Chapter 18

Wider perspectives on monetary economies

As a follow-up on Chapter 17, this chapter discusses monetary matters in a broader context. Does the relationship between money growth and long-run inflation as depicted by the Sidrauski model in the previous chapter carry over to more general settings? Does neutrality and superneutrality of money tend to hold in more general models? When does very high inflation (hyperinflation) tend to arise? What is the role of government budget deficits in this context? Is inflation always bad or can there be benefits associated with (moderate) inflation? Can the different stories about the monetary transmission mechanism in the long run and the short run be reconciled? What remains of the theory of the real interest rate as exposed in earlier chapters, when money and the need for liquidity are taken into account? What can be said about the "level of interest rates" in a monetary economy with uncertainty?

18.1 Money growth and inflation in the long run

Let the trend growth rate (per year, say) of a variable x be called g_x, that is, $g_x \equiv \dot{x}/x$. The prefix "trend" is meant to say "apart form short-run fluctuations". The general tenet from theory (whether neoclassical or Keynesian and whether in Ramsey or OLG format) and from (informal) empirical observation is that inflation on average over a long time span is closely linked to sustained money growth in excess of output growth. Thus, under "normal circumstances" the long-run inflation rate, g_P, tends roughly to satisfy

a relation like

$$g_P \approx g_M - g_Y, \tag{18.1}$$

where g_M is the trend rate of money growth and g_Y is the trend rate of *GDP* growth. Or, in order to allow the long-run elasticity of money demand w.r.t. income to differ from one, let us write

$$g_P \approx g_M - \eta g_Y. \tag{18.2}$$

This says that long-run inflation tends to equal the excess of money growth over output growth, up to the elasticity factor η.

The inflation law (18.1) is for example what the Sidrauski model predicts in the absence of technological progress (so that $g_Y = n$). As noted in the previous chapter, with Harrod-neutral technological progress at rate g, the prediction from the Sidrauski model is that $g_P \approx g_M - n - \theta g/\varepsilon$, where θ is the elasticity of marginal utility of consumption and $1/\varepsilon$ is the absolute interest elasticity of money demand. With $n \approx 0$, this corresponds to (18.2) with an elasticity factor $\eta \approx \theta/\varepsilon$.[1] If *firms'* need for liquidity is included, something similar comes up again.

The monetary aggregate M that is most relevant to (18.1) and (18.2) is not the monetary base, M_0, over which the central bank has direct control. Rather, M should be interpreted as including bank-created money, that is, the money supply in the usual meaning, M_1, or even "broader" money, thus including time deposits and other items. As long as the money multiplier (the ratio of money supply to the monetary base) is stable, this difference is not so important. And the money multiplier usually is fairly stable, as noted in Chapter 16. But there are exceptions. For example, during the early Great Depression in the US it fell drastically.

Fig. 18.1 gives some crude empirical cross-sectional evidence for 24 OECD countries 1950-1990.[2] The shown relationship between long-run inflation and excess monetary growth certainly has a flavour of the rule (18.1). Of course such a graph does not tell us whether there is a *causal* relation and, if there is one, what way it goes. It could go from the left to the right or from $g_Y + g_P$ to g_M (a form of accommodating monetary policy). When one considers country samples including very high inflation countries, the association between

[1] Goldfeld (1973) estimated the long-run elasticity w.r.t. the nominal interest rate to be around $-0,15$. He found the long-run elasticity (i.e., taking time lags into account) of money demand (M_1) w.r.t. Y to be around $2/3$. The respective estimated short-run elasticities are considerably smaller in absolute value. Anyway, with $\theta = 4.5$ and $1/\varepsilon = 0.15$ we get $\theta/\varepsilon \approx 2/3$.

[2] Because of differences in financial institutions, the measurement of M_1 and M_2 varies across countries more than does that of M_0. Therefore the data source uses the growth rate of M_0.

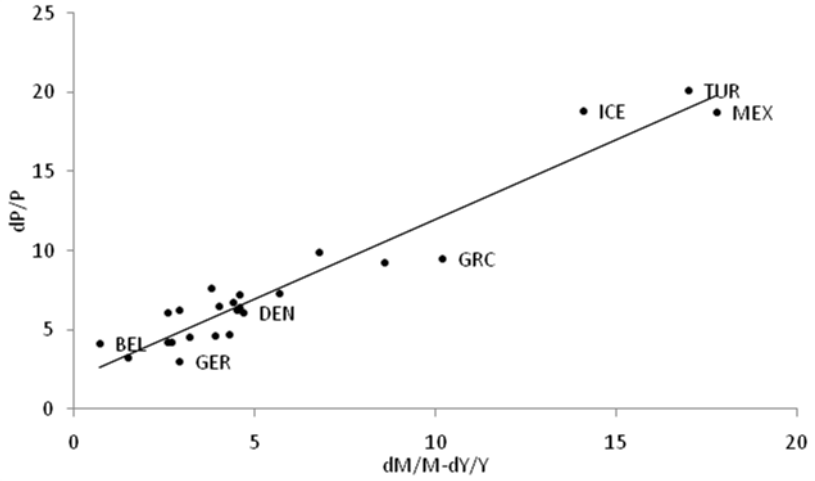

Figure 18.1: Inflation and the excess of money growth over output growth 1950-1990 for 24 OECD countries. Note: average annual rates are in per cent, P is the consumer price index, M is currency, and Y is real GDP. Data source: Barro and Grilli (1994).

money creation and inflation becomes even more striking; at least for the high inflation countries there is relatively clear evidence that the fierce money creation, or rather the underlying strive for seigniorage to the government, is the explanatory factor (see, e.g., Dornbusch et al., 1990).

Fig. 18.2 is based on *longitudinal* data (i.e., across decades) on inflation and growth of M_1/Y in Denmark, again indicating a relation like (18.1) or (18.2).

This kind of coarse evidence is sometimes (wrongly) taken to be supportive of the Quantity Theory of Money. That is the theory based on the classical supposition that the velocity of money is a constant. To clarify the issue, let V denote the income velocity of money, i.e., $V \equiv (P \cdot Y)/M$, where M is money supply. With M^d denoting money demand, the Quantity Theory of Money claims that

$$M = M^d = \frac{P \cdot Y}{\bar{V}},$$

where $\bar{V}$ is a constant and thus independent of the nominal interest rate. The above graphs seem consistent with this supposition. Yet, this sort of long-run data does not exclude that velocity fluctuates with the nominal interest rate, if such variations are short-run in character. A simple alternative to the Quantity Theory is the Keynesian money demand hypothesis,

$$M^d = P \cdot L(Y, i), \qquad L_Y > 0, L_i < 0,$$

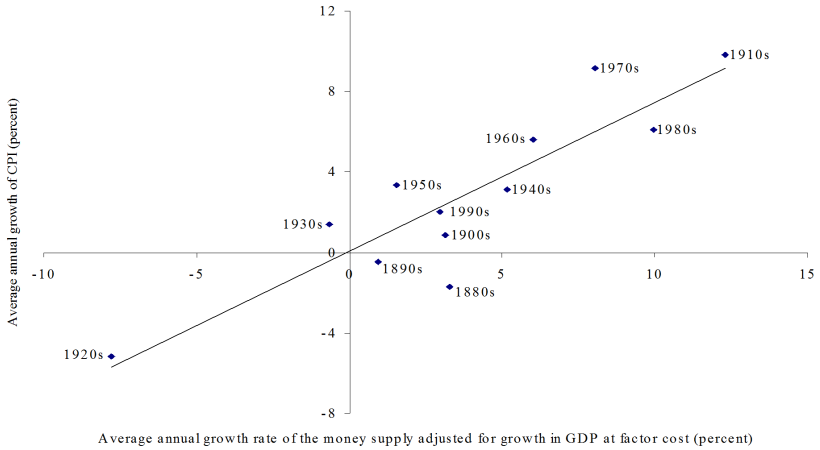

Figure 18.2: Inflation in decades since 1880 against excess of money growth over output growth in Denmark (average annual rates in per cent). Data source: Statistics Denmark, IMF (International Financial Statistics), and Kærgård (1991).

where $L(\cdot)$ is the real money demand function ("L" for liquidity) and i is the nominal short-term interest rate. Then

$$V \equiv \frac{P \cdot Y}{M} = \frac{P \cdot Y}{M^d} = \frac{P \cdot Y}{P \cdot L(Y,i)} = \frac{Y}{L(Y,i)},$$

so that velocity depends positively on i for a given income Y. Yet, if i does not show much trend, long-run data may still indicate a more or less constant velocity of money, namely if the long-run elasticity of $L(\cdot)$ w.r.t. Y is relatively close to one. Mehra (1993) finds for M_2 this elasticity to be 0.98 and the long-run interest elasticity of money demand to be -0.08.[3]

On the other hand, the data in fact shows that velocity for M_1 in USA since the Second World War has been systematically increasing, although with a temporary halt in the 1980s (Stock and Watson, 1999); also the M_2 velocity has increased somewhat (DeLong, 2006). Such trends need not reflect an income elasticity of money demand below one, but can simply be due to *changes* in the money demand function caused by improvements in the "payment technology" (say use of payment cards, credit cards, electronic home banking, and other financial innovations). A related issue in macroeconomics is to what extent the money demand is a stable function of only the short-term nominal interest rate and the transaction volume, as measured by income. Keynesian oriented economists emphasize shifts in the money demand function (due to "liquidity preference shocks"). In contrast, monetarists hold the view that the money demand function is not only relatively

[3] See also Hetzel and Mehra (1989).

insensitive to the interest rate but also relatively stable. On the basis of this view, monetarists favor a "passive" k per cent money growth rule, k being close to the trend growth rate of GDP.

18.2 Are neutrality and superneutrality of money theoretically robust properties?

Money neutrality far from robust

In line with the general tenet of classical monetary theory, the Sidrauski model features *money neutrality*: resource allocation in the economy is independent of the *level* of the money supply even in the short run. This property is a consequence of the assumption in these models that prices are perfectly flexible. Instantaneous price adjustment constitutes the mechanism through which supply and demand are matched. Lucas (1972, 1975) generalized the concept of money neutrality by emphasizing the distinction between anticipated and unanticipated changes in money supply. The idea is that money neutrality should hold only for anticipated changes in money supply (on which more in Chapter 26).

Today most macroeconomists seem to agree that the tenets of short-run price flexibility and short-run neutrality of money do *not* provide acceptable approximations to reality. Indeed, empirical studies generally conclude that in the short run, an increase in the money supply tends to decrease the short-term nominal and real interest rates rather than increasing the price level and leaving the real interest rate unchanged as the neoclassical models predict. This also holds for anticipated changes in money supply.[4]

Superneutrality?

What about superneutrality of money? In the Sidrauski model the growth rate of the money supply does not affect capital accumulation and consumption in a steady state. Hence, in that model, which is an example of a representative agent model, money is *superneutral*. This is, however, not a theoretically robust property. Even remaining within a representative agent framework, slightly more *general* specifications may show absence of superneutrality. And in overlapping generation models with life-cycle behavior this absence is the rule rather than an exception. It is another question how far away, quantitatively, the real world is from superneutrality. On this

[4] See Blanchard (1990), Walsh (2003), Gali (2008). This is not to deny that there is room for disagreement about how long the "short run" lasts.

question there seems to be less agreement. Some will argue that superneutrality often can be used as an acceptable approximation. Others emphasize that, for example, an important deviation from superneutrality arises because inflation shapes the impact of taxes, when the tax system is based on nominal income as it is in most countries.

Before looking more closely at this, notice that the basic reason that superneutrality appears in the Sidrauski model is that the Keynes-Ramsey rule holds not only at the individual level, but also at the aggregate level. Capital accumulation (increasing capital intensity) continues as long as the interest rate, r, exceeds the rate of time preference, ρ, or, in a model with technological progress at the rate g, as long as r exceeds $\rho+\theta g$ (where θ is the elasticity of marginal utility of consumption). As a consequence, the system settles down in a steady state where $r = \rho$ (or, more generally, $r = \rho + \theta g$) whatever the money growth rate.

More general representative agent models If the Sidrauski model is extended with endogenous labor supply or if we introduce "money in the production function" (firms also need cash), then there is scope for money growth affecting capital accumulation and consumption through these channels even in steady state.

However, a quantitatively more important factor is probably that tax systems are often not neutral with respect to inflation. If taxes are based on nominal incomes, the allocation of resources tends to be affected by shifts in the rate of inflation. Example: suppose there is a given proportional tax rate $\tau \in (0, 1)$ on nominal capital income. Then, when higher inflation leads to a higher nominal interest rate (given the real interest rate $\bar{r}$), capital income is taxed more heavily. The result is that the real *after tax* interest rate, $\tilde{r}$, becomes lower:

$$\tilde{r} = (1-\tau)i - \pi = (1-\tau)(\bar{r}+\pi) - \pi = (1-\tau)\bar{r} - \tau\pi. \tag{18.3}$$

Thus, an increase in the inflation rate, π, caused, say, by an increase in money growth, decreases the real interest rate after tax. And a volatile inflation rate implies a volatile after-tax real interest rate.

Overlapping generations models and the Tobin effect Let us turn to overlapping generations (OLG) models. In an OLG framework the Keynes-Ramsey rule holds only at the individual level, not at the aggregate level. Therefore, even without leisure in the utility function or money in the production function, superneutrality of money generally does not hold.

Two effects are involved that, depending on circumstances, supplement or counteract each other: the Tobin effect and the transfers effect. By creating higher inflation, more rapid money growth tends to increase the nominal interest rate and thereby the opportunity costs of holding money. This induces a larger fraction of private wealth to be held in capital. The resulting possible stimulation of capital accumulation in steady state is called the *Tobin effect* (Tobin 1965).

On the other hand, to the extent that income transfers are financed by money growth, the real value of transfers (x in the Sidrauski model) depends on the rate of monetary expansion. Through this channel there may be an additional effect on capital accumulation. This is called the *transfers effect*. As shown in the next section, if the (absolute) interest elasticity of money demand is not too high, this effect is positive, hence stimulating consumption. Thereby, saving and capital accumulation become smaller, thus counteracting the Tobin effect. If, however, money demand is sufficiently interest elastic, money growth affects the transfers negatively, thus decreasing consumption and stimulating saving and capital accumulation. In this case the transfers effect on capital accumulation is positive and adds to the Tobin effect.

Though, perhaps, not likely to be large, the net effect of the two effects will in general be different from zero in an OLG model. Using U.S. data for more than a century, Ahmed and Rogers (2000) find the Tobin effect and the absence of superneutrality to be statistically significant, but of limited size.

The overall picture is that the *classical dichotomy* − the notion that real variables (employment, production, consumption, capital accumulation) are determined separately from money variables − can easily break down even in neoclassical models with perfect price flexibility.[5] Thus, neoclassical theory should not be identified with the position that "money is a veil" (a position which is almost self-contradictory in view of money's enormous role in facilitating trade).

Inflation and deflation bubbles At the theoretical level there is yet another circumstance that may cause both superneutrality of money and the inflation law as described above to break down. We saw in the previous chapter that even maintaining the assumption of rational expectations, neither negative nor positive bubbles in the real value of money, that is, expectations-driven hyperinflation or hyperdeflation, can be theoretically ruled out. For

[5] Here, we think of the "classical dichotomy" in a broad sense. It is otherwise if "classical dichotomy" is interpreted narrowly as synonymous with money neutrality as distinct from superneutrality. The neoclassical models confirm the classical dichotomy in this narrow sense.

the Sidrauski model this opens up for absence of convergence towards the steady state. This issue notwithstanding, it is a fact that there is no known historical hyperinflation in which money growth itself did not become extremely high. This takes us to the issue of seigniorage and hyperinflation to which we now turn.

18.3 Inflationary public finance

One of the pioneers in the study of high inflation episodes is the economist Phillip Cagan from the University of Chicago. In his famous study of seven historical cases of very high inflation, he defines *hyperinflation* as occurring when inflation is running at 50 per cent per *month* or more.[6] This exact borderline may be practical in an empirical study, but is rather arbitrary in a theoretical context. So we will use the term hyperinflation as synonymous with just "very high inflation". Cagan and other observers of high inflation episodes emphasize the key role of base money creation as a source of government revenue in countries and periods with large government deficits.

Let

G = denote real government spending on goods and services,
T = real net tax revenue (i.e., gross tax revenue minus transfer payments),
M = the monetary base at time,
P = price level measured in money (i.e., the GDP deflator),
B = real government debt, and
S = seigniorage,

all at time t. The *real government budget deficit* is defined as the excess of real government spending over real government revenues, i.e., $rB + G - T$. The deficit can be financed by debt issue and by base money creation:

$$\dot{B} + \frac{\dot{M}}{P} = rB + G - T.$$

The consolidated public sector consists of the fiscal authority and the central bank. The term $\dot{M}/P$ in the above equation represents *seigniorage*, that is, the revenue per time unit obtained by the consolidated public sector by printing base money; it is virtually cost-less for the central bank to print more notes.[7] In the Sidrauski model of the previous chapter we considered

[6] Cagan (1956).

[7] Notice that when discussing seigniorage, it is the monetary base ("outside money") that is relevant.

the special case where $\dot{B} = B = G = 0$ and $T = -X$, where X are income transfers from the government to the private sector.

Denoting the real seigniorage S, we have

$$S \equiv \frac{\dot{M}}{P} = \frac{\dot{M}}{M}\frac{M}{P} = \mu\frac{M}{P}, \tag{18.4}$$

where μ is the growth rate of base money. We see that besides μ, the real money stock is a determinant of seigniorage. The real money stock depends negatively on the price level, which in turn tends to increase fast, if μ is high. This indicates that there may be an upper bound for seigniorage.

18.3.1 The seigniorage Laffer curve

To understand the argument in detail, let $L(Y, i)$ be the real money demand function, $L_Y > 0, L_i < 0$. Ignoring the commercial banks (or simply assuming a constant money multiplier), clearing in financial markets implies

$$\frac{M}{P} = L(Y, i). \tag{18.5}$$

The nominal interest rate is

$$i = r^e + \pi^e, \tag{18.6}$$

where π^e is the expected inflation rate. We follow Cagan who argued that during a hyperinflation, expected inflation swamps the influence of Y and $r^e \approx r$ on real money demand. Thus, to focus on the fast-changing nominal variables, we let output and the expected real interest rate be given at constant levels, $\bar{Y}$ and $\bar{r}$, respectively. In the context of hyperinflation, where capacity utilization is typically close to 100%, changes in Y and probably also r are slow anyway, relative to the nominal changes.

Effectively, we thus assume the *Fisher equation* (sometimes called the *Fisher hypothesis*) claiming that the nominal interest rate changes one-to-one with expected inflation.[8] Furthermore we assume rational expectations,

[8]The name Fisher equation refers to the American economist Irving Fisher (1867-1947). The Fisher equation, Fisher (1930), should be distinquished from *Fisher's identity:* $i \equiv r^e + \pi^e$. This identity just reflects the definition of the expected real interest rate, r^e, in continuous time with continuous compounding. In contrast, the *Fisher equation* is the testable proposition that the real interest rate is unaffected by changes in expected inflation. Nowadays, the consensus view seems to be that Fisher's equation fails for the short run, but may be roughly applicable for the long run.

which here means perfect foresight, i.e., $\pi^e = \pi$. By taking logs and differentiating w.r.t. t on both sides of (18.5) we then get

$$\begin{aligned} \mu - \pi &= \frac{L_i(\bar{Y}, \bar{r}+\pi)}{L(\bar{Y}, \bar{r}+\pi)}\dot{\pi} \qquad \text{or} \\ \dot{\pi} &= \varphi(\pi)(\pi - \mu), \quad \text{where} \quad \varphi(\pi) \equiv -\frac{L(\bar{Y}, \bar{r}+\pi)}{L_i(\bar{Y}, \bar{r}+\pi)} > 0. \end{aligned} \tag{18.7}$$

For simplicity, let the rate of monetary expansion, μ, be a given constant. We see that the first-order differential equation (18.7) in the inflation rate is unstable in the sense that $\dot{\pi} \gtreqless 0$ for $\pi \gtreqless \mu$, respectively. The initial inflation rate π_0 is *not* predetermined, however, since both expected and actual inflation are forward-looking. Therefore, on the face of it there is a continuum of solutions to (18.7). Except one, all these solutions are divergent (implying accelerating inflation or deflation) in spite of the rate of monetary expansion and the volume of transactions being constant. The divergent solutions are examples of purely expectations-driven hyperinflation or -deflation. We will restrict our attention to cases where such inflation or deflation bubbles do not occur. As argued in the previous chapter, it is not clear how this restriction can be defended on purely theoretical grounds. We consider ruling out bubbles as just a simplifying assumption in a first approach to questions like: What shape does the seigniorage Laffer curve have in the absence of inflation or deflation bubbles? How can hyperinflations arise, even if they are not generated by self-fulfilling expectations?

Ruling out bubbles, there is only one solution left, namely $\pi_0 = \mu$, and so

$$\pi = \mu \quad \text{for all } t \geq 0. \tag{18.8}$$

As an implication, $i = \bar{r} + \mu$, and inserting this and (18.5) into (18.4) gives

$$S = \mu \frac{M}{P} = \mu L(\bar{Y}, \bar{r} + \mu).$$

We see that

$$\begin{aligned} \frac{\partial S}{\partial \mu} &= L(\bar{Y}, i) + \mu L_i(\bar{Y}, i) \qquad (\text{since } i = \bar{r} + \mu \Rightarrow \frac{\partial i}{\partial \mu} = 1) \\ &= (1 + \frac{\mu}{\bar{r}+\mu}\frac{i}{L(\bar{Y}, i)}L_i(\bar{Y}, i))L(\bar{Y}, i) \\ &= (1 - \frac{\mu}{\bar{r}+\mu}E_{L,i})L(\bar{Y}, i) \gtreqless 0 \quad \text{for} \quad E_{L,i} \lesseqgtr \frac{\bar{r}+\mu}{\mu}, \end{aligned} \tag{18.9}$$

where $E_{L,i}$ is the absolute elasticity of real money demand w.r.t.. the nominal interest rate.

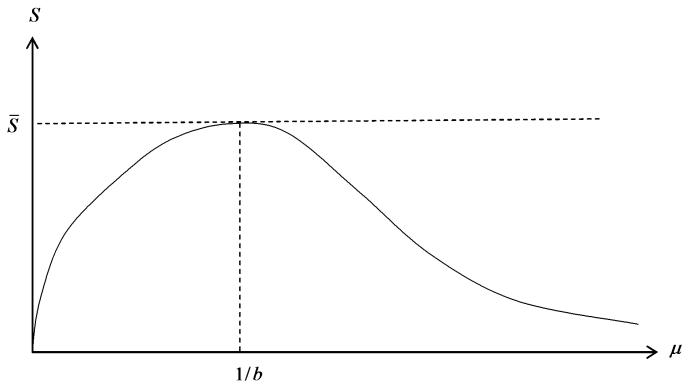

Figure 18.3: The seigniorage Laffer curve.

Since $E_{L,i}$ is under normal circumstances estimated to be below 1, with $\bar{r} > 0$ and $\mu > 0$ this analysis might seem to imply that S has no maximum. Situations of hyperinflation are not "normal circumstances", however. Empirical studies by Cagan (1956) and others indicate that when inflation is very high, so is the interest elasticity of money demand. The money demand function favoured by Cagan is

$$L(Y, i) = \alpha Y^{\eta} e^{-\beta i},$$

where α, η, and β are positive parameters (α just depends on measurement units, η is the output elasticity of money demand, and β is the semi-elasticity of money demand w.r.t. the nominal interest rate, i.e., $\beta = -L_i/L$). This gives

$$E_{L,i} = \beta i = \beta(\bar{r} + \mu).$$

Combining with (18.9) we see that

$$\frac{\partial S}{\partial \mu} \gtreqless 0 \quad \text{for } \mu\beta \lesseqgtr 1, \text{ i.e., } \mu \lesseqgtr \frac{1}{\beta},$$

where $1/\beta$ is a measure of the sensitivity of the nominal interest rate w.r.t. the real money supply.

With $\beta = 1/2$ (close to Cagan's estimates), when $\mu > 200$ per cent per year we are at the backward-bending part of the seigniorage Laffer curve

shown in Fig. 18.3.[9] The curve reflects that when μ increases, there is both a direct effect on S and an indirect effect, via $M/P = L(\bar{Y}, i)$. These effects are in opposite directions. Let $\mu = 0$ initially and let μ increase. To begin with, the direct effect on S dominates so that S increases with μ, cf. (18.4). But ultimately the increase in μ is more than offset by the ensuing decrease in real money demand due to the spurred inflation. (A reservation should be added to this exposition. Cagan's estimates are based on the hypothesis of adaptive inflation expectations and may not be transferable to a context with rational expectations, as here. And, in fact, econometric estimates of the interest elasticity of money demand under hyperinflation vary considerably.)

18.3.2 Hyperinflation

The above analysis does not in itself explain episodes with inflation rising to exorbitant levels as during the German hyperinflation Aug. 1922 - Nov. 1923 (322% per month), the Hungarian hyperinflations Mar. 1923 - Feb. 1924 (46% per month) and, again, Aug. 1945 - Jul. 1946 (19,800% per month) or the Latin American hyperinflations in the 1980s and early 1990s with inflations running at 20-40% per month (Bolivia 1984-85, Nicaragua 1987-91, Argentine and Peru 1989-90, Brazil 1989-94).[10] Such extreme inflations typically arise when the government runs into a budget crisis and attempts to get more seigniorage than the maximum sustainable seigniorage, $\bar{S}$. Thus, for each of the seven hyperinflations studied by Cagan (1956),[11] the actual rate of money growth far exceeded the rate of money growth required for maximum sustainable seigniorage (as estimated by Cagan). The underlying reason was the endeavour to monetize huge government budget deficits.

The background for a budget crisis leading to hyperinflation can be:

- the need for reconstruction after a war combined with "war reparations" to the victorious former enemy (Germany in 1922-23),
- revolution, civil war and other social conflicts reducing the ability to collect taxes (Nicaragua in the 1980s, Zimbabwe 2006-),

[9] As noted in Chapter 6, originally, "Laffer curve" (after the American economist Arthur Laffer) referred to a hump-shaped relationship between the income tax rate and the tax revenue.

[10] The reported monthly inflation rates are averages over the specified periods. The *maximum* monthly rate is much larger, for example, 32,400% in the German case and 261% in the Nicaraguan case. The numbers reported are from Sachs and Larrain (1993) and Blanchard (2003).

[11] Austria 1921-22, Germany 1922-23, Greece 1943-44, Hungary 1923-24 and 1945-46, Poland 1923-24 and Russia 1921-24.

- a substantial decline in the price of some raw material on which the country's income and revenues for the government depend heavily (as with indebted Bolivia in 1984-85, where the principal export good, tin, fell sharply in price),

- a sharp reversal of the relationship between the real interest rate in the world market and the output growth rate in a highly indebted country (this was the situation in many of the Latin American countries after the second oil price shock 1979-80; the governments had in the 1970s, where $r < g_Y$, borrowed heavily from foreign banks or had guaranteed private borrowing from these; in the 1980s, where $r > g_Y$, the governments had serious debt-service difficulties and lost most of their international credit worthiness).

Budget deficits and accelerating inflation

Whatever the background, at the start of hyperinflations there is typically a situation with full capacity utilization and a large budget deficit, which the government/central bank then attempts to finance by base money creation. Typically, the ability of the government to finance the deficit by borrowing is limited. The government may already be heavily indebted so that issue of new debt would tend to raise the interest rate on government bonds, since lenders take the risk of debt default into account. This magnifies the budget deficit. For simplicity, we shall assume that the total budget deficit is financed by money creation so that B is constant. Further, we assume that the primary budget deficit, $G - T$, is constant. This is also a simplification since, in practice, during high inflation the budget deficit and the needed seigniorage typically *rise.* This is because the real tax revenue is eroded as inflation accelerates due to the *time lag in tax collection* combined with the usually nominally defined tax rules. For the case of Bolivia Fig. 18.4 shows the rising budget deficit-income ratio up to the stabilization in August 1985. This interaction between worsening fiscal conditions and accelerating inflation in fact only speeds up the soaring instability. As we shall see, even with a constant, but large, budget deficit the essential mechanism behind hyperinflation can be envisioned.

Assuming that also r is constant, we consider a situation where the needed seigniorage is $S^* = rB + G - T$, a constant. When $S^* > \bar{S}$, hyperinflation is likely to develop. To describe the mechanism, we allow the price level and the expected inflation rate to deviate in the very short run from their equilibrium values, because adjustment takes time and to begin with the intention of the government/central bank is not transparent. Thus, real money supply may

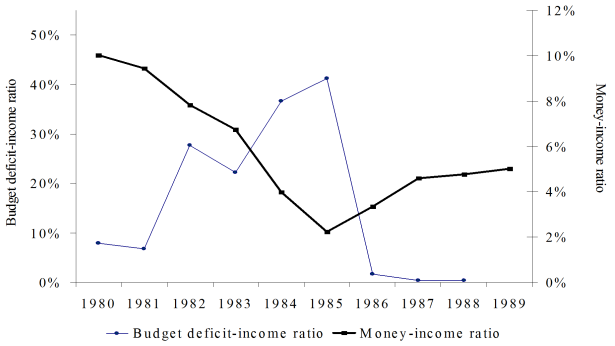

Figure 18.4: Budget deficit-income ratio and money-income ratio in Bolivia 1980-1989. Note: income is nominal GDP. Data source: IMF, International Financial Statistics.

in the very short run deviate from desired real money holdings. Suppose the government has already set μ at $1/\beta$ and realizes that the obtained seigniorage is below the desired. In an attempt to reap a temporary benefit, the government now increases μ to $\mu' > 1/\beta$. As long as the price level is not fully adjusted, M/P is larger than money demand, so that seigniorage, $S = \mu' M/P$, is for some time larger than before, due to the high μ'. To the extent that *expected* inflation temporarily lags behind actual inflation, this amounts to an additional source of seigniorage that is not sustainable; this seems plausible in the situation and was indeed supported by Cagan's study.[12] Using (18.5) and (18.6), the scenario therefore is

$$\frac{M}{P} \geq L(\bar{Y}, \bar{r} + \pi^e) \geq L(\bar{Y}, \bar{r} + \mu').$$

Strict inequalities rule for a short period after the rise of μ to μ': the first inequality because the price level path, P, has not yet risen enough, the second because expected inflation is likely to lag behind the new, higher equilibrium level equal to μ'. So, for some time we have $S = \mu' \frac{M}{P} > \mu' L(\bar{Y}, \bar{r} + \mu')$. But people rush to spend their money[13] or convert them into some other form (for example foreign currency, if possible). The flight away from the home currency swiftly reduces its real value, $1/P$. When the upward adjustment of the price level and inflation expectations is completed, seigniorage is again

[12] Cagan (1956) found for the seven hyperinflations after the first and second world wars that the hypothesis of adaptive inflation expectations, $\dot{\pi}^e = \lambda(\pi - \pi^e)$, $\lambda > 0$, gave a good fit.

[13] The legend attributes this remark to Keynes: "During high inflation you order *two* beers at a time." As inflation speeds up, we may add, you order three, four, five ..... at a time and clearly the situation is not sustainable.

too low, namely at the level $\mu' L(\bar{Y}, \bar{r} + \mu') < S^*$. This motivates a new increase in μ, a higher seigniorage is obtained, but only for a short while and thus a further increase in μ is fuelled − and so on. The ensuing hyperinflation continues until the lack of sustainability of the situation is fully realized and the chaos resulting from the breakdown of the transaction system is felt too troublesome. The way out is a fiscal reform drastically reducing the budget deficit.

Stopping hyperinflation

A stabilization program will typically consist of a package of both fiscal and monetary policy elements. This may involve public sector price increases, decrees reducing government expenditures, initiatives aiming at more effective collection of taxes, central bank independence, alignment of the exchange rate with a foreign currency with low inflation and similar attempts at commitment. Several of the Latin American programs also included wage and price controls in order to ease coordination around a lower inflation rate. A stabilization program can not be successful unless the cause underlying hyperinflation, the need for large seigniorage, is eliminated. And it is important that the stabilization program is understood and believed by the public. Otherwise, inflation − expected and actual − and the nominal interest rate will not fall significantly. And if inflation decreases by less than money growth, real money supply falls and drives the real interest rate up which can set off a recession and unemployment.[14] There may also be costs in the form of redistribution of wealth that is felt injurious. On the other hand, *if* the program is credible, then there is in fact a temporary bonus for the government in the form of a high seigniorage for some time before its final removal.

Fig. 18.5 gives a stylized picture of this, assuming the idealized case of perfect credibility and perfectly flexible prices. Until time t_0 the price level is rising fast. We imagine the credibly announced plan is to bend the price line at time t_0, making it horizontal thereafter. The time between announcement and implementation of the program is assumed to be short.[15] In line with the program, at time t_0 expected and actual inflation jumps from μ' to zero and the nominal interest rate jumps from $\bar{r} + \mu'$ to $\bar{r}$. This increases real money demand to $L(\bar{Y}, \bar{r})$. This could lead to excess money demand and thereby an

[14] Bolivia as well as the other hyperinflation countries went through several unsuccessful stabilization programs before they succeeded. In fact, an element of self-fulfilling prophecy seems to be involved. Even a well-conceived plan may lead to failure if it is expected to.

[15] This is similar to the recommendation of the "credibility doctrine" put forward by Sargent (1982).

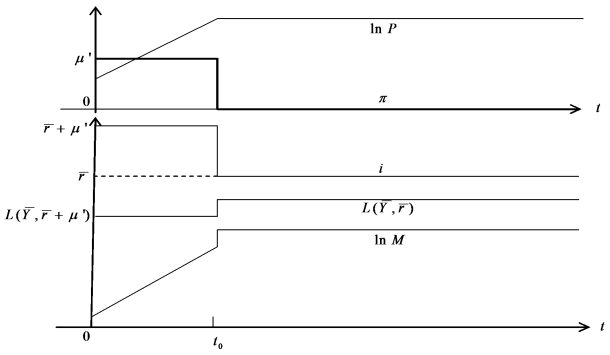

Figure 18.5: Ending hyperinflation in the idealized case of perfect credibility and perfectly flexible prices.

immediate rise in the real value of money, $1/P$, that is, a *drop* in P, contrary to the desired price path (and with devastating consequences for agents with debt fixed in nominal terms). To avoid this, the central bank creates exactly the required extra amount of money at time t_0 and leave the money supply constant at that higher level forever. This amounts to a welcome last dose of seigniorage − the temporary bonus mentioned above − without inflationary consequences if the public has confidence in the program. Indeed, at the time of stabilization, Autumn 1985, in Bolivia a small pinnacle on the money supply curve is actually visible in Fig. 18.6.[16] From the solid curve in the figure we also see that the inflation was indeed stopped quite abruptly as has been the case with all observed hyperinflations when finally a stabilization program succeeds.

Unfortunately, however, the actual course of events is certainly never as easy and straight as the perfect credibility-perfect flexibility diagram in Fig. 18.5 suggests. True, the chaos caused by hyperinflation is unbearable, but stopping hyperinflation is in practice not without economic and social costs. There is some controversy about the size of these costs and how to minimize them. Yet, most analysts agree that the radical disinflations associated with ending hyperinflation have been associated with less costs, that is, smaller output and employment reductions, than would be expected on the basis of simple extrapolation of the experience from disinflations in countries with only moderate inflation. Examples of the latter are the Reagan-Volcker and the Thatcher disinflations in the U.S. and U.K., respectively, in the late 1970s and early 1980s or the continental European disinflations during the 1980s.

Thus, there seems to be a basic difference between disinflation under

[16] But in fact a renewed inflation immediately after is also visible. It lasted only a couple of months. For details, see Morales (1988) and Morales and Sachs (1989).

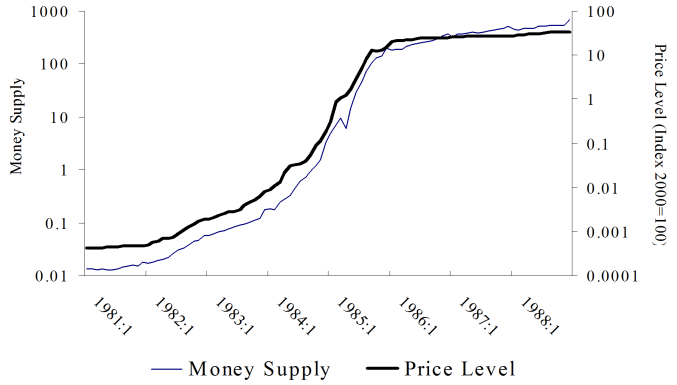

Figure 18.6: The end of hyperinflation in Bolivia (log scale). Data source: IMF, International Financial Statistics.

hyperinflation and disinflation under "normal circumstances". Why is it so? Related to this is the question: why is even under "normal circumstances" the role attributed to money so different in short-run and long-run analysis?

Normal circumstances and hyperinflation compared

The general wage level and the general price level are not asset prices in centralized asset markets, but averages of millions of distinct wages and prices, respectively, set by the agents at discrete points in time.

Under normal circumstances, say when annual inflation is less than 20%, wages are usually set or negotiated in nominal terms in advance for some period of time. And the contract periods are typically staggered across the different labor markets (this is referred to as asynchronous wage setting). Similarly, contracts between firms may fix prices for some period of time, also in a staggered pattern. Thus, when the government or central bank decides to disinflate, there is an *overhang* of past nominal contracts, embodying past expectations of high inflation. Renegotiations take time and the *nominal rigidities* lead to *inflation inertia.* Abrupt disinflation policy will then change the situation from excess demand and a relatively low ex ante real interest rate to one with high ex post real interest rate. Debtors will face liquidity problems. Debt defaults together with a fall in aggregate demand, output, and employment are likely. Although the recession − via the Phillips curve − reduces inflation, it may take several years to bring it down to the desired

level. In this situation a "gradualist" disinflation policy may achieve the goal of lower inflation with less costs in terms of unemployment than the "cold turkey" recommended by the new-classicals. The difficulty of a gradualist approach is, however, that it may not be credible (see Carlson 2002).

But when an economy has had high inflation for some time, such durable contracts in nominal terms cease to be made. Instead very frequent negotiations take place. As a consequence a well-conceived credible disinflation program should be able to stop hyperinflation quite fast and result in a much smaller "sacrifice ratio". The *sacrifice ratio* is an indicator of the costs of disinflation and can be defined as the cumulative percentage loss in output needed to reduce trend inflation by one percentage point.

As we shall see in the next section, it is in some sense similar forces that explain that even within the confines of "normal circumstances", the *short-run* monetary transmission mechanism is different from the *long-run* transmission mechanism.

18.4 Bridging the gap between the short and the long run

In the next chapters we address in detail the short-run adjustment mechanisms in the economy. As a prelude to that, this section gives a broad overview of how to bridge the gap between long-run and short-run analysis.

18.4.1 The monetary transmission mechanism in the short and the long run

A general tenet of mainstream macroeconomics is that, under "normal circumstances", real effects of changes in money supply dominate in the short run, whereas nominal effects dominate in the long run. Thus, a shift to a higher money growth rate will in the short run have positive effects on output and employment and little effect on inflation. But after some time the real effects tend to disappear and the higher money growth rate just ends up in a correspondingly higher inflation rate. What happens between the short and the long run?

Admittedly in a stylized way, Fig. 18.7 shows a diagram (inspired by Blanchard, 2003, p. 304) that may help clarifying the adjustments involved according to mainstream macroeconomics. We consider an economy with constant technology and constant labor force. Until time t_0 the annual money growth rate and inflation rate have for some time been μ (say, 2% per year)

and output and employment have been at their "natural" (or NAIRU) level. Thus for $t < t_0$ real money supply, M/P, is constant and the nominal interest rate is $i = \bar{r} + \mu \equiv i_0$, where $\bar{r}$ is the "natural" real rate of interest, i.e., the rate consistent with steady state.

At time t_0 the rate of money growth is unexpectedly increased to μ' $= \mu + 5\%$ and credibly announced to be maintained at this higher level. Owing to nominal rigidities, inflation, π, react only slowly. So real money supply goes up. The higher liquidity drives the nominal interest rate down and the real interest rate, $r = i - \pi^e$, follows suite. To the extent *expected* future inflation, π^e, goes up in response to the agents' belief that the higher growth rate in money supply will be maintained, r begins departing in a downward direction from the nominal interest rate. The fall in r stimulates aggregate demand so that output and employment go up. Hereby, wage and price inflation gradually rise. The rising inflation slows down the upward movement of real money supply; and the rising level of output increases the volume of transactions and thereby the demand for money. So the downward movement of the nominal interest rate comes to a halt and is reversed. When the boom has pushed inflation above μ', real money supply begins to fall so that the nominal interest rate increases further and pulls the real interest rate up again. This dampens output demand and eventually the economy settles down with inflation and nominal interest rate increased by 5 percentage points, output and real interest rate back at their "natural" level and the real money supply reduced to a lower level.

The "elephant trunk" in the lower part of Fig. 18.7 illustrates the basic feature of the adjustment process: the increased money growth rate leads to a fall in the nominal interest rate in the short run, but to a rise in the long run.[17] The background for the different effects is that the nominal rigidities, caused by the *overhang* of past contracts embodying past expectations, peter out as time passes by. The contract overhang is here gradually eliminated by the mere passage of time (whereas under hyperinflation medium-term contracting is simply avoided).

Another issue at the borderline between the long run and the short run is the question about the costs and benefits of inflation.

[17] In addition to the interest rate channel emphasized here, there are other channels (such as the asset price channel, the credit channel and the exchange rate channel), through which an expansion of money has real effects in the short run, but only nominal effects in the long run. Tjese are taken up in Part V-VII in this book.

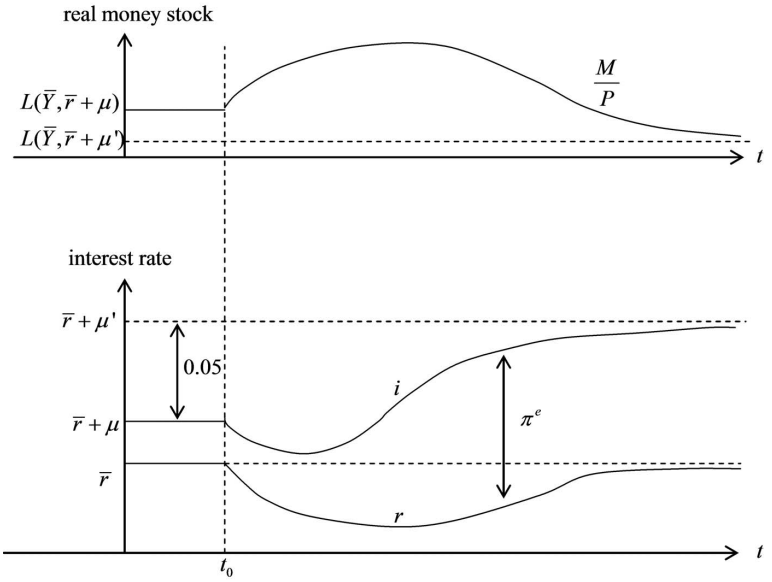

Figure 18.7: The elephant trunk. Stylized illustration of the transmission of a 5 percentage-point rise in the money growth rate.

18.4.2 Inflation - social costs and benefits

The monetarists, lead by Milton Friedman, as well as new-classicals, like Lucas, Sargent, and Barro, are of the opinion that inflation is an evil and should be avoided. There is no general agreement among economists about this matter, however. Keynesians and new-Keynesians put forward several circumstances that seem important for the role of money in the short-to-medium run, but are ignored by standard neoclassical models, e.g., the Sidrauski model. Indeed,

> "Inflation hurts, but deflation could be worse" (New York Times 1982).

Below we list main social costs of inflation. This is followed by a list of social benefits of (moderate) inflation that economists have identified.

Social costs of inflation

1. "Shoe-leather costs" In the medium term, say 7-30 years, the real interest rate does not fluctuate much. Thus, suppose it is a constant, $\bar{r}$. Then, in the medium term, a higher rate of inflation leads to a higher nominal interest rate, $i = \bar{r} + \pi$, that is, higher opportunity costs of holding money.

As a result people decrease their average money holding and make smaller and more frequent adjustments of the liquid component in their portfolios. They have to "go to the bank more often" (or, in modern times, do the home banking more often). This is cumbersome and "the shoe leather wears down" (a metaphor for the increased transaction costs). An answer to the problem is Milton Friedman's zero-interest rate rule and this requires low inflation, indeed *deflation*.

How large can the "shoe-leather costs" be? The opinion is divided − new classical economists tend to estimate them as higher than do Keynesians. Based on time series data for USA 1900-94 Lucas (2000) suggested that the per year welfare gain by reducing inflation from 10 per cent per year to zero is approximately of the same size as the welfare gain obtained by GNP being 1 per cent higher than otherwise every year. According to other authors the applied method of calculation is debatable and the estimate much too high, see, e.g., Sinn (1999) and Attanasio et al. (2002).

2. Variability of inflation Both theory and data suggest that with higher inflation follows higher volatility of the inflation rate. The implication is larger uncertainty, which is welfare-reducing when people are risk-averse.

3. Yardstick costs Inflation reduces the usefulness of money as a *numeraire* (a yardstick of variable length is not convenient).

4. Tax distortions In many countries taxation is primarily based on *nominal* income. Then fluctuating inflation creates fluctuating effective tax rate on real income. Consider the following example. Let nominal capital income be taxed at a *given* constant rate τ. Since in the long run the real interest rate is approximately a constant, $\bar{r}$, then a higher inflation rate implies a lower after-tax real interest rate as in (18.3) above. That is, under a nominal tax system the after-tax real interest rate tends to fluctuate with the inflation rate (though in the opposite direction). This goes against the principle of tax smoothing. Moreover, if nominal capital gains on some assets are taxed but on other assets (for example owner-occupied houses) not taxed, then distortions in the allocation of investment arise.

5. Menu costs Another problem of inflation is the implied need to change price lists or negotiate wages more often or to adopt indexing schemes. Zero inflation will remove the need for such endeavors.

6. Money illusion We say the market participants suffer from money illusion when they have difficulties distinguishing between nominal changes and real changes in the environment (changes in the purchasing power etc.). To the extent that the market participants have money illusion, inflation implies confusion and distorted decisions are made.

Social benefits of (moderate) inflation

We can identify at least four benefits of inflation. The first-mentioned benefit below is recognized by economists irrespective of their Keynesian or non-Keynesian persuasion, whereas the other three are emphasized by Keynesians.

1. Seigniorage As already noted, government revenue from base money creation − seigniorage − may be a non-negligible source for financing public expenditure. This pertains to some developing countries with weak taxation ability. And historically seigniorage has been important for countries at war. Strictly speaking, seigniorage as such need not involve inflation, but *high* seigniorage induces inflation. The seigniorage S obtained by the (consolidated) public sector by printing money is given in (18.4) above. Assume the money multiplier is constant. So the growth rate of base money, M, equals μ. Further, assume real money demand is given by $(M/P)^d = Y^\eta L(i)$, where η is a positive parameter and i is the nominal interest rate which we assume constant in the long run as long in line with the constant μ. Then, inflation in the long run tends to be

$$\pi \approx \mu - \eta g_Y = S\frac{P}{M} - \eta g_Y, \tag{18.10}$$

according to (18.2) and (18.4). It follows that S *can* be positive without π being positive. But *high* S is associated with $\pi > 0$, i.e., positive inflation.

As we saw in Section 15.3, depending on the money demand function there is likely to be an upper bound on how large seigniorage S can be. This was because, when μ increases, M/P tends to *decrease* and ultimately the increase in μ is more than offset by this.

A benefit of seigniorage is that it may allow lowering of distortionary taxation. And for countries that face difficulties collecting enough tax revenue, seigniorage may be a useful supplementary source of public finance. In practice, for the more developed countries seigniorage is of limited importance. In USA the ratio of the monetary base to nominal GDP is about 6% ($\frac{M}{PY} = 0.06$). A rate of monetary expansion at 4% per year implies a ratio of

seigniorage to GDP at about $\mu\frac{M}{P}/Y = \mu \times \frac{M}{PY} = 0,04 \times 0,06 = 0,24\%$ of GDP.

Seigniorage is related, but not identical, to the so-called inflation tax. The *inflation tax* is the proportionate decrease (per time unit) in the real value of the private sector's holding of base money, caused by inflation. Given the inflation rate π, the inflation tax is $\pi\frac{M}{P}$.[18] Reordering (18.10), we get

$$\pi\frac{M}{P} \approx (\mu - \eta g_Y)\frac{M}{P} = \frac{\dot{M}}{P} - \eta g_Y \frac{M}{P} = S - \eta g_Y \frac{M}{P}.$$

With $\mu = \eta g_Y > 0$, there tends to be no inflation, hence no inflation tax, but seigniorage is positive, since μ is positive. But if $\mu > \eta g_Y$, there will be inflation and the implied inflation tax can be seen as *one* component of seigniorage, hence, a source of finance for the public sector. Criminals often have a large share of cash in their portfolio. Then an additional argument for the desirability of inflation is that the inflation tax is one of the few taxes criminals can not evade.

In contrast to this observation, the next three points are related specifically to business cycle phenomena.

2. Smooth real wage adjustment If there is downward nominal rigidity (which seems to be typical for industrialized countries), then a system with some ongoing inflation is less vulnerable to cost push shocks (say, a negative supply shock, an oil price shock, etc.). The ongoing inflation simply makes the *real* wage more downward flexible than otherwise, and firm's profitability may soon be re-established. Moderate inflation is said to "grease the wheels".

3. Inflation makes a negative real interest rate possible For an economy in recession a negative real interest rate may be needed to stimulate investment and consumption in order that the economy can recover. Because the nominal interest rate, i, cannot become negative, the real interest rate, $r \equiv i - \pi^e$, can be negative only if expected inflation, π^e, is positive. Hence, the option of a negative real interest rate is available only if positive inflation expectations are induced, say by an expansionary monetary policy.[19] In May

[18] From $-\{[d(1/P(t))/dt]/(1/P(t))\}\frac{M(t)}{P(t)} = \left[P^{-2}\dot{P}/(1/P)\right]\frac{M}{P} = (\dot{P}/P)\frac{M}{P}$. Often, the inflation tax is in fact higher than this, because inflation also erodes the real value of outstanding nominally denominated government debt.

[19] The problems in Japan with more than a decade's stagnation since the early 1990s, illustrate this. The proposal by Krugman (1998) and Svensson (2003) was that the monetary authorities in Japan should commit to a large depreciation of the yen and a crawling peg until an announced desired upward-sloping price-level path is reached. At that time

2003 the ECB (the European Central Bank) changed its definition of "price stability" from "below 2%" to "close to 2%". Some economists argue that 3% would be better. The Bank of England and the Reserve Bank of Australia set their inflation target at 2.5%.

4. Reduced risk of deflation An economy with zero inflation faces the risk that an adverse shock can lead to *deflation,* which may fortify and prolong a tendency to recession or depression. The real burden of nominal debt increases and at the same time firms' and house-owners' expected revenues decrease. The deteriorating balance sheets trigger debt defaults and a financial crisis threatens. Both Irving Fisher (1933) and Keynes (1936) feared that these consequences could generate a general slump in economic activity, thereby reinforcing the deflationary spiral.

Deflation may also lead to a too high *real* interest rate, because of the zero lower bound on the nominal interest rate, cf. point 3 above. During the Great Depression, deflation in the US reached 9.2 % in 1931 and 10.8 % in 1932. In Japan there has been deflation for a decade since 1995 (as measured by the GDP deflator), though at a considerably lower level (0-2 % per year). The problem of deteriorating balance sheets also showed up in the crises in East Asian countries in 1997-98, though in a different form. In Thailand, Malaysia, South Korea and Indonesia the values of domestic currencies fell and the real burden of debt denominated in dollars or yen was increased. A speculative run on the currencies by foreign investors lead to further loss of value and bankruptcies resulted.

18.5 Theory of "the level of interest rates"

In the real world there are many different rates of return. What circumstances lie behind these differences? What can macroeconomics say about the general level around which rates of return fluctuate.

In non-monetary models without uncertainty there is in equilibrium only one rate of return, r. Under certain conditions (perfect competition in all markets, the consumption good is physically indistinguishable from the capital good, and there are no capital adjustment costs), assumed in simple neoclassical models (like the Diamond OLG model and the Ramsey model), the equilibrium real interest rate is at any time equal to the current net marginal productivity of capital ($r = \partial Y/\partial K - \delta$, standard notation). Moreover, under conditions ensuring "well-behavedness" of these models, they predict

monetary policy should shift to permanent inflation targeting with an average inflation rate equal to, say, at least 2% per year.

that the technology-corrected capital-labor ratio, and thereby the marginal productivity of capital, adjusts over time to some constant long-run level (on which more below).

Different rates of return

In simple neoclassical models with perfect competition and no uncertainty, the equilibrium short-term real interest rate is at any time equal to the net marginal productivity of capital ($r = \partial Y/\partial K - \delta$). In turn the marginal productivity of capital adjusts over time, via changes in the capital intensity, to some long-run level (on this more below). As we saw in Chapter 14, existence of convex *capital installation costs* loosens the link between r and $\partial Y/\partial K$. The convex adjustment costs create a wedge between the price of investment goods and the market value of the marginal unit of installed capital. Besides the marginal productivity of capital, the possible capital gain in the market value of installed capital as well as the effect of the marginal unit of installed capital on future installation costs enter as co-determinants of the current rate of return on capital.

When *imperfect competition* in the output markets rules, prices are typically set as a mark-up on marginal cost. This implies a wedge between the net marginal productivity of capital and capital costs (see Section 2.5 in Chapter 2). And when *uncertainty* and limited opportunities for risk spreading are added to the model, a wide spectrum of expected rates of return on different financial assets and expected marginal products of capital in different production sectors arise, depending on the risk profiles of the different assets and production sectors. On top of this comes the presence of taxation and this may complicate the picture because of different tax rates on different asset returns.

Table 18.1 reports the nominal and real average annual rates of return on a range of US asset portfolios for the period 1926–2001. The portfolio of small company stocks had an average annual real return of 13.8 per cent (the arithmetic average throughout the period).[20] This is more than that of any of the other considered portfolios. Small company stocks are also seen to be the most volatile. The standard deviation of the annual real rate of return of the portfolio of small company stocks is almost eight times higher than that of the portfolio of U.S. Treasury bills (government zero coupon bonds with 30 days to maturity), with an average annual real return of only 0.8 per cent

[20] In contrast to the arithmetic average, the geometric average takes compound interest into account.

	Arithmetic average	Standard deviation	Geometric average
	Percent		
Nominal rates			
Small Company Stocks	17,3	33,2	12,5
Large Company Stocks	12,7	20,2	10,7
Long-Term Corporate Bonds	6,1	8,6	5,8
Long-Term Government Bonds	5,7	9,4	5,3
Intermediate-Term Government Bonds	5,5	5,7	5,3
U.S. Treasury Bills	3,9	3,2	3,8
Cash	0,0	0,0	0,0
Inflation rate	3,1	4,4	3,1
Real rates			
Small Company Stocks	13,8	32,6	9,2
Large Company Stocks	9,4	20,4	7,4
Long-Term Corporate Bonds	3,1	9,9	2,6
Long-Term Government Bonds	2,7	10,6	2,2
Intermediate-Term Government Bonds	2,5	7,0	2,2
U.S. Treasury Bills	0,8	4,1	0,7
Cash	-2,9	4,2	-3,0

Table 18.1: Average annual rates of return on a range of US asset portfolios, 1926-2001. Source: Stocks, Bonds, Bills, and Inflation: Yearbook 2002, Valuation Edition. Ibbotson Associates, Inc.

throughout the period. Explanation in terms of risk aversion is in line with the displayed positive relation between high returns and high volatility. Yet, interpreting volatility as a rough measure of risk, the pattern is not without exceptions. The portfolio of long-term corporate bonds has performed better than the portfolio of long-term government bonds, although they have been slightly less volatile as here measured. But the data is historical, expectations are not always met, and risk depends significantly on the *correlation* of the asset's return with the business cycle, a feature about which Table 18.1 has nothing say; share prices are in fact very sensitive to business cycle fluctuations.

Nominal and real average annual rates of return on a range of U.S. asset portfolios for the period 1926–2001 are reported in Table 1. By a *portfolio* of n assets, $i = 1, 2, \ldots, n$ is meant a "basket", $(v_1, v_2, \ldots, v_n)$, of the n assets in value terms, that is, $v_i = p_i x_i$ is the value of the investment in asset i, the price of which is denoted p_i and the quantity of which is denoted x_i. The total investment in the basket is $V = \sum_{i=1}^{n} v_i$. If R_i denotes the gross rate of return on asset i, the overall gross rate of return on the portfolio is

$$R = \frac{\sum_i^n v_i R_i}{V} = \sum_{i=1}^{n} w_i R_i,$$

where $w_i \equiv v_i/V$ is the *weight* or *fraction* of asset i in the portfolio. Defining

$R_i \equiv 1+r_i$, where r_i is the net rate of return on asset i, the net rate of return on the portfolio can be written

$$r = R - 1 = \sum_{i=1}^{n} w_i(1+r_i) - 1 = \sum_{i=1}^{n} w_i + \sum_{i=1}^{n} w_i r_i - 1 = \sum_{i=1}^{n} w_i r_i.$$

The net rate of return is often just called "the rate of return".

In Table 1 we see that the portfolio consisting of small company stocks had an average annual real rate of return of 13.8 per cent (the arithmetic average) or 9.2 per cent (the geometric average) throughout the period 1926-2001. This is more than the annual rate of return of any of the other considered portfolios. Small company stocks are also seen to be the most volatile. The standard deviation of the annual real rate of return of the portfolio of small company stocks is almost eight times higher than that of the portfolio of U.S. Treasury bills (government zero coupon bonds with 30 days to maturity), with an average annual real return of only 0.8 per cent (arithmetic average) or 0.7 per cent (geometric average) throughout the period. The displayed positive relation between high returns and high volatility is not without exceptions, however. The portfolio of long-term corporate bonds has performed better than the portfolio of long-term government bonds, although they have been slightly less volatile as here measured. The data is historical and expectations are not always met. Moreover, risk depends significantly on the *covariance* of asset returns within the total set of assets and specifically on the correlation of asset returns with the business cycle, a feature that can not be read off from Table 1. Share prices, for instance, are very sensitive to business cycle fluctuations.

The need for means of payment − money − is a further complicating factor. That is, besides dissimilarities in risk and expected return across different assets, also dissimilarities in their degree of liquidity are important, not least in times of financial crisis. The expected real rate of return on cash holding is minus the expected rate of inflation and is therefore negative in an economy with inflation, cf. the last row in Table 18.1. When agents nevertheless hold cash in their portfolios, it is because the low rate of return is compensated by the *liquidity* services of money. In the Sidrauski model of Chapter 17 this is modeled in a simple way, albeit ad hoc, by including real money holdings directly as an argument in the utility function. Another dimension along which the presence of money interferes with returns is through inflation. Real assets, like physical capital, land, houses, etc. are better protected against fluctuating inflation than are nominally denominated bonds (and money of course).

Without claiming too much we can say that investors facing such a spectrum of rates of return choose a portfolio composition so as to balance the

need for liquidity, the wish for a high expected return, and the wish for low risk. Finance theory teaches us that adjusted for differences in risk and liquidity, asset returns tend to be the same. This raises the question: at what level? This is where macroeconomics − as an empirically oriented theory about the economy as a whole − comes in.

Macroeconomic theory of the "average rate of return"

The point of departure is that market forces by and large tend to anchor the rate of return of an average portfolio of interest-bearing assets to the net marginal productivity of capital in an aggregate production function. Some popular phrases are:

- the net marginal productivity of capital acts as a centre of gravitation for the spectrum of asset returns; and
- movements of the rates of return are in the long run held in check by the net marginal productivity of capital.

Though such phrases seem to convey the right flavour, in themselves they are not very informative. The net marginal productivity of capital is not a given, but an endogenous variable which, via changes in the capital intensity, adjusts through time to more fundamental factors in the economy.

The different macroeconomic models we have studied in previous chapters bring to mind different presumptions about what these fundamental factors are.

1. Solow's growth model The Solow growth model leads to the fundamental differential equation (standard notation)

$$\dot{\tilde{k}}_t = sf(\tilde{k}_t) - (\delta + g + n)\tilde{k}_t,$$

where s is an exogenous and constant aggregate saving-income ratio, $0 < s < 1$. In steady state

$$r^* = f'(\tilde{k}^*) - \delta, \tag{18.11}$$

where $\tilde{k}^*$ is the unique steady state value of the (effective) capital intensity, $\tilde{k}$, satisfying

$$sf(\tilde{k}^*) = (\delta + g + n)\tilde{k}^*. \tag{18.12}$$

In society there is a debate and a concern that changed demography and less growth in the source of new technical ideas, i.e., the stock of educated

human beings, will in the future result in lower n and lower g, respectively, making financing social security more difficult. On the basis of the Solow model we find by implicit differentiation in (18.12) $\partial\tilde{k}^*/\partial n = \partial\tilde{k}^*/\partial g = -\tilde{k}^*\left[\delta + g + n - sf'(\tilde{k}^*)\right]^{-1}$, which is negative since $sf'(\tilde{k}^*) < sf(\tilde{k}^*)/\tilde{k}^* = \delta + g + n$. Hence, by (18.11),

$$\frac{\partial r^*}{\partial n} = \frac{\partial r^*}{\partial g} = \frac{\partial r^*}{\partial \tilde{k}^*}\frac{\partial \tilde{k}^*}{\partial n} = f''(\tilde{k}^*)\frac{-\tilde{k}^*}{\delta + g + n - sf'(\tilde{k}^*)} > 0,$$

since $f''(\tilde{k}^*) < 0$. It follows that

$$n \downarrow \text{ or } g \downarrow \Rightarrow r^* \downarrow . \tag{18.13}$$

A limitation of this theory is of course the exogeneity of the saving-income ratio, which is a key co-determinant of $\tilde{k}^*$, hence of r^*. The next models are examples of different ways of integrating a theory of saving into the story about the long-run rate of return.

2. The Diamond OLG model In the Diamond OLG model, based on a life-cycle theory of saving, we again arive at the formula $r^* = f'(\tilde{k}^*) - \delta$. Like in the Solow model, the long-run rate of return thus depends on the aggregate production function and on $\tilde{k}^*$. But now there is a logically complete theory about how $\tilde{k}^*$ is determined. In the Diamond model $\tilde{k}^*$ depends in a complicated way on the lifetime utility function and the aggregate production function. The steady state of a well-behaved Diamond model will nevertheless have the same qualitative property as indicated in (18.13).

3. The Ramsey model Like the Solow and Diamond models, the Ramsey model implies that $r_t = f'(\tilde{k}_t) - \delta$ for all t. But unlike in the Solow and Diamond models, the net marginal productivity of capital now converges in the long run to a specific value given by the *modified golden rule* formula. In a continuous time framework this formula says:

$$r^* = \rho + \theta g, \tag{18.14}$$

where the new parameter, θ, is the (absolute) elasticity of marginal utility of consumption. Because the Ramsey model is a representative agent model, the Keynes-Ramsey rule holds not only at the individual level, but also at the aggregate level. This is what gives rise to this simple formula for r^*.

Here there is no role for n, only for g. On the other hand, there is an alternative specification of the Ramsey model, namely the "average utilitarianism" specification. In this version of the model, we get $r^* = f'(\tilde{k}^*) - \delta = \rho + n + \theta g$, so that not only a lower g, but also a lower n implies lower r^*.

Also the Sidrauski model, i.e., the monetary Ramsey model of Chapter 17, results in the *modified golden rule* formula.

4. Blanchard's OLG model A continuous time model with OLG structure and emphasis on life-cycle aspects is Blanchard's OLG model (Blanchard 1985). In that model the net marginal productivity of capital adjusts to a value where, in addition to the production function, technology growth, and preference parameters, also demographic parameters, like birth rate, death rate, and retirement rate, play a role. One of the results is that when $\theta = 1$,

$$\rho + g - \lambda < r^* < \rho + g + b,$$

where λ is the retirement rate (reflecting how early in life the "average" person retire from the labor market) and b is the (crude) birth rate. The population growth rate is the difference between the birth rate, b, and the (crude) mortality rate, m, so that $n = b - m$. The qualitative property indicated in (18.13) becomes conditional. It still holds if the fall in n reflects a lower b, but not necessarily if it reflects a higher m.[21]

5. What if technological change is embodied? The models in the list above assume a neoclassical aggregate production function with CRS and *disembodied* Harrod-neutral technological progress, that is,

$$Y_t = F(K_t, T_t L_t) \equiv T_t L_t f(\tilde{k}_t), \qquad f' > 0, f'' < 0. \tag{18.15}$$

This amounts to assuming that new technical knowledge advances the combined productivity of capital and labor *independently* of whether the workers operate old or new machines.

In contrast, we say that technological change is *embodied* if taking advantage of new technical knowledge requires construction of new investment goods. The newest technology is incorporated in the design of newly produced equipment; and this equipment will not participate in subsequent technological progress. Both intuition and empirics suggest that most technological progress is of this form. Indeed, Greenwood et al. (1997) estimate for the U.S. 1950-1990 that embodied technological change explains 60% of the growth in output per man hour.

So a theory of the rate of return should take this into account. Fortunately, this can be done with only minor modifications. We assume that the link between investment and capital accumulation takes the form

$$\dot{K}_t = Q_t I_t - \delta K_t, \tag{18.16}$$

[21] See Section 12.4 of Chapter 12.

where I_t is gross investment ($I = Y - C$) and Q_t measures the "quality" (productivity) of newly produced investment goods. Suppose for instance that

$$Q_t = Q_0 e^{\gamma t}, \qquad \gamma > 0.$$

Then, even if no technological change directly appears in the production function, that is, even if (18.15) is replaced by

$$Y_t = F(K_t, L_t) = K_t^{\alpha} L_t^{1-\alpha}, \qquad 0 < \alpha < 1,$$

the economy will still experience a rising standard of living.[22] A given level of gross investment will give rise to a greater and greater additions to the capital stock K, measured in efficiency units. Since at time t, Q_t capital goods can be produced at the same cost as one consumption good, the price, p_t, of capital goods in terms of the consumption good must in competitive equilibrium equal the inverse of Q_t, that is, $p_t = 1/Q_t$. In this way embodied technological progress results in a steady decline in the relative price of capital equipment.

This prediction is confirmed by the data. Greenwood et al. (1997) find for the U.S. that the relative price of capital equipment has been declining at an average rate of 0.03 per year in the period 1950-1990, a trend that has seemingly been fortified in the wake of the computer revolution.

Along a balanced growth path the constant growth rate of K will now exceed that of Y, and Y/K thus be falling. The output-capital ratio in value terms, $Y/(pK)$, will be constant, however. Embedding these features in a Ramsey-style framework, we find the long-run rate of return to be[23]

$$r^* = \rho + \theta \frac{\alpha\gamma}{1-\alpha}.$$

This is of exactly the same form as (18.14), if we define $g = \alpha\gamma/(1-\alpha)$.

6. Adding uncertainty and risk of bankruptcy Although absent from many simple macroeconomic models, uncertainty and risk of bankruptcy are significant features of reality. Bankruptcy risk may lead to a conflict of interest between share owners and managers. Managers may want less debt and more equity than the share owners because bankruptcy can be very costly to managers who loose a well-paid job and a promising carrier. So managers are unwilling to finance all new capital investment by new debt in spite of

[22] We specify F to be Cobb-Douglas, because otherwise a model with embodied technical progress in the form (18.16) will not be able to generate balanced growth and comply with Kaldor's stylized facts.

[23] See Appendix.

the associated lower capital cost (there is generally a lower rate of return on corporate bonds than on equity). In this way the excess of the rate of return on equity over that on debt, the equity premium, is sustained.

A rough, behavioral theory of the equity premium goes as follows.[24] Firm managers prefer a payout structure with a fraction, s_f, going to equity and the remaining fraction, $1 - s_f$, to debt (corporate bonds). That is, out of each unit of expected operating profit, managers are unwilling to commit more than $1 - s_f$ to bond owners. This is to reduce the risk of a failing payment ability in case of a bad market outcome. And those who finance firms by loans definitely also want debtor firms to have some equity at stake.

We let households' preferred portfolio consist of a fraction s_h in equities and the remainder, $1 - s_h$, in bonds. In view of households' risk aversion and memory of historical stock market crashes, it is plausible to assume that $s_h < s_f$.

As a crude adaptation of for instance the Blanchard OLG model to these features, we interpret the model's r^* as an average rate of return across firms. Let time be discrete and let aggregate financial wealth be $A = pK$, where p is the price of capital equipment in terms of consumption goods. In the frameworks 1 to 4 above we have $p \equiv 1$, but in framework 5 the relative price p equals $1/Q$ and is falling over time. Anyway, given A at time t, the aggregate gross return or payout is $(1 + r^*)A$. Out of this, $(1 + r^*)As_f$ constitutes the gross return to the equity owners and $(1 + r^*)A(1 - s_f)$ the gross return to the bond owners. Let r_e denote the rate of return on equity and r_b the rate of return on bonds.

To find r_e and r_b we have

$$\begin{aligned} (1 + r_e)As_h &= (1 + r^*)As_f, \\ (1 + r_b)A(1 - s_h) &= (1 + r^*)A(1 - s_f). \end{aligned}$$

Thus,

$$\begin{aligned} 1 + r_e &= (1 + r^*)\frac{s_f}{s_h} > 1 + r^*, \\ 1 + r_b &= (1 + r^*)\frac{1 - s_f}{1 - s_h} < 1 + r^*. \end{aligned}$$

We may define the *equity premium,* π, by $1 + \pi \equiv (1 + r_e)/(1 + r_b)$. Then

$$\pi = \frac{s_f(1 - sh)}{s_h(1 - s_f)} - 1 > 0.$$

[24] This draws on Baker et al. (2005).

Of course these formulas have their limitations. The key variables s_f and s_h will depend on a lot of economic circumstances and should be endogenous in an elaborate model. Yet, the formulas may be helpful as a way of organizing one's thoughts about rates of return in a world with asymmetric information and risk of bankruptcy.

There is evidence that in the last decades of the twentieth century the equity premium had become lower than in the long aftermath of the Great Depression in the 1930s.[25] A likely explanation is that s_h had gone up, along with rising confidence. The computer and the World Wide Web have made it much easier for individuals to invests in stocks of shares. On the other hand, the recent financial and economic crisis, known as the Great Recession 2007- , and the associated rise in mistrust seems to have halted and possibly reversed this tendency for some time (source??).

18.6 Literature notes

1. The discussion in Section 18.2 about money in an OLG framework refers to OLG models where money is demanded because of its role as a means of payment. In Paul Samuelson's original OLG model with money, however, the demand for money relies on money being the *only* asset in the model (Samuelson 1958). As inflation affects the rate of return on holding money (negatively), inflation affects thereby the rate at which agents can transfer value over time. Hence, saving and wealth formation is affected by money growth − money is not superneutral. This is not a convincing theory, however, since in the real world money is, as a store of value, clearly dominated by other assets paying a higher rate of return.

2. The account of macroeconomic theories of the rate of return in the last part of Section 15.4.3 is inspired by Baker, DeLong, and Krugman (2005). These authors go more into detail with the implied predictions for U.S. rates of return in the future and with implications for the debate on social security reform.

18.7 Appendix

(no text yet available)

18.8 Exercises

[25] Blanchard (2003, p. 333).