Notes on a toy model of Baumol's cost "disease"

Richard A.L. Jones

14 April 2020

@RichardALJones,

www.softmachines.org

1 Introduction to Baumol's cost disease

The economist William Baumol identified a long run tendency in the economy for the cost of services to increase relative to the cost of goods. To explain this tendency, he pointed to the fact that it is much easier to increase productivity in the manufacturing sector than in the service sector - in manufacturing, advances in technology and automation have, over the centuries, increased the amount of stuff a single worker can manufacture in an hour's labour by orders of magnitude. Yet for many services, it is the time taken by the worker delivering the service that actually constitutes the product, so productivity almost by definition cannot increase. A string quartet playing an hour of music takes the same amount of labour now as it did in Mozart's time, so its productivity has remained constant. But a professional musician needs to be paid a wage that bears some relationship to the overall wage levels in an economy, so as the cost of goods falls because of technological improvements, the cost of providing a service must increase in relationship to the cost of goods. This relative increase in the cost of services relative to goods is known - somewhat inappropriately, in my view, for reasons that should become clear - as *Baumol's cost disease*.

Baumol's cost disease will arise in any situation in which different sectors of the economy have different rates of productivity growth, while there exists a reasonably efficient labour market which ensures some comparability between the wages earned by workers of comparable levels of skill in different industries. In effect, it is a mechanism by which the productivity gains in one, fast-advancing sector, are spread across the economy as a whole. It's for this reason that I think it is wrong to frame it as a "disease" - it is a positive feature of an economy that allows everyone to benefit from the necessarily uneven advance of technology.

It's worth stressing at this point just how uneven these advances are. We've seen in the last half-century, for example, technological advances in the information technology industry that have led to a drop in the cost of computer power

of many orders of magnitude, while other areas of technology - like the way we build houses - have remained largely static. Understanding Baumol's cost disease provides the crucial bridge between the real world, with these grossly varying rates of technological progress across technologies and sectors, and theories of economic growth, where the diversity of progress is folded into a single variable - *"total factor productivity"*.

2 Baumol's "cost disease" - a toy model

It's often instructive to look at "toy models" - a theory which describes a grossly simplified version of the world, which, while remaining mathematically tractable and transparent, still might provide insight about the mechanisms at work in more complicated systems. Here I elaborate a very simple model described in Dietrich Vollrath's recent book "Fully Grown"[1].

We imagine that our economy has only two sectors. We have a service sector, in which the output is essentially proportional to the number of hours of labour put in, and a goods sector, where technological progress allows us, as time goes on, to produce more goods for a given amount of labour. We write the output in each sector, Y_g, and Y_s, as the product of the hours of labour we put in - L_g for goods, and L_s for services - and the productivity of each sector - A_g and A_s. So for the quantity of goods we have:

$$Y_g = A_g L_g \tag{1}$$

and for the quantity of services:

$$Y_s = A_s L_s \tag{2}$$

Note that the productivities A_g and A_s are defined in terms of how much output is produced, and not to its monetary value. We're talking about the number of cars or TV sets, or the number hours spent with the physiotherapist or having piano lessons, not how much we'd have to pay for them. This is different to the usual economic definition of productivity, which is defined in terms of the monetary value of goods or services. This distinction is important and we will come back to it later.

Let's assume that the productivity of services A_s is constant with time, but that we get better at making goods - our productivity for goods rises by a fixed proportion each year. That means this productivity for goods produced is an exponential function of time.

$$A_g(t) = A_g^0 \exp(\frac{t}{\tau}) \tag{3}$$

Here the rate of growth of productivity is expressed through τ, the number of years it would take to increase the material productivity of goods by a factor of e.

We want to know much one unit of goods, or one unit of services, costs. As we'll see, the relative cost of goods and services change with time. We write the price of one unit of goods as $P_g(t)$, and the price of one unit of services as $P_s(t)$.

How much do people get paid for delivering services or making things? Here we make a big assumption - there's just one wage rate for goods and services, and people get paid the marginal value of what they produce. So the wage of a goods industry worker is $w_g = P_g(t)A_g(t)$, and the wage of a services industry worker is $w_s = P_s(t)A_s$ (remember that we've assumed that the productivity of services doesn't change with time).

The next crucial assumption is that we have a labour market, which means that service workers and goods workers get paid the same amount. The goods workers might think they deserve to get more, because they're producing more stuff, but if they did, then service workers would move into the goods trade until the prices adjusted to equalise wages. This means we can work out how the prices adjust in response to the change in productivity.

$$w_g = w_s = P_g(t)A_g(t) = P_s(t)A_s \tag{4}$$

So we find the ratio of prices

$$\frac{P_s(t)}{P_g(t)} = \frac{A_g^0}{A_s}\exp(\frac{t}{\tau}) \tag{5}$$

The price of services relative to goods increases in proportion to the increasing productivity of making goods: this is Baumol's cost disease.

3 Baumol's cost disease isn't a disease

We shouldn't call this a disease, though, because as a result of the increase in productivity for goods, the economy as a whole is more productive. If the same amount of labour went into producing goods as before, we'd have more and more goods to consume, while still being able to consume the same amount of services. But this seems to be an unlikely response to the scenario - we might expect that people would respond to the growing prosperity of the economy by increasing the amount of services they consume, while still being able to consume more goods. The economy can adjust to respond to this preference by switching some labour from making goods into delivering services.

It's not immediately obvious how these preferences might unfold, but we can identify two plausible limits. We might, at one extreme, imagine that people would want to keep their consumption of goods and services in constant proportion K (this means that they do not respond at all to their changing relative prices). Writing the time dependent consumption of goods and services as $C_g(t)$ and $C_s(t)$ respectively, this gives us

$$\frac{C_g(t)}{C_s(t)} = K \tag{6}$$

Alternatively, people might budget a constant amount of money for buying each of goods and services, so the ratio of their expenditure on goods and services was a constant $K^{'}$:

$$\frac{C_g(t)}{C_s(t)} = K^{'}\left(\frac{P_s(t)}{P_g(t)}\right) \tag{7}$$

More generally, we can describe an intermediate situation by

$$\frac{C_g(t)}{C_s(t)} = K\left(\frac{P_s(t)}{P_g(t)}\right)^{\sigma} \tag{8}$$

where the propensity to substitute services for goods is described by the parameter σ, where $0 < \sigma < 1$.

Now our economy adjusts so that production equals consumption: $C_g(t) = Y_g(t)$ and $C_s(t) = Y_s(t)$. With a bit of algebra, we can deduce from this how the split of labour between goods and services evolves:

$$\frac{L_g(t)}{L_s(t)} = K\left(\frac{A_g^0}{A_s}\right)^{-\gamma} \exp\left(\frac{-\gamma t}{\tau}\right) \tag{9}$$

where I have written $\gamma = 1 - \sigma$, i.e. a positive number between 0 and 1, expressing the propensity of people to switch some of their spending from rapidly cheapening goods to more expensive services. If $\gamma = 0$ people keep the amount of money they spend on goods and services respectively constant - i.e., they just buy more and more goods with the same money as they get cheaper, but if $\gamma = 1$, people increase both the amount of goods and the number of services they buy in the same ratio.

4 Rising output of both goods and services

How many goods and services are produced? Let's assume our economy has a constant total supply of labour, L_0, so $L_g(t) + L_s(t) = L_0$

So we have for the amount of labour going into services $L_s(t)$

$$L_s(t) = \frac{1}{1 + K\left(\frac{A_g^0}{A_s}\right)^{-\gamma} \exp\left(\frac{-\gamma t}{\tau}\right)} \tag{10}$$

and the amount of labour going into making goods $L_g(t)$

$$L_g(t) = \frac{L_0 K\left(\frac{A_g^0}{A_s}\right)^{-\gamma} \exp\left(\frac{-\gamma t}{\tau}\right)}{1 + K\left(\frac{A_g^0}{A_s}\right)^{-\gamma} \exp\left(\frac{-\gamma t}{\tau}\right)} \tag{11}$$

Now we can work out the total production - and consumption - of goods and services. For goods

$$Y_g(t) = \frac{L_0 K A_g^0 \left(\frac{A_g^0}{A_s}\right)^{-\gamma} \exp\left(\frac{(1-\gamma)t}{\tau}\right)}{1 + K\left(\frac{A_g^0}{A_s}\right)^{-\gamma} \exp\left(\frac{-\gamma t}{\tau}\right)} \tag{12}$$

Even though the amount of labour devoted to making goods decreases with time, this is outweighed by the increasing productivity, and so the total number of goods produced increases exponentially.

For services we have

$$Y_s(t) = \frac{A_s^0 L_0}{1 + K\left(\frac{A_g^0}{A_s}\right)^{-\gamma} \exp\left(\frac{-\gamma t}{\tau}\right)} \tag{13}$$

The amount of services increases too, but more slowly than the increase in production of goods, as the increase is driven solely by the movement of workers from producing goods to producing services.

The larger the value of γ, the more switching from goods into services. For a value of $\gamma = 0.6$, the outputs are shown in figure 1.

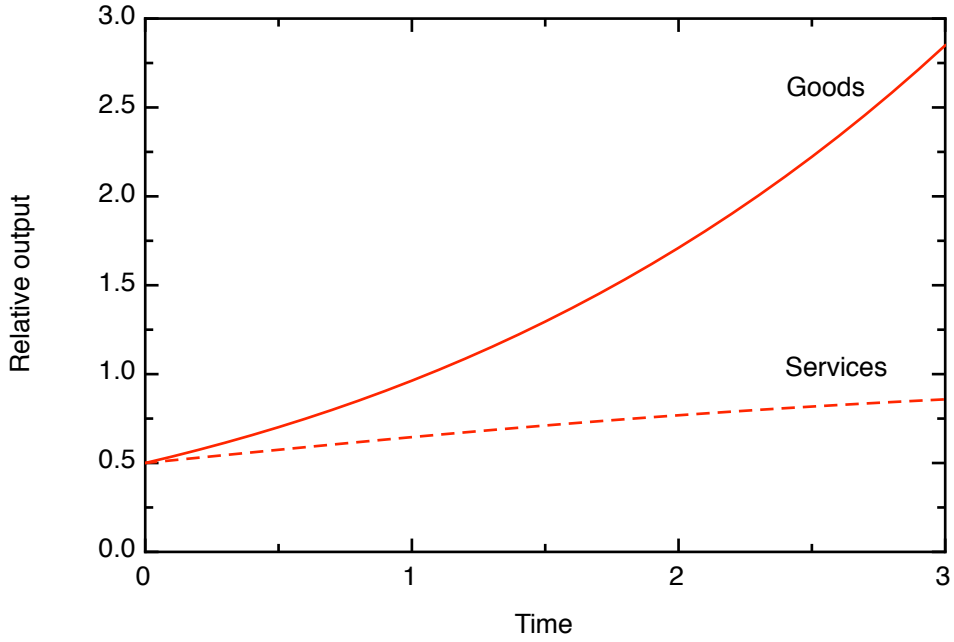

Figure 1. Relative output of goods and services in a toy model of Baumol's cost disease, with the parameter $\gamma = 0.6$. Output is relative to output at time $t = 0$, when the labour share devoted to goods and services is equal. The time scale is in units of the characteristic time for the growth of goods productivity, τ.

5 While the number of goods produced increases, their total monetary value actually decreases

As time goes on in this economy, people become richer - they produce (and consume) both more goods and more services. But, if we measure, not the amount of stuff produced, but its monetary value, the situation is more complicated, because the cost of goods relative to services is falling. This raises some interesting questions about how we measure the total output of an economy in the presence of sectors with widely differing rates of productivity growth.

To explore this, let's first simplify the mathematics by noting that we can define units such that all the constants apart from the exponent γ are unity. In these units, we have for labour devoted to goods

$$L_g(t) = \frac{\exp(-\gamma t)}{1 + \exp(-\gamma t)} \tag{14}$$

and to services

$$L_s(t) = \frac{1}{1 + \exp(-\gamma t)} \tag{15}$$

The output of goods and services comes simply from multiplying the labour by the productivity in each case, giving us for goods:

$$Y_g(t) = \frac{\exp((1-\gamma)t)}{1 + \exp(-\gamma t)} \tag{16}$$

and for services

$$Y_s(t) = \frac{1}{1 + \exp(-\gamma t)} \tag{17}$$

.

How do we attach a value to these outputs of goods and services? We know that the ratio of the prices of a unit of output of services to a unit of output of goods changes with time:

$$\frac{P_s(t)}{P_g(t)} = \exp(t) \tag{18}$$

What should our currency unit be for defining the overall output of the economy? If we pegged our currency to the value of a unit of goods, the total value of economy would appear to increase massively, because of the exponential increase in relative value of services. But if we pegged our currency to the value of a unit of services, in these units the economy would actually appear to shrink, because of the relative fall in value of the goods that are produced. The standard way to account for this would be to correct for the inflation or deflation implicit in either of these two approaches by calculating the change in apparent price of a basket of goods and services.

If we start the clock at $t = 0$, goods and services have equal weight; we could calculate an analogue of the "consumer price index" to account for the effect of inflation or deflation. If our currency unit is pegged to the price of goods, we can write our price index from the base year at $t = 0$ as

$$CPI(t) = 0.5(1 + \exp(t)) \tag{19}$$

In these units we can write the value of output of goods $V_g^{cpi}(t)$ and services $V_s^{cpi}(t)$ as

$$V_g^{cpi}(t) = \frac{\exp\left[(1-\gamma)\,t\right]}{CPI(t)\,(1+\exp(-\gamma t))} \tag{20}$$

and

$$V_s^{cpi}(t) = \frac{\exp\left[t\right]}{CPI(t)\,(1+\exp(-\gamma t))} \tag{21}$$

If we now look at how the total value of output of goods and services evolve, we see that while the value of services increases, the value of goods starts out relatively constant, and then begins to fall. The increase in production of goods cannot keep up with the fall in their cost. The total economy grows, but with a growth rate that decreases with time, saturating at a size exactly twice its initial value.

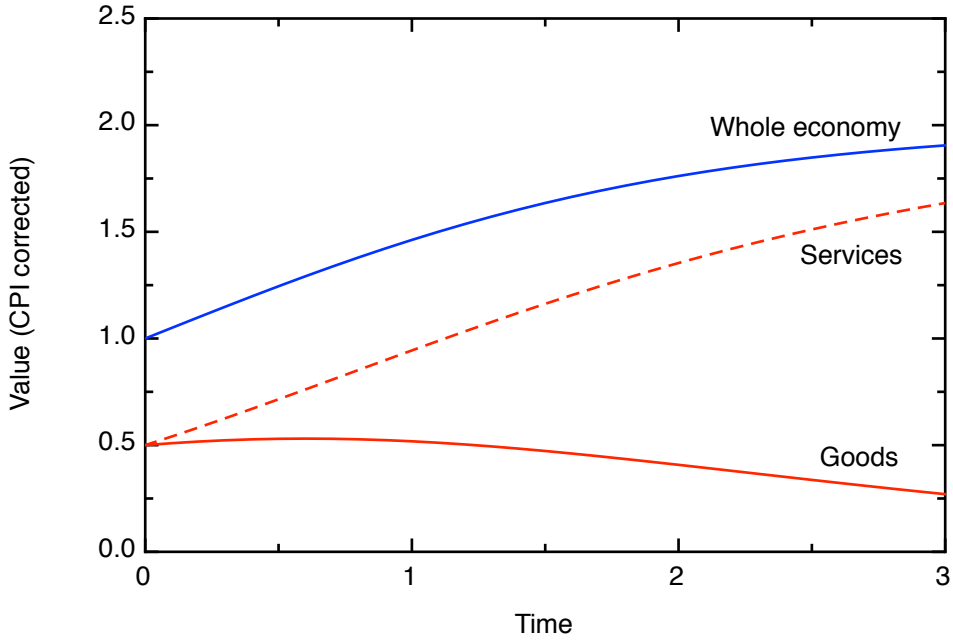

Figure 2. The total value of goods and services produced in a toy model of Baumol's cost disease, with the parameter $\gamma = 0.6$. Value is relative to value at time $t = 0$, when the labour share devoted to goods and services is equal, and is

corrected for inflation by a factor giving equal weight to price changes in goods and services.

But this way of correcting for inflation doesn't account for the fact that the relative weights of goods and services in people's consumption themselves change with time. We could instead use a changing weight in the basket of goods and services to reflect this. This corresponds more closely to how GDP is calculated for the real economy; the price of a basket of goods and services weighted by their proportional contribution to the total economy is what an economist would call a GDP deflator. The relative weight by value of goods and services is the same as the relative weight by labour share (which changes with time), so in currency units pegged to the value of a unit of goods (and remembering we have defined units such that the total amount of labour in the economy is unity) we can write the GDP deflator $D\left(t\right)$ as

$$D(t) = L_s(t)\exp(t) + L_g(t) \tag{22}$$

Now the value in real money of goods $V_g(t)$ can be written

$$V_g(t) = \frac{L_g \exp(t)}{D(t)} \tag{23}$$

and of services

$$V_S(t) = \frac{L_s \exp(t)}{D(t)} \tag{24}$$

Knowing the ratio of labour devoted to goods and services, we can rewrite these expressions as

$$V_g(t) = \frac{\exp(t)}{1 + \exp\left[\left(1+\gamma\right)t\right]} \tag{25}$$

and

$$V_s(t) = \frac{\exp\left[\left(1+\gamma\right)t\right]}{1 + \exp\left[\left(1+\gamma\right)t\right]} \tag{26}$$

with the total real GDP of our economy being

$$V_g(t) + V_s(t) = \frac{\exp(t) + \exp\left[\left(1+\gamma\right)t\right]}{1 + \exp\left[\left(1+\gamma\right)t\right]} \tag{27}$$

This is illustrated, again for a value of the parameter $\gamma = 0.6$, in figure 3.

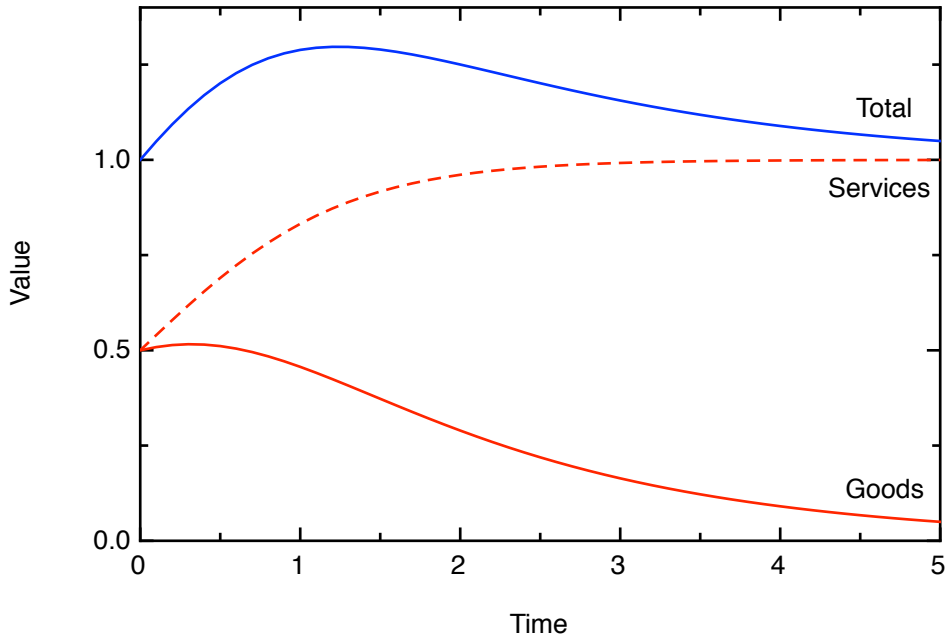

Figure 3. The total value of goods and services produced in a toy model of Baumol's cost disease, with the parameter $\gamma = 0.6$. Value is relative to value at time $t = 0$, when the labour share devoted to goods and services is equal, and is corrected for inflation by a factor giving equal weight to price changes in goods and services.

This is paradoxical. At the earliest times, the value of both goods and services outputs increases, but as before the increase in production of goods cannot keep up with the fall in their cost, so their total value diminishes. The value of services increases, but after the productivity of goods has roughly trebled the total size of the economy starts to decline (the position of the maximum varies with the value of γ, between about 3τ and τ for values of γbetween 0.1 and 0.9). In the limit of long times, the real value of services output saturates at twice its initial value, reflecting the fact that the economy is devoting twice as much labour to it, but the value of the goods output dwindles to zero. The service sector has doubled in size, but the goods sector has (in terms of its contribution to the economy) gone from making up half the economy to making no contribution to it at all (in spite of the fact that we are able to consume an unlimited number of goods). The net result is that the overall size of the economy appears unchanged. Goods, now essentially being free, make no contribution at all to GDP.

6 Defining productivity

To an economist, labour productivity is defined as the amount of value created, on average, by a fixed amount of labour, for example in an hour of work. A factory owner wishing to increase productivity will do this by trying to increase the amount of products that the factory makes for a given input of labour. If the price of his products is constant, then the two definitions of productivity are the same. But, as we've seen, at the scale of the whole economy, increasing the number of products made per hour of labour reduces their price, while indirectly increasing the price in other sectors even where those sectors has seen no improvement at all in their output for a given amount of labour input. So we need to distinguish between two types of productivity.

What we might call *output productivity* refers to the total amount of output - measured as the amount of goods produced or services delivered - for a given labour input. But if we look at, not the output itself, but the real money value of the output, we might talk about *real money productivity*. Our toy model makes clear that these are not the same.

We have assumed in the model that output productivity for goods - $A_g(t)$ - enjoys a fixed fractional rate of growth, while output productivity for services - $A_s(t)$ is constant. But in our model, we can see that the real money productivity for goods and services (which we write as $B_g(t)$ and $B_s(t)$) are identical:

$$B_g(t) = B_s(t) = \frac{\exp(t)}{D(t)} \tag{28}$$

where $D(t)$ is our GDP deflator. In fact, the real money productivity of both goods and services has the same form as the total GDP:

$$B_g(t) = B_s(t) = \frac{\exp(t) + \exp\left[(1+\gamma)\, t\right]}{1 + \exp\left[(1+\gamma)\, t\right]} \tag{29}$$

Baumol's cost disease is in fact a mechanism for redistributing the very different output productivities of our two different sectors, so that the real money productivities of the two sectors come out the same.

7 What can this toy model tell us about the real world?

Can such a simple model tell us anything about a real economy? The best one can hope for is that it yields some qualitative insights, and perhaps generates some conjectures to be tested. It's worth distinguishing between two kinds of faults in toy models of this kind. Obviously the model is much too simple; no real economy can be described in terms of just two sectors. It's fairly easy to see how the model could be extended to account for multiple sectors, each with their own rate of growth of output productivity, but this extension wouldn't account for the complex way in which the sectors interact. Technological progress in one sector will enable output productivity gains in another, the

effect rippling out through the whole economy. One can also identify failures in its assumptions, and locating these might tell us important things about the operation of economies in the real world. So, with these reservations, here are some conclusions and conjectures.

The most important conclusion I believe is robust - it is that Baumol's cost disease isn't a disease at all, but the essential way by which an economy redistributes the gains made in the fastest developing sectors across the whole economy. I've framed the model in the way Baumol introduced it, contrasting goods with services. What the model suggests is that the apparent gains one measures in real money productivity in sectors with low or zero output productivity growth actually reflect a redistribution of value from sectors with high output productivity growth - and that, conversely, the growth in real money productivity one measures in fast developing sectors actually substantially underestimates their rate of output productivity growth.

In fact, in our toy model, the real money productivity of both sectors ends up being equalised. This reflects the understanding that in a truly competitive economy, any advantage that technological innovation gives its inventor is instantaneously competed away. One can conjecture that measured differences in real money productivity between sectors actually reflect frictions and market failures. Some of these frictions are actually societally beneficial, as they provide incentives to innovate, but others may reflect power imbalances and the resulting differential abilities of different actors in the economy to extract rent.

Rather than framing our model in terms of idealised and homogenous goods and services sectors, perhaps a better way of thinking about it is as a cartoon of what happens when any kind of new, fast developing invention is introduced into an otherwise static economy. What we see is an initial period of growth, after which the new invention becomes so cheap we don't count it in the economy any more. This perhaps suggests a lesson about how we should think about GDP changes over the long run - an exponential growth in output productivity in one part of the economy doesn't create exponential growth in the economy as a whole - instead it gives the economy a temporary, one-off boost, but leaves a lasting legacy of increased utility that is not captured in the GDP figures.

It's worth stressing how common these kinds of massive changes in output productivity, driven by techological progress, are in the history of technology. The history of computing since the mid-1970's is the most obvious example. For a less familiar, but no less instructive, story, we can look at the humble nail, the subject of a fascinating recent study [2]. In the mid-1700's, it would take a blacksmith about a minute to make a single nail. Now a machine can make 350 nails a minute, and a single worker can operate four machines, leading to a 1400-fold increase in output productivity. The industrial revolution and 19th century urbanisation created a huge demand for nails; and satisfying this demand initially formed a significant part of the economy - estimated as 0.4% of US GDP in 1810. It is precisely the sector's huge increase in output productivity that means we no longer think of nails as an important part of the economy - their current share of the US economy is estimated as only 0.01% in 2002. But it's on the availability, at very little cost, of these and many other types of

formerly expensive hardware, that a big part of our current prosperity rests.

Economic growth needs technological innovation, but this innovation is necessarily uneven. Baumol's cost mechanism spreads out the benefits of rapid progress in a few areas across the economy as a whole, creating the impression of steady progress across the whole economy from the introduction and rapid improvement of a continuous sequence of new inventions.

References

[1] Dietrich Vollmath. *Fully Grown: why a stagnant economy is a sign of success*. University of Chicago Press, 2017

[2] Daniel Sichel, *The Price of Nails since 1700: Even Simple Products Experienced Large Price Declines*. Preprint, April 2017.