# On Vickrey's Income Averaging

**Stefan Steinerberger**[*]   **Aleh Tsyvinski**

Yale University               Yale University[†]

April 13, 2020

### Abstract

We consider a small set of axioms for income averaging – recursivity, continuity, and the boundary condition for the present. These properties yield a unique averaging function that is the density of the reflected Brownian motion with a drift started at the current income and moving over the past incomes. When averaging is done over the short past, the weighting function is asymptotically converging to a Gaussian. When averaging is done over the long horizon, the weighing function converges to the exponential distribution. For all intermediate averaging scales, we derive an explicit solution that interpolates between the two.

1

# 1   Introduction

How to average over the past is one of the most basic questions that arises in a variety of economic fields. Our particular focus is on a classic public economic issue – how to average income for the tax purposes – but the answer is broadly applicable to many other topics. The question of income averaging is typically attributed to Vickrey (1939) and can be summarized as follows.[1] Consider a progressive tax system and two taxpayers with the same income over a period of time. The person with the constant income pays a lower total amount of taxes than the person with the fluctuating income. Averaging income to equalize the tax burden then may be desirable.[2]

In this paper, we want to abstract from the desirability of averaging. Our goal is to formalize the question of averaging and to propose a small set of reasonable axioms that an averaging mechanism may possess. Given this set of the axioms, we are interested in what averaging functions may arise. More broadly, the question of averaging over the past, given a small set of assumptions, may be of use in a variety of other applications such as behavioral economics or dynamic contracting.

We chose two main axioms that the averaging function that weights income at different periods should satisfy. The first is recursivity – averaging should treat various scales of smoothing in a unified way. In this context, the assumption implies that averaging at some scale and then at another scale is equivalent to averaging at the combined scale. This condition ensures that all these outcomes agree: there is no difference in averaging over, say, a year, twelve units of a month or 52 weeks. Alternatively, one can think of this assumption as stating that no scale of averaging is singled out and all of them are treated equally. It is natural that a reasonable multiscale averaging method has the scales connected with this intrinsic compatibility condition. The second is a continuity or locality assumption that requires that the very distant

---

[1]See also Simons (1938).

[2]A related issue is the choice of the reference period for taxation. Most commonly the taxes are assessed on the annual basis. However, in principle, a government may use a shorter or a longer accounting period.

incomes do not play a disproportionate role in averaging. We also need to postulate how we treat the present, that is, to set the boundary condition of the averaging rule. In addition, we impose some other, less essential conditions, such as various normalizations.

In principle, many weighting functions are possible: equal weights for all income, assigning lower weights to the more distant past, singling out some incomes, etc. We show that the small number of assumptions that we make result in a definite general weighting scheme. Specifically, the smoothed incomes and the weighting function are a solution to the second order parabolic partial differential equation. The easiest way to describe the intuition behind this result is using a probabilistic argument. One can think of averaging being done using a stochastic process. Fix a given time and consider a stochastic process originating in that period. The probability of the process reaching some other time is then the weight that the averaging assigns to that income. The recursivity (or semi-group) assumption implies that such a process is Markovian. The locality assumption implies that the process has continuous paths. The classic result is that the process is a diffusion. Since we are averaging over the past and given the behavior on the boundary, the resulting process is then a reflected Brownian motion with a drift. The weighting function is the density of this process. The smoothed incomes in turn satisfy a second order parabolic partial differential equation.[3]

The density of such process that defines a weighting function is known in a simple closed form. From the point view of present, the averaging function has particularly meaningful properties. Consider averaging over the very short period in the past. In this case, the drift does not have any substantial effect and the averaging is done mainly via the normal density with the nonzero mean determined by the drift. Consider now averaging over the long horizon. A remarkable fact in probability theory is that for any constant positive drift, the weighting function converges to an exponential. For all the averaging scales in between these two, there is an explicit solution that interpolates

---

[3]In fact, we do not have to run average on the pre-tax income and could rather consider smoothing or averaging the tax contributions or the post tax income directly.

between them. We also show that the smoothed income has a particular strong smoothing structure of a gradient flow.

Finally, we want to remark that our work connects the economic question of averaging to two literatures that previously have not been represented in economics. The first is the mathematical imaging and vision literature that considers representation of images over various smoothing scales (see, e.g. Aubert and Kornprobst 2006 and Lindeberg 2013). The second, is the Schönberg's theory of variation diminishing transformations and Polya's frequency functions (Schönberg 1948, Steinerberger 2019).[4] Our results nest and interpolate between the results on averaging obtained in these two approaches.

## 2    Literature

The question of Vickrey's income averaging is classic and familiar to any student of public finance. While this mechanism is not widely used in fiscal practice today (there are some provisions for income averaging for artists and farmers), it was extensively employed in the past. Great Britain applied a progressive tax schedule to the average of the individual income of the previous three years from 1799 to 1926. Between 1923 and 1938 Australia used a five-year moving average of income over the past five years (Holt 1949). Gordon and Wen (2018) describe a more recent experience the summary of which we present: The United States introduced general income averaging in 1964 and it was repealed in the 1986 Tax Reform Act. In Canada, a similar policy to that in the United States was introduced in 1972 together with forward averaging of the income-averaging annuity contracts and had been in effect until 1988. The impact of progressive tax rates on realized capital gains was one reason for setting of low tax rates on capital gains. There are also several prominent recent proposals to reintroduce income averaging. Batchelder (2003) proposes targeted averaging for the poor in the context such as EITC in the US. In Canada, Mintz and Wilson (2002) for primarily retirement savings plans and

---

[4]The literature on total positivity that builds on this work (Karlin 1968) is used more extensively in economics.

Gordon and Wen (2018) more broadly argue for income averaging.

The question of averaging regularly appears in models of dynamic taxation in which agents experience idiosyncratic shocks. The Mirrlees review (Mirrlees and Adam 2010) that analyzes the theoretical foundations behind the practical tax design devotes a signficant amount of space to the question of lifetime earning variability and lifetime income as the tax basis. Diamond (2009) discusses Vickrey's income averaging in the context of design of pensions systems. Huggett and Para (2010) study optimal lifetime tax in a model of social insurance. Kapicka (2017) extends the model of Heathcote, Storesletten, and Violante (2014) to allow for history-dependent taxes. He finds that the optimal weights on past incomes decline geometrically at a rate equal to the discount rate and that the gains from the history dependence are large. Jacobs (2017) reviewing digitization and taxation states increased feasibility of practical implementation of the income averaging rules.

There are relatively few recent studies of the empirical effects of income averaging.[5] The most comprehensive is Gordon and Wen (2018) which also contains a review of the older literature on the topic. For the Canadian data, they find that while fluctuation penalty is low on average, 10 percent of taxpayers faced annual tax penalties of 1 percentage point of their income and 1 percent of taxpayers paid 4 percentage points. That is, those in the top 1 percentile of the penalty paid 4 percent of their average annual income more in taxes per year if they had been able to perfectly average. The highest percentile is composed largely of the self-employed or those with the realized capital gains. What is more, 57 percent of taxpayers located between the 95th and 100th percentiles of the penalty are from the bottom income quintile. Similarly, in the US, Batchelder (2003) finds that families the bottom quartile of families ranked by the annual income faced an additional effective tax rate of 2.0 percentage points higher under annual income measurement than it would be if income were fully averaged, whereas for the top quartile's rate it is only 0.5 percentage points higher. Bargain, Trannoy, and Pacifico (2017) examine

---

[5]In terms of the assessment of the practical implementation of income averaging, there is an extensive literature in law (see, e.g., a summary in Buchanan 2005).

French administrative data and show that increasing the tax frequency can lead to substantial social welfare gains, coming to an important degree from the bottom of the income distribution.[6]

There are two related answers in the mathematical literature to the question of averaging.

The first is given by the literature on scale-space in mathematical imaging and vision analysis (see e.g., books by Aubert and Kornprobst 2006 and Lindeberg 2013). This literature studies image representations at various scales – from the finest scale that represents the original image to the coarser scales of the smoothed versions of the images. Smoothing is conducted at various scales which are tightly related to each other. The analysis there is mainly concerned with two-sided averages and derives a deep and substantial result – a Gaussian kernel arises as a unique averaging object based on a small set of assumption when averaging is done over the whole line. The Gaussian kernel has many properties and appears in a variety of fields of mathematics: in the scale-space literature it is derived primarily using two main assumptions: varying forms of smoothing and the semigroup (recursivity) structure.[7] For example, Lindeberg (1997) writes "A notable coincidence between the different scale-space formulations that have been stated is that the Gaussian kernel arises as a unique choice for a large number of different combinations of underlying assumptions (scale-space axioms)." This is very important point: given a variety of reasonable assumptions you make (there is quite a lot of natural things one could assume), you usually end up with the Gaussian. At the same time, the one-sided question of averaging over the past is much less studied and the answer is less canonical in that literature.[8]

---

[6]Saez (2002) considers the question for understanding the optimal period for computing the time liability.

[7]The literature identified also some other possible results where under different assumptions one may get the stable distributions from probability theory (see e.g., Pauwels, Van Gool, Fiddelaers, and Moons 1995) or a nonlinear diffusion (see e.g., Alvarez, Guichard, Lions, and Morel 1993) as the averaging principle – yet the Gaussian is a canonical and most widely used answer.

[8]See Fagerström (2005), Lindeberg and Fagerström (1996), Lindeberg (2011), and Salden, ter Haar Romeny, and Viergever (1998).

The second literature (Steinerberger 2019) considers one-sided averages without the semi-group property. The main insight is based on the Schönberg's theory of variation diminishing transformations and Polya's frequency functions (Schönberg 1948). The variation diminishing property (total positivity) is a particular way to define uniform smoothing at all scales, and it requires that the number of the function's crossings at any levels is decreased.[9] The second assumption that is imposed is monotonicity where the more recent past is weighed heavier than the more distant path. Steinerberger (2019) shows that total positivity and monotonicity on the half line lead to the unique weighting given by the exponential distribution. That is, the "exponential smoothing" classical in time series analysis (Brown 1957 and Holt 1957 ) arises from a small set of axioms.

Our result can generate the conclusions of both of these approaches from a reasonable set of assumptions as well as a range of intermediate results. This leads to the weighting scheme that interpolates between the Gaussian (similar in spirit to the scale-space theories) for a short horizon of averaging and the exponential smoothing for a positive drift case if one considers the long time horizon.

# 3   A question of averaging

In this section, we define a question that we aim to study. Let a bounded measurable income function $f : \mathbb{R} \to \mathbb{R}$ be defined on time $x \in (-\infty, \infty)$. That is, $f(x)$ is income at time $x$. We are interested, at a given time, in finding the average of the income in the past. We want this process to be translation invariant: the way we average over the past should not depend on whether it is, for example, currently January or July. Moreover, we would like the process to be linear in the income: the sum of two averaged incomes should be the average of the sum of the two incomes. The canonical setting

---

[9]Karlin (1968) shows the unique averaging kernel that is variation diminishing and is a semigroup on the whole line is the Gaussian.

for this is to average by

$$g(x) = \int_0^\infty f(x-y)h(y)dy,$$

where $h : [0, \infty) \to \mathbb{R}$ is a (not necessarily continuous) weighting function, assigning weight $h(y)$ to the income $y$ units of time in the past. Many different weighting functions are conceivable. For example, one could simply average the income incurred over the last $a$ units of time - this would correspond to $h$ being a step function on $[0, a]$ having constant value $1/a$.

Our main question is what kind of averaging functions would arise given a small number of reasonable axioms. We emphasize that while we give one possible answer to this question under reasonable axiomatic assumptions, we believe that this question is well worth of further study. In particular, it is quite conceivable that other sets of assumptions would lead to other natural functions $h(y)$. We recall that in the mathematical imaging literature, the Gaussian arises naturally from a wide variety of very different assumptions. No such analogous way of forming averages seems to be known for the cases of one-sided averages; both the one-sided Gaussian and the exponential distribution appear in very different settings but the existing literature is very sparse and not as comprehensive as the two-sided case. We believe this to be an exciting avenue for further work.

The second question, that we are interested in, is the issue of scale. We would like the weighting function and averaging to apply at different scales. That is, we want the function $h$ to be in fact a family of functions parameterized by a scale parameter $t$. Intuitively, the scale parameter measures both the range of averaging (the effective length of the time period of averaging) and, as it will turn out, the degree of smoothing.[10] The reason for considering this parameter is that income fluctuations may occur at different time scales – from the weekly earnings of a restaurant worker to multi-year royalties of a songwriter. Having a range of scale parameters allows to consider different

---

[10]That is, considering longer intervals allows for overall smoother averages – if I take my daily income, it may fluctuate significantly; if I average over a week, it will be smoother; and if I average over a year it will be smoother still.

averaging requirements that varying circumstances necessitate. If no assumptions are made on the range of income fluctuations a priori then all scales should be considered simultaneously. A challenge is to model and understand the representation of averaging not as an unrelated rules for different scales but rather to have a unifying principle that operates at all scales.[11]

Without loss of generality, we fix the initial time to 0 and call it the present. Let $x \in (-\infty, 0]$ denote some time in the past and $f(x)$ denote income at time $x$. We further introduce another parameter – scale $t$. We are interested in the transformations of the income function $f(x)_{x \in (-\infty, 0]}$ at different scales $t$: $u(t, x)$. The function $u(t, x)$ is the smoothed income at time $x$ where the scale of smoothing is $t$. Specifically,

$$u(t, x) = \int_{-\infty}^{0} f(y) p_t(x, y) dy, \tag{1}$$

where $p_t(x, y) \geq 0$ is an averaging or weighting function.[12]

The scale $t$ in what follows also determines the intensity of smoothing. For a given scale $t$, this operation takes an initial income function $f$ and transforms it into a new function $u(t, x)_{x \in (-\infty, 0]}$ by convolving with the function $p_t$. That is, for a given $x$, $u(t, x)$ is a weighted average of incomes $f(y)$ with the weights $p_t(x, y)$. We are particularly interested in the value of smoothed income at the present time $x = 0$ at various scales $t$:

$$g(0) = u(t, 0) = \int_{-\infty}^{0} f(y) p_t(0, y) dy.$$

---

[11]This is exactly the foundations of the scale-space theory in mathematical imaging and vision where an image has to be represented at different scales simultaneously – from the minute details at the close inspection to the outlines of the main features when viewed from a distance (see, e.g., Koenderink 1994 and Lindeberg 2013).

[12]Instead of making the assumptions on the form of (1) we could have more abstractly considered a family of linear operators $T_t$ acting on bounded and continuous functions $f(x)_{x \in (-\infty, 0]}$. Riesz-Markov-Kakutani representation theorem implies that a (positive) continuous linear functional $f \to T_t f$ is represented by a measure $T_t f(x) = \int_{-\infty}^{0} f(y) P_t(x, dy)$. Further assuming $T_t 1 = 1$ implies that $P_t(x, dy)$ is a probability measure.

# 4   Main Assumptions

At this stage, the weighting function $p_t(x, y)$ and the corresponding smoothed income $u(t, x)$ can be very general, and we now state further assumptions that allow us to specifically determine it.

A natural condition, that is often not even mentioned, is that if the income function is constant, $f(x) \equiv c$, then the averaged income function should also be constant and equal to the same numerical value. We also normalize the weighting function: for all $x, t$

$$\int_{-\infty}^{0} p_t(x, y) dy = 1.$$

The natural averaging after 0 units of time have passed is to simply return the original value at the point since nothing has yet happened. We assume then the initial condition that $p_0(x, \cdot) = \delta_x$, where $\delta_x$ is the Dirac delta function.

**Assumption 1. [Recursivity]** *For any* $x, y \in (-\infty, 0]$ *and* $t, s \geq 0$,

$$p_{t+s}(x, y) = \int_{-\infty}^{0} p_s(x, z) p_t(z, y) dz.$$

This is a natural assumption that connects different scales of averaging.[13] It is similar to many other recursive formulations common in economics. In this context, the assumption implies that averaging at the scale $s$ and then at the scale $t$ is equivalent to averaging at scale $t+s$. Alternatively, one can think of this assumption as stating that no scale of averaging is singled out and all of them are treated equally. One could thus interpret it as a statement about the internal consistency: if such an averaging method were to be implemented, then a citizen could conceivably ask to have their income averaged twelve times over the scale of a month as well as over the scale of a year and then pick the more favorable outcome. This condition ensures that all these outcomes agree: there is no difference in averaging over a year, twelve units of a month or 52

---

[13]More abstractly, we could have posed the semi-group property of the operator $(T_t)$ : $T_0 f = f, T_s \circ T_t = T_{s+t}$, for all $s, t \geq 0$.

weeks. It is natural that a reasonable multiscale averaging method has the scales connected with this intrinsic compatibility condition.

We now turn to the third assumption – locality.

**Assumption 2. [Locality]** *For each $x \in (-\infty, 0)$ and $\varepsilon > 0$*

$$\int_{|y-x|>\varepsilon} p_t(x,y)dy = o(t).$$

*Furthermore, there exist the infinitesimal characteristics $a$ and $r$:*

$$\int_{|y-x|\leq\varepsilon} (y-x)p_t(x,y)dy = rt + o(t),$$

$$\int_{|y-x|\leq\varepsilon} (y-x)^2 p_t(x,y)dy = at + o(t).$$

In essence, this assumption states a form of continuity for the averaging operator that for a given $x$, only the local values $y$ matter for the resulting average (for small $t$). There is also an additional assumption built in here – time and scale independence of the coefficients – which we chose not to state separately. We could have instead assumed that $a(t,x)$ and $r(t,x)$ with the results straightforwardly extending.[14] There is also one symmetry and, without loss of generality, we can set $a = 1$. [15]

There is also a probabilistic interpretation of this assumption. One can think of the weight $p_t(x,y)$ as the probability that a process travels from $x$ to $y$ in time $t$. The first part of the above assumption then states that the probability of non-local jumps is vanishingly small in time. This assumption together with the recursivity assumption then ensures the continuity of the paths of the stochastic process (see, e.g., Feller 1954).

Finally, we need an assumption on the behavior of the weighting function $p_t(x,y)$ at the boundary. There are two canonical ways of dealing with a

---

[14]More broadly, it may be of interest to also incorporate some assumptions related to the time-value of money which would determine the value of $r(x)$.

[15]Different $a$ would correspond to speeding up the time; however, since time will actually be one of the parameters in our solution formula, it can be recovered from there.

boundary: to impose Dirichlet or Neumann conditions. Dirichlet conditions are not suitable for our application because we would not want to impose a zero weight being assigned to the present income at $x = 0$.[16] This leaves Neumann conditions as a natural choice:[17]

**Assumption 3.** *[**Boundary Conditions**] For any $y < 0$,*

$$\frac{\partial}{\partial x} p_t(x, y)\big|_{x=0} = 0.$$

*Remark.* Finally, it is important to note that clearly there are many possible choices for the assumptions on averaging. One goal of our work is to formalize the question and open the venues to exploring possibly other choices of assumptions.

# 5 Results

We now state the result characterizing the weighting function $p_t(x, y)$.

**Proposition.** *The evolution equation that follows from Assumptions 1-3 is*

$$\frac{\partial}{\partial t} p_t(x, y) = \frac{1}{2} \frac{\partial^2}{\partial x^2} p_t(x, y) + r \frac{\partial}{\partial x} p_t(x, y). \tag{2}$$

*Furthermore, $p_t(x, y)$ has an explicit closed form given by the probability distribution function of the Reflected Brownian Motion:*

$$p_t(x, y) = 2re^{2ry} \Phi\left(\frac{rt + x + y}{\sqrt{t}}\right) + \frac{1}{\sqrt{t}} \phi\left(\frac{-rt - x + y}{\sqrt{t}}\right) + \frac{e^{2ry}}{\sqrt{t}} \phi\left(\frac{rt + x + y}{\sqrt{t}}\right), \tag{3}$$

---

[16]Also Dirichlet condition contradicts the fact that the total mass is preserved.

[17]There are also Robin conditions that one could impose, these are of the form

$$p_t(x, y) + \alpha \frac{\partial}{\partial x} p_t(x, y)\big|_{x=0} = 0 \qquad \text{for some fixed } \alpha \in \mathbb{R}.$$

However, since this also involves the value of the function at the boundary (the quantity of interest), it seems unnatural to force it to be of any particular form.

*where $r \in \mathbb{R}$, $x, y < 0$, $\phi$ is the probability density function of the standard $\mathcal{N}(0, 1)$ Gaussian distribution, and $\Phi$ is its cumulative density function.*

The first part of the result follows from the classic paper of Kolmogorov (1931) on the connection of the diffusion processes with the second order parabolic partial differential equations. The assumptions of the recursivity (semigroup) and locality (continuity) assure that the associated process is a diffusion with the density characterized by (2).[18]

This particular partial differential equation 2 is actually quite simple and is easy to solve on the whole line $\mathbb{R}$. What is different in our setting is that we are working on the half-line $\mathbb{R}_{\leq 0}$ and have reflecting boundary conditions which is more challenging. For any fixed $x < 0$ and any $t > 0$, we can interpret $p_t(x, \cdot)$ as a probability distribution function. This probability distribution describes the position of a Brownian particle started in $x$ and moving with constant drift in direction $r$ (which points towards the origin for $r > 0$ and away from the origin for $r < 0$). Specifically, the relevant process $Z_t$ is defined as follows – this is the Skorohod reflection problem (Skorohod 1961, 1962). Let $(B_t)_{t \geq 0}$ be a Brownian motion, and consider the process

$$X_t = x + rt + B_t.$$

There exists a unique increasing continuous function $L_t$ such that $L_0 = 0$, $Z_t = X_t - L_t \leq 0$ and $L_t$ grows only at the points where $Z_t = 0$. Precisely, $L_t = \sup_{0 \leq s \leq t} X_s^+$.

The second part of the result and the explicit form of weighting $p_t(x, y)$ function in (3) follows from the results in the queueing theory of Harrison (2013, p. 48) and Glynn and Wang (2018).

---

[18] In fact, one would need a weaker set of assumptions to ensure that $p_t(x, y)$ is represented by a second-order differential equation. Specifically, assuming that the operator $T_t$ is a semigroup, preserves unity ($T_t 1 = 1$), and is non-negative ($f \geq 0 \Rightarrow T_t f \geq 0$) would suffice. This can be proven modifying Stroock (2008, Lemma 1.1.6, p.2) and is a consequence of a more general result of Peetre (1959) that local operators are differential operators of finite order.

Moreover, the smoothed income at scale $t$

$$u(t, x) = \int_{-\infty}^{0} p_t(x, y) f(y) dy$$

is the solution $u(t, y)$ of the initial-boundary value problem

$$\frac{\partial}{\partial t} u(t, x) = \frac{1}{2} \frac{\partial^2}{\partial x^2} u(t, x) + r \frac{\partial}{\partial x} u(t, x) \qquad \text{in } (-\infty, 0),$$

$$\frac{\partial}{\partial x} u(t, 0) = 0; u(0, x) = f(x).$$

In particular, the primary object of our interest – smoothed income at the present time $(x = 0)$ at scale $t$ is given by

$$g(0) = u(t, 0) = \int_{-\infty}^{0} p_t(0, y) f(y) dy$$

This is the setting that we originally set out to study, and we have identified an averaging function $h(t, y) = p_t(0, y)$.

We now derive two important properties of the probability distribution function function $p_t(x, y)$ – the behavior at the small and large scales.

**Lemma.** *We have:*

*(1) for any fixed $x < 0$ and $y < 0$*

$$p_t(x, y) \sim \frac{1}{\sqrt{t}} \phi \left( \frac{y - x - rt}{\sqrt{t}} \right), t \to 0.$$

*(2) if $r > 0$, then, for all $x < 0$ and $y < 0$, we have*

$$\lim_{t \to \infty} p_t(x, y) = 2r e^{2ry},$$

*if $r < 0$, then $p_t(x, \cdot)$ converges to 0 on every compact interval as $t \to \infty$.*

*Proof.*

Part (1): Smoothing at small scale: $t \to 0$.

For any $x, y < 0$ consider the ratio

$$\frac{p_t(x, y)}{\frac{1}{\sqrt{t}}\phi\left(\frac{y-x-rt}{\sqrt{t}}\right)} = 1 + \frac{\sqrt{t}2re^{2ry}\Phi\left(\frac{y+x+rt}{\sqrt{t}}\right)}{\phi\left(\frac{y-x-rt}{\sqrt{t}}\right)} + e^{2ry}\frac{\phi\left(\frac{y+x+rt}{\sqrt{t}}\right)}{\phi\left(\frac{y-x-rt}{\sqrt{t}}\right)}.$$

We observe that, as $t \to 0$,

$$\frac{\phi\left(\frac{y+x+rt}{\sqrt{t}}\right)}{\phi\left(\frac{y-x-rt}{\sqrt{t}}\right)} = \exp\left(\frac{(y-x-rt)^2 - (y+x+rt)^2}{2t}\right)$$

$$= \exp\left(\frac{-2y(x+rt)}{t}\right) \to 0$$

Also, as $t \to 0$,

$$\frac{\Phi\left(\frac{y+x+rt}{\sqrt{t}}\right)}{\phi\left(\frac{y-x-rt}{\sqrt{t}}\right)} = e^{\frac{(y-x-rt)^2}{2t}} \int_{-\infty}^{\frac{y+x+rt}{\sqrt{t}}} e^{-u^2/2}du$$

$$\leq \sqrt{\frac{\pi}{2}}\exp\left(\frac{(y-x-rt)^2 - (y+x+rt)^2}{2t}\right) \to 0.$$

So, for all $x, y < 0$ and $t \to 0$,

$$\frac{p_t(x, y)}{\frac{1}{\sqrt{t}}\phi\left(\frac{y-x-rt}{\sqrt{t}}\right)} \to 1.$$

and the equivalence $p_t(x, y) \sim \frac{1}{\sqrt{t}}\phi\left(\frac{y-x-rt}{\sqrt{t}}\right), t \to 0$, is established.

Part (2): Smoothing at large scale: $t \to \infty$. This result shows that for the positive drift (towards the origin) at large scale $t \to \infty$, the function $p_t(x, y)$ converges to a universal (not depending on time $x$) limiting object

$$p_t(x, y) \to 2re^{2ry}$$

which is a negative of the exponential distribution. We consider the steady-

15

state for the Kolmogorov forward equation:

$$\frac{1}{2}f''(y) = rf'(y).$$

Clearly,

$$f(y) = Ae^{2ry} + B$$

Boundary condition $f'(0) = 2rf(0)$ implies that $B = 0$. Normalization to the mass equal to one gives $A = 2r$. □

The lemma above has the following meaning. Part (1) considers averaging over small time, that is, over a very short effective range. The main idea is that we average over very short windows of time, the boundary condition has no effect, the drift is still presented and we get averaging with the (non-centered) Gaussian distribution). Part (2) is in fact a rather remarkable fact in probability theory. The dynamical situation is as follows: we have a Brownian particle on $\mathbb{R}_{\leq 0}$ that is reflected at the origin. A particle such as this would slowly start exploring the space and be spread out more and more (roughly at scale $\sim \sqrt{t}$ after $t$ units of time). Without the drift $r$ (or with the drift away from zero, $r < 0$), there is no interesting limit as $t \to \infty$, the probability distribution function of Brownian motion goes to $0$ (because the particles are spread out more and more). Here (when $r > 0$), we have a slightly different situation resulting in a very different outcome: we have a constant drift (of strength $r$) moving the particles back to the origin. As time becomes large there is a limiting profile resulting as the balance of two forces: the constant drift trying to move everything to the origin and the Brownian particle moving around randomly. This limiting profile is given by the exponential distribution (which we encountered previously for very different reasons). The parameter $r > 0$, drift towards the origin, can be thought of modeling a form of monotonicity where from the point of view of the present $x = 0$ the more recent observations receive a higher weight.

The explicit form for the density $p_t(x, y)$ in equation (3) thus gives an interpolation of the weighting function between the Gaussian and the exponential
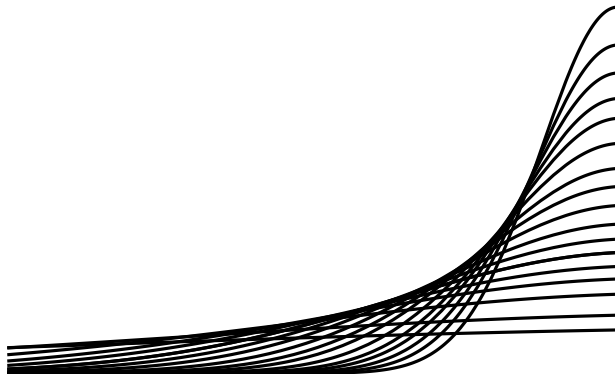
distribution. We plot it in Figure 1.



Figure 1: Weighting function $p_t(0, y)$ interpolates between Gaussian and exponential distributions $(r > 0)$.

Figure 2 plots smoothed income $g(x)$ at different scales $t$. Higher scales imply a larger effective range of averaging and result in smoother profile of income.
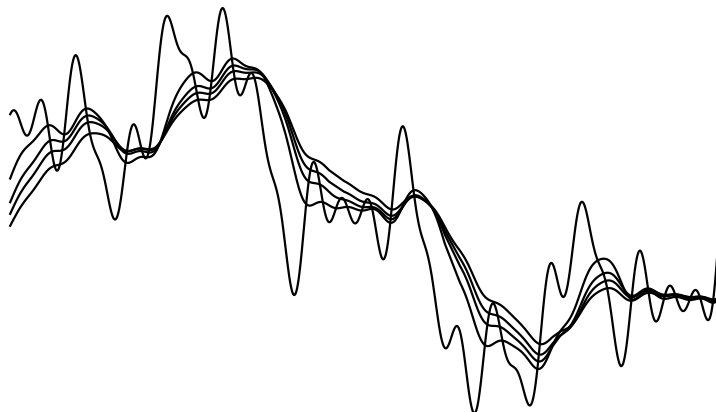


Figure 2: Smoothed income, $g(x)$, at different scales $t$.

Finally, we are interested in the smoothing properties of the weighting function that we found. Specifically, we are interested in knowing whether our equation

$$\frac{\partial}{\partial t} u(t, x) = \frac{1}{2} \frac{\partial^2}{\partial x^2} u(t, x) + r \frac{\partial}{\partial x} u(t, x)$$

17

arises as a gradient flow on some space. The gradient flow is the analogue of the usual gradient descent process but for the space of the functions. That is, we evolve the whole function in the direction of the steepest increase in some objective function.[19]

Consider the Hilbert space $L^2((-\infty, 0], \mu)$, where $\mu(dx) = e^{2rx}dx$. Let $M$ be the subspace consisting of all continuously differentiable functions $u : (-\infty, 0] \to \mathbb{R}$, such that $u'(0) = 0$. On $M$ we define a functional

$$I(u) = \int_{-\infty}^{0} u_x^2 d\mu,$$

where $\mu = e^{2rx}dx$.

The corresponding gradient flow is a function $u : [0, \infty) \to M$ such that $\partial_t u = -\nabla I(u)$, where $\nabla I$ is the gradient of the functional $I$. We compute the gradient:

$$I(u + \epsilon v) - I(u) = \frac{\epsilon}{2} \int_{-\infty}^{0} u'(x)v'(x)e^{2rx}dx + o(\epsilon).$$

So,

$$\begin{aligned}
(\nabla I(u), v) &= \frac{1}{2} \int_{-\infty}^{0} u'(x)v'(x)e^{2rx}dx \\
&= -\frac{1}{2} \int_{-\infty}^{0} v(x)(u''(x) + 2ru'(x))e^{2rx}dx.
\end{aligned}$$

It follows that
$$\nabla I(u) = -\frac{1}{2}u''(x) - ru'(x),$$

and the gradient flow coincides with the equation

$$\frac{\partial}{\partial t}u(t,x) = \frac{1}{2}\frac{\partial^2}{\partial x^2}u(t,x) + r\frac{\partial}{\partial x}u(t,x).$$

Let us consider $u(t + \varepsilon, x) = u(x) + \varepsilon v(x)$. We then construct the flow in

---

[19]See Steinerberger and Tsyvinski (2019) for a detailed description of gradient flows arising in the context of optimal taxation.

the direction opposite to the gradient of $I$. From this point of view, the PDE smoothes functions as it reduces the $L^2((-\infty, 0], \mu)$-norm of the derivative in $x$)

$$\int_{-\infty}^0 (u_x + \varepsilon v_x)^2 d\mu = \int_{-\infty}^0 (u_x + \varepsilon v_x)^2 e^{2rx} dx$$

$$= \int_{-\infty}^0 u_x^2 e^{2rx} dx + 2\varepsilon \int_{-\infty}^0 u_x v_x e^{2rx} dx + \mathcal{O}(\varepsilon^2)$$

At the same time, integration by parts (and applying Neumann conditions on the boundary) results in

$$2\varepsilon \int_{-\infty}^0 u_x v_x e^{2rx} dx = -2\varepsilon \int_{-\infty}^0 v \frac{\partial}{\partial x}\left(u_x e^{2rx}\right) dx.$$

We have

$$\frac{\partial}{\partial x}\left(u_x e^{2rx}\right) = u_{xx} e^{2rx} + 2u_x r e^{2rx} = (u_{xx} + 2ru_x)e^{2rx}$$

Therefore,

$$-2\varepsilon \int_{-\infty}^0 v \frac{\partial}{\partial x}\left(u_x e^{rx}\right) dx = -2\varepsilon \int_{-\infty}^0 v(u_{xx} + 2ru_x)e^{2rx} dx$$

$$= -4\varepsilon \int_{-\infty}^0 v(\frac{1}{2}u_{xx} + ru_x)d\mu$$

By $L^2-$duality (or Cauchy-Schwarz), this quantity is made as small as possible when

$$v = \frac{1}{2}u_{xx} + ru_x.$$

That is, $\frac{1}{2}u_{xx} + ru_x$ is the gradient flow that maximally smoothes income in the sense of maximally decreasing the present value of the sum of $u^2$.

# 6 Conclusion

We examine a classic public finance question from a new perspective and propose an averaging rule based on a small set of assumptions.

Anticipating potential criticism, we now address some of the broad issues with this approach. First, the assumptions that we used, while reasonable, are certainly not the only ones one can use and, hence, derive a different averaging and smoothing rule. A good parallel to make is a discussion in the mathematical imaging literature that examines how various sets of assumptions generate different smoothing mechanisms. Moreover, the focus there is exactly the one we take here – how a small set of assumptions generate reasonable results and what the consequences are of relaxing or changing some of those. In particular, we believe it could be quite desirable to have the same question addressed from various different perspectives and see what kind of averaging methods may arise from completely different sets of axioms. A fascinating question is whether the universal appearance of the Gaussian in the two-sided case has an analogous "universal" averaging scheme. Both the half-sided Gaussian and the exponential distribution are natural candidates. Second, the question of the practicality of the results. While the exponential weighting, Gaussian and the explicit form of the density of the reflected Brownian motion for the intermediate case are simple mathematically, it is more difficult (with the exception of the exponential case) to imagine them being implemented in practice. The abstract formulation of the problem and the explicit solution that we derive allow, however, to both precisely state and solve the question rather than rely on the perhaps more useful heuristics. At the same time, with increased digitization one can imagine that some of the theoretical insights presented here to be implemented in practice. This is the main point of an excellent discussion of the broad range of practical implementation topics of the theoretical taxation literature, including income averaging, in Jacobs (2017).

# References

Alvarez, Luis, Frederic Guichard, Pierre-Louis Lions, and Jean-Michel Morel. "Axioms and fundamental equations of image processing." Archive for Rational Mechanics and Analysis 123, no. 3: 199-257. 1993.

Aubert, Gilles, and Pierre Kornprobst. Mathematical problems in image processing: partial differential equations and the calculus of variations. Vol. 147. Springer Science & Business Media. 2006.

Bargain, Olivier, Alain Trannoy, and Adrien Pacifico. "The Impact of Tax Frequency: Theoretical and Empirical Investigations". Working paper. 2017.

Batchelder, Lily L. "Taxing the poor: Income averaging reconsidered." Harvard Journal on Legislation. 40: 395. 2003.

Brown, Robert G. "Exponential smoothing for predicting demand." In Operations Research, vol. 5, no. 1, pp. 145-145. 1957.

Buchanan, Neil H. "The Case Against Income Averaging." Virginia Tax Review. 25: 1151. 2005.

Diamond, Peter. "Taxes and pensions." Southern Economic Journal 76, no. 1 (2009): 2-15

Fagerström, Daniel. "Temporal scale spaces." International Journal of Computer Vision 64, no. 2-3: 97-106. 2005.

Feller, William. "Diffusion processes in one dimension." Transactions of the American Mathematical Society 77, no. 1: 1-31. 1954.

Glynn, Peter W., and Rob J. Wang. "On the rate of convergence to equilibrium for reflected Brownian motion." Queueing Systems 89, no. 1-2 (2018): 165-197.

Gordon, Daniel V., and Jean-Francois Wen. "Tax penalties on fluctuating incomes: estimates from longitudinal data." International Tax and Public Finance 25, no. 2: 430-457. 2018.

Harrison, J. Michael. Brownian models of performance and control. Cambridge University Press, 2013.

Heathcote, Jonathan, Kjetil Storesletten, and Giovanni L. Violante. "Con-

sumption and labor supply with partial insurance: An analytical framework." American Economic Review 104, no. 7: 2075-2126. 2014.

Holt, Charles C. "Averaging of income for tax purposes: Equity and fiscal-policy considerations." National Tax Journal 2, no. 4 : 349-361. 1949.

Holt, Charles C. "Forecasting seasonals and trends by exponentially weighted moving averages" Office of Naval Research. Research Memorandum No. 52., 1957. and reprinted in Holt, Charles C. "Forecasting seasonals and trends by exponentially weighted moving averages." International journal of forecasting 20, no. 1: 5-10. 2004.

Huggett, Mark, and Juan Carlos Parra. "How well does the US social insurance system provide social insurance?." Journal of Political Economy 118, no. 1: 76-112. 2010.

Jacobs, Bas. "Digitalization and Taxation", in: Sanjeev Gupta, Michael Keen, Alpa Shah, and Genevieve Verdier (eds), Digital Revolutions in Public Finance, Washington-DC: International Monetary Fund, Ch. 2. 2017.

Karlin, Samuel. Total positivity. Vol. 1. Stanford University Press. 1968.

Kapicka, Marek. " Quantifying the Welfare Gains from History Dependent Income Taxation." ADEMU Working paper series. 2017.

Kolmogorov, Andrei. "Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung." Mathematische Annalen 104, no. 1: 415-458. 1931.

Koenderink, Jan J. "The structure of images." Biological cybernetics 50, no. 5: 363-370. 1984.

Lindeberg, Tony. "On the axiomatic foundations of linear scale-space." In Gaussian scale-space theory, pp. 75-97. Springer, Dordrecht. 1997.

Lindeberg, Tony. "Generalized Gaussian scale-space axiomatics comprising linear scale-space, affine scale-space and spatio-temporal scale-space." Journal of Mathematical Imaging and Vision 40, no. 1: 36-81. 2011.

Lindeberg, Tony. Scale-space theory in computer vision. Vol. 256. Springer Science & Business Media. 2013.

Lindeberg, Tony, and Daniel Fagerström. "Scale-space with casual time direction." In European Conference on Computer Vision, pp. 229-240. Springer, Berlin, Heidelberg. 1996.

Mintz, Jack M., and Thomas A. Wilson. "Saving the future: restoring fairness to the taxation of savings." Commentary-CD Howe Institute 176. 2002.

Mirrlees, James A., and Stuart Adam. Dimensions of tax design: the Mirrlees review. Oxford University Press, 2010.

Pauwels, Eric J., Luc J. Van Gool, Peter Fiddelaers, and Theo Moons. "An extended class of scale-invariant and recursive scale space filters." IEEE Transactions on Pattern Analysis and Machine Intelligence 17, no. 7: 691-701. 1995.

Peetre, Jaak. "Une caracterisation abstraite des operateurs differentiels." Mathematica Scandinavica: 211-218. 1959.

Salden, Alfons H., Bart M. ter Haar Romeny, and Max A. Viergever. "Linear scale-space theory from physical principles." Journal of Mathematical Imaging and Vision 9, no. 2: 103-139. 1998.

Saez, Emmanuel. "Optimal income transfer programs: intensive versus extensive labor supply responses." The Quarterly Journal of Economics 117, no. 3 1039-1073. 2002.

Skorokhod, Anatoliy V. "Stochastic equations for diffusion processes in a bounded region." Theory of Probability & Its Applications 6, no. 3: 264-274. 1961.

Skorokhod, Anatoliy V. "Stochastic equations for diffusion processes in a bounded region. II." Theory of Probability & Its Applications 7, no. 1: 3-23. 1962.

Schönberg, Isaac Jacob. "On variation-diminishing integral operators of the convolution type." Proceedings of the National Academy of Sciences of the United States of America 34, no. 4, p. 164-169. 1948.

Simons, Henry C. Personal income taxation: The definition of income as a problem of fiscal policy. Chicago University, Chicago. 1938.

Steinerberger, Stefan and Aleh Tsyvinski. "Tax mechanisms and gradient flows". National Bureau of Economic Research Working paper, No. w25821. 2019.

Stroock, Daniel W. Partial differential equations for probabilists. Cambridge Univ. Press. 2008.

Vickrey, William. "Averaging of income for income-tax purposes." Journal of Political Economy 47, no. 3: 379-397. 1939.

Weickert, Joachim, Seiji Ishikawa, and Atsushi Imiya. "On the history of Gaussian scale-space axiomatics." In Gaussian scale-space theory, pp. 45-59. Springer, Dordrecht. 1997.