching models by Chiappori and Salanie (2016)). Bolton and Harris (2010) associate a dynamic risk sharing rule with that of the static problem and obtain an elegant asymptotic expansion of the dynamic problem around a myopic optimum showing how the static problem is modified by the dynamic correction terms.

[1] See, e.g., Saez (2001), Kleven and Kreiner (2006), and Golosov, Tsyvinski, and Werquin (2014) for the methodology; Kleven, Kreiner, and Saez (2009) and Jacquet and Lehmann (2015) for the analysis of the multidimensional types; Saez and Stantcheva (2016) and Bierbrauer and Boyer (2018) for the political economy context; Sachs, Tsyvinski, and Werquin (2016) and Scheuer and Werning (2016) for the analysis in general equilibrium; Saez and Stantcheva (2018) for capital income taxation; Garrett and Pavan (2015) and Moser and de Souza e Silva (2019) for examples of a variational approach in mechanism design settings.

1 Environment

We start by presenting a standard economic environment of taxation with heterogenous agents.

Agents are characterized by an exogenous and fixed productivity type $\theta \in \Theta \subset \mathbb{R}_+$. Preferences over consumption c and labor effort l are represented by the utility function $U(c,l) = u(c) - v(l)$, where u is utility of consumption and is twice continuously differentiable, increasing and strictly concave; disutility of labor effort v is twice continuously differentiable, increasing and strictly convex. The government levies a tax liability $T : \mathbb{R}_+ \to \mathbb{R}$ which can be an arbitrarily non-linear function of the individual's labor income $y = \theta l$. The agent's budget constraint is $c = y - T(y)$.

The optimization problem of an individual with type θ reads:

$$\max_{y \geq 0} \quad u(y - T(y)) - v\left(\frac{y}{\theta}\right). \tag{1}$$

We denote the argmax of this problem by $y(\theta, T) \in \mathbb{R}_+$. For ease of notation, when there is no ambiguity we remove the argument T from this variable and write it as $y(\theta)$.

Assuming that the tax function T is continuously differentiable, labor income $y(\theta)$ is characterized by the first-order condition:

$$u'(y(\theta) - T(y(\theta)))\,(1 - T'(y(\theta))) \quad = v'\left(\tfrac{y(\theta)}{\theta}\right)\tfrac{1}{\theta}. \tag{2}$$

We assume that no individual θ is indifferent between two or more incomes in the initial equilibrium: for all θ, the individual problem (1) has a unique global maximum given the tax system T. It is straightforward then to show that there is a one-to-one map between productivity types θ and pre-tax incomes $y(\theta)$.

We denote by $H(\theta)$ the c.d.f. of $\theta \in \Theta$, and by $h(\theta)$ the corresponding density function. We assume that the set Θ is a compact interval of $\mathbb{R}_+$, and that the density of types h is equal to zero at the boundaries of Θ. We also denote by $\Phi(y)$ and $\phi(y)$ the c.d.f. and the p.d.f. of incomes $y \in Y \subset \mathbb{R}_+$. We assume that the density of incomes ϕ is continuous and bounded away from zero on any finite interval $[\underline{y}, \bar{y}] \subset Y$ with $\underline{y} > 0$.[2]

We define social welfare given the tax function T as follows:

$$W(T) = \int_\Theta (u(y(\theta) - T(y(\theta))) - v(\frac{y(\theta)}{\theta}))h(\theta)\, d\theta + \int_\Theta T(y(\theta))h(\theta)\, d\theta, \quad (3)$$

The first integral on the right-hand side of (3) is the social objective, given by the sum of individual utilities. We are assuming that the government is utilitarian and weighs the utility of each agent equally. The second integral is the government tax revenue. We are assuming that the marginal cost of public funds is equal to one and that any extra government revenue is redistributed (e.g., lump-sum) raising the social welfare by one. All of the terms in this expression take into account the optimization behavior of individual agents and are defined over $y(\theta)$, the solution to (2).

2 Tax reforms and optimal taxation

In this section we define a notion of local tax reforms, and derive their effects on individual behavior and social welfare. We then derive the formula for the optimal tax. All of the results in this section are standard in the literature.

[2]Our results can be straightforwardly generalized to the case of types and incomes in the whole space $\mathbb{R}_+$ by using an increasing sequence of compact sets $\Theta, Y \subset \mathbb{R}_+$.

2.1 Variations of taxes

We define a direction of the reform of the tax function T as a continuously differentiable function $\hat{T} : \mathbb{R}_+ \to \mathbb{R}$. The perturbed tax function is then $T + \mu\hat{T}$, where $\mu > 0$ is the size of the reform in the direction $\hat{T}$.

We now derive the first-order changes in individual labor income $y(\theta)$ and social welfare $W(T)$ in response to the tax reform $\mu\hat{T}$ as $\mu \to 0$. That is, we compute the Gateaux derivatives of the functionals $T \mapsto y(T), W(T)$, defined (for a generic functional $\Psi(T)$) by:

$$\delta\Psi(T, \hat{T}) \equiv \lim_{\mu \to 0} \frac{\Psi(T + \mu\hat{T}) - \Psi(T)}{\mu}.$$

Lemma 1 describes the impact of a tax reform $\hat{T}$ of the tax schedule T on individual income choices. Lemma 2 describes the impact of a tax reform $\hat{T}$ of the tax schedule T on social welfare. The proofs of both of these results are standard in the literature.

Lemma 1. *The Gateaux derivative of the individual income $y(\theta)$ in the direction $\hat{T}$ is given by:*

$$\delta y(\theta) = -\varepsilon(y(\theta))\hat{T}'(y(\theta)) - \eta(y(\theta))\hat{T}(y(\theta)), \tag{4}$$

with the expressions for $\varepsilon(y)$ and $\eta(y)$ given in the Appendix.

Formula (4) shows that the shift in the choice of income of individual θ due to the tax reform is given by the sum of two terms. The first is the change in the agent's marginal tax rate implied by the reform, $\hat{T}'(y)$, multiplied by his labor income adjustment in response to this tax change (taking into account the non-linearity of the initial tax schedule), $\varepsilon(y)$. The second term in (4)

is the change in the absolute tax payment faced by the agent due to the reform, $\hat{T}(y)$, multiplied by his labor income change due to the income effect, $\eta(y)$. The labor elasticity ε and the income effect η along the nonlinear budget constraint are discussed in details in Jacquet and Lehmann (2015) and Scheuer and Werning (2017) and the expressions for them are given in the Appendix.

The next proposition describes the impact of a tax reform $\hat{T}$ of the tax schedule T on the social welfare.

Lemma 2. *The Gateaux derivative of the social welfare functional $W(T)$ in the direction $\hat{T}$ is given by*

$$\delta W(T,\hat{T}) = \int_Y (1-\gamma(y))\,\phi(y)\,\hat{T}(y)\,dy - \int_Y T'(y)\,\varepsilon(y)\,\phi(y)\,\hat{T}'(y)\,dy, \quad (5)$$

with the expression for $\gamma(y)$ given in the Appendix.

Formula (5) shows that the total first-order effect of the tax reform $\hat{T}$ on social welfare W is given by the sum of two terms. The first integral is the social utility gain. Specifically, for each dollar of revenue raised by the increase in the tax payment $\hat{T}(y)$ at income y, social welfare is lowered by the social marginal utility of income $\gamma(y)$. The social marginal utility of income is defined in the appendix and is given by the social marginal welfare weight (Saez and Stantcheva 2015) net of the revenue loss due to the income effect. Summing over all incomes y using the density of income ϕ yields the first term in the right-hand side of (5). The second term is the excess burden, or deadweight loss, from the tax reform. Specifically, an increase in the marginal tax rate at income y by $\hat{T}'(y)$ lowers the labor income of these agents by $\varepsilon(y)$, by construction of the elasticity of labor supply, which in turn reduces government revenue by the fraction $T'(y)$ of this income loss. Summing over all incomes y yields the second integral in (5).

2.2 Optimal Tax

Proposition 1 provides a formula for the welfare effects of any tax reform $\hat{T}$ in the economy starting from any, optimal or suboptimal, tax schedule T. As a by-product, we obtain a characterization of the optimal tax schedule T_* by imposing that no tax reform has a positive first-order effect on social welfare, i.e. $\delta W(T_*, \hat{T}) = 0$ for all $\hat{T} : \mathbb{R}_+ \to \mathbb{R}$. Let $\varepsilon_*(y)$, $\gamma_*(y)$ and $\phi_*(y)$ denote the elasticity of labor supply, the social marginal value of income and the density of incomes given that the optimal tax schedule T_* is implemented. Assume furthermore the boundary conditions $T'\left(\underline{y}\right)\varepsilon\left(\underline{y}\right)\phi\left(\underline{y}\right) = T'(\bar{y})\,\varepsilon(\bar{y})\,\phi(\bar{y}) = 0$. Integrating by parts the Gateaux differential of the social welfare (5)

$$\delta W(T, \hat{T}) = \int_Y \hat{T}(y)\,(1 - \gamma(y))\,\phi(y)\,dy + \int_Y \hat{T}(y)\,\frac{d}{dy}\left(T'(y)\,\varepsilon(y)\,\phi(y)\right)dy,$$

and setting it to zero for any $\hat{T}(y)$ yields the equation for the optimal tax:

$$0 = (1 - \gamma_*(y))\,\phi_*(y) + \frac{d}{dy}\left(T_*'(y)\,\varepsilon_*(y)\,\phi_*(y)\right). \tag{6}$$

In the Appendix, we show that equation (6) is identical to the optimal tax formulas of Diamond (1998) and Saez (2001).

3 Smoothing properties of optimal taxes

The optimal tax in (6) is given by a solution to the second order differential equation in the tax T as the derivative $\frac{d}{dy}(T'(y))$ appears in the equation for the optimum. It is known that the properties of the equations of this type can be studied using an object called the heat kernel or the fundamental solution to the heat equation (Grigor'yan 2009).

Definition. The *heat kernel* $q_t(x, y)$ is given by the solution to the heat equation

$$\frac{\partial}{\partial t} q_t(x, y) = \frac{\partial}{\partial y}\left(\varepsilon_*(y)\,\phi_*(y)\,\frac{\partial}{\partial y} q_t(x, y)\right), \tag{7}$$

and $\lim_{t\to 0} q_t(x, y) = \delta(x - y)$, where δ is a Dirac delta function.

What is the heat kernel and why it may be useful for deriving the properties of the optimal taxes? Originally, the heat equation (7) was motivated and used for studying of the evolution of heat in a medium over time. However, for our application, we want to think of it somewhat differently – as a family of averaging functions over different scales of incomes. Consider, for example, the averages of the optimal taxes $\int_Y q_t(x, y)\, T_*(y)\, dy$ at different scales t for a given income x. Here, the average is taken from the point of view of income x, weighting all different incomes $y \in Y$ by the weight $q_t(x, y)$. We will show that the parameter t of the heat kernel determines the scale of averaging. Intuitively one can think of the scale as determining how much the average is weighting incomes y which are more distant from a given income x. As a concrete example, suppose the income x is \$100 thousand dollars and all the taxes are 100 percent. At scale $t = 0$, the kernel, being a Dirac delta function at x, puts the weight $q_{t=0}(x = \$100, y = \$100) = 1$ on income x, and the average is just $x = \$100$ thousand. The larger scales may average and encompass a wider set of incomes around x. For example, at scale $t = 1$, the kernel may put the weight 0.5 on income \$100 thousand and the weights of 0.25 each on incomes \$50 thousand and \$150 thousand: $q_{t=1}(x = \$100, y = \$100) = 0.5$, $q_{t=1}(x = \$100, y = \$50) = 0.25$, $q_{t=1}(x = \$100, y = \$150) = 0.25$. The average is then given by $0.25 \times \$50 + 0.5 \times \$100 + 0.25 \times \$150 = \$100$. The larger scales t may put more weight on more distant incomes. We can thus, for a given point x, understand $q_t(x, y)$ as a family of weighting functions in

the variable y parametrized by the scale t.

There are three basic properties of the heat kernel: (1) $q_t(x,y) \geq 0$, (2) $q_t(x,y) = q_t(y,x)$, and (3) preservation of integral mass $\int_Y q_t(x,y)dy = 1$. At this stage, the weighting function $q_t(x,y)$ may be rather arbitrary. For example, it can put some very high weight on a particular distant income, say, \$1 million. Moreover, the functions $q_t(x,y)$ may be completely disconnected from each other for different scales t. For example, at scales $t = 0$ and $t = 1$ they may behave as in the example above, putting more weight on the distant incomes y, but at scale $t = 3$ it may put weight on some completely different incomes.

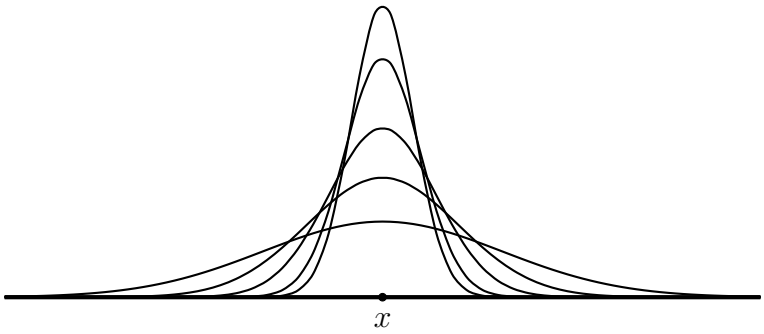

Figure 1: The heat kernel $q_t(x,y)$ for various scales t (smaller scales correspond to larger maxima).

The main result of this section is that, because the optimal tax T_* solves the equation (6), the weighting function $q_t(x,y)$ and the averages of the taxes have particularly well-defined properties. First, we show that the weighting function is not arbitrary and behaves similarly to the Gaussian as in Figure 1. That is, for a given t, the weighting function puts the bulk of the weight on some neighborhood of income x, exponentially downweghting more distant incomes – thus having the interpretation of scale. Second, the kernels at different scales t are tightly connected to each other. Overall, the key property of this family

of Gaussian weighting functions is the strong smoothing behavior over different income ranges.

We now show these properties of the optimal taxes.

Proposition 1. *(Smoothing for the optimum). (1) The optimal tax* $T_*(y)$ *is invariant under the heat kernel* $q_t(x,y)$ *given by (7), for any* $x \in Y$ *and any* $t > 0$*:*

$$T_*(x) = \int_0^t \int_Y q_s(x,y)\,(1-\gamma_*(y))\,\phi_*(y)\,dyds + \int_Y q_t(x,y)\,T_*(y)\,dy. \quad (8)$$

(2) The heat kernel (7) satisfies a Gaussian upper bound, for any $t > 0$*:*

$$q_t(x,y) \leq \frac{c_1}{\sqrt{t}} \exp(c_2 t) \exp\left(-c_3 \frac{(x-y)^2}{t}\right),$$

for some positive constants c_1, c_2, c_3*;*

(3) For $t \to 0$,

$$q_t(x,y) \sim \frac{1}{\sqrt{4\pi\sigma_*(x)\,t}} \exp\left(-\frac{(y-x-\sigma_*'(x)\,t)^2}{4\sigma_*(x)\,t}\right),$$

where $\sigma_*(x) = \varepsilon_*(x)\,\phi_*(x)$*;*

(4) For all points x, y *and all scales* $t, s > 0$*, the heat kernels are connected across all scales and satisfy the semigroup property:*

$$q_{t+s}(x,y) = \int_Y q_t(x,z) q_s(z,y) dz.$$

This proposition states that the central tradeoff of the optimal Mirrlesean taxation – the equity versus efficiency tradeoff is represented in a particular way. The redistributional (equity) considerations and the deadweight loss

(efficiency) are averaged and, hence, smoothed over all incomes with the specific weights given by the Gaussian. Intuitively, heat spreads evenly across all ranges of the underlying medium in the steady state. In our taxation context, the planner wants to spread the benefits and the costs of the optimal taxes across all possible income ranges.

Part (1) of the proposition describes the optimal tax $T_*(x)$ as resolving the equity-efficiency tradeoff in a particular way. The first integral in (8) represents a weighted average at all incomes y of the impact of taxes on social welfare. The second term in (8) arises from the behavioral effect of taxation and is a weighted average of the distortions or the behavioral effects of taxes at all incomes. In other words, that effect ensures that an agent at a given income x is paying the weighted average of the amount of taxes paid by people of other incomes. Importantly, this proposition shows that this averaging of equity and efficiency happens at every scale t with the weighting function $q_t(x, y)$.

It may seem that the representation result of part (1) of Proposition 1 is trivial. Indeed, for any function $f \in C^1$, f at a point is the local average of its neighboring values (this follows from continuity and boundedness and does not require differentiability)

$$\lim_{t \to 0^+} \int_Y q_t(x, y) f(y) dy = f(x).$$

The optimal tax satisfies, however, a much stronger relationship for *all* scales $t > 0$ and not only in the limit $t \to 0$. Moreover, Parts (2)-(4) of the Proposition give the specific form of the weighting function and its determinants – that the heat kernel behaves similarly to the Gaussian average.[3] We also show

[3]Of course, there exist many other second-order differential equations for which the fundamental solution is not the heat kernel and, hence, they do not possess, for example, smoothing properties.

that the weighting functions at different scales are tightly linked to each other representing one unified smoothing mechanism.

Part (2) of the Proposition shows that the heat kernel satisfies a Gaussian upper bound for all times t. In other words, the kernel exponentially downweighs more distant incomes and incorporates more income as the scale t increases. For the small scale t, only the incomes nearby matter. For the larger scale t, more distant incomes matter.

Part (3) of the proposition shows that for small t, where averaging is done over the nearby incomes, the heat kernel is exactly the Gaussian. In particular, its two key determinants are the labor elasticity and the density of incomes at the optimum. Just as the way the heat spreads in a medium depends on the material's conductivity or sensitivity to heat $\sigma_*(x) = \varepsilon_*(x)\,\phi_*(x)$, the incomes where either the disincentive effect of taxes, represented by $\varepsilon_*(x)$, or where more people are affected, represented by $\phi_*(x)$, play an important role in determining the Gaussian.

Part (4) of the proposition shows that the heat kernels are tightly linked at all time scales. Averaging at scale t and then averaging at scale s is equivalent to averaging over scale $t + s$. In other words, there is one unified averaging scheme and welfare gains and deadweight losses matter across all income scales.

Our results so far pertain to the level of the tax and it is not clear whether they have something to say about the marginal tax. Given any function $f_1 \in C^1(\mathbb{R})$, it is possible to change it slightly into a function $f_2 \in C^1(\mathbb{R})$ such that f_1 and f_2 give almost the same values everywhere but f_2 has a very different derivative. Put differently, even a very good understanding of the optimal tax code T_* need not a priori translate into a good understanding of the marginal tax T_*'. We show in Corollary 1 that this is not the case, the marginal tax is determined by the global behavior of the optimal tax code.

Corollary 1. *The optimal marginal tax is given by*

$$\frac{\partial}{\partial x}T_*(x) = \int_0^t \int_Y \frac{\partial}{\partial x}(q_s(x,y))\,(1-\gamma_*(y))\,\phi_*(y)\,dyds \qquad (9)$$
$$+ \int_Y \left(\frac{\partial}{\partial x}q_t(x,y)\right) T_*(y)\,dy,$$

for any $x \in Y$ *and any* $t > 0$

We plot derivative $\partial_x q_t(x,y)$ which is also a weighting function in Figure 2 with the thick line (together with the second-order derivatives given by the dotted line).

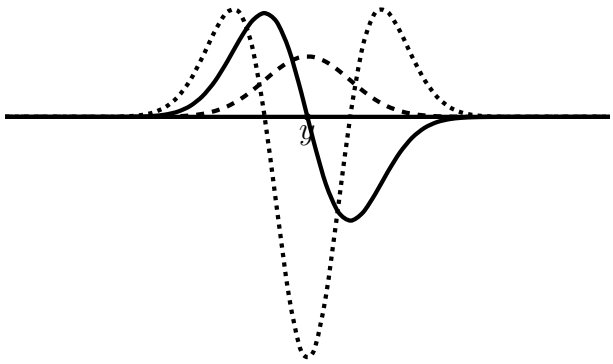

Figure 2: q_t (dashed) and its derivative q_t' (bold) and q_t'' (dotted).

The integral in (9) evaluates a weighted average of the welfare gains and of taxes paid by individuals with higher incomes, subtracts a weighted average of the welfare gains and of taxes for individuals with lower incomes, and this results in the quantity determining the size of the marginal tax $T_*'(x)$. Moreover, this holds for all $t > 0$ and the weighting function is a derivative of a Gaussian and, hence, has a very specific form. The result of the Corollary is interesting even as a stand-alone result as it gives a new relationship of the optimal marginal tax at a given income to the optimal tax levels at *all* other income and *all* scales, and not just for $t \to 0$.

4 Short-term tax reform as a gradient flow

We show that the results on smoothing that we derived for the optimal tax also hold for the short-term gradient flow tax reform and determine the areas of the current tax system that such reform aims to affect the most.

4.1 A gradient flow of taxes

We construct a dynamic system of tax reforms, a gradient flow, which starts at any (optimal or suboptimal) tax function and then changes the tax system in the direction of the steepest increase in social welfare. The optimal tax is a stationary point of this system. We start by formally defining the gradient flow.[4] In this section, we use the straightforward adaptation of notation in Section 1 to index the relevant variables by time.

Definition 1. For all $t \geq 0$ and $y \in Y$, the *gradient flow* of the social welfare functional $W(T_t)$ is defined as the dynamical system:

$$\frac{\partial T_t(y)}{\partial t} = (1-\gamma_t)\,\phi_t(y) + \frac{\partial}{\partial y}\left(T_t'(y)\,\varepsilon_t(y)\,\phi_t(y)\right), \tag{10}$$

where ϕ_t is governed by T_t according to the change of variables $\phi_t(y(\theta)) = (y_t'(\theta))^{-1} h(\theta)$ and by the first order conditions to the problem (1).

One can think of the gradient flow as the equation describing the familiar steepest descent. Doing so for a linear tax would be straightforward, as only one parameter has to be evolved, and this would lead to an ordinary differential equation (Tirole and Guesnerie 1981). For the nonlinear tax schedule, the whole tax function changes and this leads to the partial differential equation.

[4] Additional mathematical details on gradient flows are given in the Appendix.

We now turn to the derivation of the the gradient flow (10). Assume the boundary conditions $T_t'\left(\underline{y}\right)\varepsilon_t\left(\underline{y}\right)\phi_t\left(\underline{y}\right) = T_t'\left(\bar{y}\right)\varepsilon_t\left(\bar{y}\right)\phi_t\left(\bar{y}\right) = 0$. An integration by parts in the second integral of equation (5) implies that the impact of the tax reform $\hat{T}_t$ on social welfare can equivalently be rewritten as

$$\delta W(T_t, \hat{T}_t) \;=\; \int_Y \Lambda_t\left(y\right)\hat{T}_t\left(y\right)dy, \text{ with } \Lambda_t \equiv \left(1-\gamma_t\right)\phi_t + \frac{\partial}{\partial y}\left(T_t'\varepsilon_t\phi_t\right).$$

Assuming that the space of functions $\mathcal{C}^2\left(\mathbb{R},\mathbb{R}\right)$ is endowed with the L^2 norm $\|T\|^2 = \int\left(T\left(y\right)\right)^2 dy$, this can be expressed as $\langle\Lambda_t, \hat{T}_t\rangle$.[5] Therefore, the gradient flow we obtain in this case can be written as the dynamical system (10).

The gradient flow (10) can be equivalently derived as the solution to the problem of choosing the trajectory of the tax schedule $t \mapsto T_t$ that maximizes at each instant t the increase in social welfare:

$$\max_{T_t} \;\; \frac{\partial}{\partial t}W(T_t),$$

subject to the quadratic cost. The use of other norms is equivalent to different specification of the cost function for changing taxes.

4.2 Short-term tax reform and smoothing

In this section, we describe the short-term evolution of the tax schedule $T_{\tilde{t}}$ under the gradient flow by solving the heat equation (10) over a short time interval $\left[t, \tilde{t}\right]$ – a short-term tax reform. We now show how the results on smoothing derived for the optimum extend to the short-term tax reform.

Proposition 2. *(Smoothing for the short-term tax reform). Consider any initial time t with the corresponding tax profile $T_t\left(y\right)$, density of incomes $\phi_t\left(y\right)$,*

[5] We discuss the use of other norms in the Appendix.

social marginal utility of income $\gamma_t(y)$*, elasticity* $\varepsilon_t(y)$*, and define* $\sigma_t(y) = \phi_t(y)\varepsilon_t(y)$*. Then, for small* $\left(\tilde{t}-t\right)$*, the tax* $T_{\tilde{t}}(y)$*, generated by the gradient flow (10), is given by:*

$$T_{\tilde{t}}(x) \sim \left(\tilde{t}-t\right)(1-\gamma_t(x))\,\phi_t(x) + \int_Y q_{t,\tilde{t}}(x,y)\,T_t(y)dy, \tag{11}$$

where

$$q_{t,\tilde{t}}(x,y) = \frac{1}{\sqrt{4\pi\sigma_t(x)\left(\tilde{t}-t\right)}} \exp\left(-\frac{\left(y-x-\sigma_t'(x)\left(\tilde{t}-t\right)\right)^2}{4\sigma_t(x)\left(\tilde{t}-t\right)}\right).$$

Proof. In the Appendix □

This equation extends the notion of tax smoothing that we derived for the optimal tax to that of the trajectory of the gradient flow of taxes. The government that considers a short-term tax reform in the direction of maximizing social welfare changes the tax such that the new, evolved tax $T_{\tilde{t}}(x)$ is equal to $(1-\gamma_t(x))\phi_t(x)$ representing the social marginal utility of income and the weighted average of the initial taxes $T_t(y)$ representing the behavioral effect. Note that the property (11) holds for any starting time t on the gradient flow trajectory and that we average over the known and given parameters $\varepsilon_t(y)$, $\phi_t(y)$, $\gamma_t(y)$ evaluated at the time t. That is, it is a closed-form expression. This characterization is valid for the short time $\tilde{t}$, as the parameters $\varepsilon_t(y)$, $\phi_t(y)$, $\gamma_t(y)$ are frozen over that short time interval.[6] The intuition for this result is as follows. The heat equation diffuses the initial distribution of temperature by spreading heat to nearby points. In our taxation context, the tax reform takes the initial tax distribution $T_t(y)$ and spreads it by making

[6]The proof of the result makes this notion specific by associating a stochastic process with the equation and studying its short-term evolution.

the point $T_{\tilde{t}}(x)$ to be equal to the Gaussian average of the points around it $\int_Y q_{t,\tilde{t}}(x,y)\, T_t(y)dy$ thus spreading the behavioral effect of taxes and also by adding the welfare gains $\left(\tilde{t}-t\right)(1-\gamma_t(x))\,\phi_t(x)$ that such tax reform generates. This characterization also holds for the optimal tax which is the stationary point of the gradient flow and thus is true for all income ranges and not just for the short-term asymptotics.

4.3 Most important areas of the short-term tax reform

This section provides a further characterization of the smoothing properties of the gradient flow for the short-term tax reform. Proposition 3 shows which parts of the current tax are the most important for the short-term gradient flow tax reform.

The initial tax system T_t can be very far from the optimal tax system T_*. Hence, the elasticities, densities and the social marginal utility gains $(\varepsilon_t(y), \phi_t(y), \gamma_t(y))$ which depend on the initial tax system T_t, may be very different from their counterparts $(\varepsilon_*(y), \phi_*(y), \gamma_*(y))$ at the optimum. Since we are considering the short-term asymptotics, that is the tax reform as a gradient flow over a short period of time, we first need to introduce the proper reference point for the analysis of smoothing of a gradient flow tax reform, starting from a given tax.

Definition. The *short-term optimum at time t,* τ_t, is the stationary point of the gradient flow as $\tilde{t} \to \infty$ starting from a tax system T_t with the fixed $\varepsilon_t(y)$, $\phi_t(y)$, $\gamma_t(y)$, that is, $\tau_t(y)$ is the solution of

$$0 = (1-\gamma(y))\,\phi(y) + \frac{\partial}{\partial y}\left(\varepsilon(y)\,\phi(y)\,\frac{\partial \tau_t(y)}{\partial y}\right), \tag{12}$$

where $\gamma(y) = \gamma_t(y)$, $\varepsilon(y) = \varepsilon_t(y)$, and $\phi(y) = \phi_t(y)$.

This short-term optimum is the stationary point to the planner's problem, conditional on keeping the density of agents' incomes, the social marginal utility of income, and elasticities fixed at their value given the tax system $T_t(y)$. Hence, τ_t serves as a proper reference point to describe various smoothing properties of the gradient flow evolving taxes for a small time $\tilde{t} - t$ in a relationship to the initial tax system T_t. The short term optimum τ_t, in principle, can be quite different from the optimal tax T_*, if the initial tax system T_t is far from the optimum. However, if the initial tax system is close to the optimal one in the sense that $(\varepsilon_t(y), \phi_t(y), \gamma_t(y)) \approx (\varepsilon_*(y), \phi_*(y), \gamma_*(y))$, then the short term optimum τ_t and the long term optimum T_* coincide.[7]

The next proposition shows which part of the current tax schedule T_t are smoothed the most.

Proposition 3. *Equation (12) coincides with the gradient flow of the functional $\mathcal{J}$:*

$$\mathcal{J}_t(T) = \frac{1}{2}\int_Y \sigma_t(y)\left(T_t'(y) - \tau_t'(y)\right)^2 dy.$$

Proof. In the Appendix. □

The result shows that the gradient flow of the short-term tax reforms is also the direction of the steepest descent of the functional that decreases the weighted squared deviations of the marginal taxes. In other words, the gradient flow also smoothes the marginal taxes. There are two types of situations where this smoothing is particularly important for the gradient flow

[7]The optimal tax T_* can be thought of as a long-term optimum which evolves both the tax system and the sufficient statistics. Section 6.9 in the Appendix discusses the use of operator splitting methods to describe a relationship between the short-term and the long-run gradient flows.

tax reform. First, when the initial marginal taxes are very suboptimal – the marginal tax $T_t'(y)$ different from the marginal tax in the short term optimum $\tau_t'(y)$. A large value of $|T'(y) - \tau_t'(y)|$ implies the existence of a large value of $(T_t'(y) - \tau_t'(y))^2$ and the gradient flow tax reform is trying to decrease this as quickly as possible. Second, the gradient flow acts more on the regions where the behavioral responses are the largest as evidenced by the high value of $\sigma_t(y) = \varepsilon_t(y)\,\phi_t(y)$. That is, if $\sigma_t(y)$ is large, then irregularities in that region count even more severely and are dampened quicker than in regions where $\sigma_t(y)$ is very small.[8] The intuition for this result is as follows. We have shown that the gradient flow of the tax reform is a heat equation. The initial tax system can be thought of as the initial distribution of the temperature. The heat equation diffuses heat starting from the initial distribution and smoothes the most uneven points of the initial distribution and affects the most conductive points. The most uneven points are given by those areas where the changes in the initial temperature are the largest, that is, where the derivative of the initial temperature distribution is the largest. In our taxation context, these uneven points are given by the areas of income where the marginal taxes under the current systems differ significantly from the short-term optimum. The heat also spreads the most across the areas of the medium which are most sensitive to heating, that is, the most conductive areas. In the taxation context, these are the areas where the behavioral responses of taxes are the largest.[9]

[8] Additionally, a classic result from the theory of parabolic equations can also be used to show that if $T_t(y)$ has large amounts of strong oscillations or maybe even discontinuous jumps, then the gradient flow acts strongest on those parts first and that $T_{\tilde{t}}(y)$ implied by the gradient flow leads is infinitely differentiable for any $\tilde{t} > 0$.

[9] In Section 6.8 of the Appendix we discuss additional smoothing results for the gradient flow of the current tax system.

5 Conclusion

We show that the heat kernel and the heat equation are intimately connected with the analysis of the classic taxation problem. The smoothing results for the optimum and for the gradient flow tax reform and its Gaussian form are likely applicable to a variety of other mechanism design and optimal taxation problems where equity and efficiency tradeoff is present.

References

Aronson, Don. Non-negative solutions of linear parabolic equations, Ann. Sci. Norm. Sup. 22, p. 607–694. 1968.

Artzner, Philippe, Carl P. Simon, and Hugo Sonnenschein. " Convergence of Myopic Firms to Long-Run Equilibrium via the Method of Characteristics." Models of Economic Dynamics. Springer, Berlin, Heidelberg, 157-183. 1986.

Bierbrauer, Felix, and Pierre Boyer. "Politically feasible reforms of non-linear tax systems." 2018.

Bogachev, Vladimir I., Nicolai V. Krylov, Michael Röckner, and Stanislav V. Shaposhnikov. Fokker-Planck-Kolmogorov Equations. Vol. 207. American Mathematical Soc., 2015.

Bolton, Patrick, and Christopher Harris. "The dynamics of optimal risk sharing." No. w16094. National Bureau of Economic Research. 2010.

Brewer, Mike, Emmanuel Saez, and Andrew Shephard. "Means-testing and tax rates on earnings." Dimensions of Tax Design: the Mirrlees Review. 2010.

Chiappori, Pierre-André, Robert J. McCann, and Lars P. Nesheim. "Hedonic price equilibria, stable matching, and optimal transport: equivalence, topology, and uniqueness." Economic Theory 42, no. 2: 317-354. 2010.

Chiappori, Pierre-André, and Bernard Salanié. "The econometrics of matching models." Journal of Economic Literature 54, no. 3 : 832-61, 2016.

Courant, Richard and David Hilbert, Methods of mathematical physics. Vol. I. Interscience Publishers, Inc., New York, N.Y., 1953.

Diamond, Peter A. " A many-person Ramsey tax rule." Journal of Public Economics 4.4: 335-342. 1975.

Diamond, Peter A. " Optimal income taxation: an example with a U-shaped pattern of optimal marginal tax rates." American Economic Review: 83-95. 1998.

Friedman, Avner. Partial differential equations of parabolic type. Courier Dover Publications, 2008.

Galichon, Alfred. Optimal transport methods in economics. Princeton University Press. 2016.

Garrett, Daniel F., and Alessandro Pavan. "Dynamic managerial compensation: A variational approach." Journal of Economic Theory 159: 775-818. 2015.

Glowinski, Roland, Stanley J. Osher, and Wotao Yin, eds. Splitting Methods in Communication, Imaging, Science, and Engineering. Springer, 2017.

Grigor'yan, Alexander. Heat Kernel and Analysis on Manifolds, AMS/IP Studies in Advanced Mathematics. 2009.

Golosov, Mikhail, Aleh Tsyvinski, and Nicolas Werquin. A variational approach to the analysis of tax systems. No. w20780. National Bureau of Economic Research, 2014.

Hörmander, Lars. The Analysis of Linear Partial Differential Operators I – IV, Springer, 2003.

Jacquet, Laurence, and Etienne Lehmann. "Optimal Income Taxation when Skills and Behavioral Elasticites are Heterogeneous." 2015.

Kleven, Henrik Jacobsen. "Sufficient Statistics Revisited." 2018.

Kleven, Henrik Jacobsen, and Claus Thustrup Kreiner. "The marginal cost of public funds: Hours of work versus labor force participation." Journal of Public Economics 90, no. 10-11: 1955-1973. 2006.

Kleven, Henrik Jacobsen, Claus Thustrup Kreiner, and Emmanuel Saez. "The optimal income taxation of couples." Econometrica 77, no. 2: 537-560.

2009.

Lörinczi, Jozsef, Fumio Hiroshima, and Volker Betz. Feynman-Kac-type theorems and Gibbs measures on path space: with applications to rigorous quantum field theory. Vol. 34. Walter de Gruyter, 2011.

MacNamara, Shev, and Gilbert Strang. In Glowinski, Roland, Stanley J. Osher, and Wotao Yin. Operator splitting. In Splitting Methods in Communication, Imaging, Science, and Engineering (pp. 95-114). Springer, Cham. 2016.

Metafune, Giorgio, El Maati Ouhabaz, and Diego Pallara. "Long time behavior of heat kernels of operators with unbounded drift terms." Journal of Mathematical Analysis and Applications 377, no. 1: 170-179. 2011.

McCann, Robert J. " Academic wages, singularities, phase transitions and pyramid schemes." Proceedings of the International Congress of Mathematicians (Seoul 2014). Vol. 3. 2014.

Mirrlees, James A. " An exploration in the theory of optimum income taxation." The Review of Economic Studies 38.2: 175-208. 1971.

Molchanov, Stanislav A. " Diffusion processes and Riemannian geometry." Russian Mathematical Surveys 30, no. 1: 1-63. 1975.

Moser, Christian, and Pedro Olea de Souza e Silva. "Optimal paternalistic savings policies." Columbia Business School Research Paper 17-51. 2019.

Musgrave, Richard A., and Tun Thin. " Income tax progression, 1929-48." Journal of Political Economy 56.6: 498-514. 1948.

Sachs, Dominik, Aleh Tsyvinski, and Nicolas Werquin. Nonlinear tax incidence and optimal taxation in general equilibrium. No. w22646. National Bureau of Economic Research. 2016.

Saez, Emmanuel. Using elasticities to derive optimal income tax rates. The Review of Economic Studies, vol. 68, no 1, p. 205-229. 2001.

Saez, Emmanuel, and Stefanie Stantcheva. "Generalized social marginal welfare weights for optimal tax theory." American Economic Review 106, no. 1: 24-45. 2016.

Saez, Emmanuel, and Stefanie Stantcheva. "A simpler theory of optimal capital taxation." Journal of Public Economics 162: 120-142. 2018.

Salanie, Bernard. The economics of taxation. MIT press, 2011.

Scheuer, Florian, and Ivan Werning. Mirrlees meets diamond-mirrlees. No. w22076. National Bureau of Economic Research. 2016.

Scheuer, Florian, and Ivan Werning. "The taxation of superstars." The Quarterly Journal of Economics 132, no. 1: 211-270. 2017.

Sonnenschein, Hugo. " Price dynamics and the disappearance of short-run profits: An example." Journal of Mathematical Economics 8.2: 201-204. 1981.

Sonnenschein, Hugo. " Price dynamics based on the adjustment of firms." The American Economic Review 72.5: 1088-1096. 1982.

Steinerberger, Stefan. " Fast escape in incompressible vector fields." Monatshefte für Mathematik 186, no. 3: 525-537. 2018.

Tao, Terence. " On the universality of potential well dynamics.", Dynamics of PDE, Vol.14, No.3, 219-238, 2017

Taylor, Michael. Partial Differential Equations I-III, Springer. 1996.

Teschl, Gerald. Ordinary differential equations and dynamical systems. Vol. 140. American Mathematical Soc., 2012.

Tirole, Jean, and Roger Guesnerie." Tax reform from the gradient projection viewpoint." Journal of Public Economics 15.3: 275-293. 1981.

Titchmarsh, Edward Charles. Eigenfunction expansions associated with second-order differential equations. Part I. Second Edition Clarendon Press, Oxford. 1962.

Varadhan, Sathamangalam R. Srinivasa. " On the behavior of the funda-

mental solution of the heat equation with variable coefficients." Communications on Pure and Applied Mathematics 20, no. 2: 431-455. 1967.

Varga, Richard S. Matrix Iterative Analysis, New Jersey: Prentice-Hall, 1962

Villani, Cedric. Topics in optimal transportation. No. 58. American Mathematical Soc., 2003.

Zettl, Anton. Sturm–Liouville Theory. Providence: American Mathematical Society, 2005.

6 Appendix

6.1 Proof of Lemma 1

Consider the perturbed first order condition of the agent (2):

$$u'\left(y\left(\theta,T+\mu\hat{T}\right)-T\left(y\left(\theta,T+\mu\hat{T}\right)\right)-\mu\hat{T}\left(y\left(\theta,T+\mu\hat{T}\right)\right)\right)\times$$
$$\times\left(1-T'\left(y\left(\theta,T+\mu\hat{T}\right)\right)-\mu\hat{T}'\left(y\left(\theta,T+\mu\hat{T}\right)\right)\right)=v'\left(\frac{y\left(\theta,T+\mu\hat{T}\right)}{\theta}\right)\frac{1}{\theta}.$$

Taking the derivative with respect to μ and evaluating at $\mu=0$:

$$u''\left(y\left(\theta\right)-T\left(y\left(\theta\right)\right)\right)\left(\delta y(\theta)-T'\left(y\left(\theta\right)\right)\delta y(\theta)-\hat{T}\left(y\left(\theta\right)\right)\right)\left(1-T'\left(y\left(\theta\right)\right)\right)$$
$$+u'\left(y\left(\theta\right)-T\left(y\left(\theta\right)\right)\right)\left(-T''\left(y\left(\theta\right)\right)\delta y(\theta)-\hat{T}'\left(y\left(\theta\right)\right)\right)=v''\left(\frac{y\left(\theta\right)}{\theta}\right)\frac{\delta y(\theta)}{\theta^2}.$$

Define

$$\eta\left(y\left(\theta\right)\right)=$$
$$=-\frac{u''\left(y\left(\theta\right)-T\left(y\left(\theta\right)\right)\right)\left(1-T'\left(y(\theta)\right)\right)}{u''\left(y(\theta)-T(y(\theta))\right)\left(1-T'\left(y(\theta)\right)\right)^2-u'\left(y(\theta)-T(y(\theta))\right)T''(y(\theta))-v''(\frac{y(\theta)}{\theta})\frac{1}{\theta^2}},$$
$$\varepsilon\left(y\left(\theta\right)\right)=$$
$$-\frac{u'\left(y(\theta)-T(y(\theta))\right)}{u''\left(y(\theta)-T(y(\theta))\right)\left(1-T'\left(y(\theta)\right)\right)^2-u'\left(y(\theta)-T(y(\theta))\right)T''(y(\theta))-v''(\frac{y(\theta)}{\theta})\frac{1}{\theta^2}},$$

and solve for

$$\delta y(\theta)=-\eta\left(y\left(\theta\right)\right)\hat{T}\left(y\left(\theta\right)\right)-\varepsilon\left(y\left(\theta\right)\right)\hat{T}'\left(y\left(\theta\right)\right).$$

6.2 Proof of Lemma 2

The effect on the social welfare from the perturbation of the taxes by $\mu\hat{T}$, evaluated at $\mu = 0$ is given by

$$\begin{aligned}
&\delta W(T,\hat{T}) = \\
&= \int_{\Theta} \left(\begin{array}{c} u'\left(y\left(\theta\right) - T\left(y\left(\theta\right)\right)\right)\left(\left(1 - T'\left(y\left(\theta\right)\right)\right)\delta y\left(\theta\right) - \hat{T}\left(y\left(\theta\right)\right)\right) - \\ -v'\left(\frac{y(\theta)}{\theta}\right)\frac{\delta y(\theta)}{\theta} \end{array} \right) h\left(\theta\right) d\theta + \\
&+ \int_{\Theta} \left(T'\left(y\left(\theta\right)\right)\delta y\left(\theta\right) + \hat{T}\left(y\left(\theta\right)\right)\right) h\left(\theta\right) d\theta = \\
&= \int_{\Theta} \left(\left(1 - u'\left(y\left(\theta\right) - T\left(y\left(\theta\right)\right)\right)\right)\hat{T}\left(y\left(\theta\right)\right) + T'\left(y\left(\theta\right)\right)\delta y\left(\theta\right)\right) h\left(\theta\right) d\theta = \\
&= \int_{\Theta} \left(\begin{array}{c} \left(1 - u'\left(y\left(\theta\right) - T\left(y\left(\theta\right)\right)\right) - \eta\left(y\left(\theta\right)\right)T'\left(y\left(\theta\right)\right)\right)\hat{T}\left(y\left(\theta\right)\right) - \\ -\epsilon\left(y\left(\theta\right)\right)T'\left(y\left(\theta\right)\right)\hat{T}'\left(y\left(\theta\right)\right) \end{array} \right) h\left(\theta\right) d\theta.
\end{aligned}$$

Define

$$\gamma\left(y\left(\theta\right)\right) = u'\left(y(\theta) - T\left(y(\theta)\right)\right) + T'\left(y\left(\theta\right)\right)\eta\left(y\left(\theta\right)\right).$$

Change of variables from θ to y completes the proof.

6.3 Equivalence of (6) and the ABC formulas

Integrating (6) with respect to y yields the familiar ABC formula due to Diamond (1998) and Saez (2001): the optimal tax schedule T_* satisfies, for all $y \in Y$:

$$T_*'\left(y\right) = \frac{1}{\varepsilon_*\left(y\right)}\frac{1 - \Phi_*\left(y\right)}{\phi_*\left(y\right)}\left\{1 - \mathbb{E}[\gamma_*\left(x\right) \mid x \geq y]\right\}, \tag{13}$$

where $\mathbb{E}[\gamma_*\left(x\right) \mid x \geq y] = \frac{\int_y^{\bar{y}} \gamma_*(x)\phi_*(x)dx}{1-\Phi_*(y)}$. Equation (13) shows that the optimal marginal tax rate at income level y is the product of three terms. First, it

is proportional to the inverse elasticity of labor supply at income y, $1/\varepsilon_*(y)$: the higher the disincentive effects of taxes, the lower the optimal tax rate. The second term is related to the hazard rate of the income distribution, $(1-\Phi_*(y))/\phi_*(y)$. This is a benefit-cost ratio that measures the fraction of agents whose tax liability increases lump-sum in response to a marginal tax rate increase at income y, relative to the fraction of agents whose labor supply is distorted. The third term accounts for the fact that an increase in the marginal tax rate at income y lowers the social welfare contribution of all agents with income $x \geq y$, by their social marginal utility of income.

6.4 Proof of Proposition 1

Part (1). Consider the derivative

$$\frac{\partial}{\partial t}\int_Y q_t(x,y)\,T_*(y)\,dy = \int_Y \frac{\partial}{\partial t} q_t(x,y)\,T_*(y)\,dy =$$

by the heat equation

$$= \int_Y \frac{\partial}{\partial y}\left(\varepsilon_*(y)\,\phi_*(y)\,\frac{\partial}{\partial y} q_t(x,y)\right) T_*(y)\,dy =$$

integrating twice by parts, using the assumption that densities are zero at the boundaries, and rearranging

$$= \int_Y q_t(x,y)\frac{\partial}{\partial y}\left(\varepsilon_*(y)\,\phi_*(y)\,\frac{\partial}{\partial y} T_*(y)\right) dy =$$

using equation (6)

$$= -\int_Y q_t(x,y)\,(1-\gamma_*(y))\,\phi_*(y)\,dy.$$

Integrating the equation

$$\frac{\partial}{\partial t}\int_Y q_t(x,y)\,T_*(y)\,dy = -\int_Y q_t(x,y)\,(1-\gamma_*(y))\,\phi_*(y)\,dy,$$

we get

$$\int_Y q_t(x,y)\,T_*(y)\,dy = T_*(x) - \int_0^t \int_Y (1-\gamma_*(y))\,\phi_* q_s(x,y)\,\phi_*(y)\,dyds.$$

Part (2). A classical result of Aronson (1968) is that the heat kernel $q_t(x,y)$ on a general n-dimensional manifold M(satisfying very mild regularity assumptions) satisfies what is called a Gaussian upper bound

$$q_t(x,y) \leq \frac{c_1}{t^{n/2}} \exp\left(-\frac{d(x,y)^2}{c_2 t}\right), \quad \forall t>0, x, y \in M,$$

where $d(x,y)$ is the geodesic distance between x and y, and c_1 and c_2 are positive constants. In particular, while the heat kernel $q_t(x,y)$ may no longer look like a Gaussian centered at y having variance t, it certainly has the same decay behavior. That is, it acts as a local averaging operator at scale $d(x,y) \sim \sqrt{t}$ and averages the nearby income, where the nearby is given by the scale $\sqrt{t}$.[10] In the proposition, we use a slightly more general result in Metafune, Ouhabaz, and Pallara (2011).

Part (3). The classical results for the short-time asymptotics (see, e.g., Varadhan (1967), Molchanov (1975), and Grigor'yan (2009)) imply that for $t \to 0$, the heat kernel is the Gaussian with the scale determined by the conductivity parameter $\varepsilon_*(x)\,\phi_*(x)$. In the next section, we provide a detailed proof that applies also to the optimum.

[10] If one is interested in the higher order expansions, those can be straightforwardly derived in closed form to any order using the parametrix method which represents the heat kernel as the sum of the Gaussian and the higher order corrections (see, e.g., Friedman (2008)).

Returning to the interpretation of q_t as creating an averaging operator at scale $\sim \sqrt{t}$, Part (4) of the proposition shows that smoothing for different scales is linked. This semi-group property follows from Grigor'yan (2009).

6.5 Mathematical foundations for gradient flows

Finite-dimensional spaces. Gradient flows are natural mathematical objects attached to functions or functionals, mapping to real numbers. For simplicity, start with a differentiable function $V : \mathbb{R}^n \to \mathbb{R}$ and define the gradient flow as a curve $x : [0, \infty] \to \mathbb{R}^n$ starting at some point $x_0 \in \mathbb{R}^n$ with the property that the curve always flows in the direction of steepest descent of V. Intuitively, this direction is determined by the gradient of V. Formally, we want to choose the vector $\hat{x} \in \mathbb{R}^n$ with $\|\hat{x}\|_{\ell^2} = 1$ that minimizes

$$\lim_{\mu \to 0} \frac{1}{\mu} [\, V(x + \mu \hat{x}) - V\,(x)\,] \;\; = \;\; \langle\, \nabla V\,,\, \hat{x}\,\rangle_{\ell^2},$$

where the equality follows from the definition of the gradient. This gives rise to an ordinary differential equation that describes the law of motion of $x_t \in \mathbb{R}^n$ for $t \geq 0$ and assuming that the rate of decrease is one:

$$\frac{d}{dt} x_t \;\; = \;\; -\nabla\, V(x_t),$$

with the property that V is decreasing along the flow of x since

$$\frac{d}{dt} V(x_t) = \langle \nabla V, \frac{d}{dt} x_t \rangle = -\|\nabla V(x_t)\|^2 < 0.$$

While this model is rather classical and the existence and uniqueness properties of the solution are well known, understanding the actual dynamical behavior

can pose considerable challenges (recent examples being given by Tao (2017), Steinerberger (2018)).

Infinite-dimensional spaces. The very same principle can be applied in settings where the underlying domain is not finite-dimensional but instead given by the space of functions. We illustrate this with a representative example. We may define a functional Ψ by assigning to any twice-differentiable *function* $f \in \mathcal{C}^2(\mathbb{R}, \mathbb{R})$, the number

$$V(f) \;=\; \frac{1}{2}\int_{\mathbb{R}} |f'(x)|^2 \, \mathrm{d}x.$$

It is easy to show that the Gateaux differential of V in the direction $\hat{f}$ is given by $\delta V(f, \hat{f}) = \int_{\mathbb{R}} f'(x)\, \hat{f}'(x)\, \mathrm{d}x$. An integration by parts implies that $\delta V(f, \hat{f}) = -\int_{\mathbb{R}} f''(x)\, \hat{f}(x)\, \mathrm{d}x$. More generally, for any function $f \in \mathcal{C}^2(\mathbb{R}^n, \mathbb{R})$, let $V(f) = \int_{\mathbb{R}^n} |\nabla f|^2$. We then have $\delta V(f, \hat{f}) = \int_{\mathbb{R}^n} \nabla f \cdot \nabla \hat{f}$. By Green's first identity, this can be represented as a functional $\hat{f} \mapsto \langle -\Delta f, \hat{f} \rangle_{L^2}$, where Δ denotes the Laplace operator, thus recovering the same structure as above. That is, in order to flow in the direction of steepest descent of the functional V, we must set

$$\frac{\partial}{\partial t} f_t \;=\; \Delta\, f_t.$$

This gives rise to a law of motion for the function f characterized by a parabolic PDE (namely, a heat equation). Needless to say, even showing that all of these operations remain valid for any time $t > 0$ is a difficult task, the theory of partial differential equations being substantially more challenging than that of ordinary differential equations.

Other norms for the graduate flow We also could have considered the weighted-L^2 norm

$$\|T_t\|^2 = \int_Y \kappa_t(y) (T_t(y))^2 \, dy,$$

for some weights $\kappa_t(y)$ and this expression can be represented as $\langle \kappa_t^{-1} \Lambda_t, \hat{T}_t \rangle$ with the resulting gradient flow

$$\frac{\partial T_t(y)}{\partial t} = (\kappa_t(y))^{-1} (1 - \gamma_t(y)) \phi_t(y) + (\kappa_t(y))^{-1} \frac{\partial}{\partial y} (T_t'(y) \varepsilon_t(y) \phi_t(y)).$$

The analysis for this case is identical. There is a re-interpretation of such a weight as simply changing the metric of the underlying manifold $\mathbb{R}$. Put differently, one can interpret everything as an equal-weight problem on a curved geometry; heat and associated processes are not very sensitive to "curving" (heat propagates on a plane and on a sphere in roughly the same sense) – see, e.g., Taylor (1996), HÃ¶rmander (2003), and Grigor'yan (2009).

We now briefly discuss a broader question of other norms. Suppose we were to define a gradient descent of V with respect to the ℓ^p-norm with $p \neq 2$ or a more general norm $\|\cdot\|_X$. This means we want to minimize

$$\arg \min_{\|h\|_X = \varepsilon} f(x+h) - f(x) \sim \arg \min_{\|h\|_X = \varepsilon} \langle \nabla f, h \rangle$$

However, this can be equivalently written as a minimization problem over

$$\arg \min_{\|h\|_{\ell^2} = \varepsilon} \left\langle g\left(\frac{h}{\|h\|_{\ell^2}}\right) \nabla f, h \right\rangle,$$

where g is a function mapping the ℓ^2-unit ball to the $X-$unit ball (this is always possible in finite dimensions since any two norms are equivalent; in particular, g is positive, bounded from above and bounded away from 0.

We see that this change in the underlying geometry thus corresponds to a gradient flow of a different function weighted by a certain directionality. The analysis can be extended to a broader set of infinitely dimensional settings by considering approximations with the finite-dimensional case.

In one dimension that we have, the body of literature on parabolic PDEs shows that for almost any modification of the problem, the heat kernel looks and behaves exactly as a Gaussian – this is true for the heat kernels on arbitrary manifolds, for very wide classes of conductivities σ, and for a very broad range of spaces (see e.g, Grigor'yan (2009) or Bogachev, Krylov, Röckner, and Shaposhnikov (2015) for extensive reviews).

6.6 Proof of Proposition 2

The proof uses the Feynman-Kac formula for path integrals (Lorinczi, Hiroshima, and Betz 2011) to study the short-time behavior of solutions of equations of the type (10). Let $B(s)$ denote the diffusion process that satisfies the SDE:

$$dB_s = \sigma_s'(B_s)\,ds + \sqrt{2\sigma_s(B_s)}dW_s,$$

where W is a Brownian motion. We then have

$$T_{\tilde{t}}(x) \quad = \mathbb{E}\left[\int_t^{\tilde{t}} (1-\gamma_s(B(s)))\,\phi_s(B(s))\,ds\right] + \mathbb{E}[T_t(B_{\tilde{t}})],$$

where the expectation runs over the diffusion process B_s, started in x at time t and running up to time $\tilde{t}$.[11] We can now perform a Taylor expansion of these quantities. Up to the first order, diffusivity is constant. The short-time

[11] Technically, we need to specify boundary conditions for the Brownian motion $B(s)$. However, since we are only using Brownian motion for very small times $\tilde{t}$, the exact form of the boundary conditions does not matter.

asymptotics for Brownian motion is then given by a Gaussian distribution $B_{\tilde{t}} \sim x+\sigma_t'(x)\left(\tilde{t}-t\right)+\sqrt{2\sigma_t(x)}\left(W_{\tilde{t}}-W_t\right) \sim N\left(x+\sigma_t'(x)\left(\tilde{t}-t\right), 2\sigma_t(x)\left(\tilde{t}-t\right)\right)$ and

$$\text{distribution of } B_{\tilde{t}} \;\sim\; \frac{1}{\sqrt{4\pi\sigma_t(x)\left(\tilde{t}-t\right)}}\exp\left(-\frac{\left(y-x-\sigma_t'(x)\left(\tilde{t}-t\right)\right)^2}{4\sigma_t(x)\left(\tilde{t}-t\right)}\right).$$

Then, up to a first order for $\tilde{t}$ small,

$$\int_t^{\tilde{t}}\left(1-\gamma_t\left(B(s)\right)\right)\phi_t\left(B(s)\right)ds \sim \left(\tilde{t}-t\right)\left(1-\gamma_t(x)\right)\phi_t(x).$$

This implies that

$$T_{\tilde{t}}(x) \sim \left(\tilde{t}-t\right)\left(1-\gamma_t(x)\right)\phi_t(x)+\mathbb{E}[T_t(B_{\tilde{t}})] \sim$$

$$\sim \left(\tilde{t}-t\right)\left(1-\gamma_t(x)\right)\phi_t(x)+\int_Y \frac{1}{\sqrt{4\pi\sigma_t(x)\left(\tilde{t}-t\right)}}\exp\left(-\frac{\left(y-x-\sigma_t'(x)\left(\tilde{t}-t\right)\right)^2}{4\sigma_t(x)\left(\tilde{t}-t\right)}\right)T_t(y)dy.$$

6.7 Proof of Proposition 3

To ease notational burden, we suppress the dependence on t in this and the next section. Let us compute the directional derivative in the direction of a function w evaluated at T

$$\delta\mathcal{J}(w)=\lim_{\varepsilon\to 0}\frac{\mathcal{J}(T+\varepsilon w)-\mathcal{J}(T)}{\varepsilon}.$$

We see that

$$\lim_{\varepsilon \to 0} \frac{\mathcal{J}(T+\varepsilon w)-\mathcal{J}(T)}{\varepsilon} = \int_Y \sigma(y)(T'(y)-\tau'(y))w'(y)\,dy = \\ = \int_Y w(y)\,\frac{\partial}{\partial y}\left(\sigma(y)\left(\tau'(y)-T'(y)\right)\right)dy.$$

This shows that the negative gradient flow is given by

$$\frac{\partial T}{\partial t} = -\frac{\partial}{\partial y}\left(\sigma(y)\left(\tau'(y)-T'(y)\right)\right) = -\frac{\partial}{\partial y}\left(\sigma(y)\tau'(y)\right) + \frac{\partial}{\partial y}\left(\sigma(y)T'(y)\right).$$

Using the equation (12) for τ, we see that $-\frac{\partial}{\partial y}\left(\sigma(y)\tau'(y)\right) = \left(1-\gamma(y)\right)\phi(y)$ and we have established the desired claim.

6.8 Additional smoothing results for section 4.2

The next proposition shows that the gradient flow smoothes a measure of variability of the tax schedule, the squared deviation from the limiting stationary solution. Moreover, such smoothing is exponential. As we discussed above, this result is not about the convergence to the optimal tax T_* but rather about the smoothing behavior with respect to the correct reference point, the stationary solution $\tau(y)$. In this section, whenever it does not cause confusion, we suppress the dependence on t.

Proposition 4. *Let $T(t,y)$ be the solution to*

$$\frac{\partial T}{\partial t} = \left(1-\gamma(y)\right)\phi(y) + \frac{\partial}{\partial y}\left(\sigma(y)\,\frac{\partial T}{\partial y}\right),$$

$T(0,y)$ be an arbitrary initial tax schedule, $\tau(y)$ be the solution to the stationary problem (12), and λ_1 be the first eigenvalue of the associated Sturm-

Liouville operator $Hz = -\frac{\partial}{\partial y}\left(\sigma(y)\frac{\partial z}{\partial y}\right)$. *Then,* $\forall t > 0$:

$$\int_Y (T(t,y) - \tau(y))^2 dy \leq e^{-2\lambda_1 t} \int_Y (T(0,y) - \tau(y))^2 dy.$$

Proof. We use the standard Sturm-Liouville theory (see Zettl (2010), Teschl (2012), Titchmarsh (1962)) to prove this result. This equation can be studied by first solving for the stationary problem

$$0 = (1 - \gamma(y))\phi(y) + \frac{\partial}{\partial y}\left(\sigma(y)\frac{\partial \tau}{\partial y}\right).$$

This leads to an equation for $z(t,x) = T(t,x) - \tau(x)$ given by

$$\frac{\partial z}{\partial t} = \frac{\partial}{\partial y}\left(\sigma(y)\frac{\partial z}{\partial y}\right).$$

It remains to study problems of this type. We will do so by studying the spectrum of the associated differential operator H given by

$$Hz = -\frac{\partial}{\partial y}\left(\sigma(y)\frac{\partial z}{\partial y}\right)$$

or, in other words, we study the problem $Hz = \lambda z$. This eigenvalue problem has a discrete sequence of admissible values λ for which the equation has a solution: these values $0 < \lambda_1 < \lambda_2 < \dots$ are the eigenvalues of this operator of Sturm-Liouville type, the corresponding solutions will be denoted by $\eta_1, \eta_2, \dots$ and are assumed to be L^2-normalized, i.e. $\|\eta_n\|_{L^2} = 1$. We note that $\lambda_0 = 0$ is a special value and $\eta_0 = \text{const}$. We see that these eigenfunctions are necessarily

orthogonal in L^2 since, again by integration by parts,

$$\begin{aligned}\int_Y \eta_k(y)\eta_\ell(y)dy =& \frac{1}{\lambda_k}\int_Y -(\sigma(y)\eta_k'(y))'\eta_\ell(y)dy \\ =& -\frac{1}{\lambda_k}\int_Y -\sigma(y)\eta_k'(y)\eta_\ell'(y)dy \\ =& \frac{1}{\lambda_k}\int_Y (-\sigma(y)\eta_\ell'(y))'\eta_k(y)dy \\ =& \frac{\lambda_\ell}{\lambda_k}\int_Y \eta_k(y)\eta_\ell(y)dy\end{aligned}$$

If $k \neq \ell$, then the factor in front of the integral is different from 1 and the integral is therefore 0. This together with the completeness of the system of eigenfunctions in L^2 allows us to expand an arbitrary initial function $T(0,x)$ into a series

$$T(0,x) = \tau(x) + \sum_{k=1}^{\infty} \langle z(0,x), \eta_k(x)\rangle \eta_k(x).$$

We will abbreviate $a_k = \langle z(0,x), \phi_k(x)\rangle$ for simplicity of exposition. We then claim that

$$T(t,x) = \tau(x) + \sum_{k=1}^{\infty} a_k e^{-\lambda_k t}\eta_k(x)$$

is a solution. This can be verified by computing

$$\left(\frac{d}{dt} - \frac{d}{dx}\sigma(x)\frac{d}{dx}\right)\sum_{k=1}^{\infty} a_k e^{-\lambda_k t}\eta_k(x) = \sum_{k=1}^{\infty} a_k \left(\frac{d}{dt} - \frac{d}{dx}\sigma(x)\frac{d}{dx}\right)e^{-\lambda_k t}\eta_k(x).$$

The separation of variables implies that

$$\left(\frac{d}{dt} - \frac{d}{dx}\sigma(x)\frac{d}{dx}\right)e^{-\lambda_k t}\eta_k(x) = e^{-\lambda_k t}\left(-\lambda_k\eta_k(x) - \frac{d}{dx}\sigma(x)\frac{d}{dx}\eta_k(x)\right) = 0$$

as desired. Since we now have a complete description of a solution, we can analyze the convergence to the limiting function arising for $t \to \infty$ at a greater

level of detail: we have

$$\int_Y (T(t,y) - \tau\,(y))^2 dy =$$

$$= \int_Y \sum_{k,\ell=1}^{\infty} a_k e^{-\lambda_k t} \eta_k(y) a_\ell e^{\lambda_\ell t} \eta_\ell(y) dy =$$

$$= \int_Y \sum_{\ell=1}^{\infty} a_\ell^2 e^{-2\lambda_l t} \eta_l(y)^2 dy = \sum_{\ell=1}^{\infty} a_\ell^2 e^{-2\lambda_k t} \leq e^{-2\lambda_1 t} \sum_{\ell=1}^{\infty} a_\ell^2.$$

We note that

$$\sum_{\ell=1}^{\infty} a_\ell^2 = \int_Y (T(0,y) - \tau(y))^2 dy$$

and that we have therefore shown that

$$\int_Y (T(t,y) - \tau\,(y))^2 dy \leq e^{-2\lambda_1 t} \int_Y (T(0,y) - \tau(y))^2 dy.$$

Appealing to the classical Rayleigh-Ritz formula, we see that

$$\lambda_1 = \inf_{\int_Y f(y)dy=0} \frac{\int_Y \sigma(y) f'(y)^2 dy}{\int_Y f(y)^2 dy} \tag{14}$$

where the last step follows from the classical Neumann eigenvalue computation for the homogeneous rod (see Courant and Hilbert (1989)). This shows that for sufficiently regular values of $\phi(y)$, we can expect $\lambda_1 > 0$ and therefore the distance to $\tau(y)$ undergoes exponential decay. □

6.9 Trajectory of the tax reform

In this section, we briefly discuss a construction of a trajectory of the tax reform beyond the short-term asymptotics.

Equation (10) changes the tax schedule in favor of increasing social welfare, letting ϕ_t and ε_t be endogenously driven by T_t – that is, taking into account the fact that the density and the elasticity change in response to the evolution of taxes. We propose to evolve the system separately (this underlying idea is a straightforward application of "operator splitting"). The simplest instance of this idea is as follows. Suppose we are given a system of ordinary differential equations given as

$$\frac{d}{dt}u(t) = (A+B)u(t),$$

then the solution is given by the matrix exponential $u(t) = e^{t(A+B)}u(0)$. A formal Taylor series expansion suggests that

$$\begin{aligned} e^{t(A+B)} &= \mathrm{Id} + t(A+B) + \mathcal{O}(t^2) \\ &= (\mathrm{Id} + tA)(\mathrm{Id} + tB) + \mathcal{O}(t^2) \\ &= e^{tA}e^{tB}u(0) + \mathcal{O}(t^2). \end{aligned}$$

These computation suggest that, at least for small values of t, we may solve the system by first evolving along the simpler system $\dot{u}(t) = Au$ and then along the system $\dot{u} = Bu$ and alternate in this manner (Varga (1962), Glowinski and Osher (2016)).

We apply the very same method in our problem: more precisely, we fix the distribution of incomes ϕ_t and the elasticity ε_t for a short period of time δt, evolve T_t, and then re-compute $\phi_{t+\delta t}$ and $\varepsilon_{t+\delta t}$ based on the new tax function $T_{t+\delta t}$.[12] In standard situations, this procedure will converge to a solution path of the dynamical system as $\delta t \to 0$ (Glowinski, Osher, and Yin (2016)). The operator splitting technique also has a natural economic meaning. The

[12]This also can be regarded as a classical numerical technique for systems of this type.

government evolves taxes in the direction of increased welfare, keeping the density of agents' incomes, marginal social welfare weights, and elasticities fixed at their value observed in the current economy. That is, the government evaluates the changes in revenues under the current information given by the exogenous sufficient statistics evaluated at a given initial time – this is our notion of the short term reform.

This implies equation (10) is a heat equation (with source term $(1-\gamma)\,\phi$ and local conductivity $\sigma = \varepsilon\phi$), i.e. a PDE of the form

$$\frac{\partial T}{\partial t} = (1-\gamma(y))\,\phi(y) + \frac{\partial}{\partial y}\left(\sigma(y)\,\frac{\partial T}{\partial y}\right)$$

and guarantees in particular that the problem always has a solution (the heat equation being well-posed). We also note that our assumption that the density tends to 0 at the boundary of the interval implies that no boundary conditions need be imposed. Since heat equations are among the most well-known and well-behaved partial differential equations, we can apply standard mathematical results to obtain theoretical properties of the evolution of the tax schedule over time.

Fixing γ, ϕ, and ε and letting T evolve for a short amount of time, then unfreezing γ, ϕ, and ε and recomputing it can be regarded as a classical example of operator splitting. While the analysis of convergence of this dynamic system is outside the scope of the paper, one can expect that for sufficiently short time steps, the solution converges to the global optimum at a great level of generality. For example, the review of Glowinski, Osher, and Yin (2016, p.13) concludes: "Last but not least, operator splitting algorithms are theoretically attractive because they converge under very few assumptions." More broadly, the splitting procedure we use is similar in spirit to the ones used in

physical sciences where the split terms correspond to different physical processes – for example, splitting convection from diffusion (see, e.g. MacNamara and Strang (2016)) or splitting fast from slow variables. Finally, one can think of the results in this section as justifying the commonly used iterative method of computing optimal taxes in such environments. That is, our analysis shows that continuous version of the iterative fixed point method commonly used for computing optimal taxes (see, e.g., Brewer, Saez, and Shepard 2010) can be represented within each step as a heat equation.