# MACROECONOMICS

## Incomplete draft: please do not distribute outside of class use

*Marina Azzimonti, Per Krusell,*
*Alisdair McKay, and Toshihiko Mukoyama*
with
*Timo Boppart, Giancarlo Corsetti, Luca Dedola, John Hassler,*
*Juan Carlos Hatchondo, Jonathan Heathcote, Andreas Hornstein,*
*Pete Klenow, Simon Lloyd, Leo Martinez, Kurt Mitman, Conny Olovsson,*
*Monika Piazzesi,Vincenzo Quadrini, Morten Ravn, Richard Rogerson,*
*Víctor Ríos-Rull, Ayşegül Şahin, Martin Schneider,*
*Kjetil Storesletten, and Gianluca Violante*

March 2024

# Contents

# Part I

# Introduction

# Preface

This document is a comprehensive textbook for first-year Ph.D. courses in macroeconomics. It started as a pandemic project, at a time when the discovery of how well zoom works allowed the authors to (re-)connect and have frequent online meetings. The present preface text is a first draft, prepared for the online launch in January of 2023. Thus, at the moment of writing, not all chapters have been finished but most of them are done or are near completion and will be added to the online site within not too long. We are very happy that we have now come this far and produced a text we are proud of. As a note of (minor) caution, though the chapters we post now are complete, they will be subject to continuous updates in response to feedback we receive.

The book contains the modeling and analytical tools of macroeconomic analysis and applies them to the main topics within the field of macroeconomics. A key feature, compared to all other textbooks in macro, is that it makes explicit connection with data throughout the book, and the macro theory is then used to interpret the data. It also emphasizes both strengths and weaknesses of our understanding of the field, making clear that research is still ongoing.

The text has the rigor of a Ph.D. course but is presented in an accessible way, making it suitable to a wide range of students, including those in master's courses. The foundational chapters have been written by the core authors (Azzimonti, Krusell, McKay, and Mukoyama). The topics chapters are written by leading subject matter experts in consultation, coordination, and in several cases coauthorships with the core authors. The book thus brings together a set of researchers who have been selected for their communication skills as well as their world-class knowledge of the field. We are immensely grateful to them.

The complete text will contain 23 chapters, each of about 30 pages. From a teaching perspective, the idea is for the text to work as a textbook where the chapters can be read in order. Clearly, though, there is more material than can be covered in a first-year course. However, since we hope that the book will be used in a variety of contexts and places around the world, we have included chapters that would not ordinarily be considered first-year material. For example, we have two chapters on international macro, one on sustainability and climate change, and one on inequality. Appendix material will be provided online, as well as computer software for solving macro models numerically.

# Chapter 1

# The subject matter

This chapter has two main goals. One is to introduce some of the central questions that macroeconomists are interested in. The second is to emphasize a central theme in the text: how our field very fundamentally is an empirical one and that measurement—the ongoing construction and improvement of data sets—and theories addressing the resulting data are intertwined and have evolved in tandem. The chapter also serves to present and motivate the methods, empirical as well as theoretical, that we use throughout the text.

There are several ways to define the field of macroeconomics, and there is a point to all of them, but they each also come with caveats. One definition is that the field is the study of aggregates. Here an important caveat is that macroeconomics, at least nowadays, is greatly concerned with the distribution of income, the distribution of wealth, and so on, which are not traditionally thought of as aggregates. A second definition is based on methods: macroeconomics is the quantitative study of general equilibrium. Then again, a full general-equilibrium analysis is often not necessary for analyzing many of our core issues. One example illustrating this statement is that most national economies, and even the U.S., are very dependent on the global economy, thus always making the study of a single economy partial to some extent. In addition, a more general point is that, very often, the key component of general-equilibrium analysis turns out to be the characterization of behavior—of consumers and firms—given prices, with the market-clearing mechanism playing a subordinate role. A third definition also emphasizes methods: macroeconomics is the study of the dynamics of the economy, for example by its emphasis on investment and on the role of expectations. However, as we will illustrate in this text, often a static analysis can shed major light on macroeconomic issues. Less of a definition but of equal importance is the fact that macroeconomics tracks current events: the field tends to follow what is perceived as the major issues for our economies at any point in time, whatever they might be. For example, the Great Recession 2007–2009 appeared to have its roots in problems in the housing and financial markets, and since then these "topics" have become central to the field and drawn in large numbers of researchers, ranging from PhD students interested in macroeconomic issues to leading macroeconomists. Recently, because of COVID-19, many macroeconomists have been busy studying the intersection of macroeconomics and epidemiology.

The list of definitions of the field is not exhaustive and we think it is useful to use all these perspectives, in addition to the fact that a driver behind macroeconomic research is not only to understand but also to fix problems, if possible, through government policy. For now, the present chapter will be organized around the last point above: the idea is to take you through some of important events in the development of our field. This is by no means meant as an attempt to write a doctrine history; rather, the main, and really only, purpose is to illustrate the continuous development of measurement and theory to, precisely, respond to the major needs of our times: our macro-needs. Also, we hope that the discussion will illustrate how the focus of macroeconomics keeps moving and touching base with—and, most of all, borrowing insights from—various other subfields of economics, such as labor economics, finance, microeconomic theory, and public economics.

## 1.1 A random walk along our macroeconomic history

We will now highlight some of the key questions in macroeconomics, along with the need for measurement and theory, and bring them up in the context of some real-world events. The focus, especially when it comes to measurement and where data is collected, is on the U.S.; for what it is worth, there is no deep motivation behind this choice, other than a practical one: most of macroeconomic frontier research has addressed U.S. data. At the same time, it is important to recognize that macroeconomic analysis typically does need to be adapted to the country under study—institutions differ, the role of foreign trade differs, the availability of data differs, and so on. In particular, perhaps, countries further from the development frontier face different macroeconomic problems and we will only touch on these later in Chapters 11 and 21. Generally, however, we very much encourage you to make comparisons to other countries—both in terms of events and measurement—as you go through the text.

### 1.1.1 The Great Depression: what is going on?

Circa one hundred years ago, there were macroeconomists—in the sense of economists who had thought long and carefully about the performance of the economy as a whole—but these macroeconomists had virtually no systematic data to relate to; there were only scattered accounts of some production figures and some prices. The birth of national income and product accounting was, in fact, precisely a response to the need for systematic measures of just how badly the economy was doing in and around the Great Depression. The emerging systematic measures allowed macroeconomists to at least obtain partial answers to the simple question **"What is going on?"**. The newly available data therefore helped John Maynard Keynes, and many others following in his footsteps, along in their analysis of the macroeconomy and of how there might be ways to use government policy actively to combat recessions. In contrast to the thinking at the time, Keynes emphasized market imperfections, in particular the sluggish adjustment of wages and prices to changes in market conditions, leaving room for a decline in the demand for consumption and investment to affect production. Simple theories connecting our now-observable main aggregates—output, consumption, and investment,

private and public—were thus constructed and they are influential still today.

Constructing our aggregate data is not an easy task, however, and many of these early measurement efforts were major research achievements. Let us now therefore briefly describe some of the results of this work.

### Some conceptual issues

National income accounts measure some central macroeconomic aggregates and although their main purpose is to measure "what's *really* going on," the main variables of course come in nominal, not real, terms. Nominal terms refer to (current) dollar values, which are obtained for any transaction since dollars, in the U.S., are a universal medium of exchange and unit of account.[1] Thus, comparisons across years is not immediate since there may be general inflation (or deflation). How are real terms obtained, so that comparisons can be made across years?

So suppose the goal is to find how real aggregate consumption has grown over time and that we have all the relevant data (prices and quantities for each year). The idea is then to construct a price index for consumption, a CPI (consumer price index), such that real consumption in each year equals that year's total nominal consumption expenditures divided by the price index for that year. The price index, in turn, is constructed from a weighted average of price changes, with the weight on each item based on the expenditure share on that item.

**More about index construction**

The details generally matter here: the weighted average could for example be arithmetic or geometric, and the expenditure shares could be taken from a base year or could be evolving over time.[a] The choice of these details are often based on microeconomic theory. For example, if the consumer's utility is based on an aggregate where goods have a constant elasticity of substitution between them (a CES function), then utility maximization delivers a closed-form expression for a weighted average of prices such that, for a given value of total expenditures, relative price changes that do not alter the weighted average will not affect the utility level. This closed form, which we will derive and use in the context of some models of consumers' "taste for variety" in Chapter 2, has implicit in it the optimally chosen shares for each good and gives direct guidance as to the details of the index construction. A number of countries have adopted so-called superlative indexes (e.g., the Fisher and Törnquist indexes) that allow for some generality in the features of the utility function, such as non-constant substitution elasticities

[1] Given the prices of all goods, it would also be possible to denote, say GDP, in units of pencils, but it would then be unclear what kind of pencil is considered, and many people don't even use pencils and therefore have a hard time relating to their value. Also, prices of individual goods go up and down in ways that would cause GDP to fluctuate in ways that would seem arbitrary.

between goods.

[a]E.g., they could be straight averages of the current and last year's shares, as in the Törnqvist index.

Different indexes are used for different aggregates; PPI (the producer price index) and the GDP deflator (which is defined by nominal divided by real output, where "real" is a quantity index) are examples. Indexes are also, depending on how they are constructed, subject to a variety of biases. They include substitution bias (the index does not fully take into account how consumers change their purchasing behavior due to relative price changes), new product bias (new products appear—and others disappear from year to year), and quality bias (where the nature of a product changes in quality over time). Also, if different consumers have different utility functions, or if utility functions depend in fundamental ways on the consumer's wealth (so-called non-homothetic functions), there may not be an obvious, perfect index; similarly, certain changes in consumers' utility functions over time can be difficult to take into account.

As we will discuss below, especially in our chapter studying economic growth, GDP is not a welfare measure in general; rather, it is a measure of how much the market economy is producing, thus, not taking many relevant sources of utility—that are not market-produced goods and services—into account.[2] At the same time, markets are very useful for measurement: the fact that goods and services are transacted on markets allows us to measure their value in interpretable units. When a good or service is not transacted in a market, it is therefore potentially very hard to measure its value. Think of trying to value a loaf of bread when it is not transacted in a market and the inputs that are used for producing the bread do not have prices either. Government goods and services usually have no market prices, but their costs are typically recorded in dollars; hence, they are often valued at cost. Leisure, along with many environmental amenities but also, of course, illegal transactions, are not counted; in fact, a positive side effect of privatization and of legalizing illegal activities is that we obtain dollar measures of how people value them.

### National income and product accounts

The initiatives toward constructing systematic national accounts were taken in the 1920s and 1930s; Colin Clark and Simon Kuznets were central figures in these efforts. Richard Stone made important contributions beginning during World War II and the first formal U.S. national accounts appeared in 1947, shortly after which many European countries followed suit.

One feature of the accounts is that they follow double-entry methods: they measure both expenditures and income. That is, someone's expenditure is always someone else's income, thus allowing double-checking. GDP can thus be seen either as the sum of all final expenditures, private and public consumption and investment (plus net exports), or as the sum of all incomes. It can also be described as the sum of all values added by all firms. This

[2]Under certain conditions, GDP can be sufficient indicator of welfare, but these conditions are very restrictive.

means that a firm's expenditures on intermediate inputs are not included. The reason for not including intermediates is obvious: if one firm produces a fully equipped car, except for its exterior paint, and the second firm buys the car, paints it, and sells it, a sum of these two transactions would be close to twice the final value of the car and, thus, not be a good measure of the economy's car output. Finally, GDP is also equal to the market value of all total production.

The two broad categories for income are labor income and capital income: wages and salaries and profits, interests, and dividends, respectively. Note here that profits for firms that, say, extract and sell oil, are counted in full; the resulting decline in the stock of natural resources is not deducted. For this reason, GDP is sometimes reported with a deduction for "natural-resource rents."

Capital (equipment or structures) is an intermediate good in some sense—it is bought as an investment good by a firm and used as an input in future production—but expenditures on capital are treated as final expenditures and thus counted in GDP. This observation actually suggests that when GDP is measured for longer time periods, it should perhaps treat capital differently. Capital income should, then, be a smaller share of total income, and GDP would be correspondingly smaller since it would count some of the investments—at least those that depreciate fully within the now longer period—as intermediate goods instead.[3] With this perspective, over longer time periods, GDP averages should perhaps receive less focus than average consumption.

*Gross output*, in contrast to value added, in any given firm is the value of its total production, not deducting the values of inputs purchased elsewhere; gross output in an industry should, conceptually, be a sum across firms in the industry, deducting the values of intermediate inputs bought from within the industry.[4]

### Who does what?

The Bureau of Economic Analysis (BEA) is in charge of the national income and product accounts (NIPA). They collect annual data from all firms selling products and services and these are then categorized into expenditure groups.[5] GDP from the expenditure side is thus written as $C + I + G + NX$ in textbooks, where $C$ denotes aggregate consumption, $I$ aggregate investment, $G$ government spending, and $NX$ net exports (e.g., the difference between exports and imports). Here, $G$ and $NX$ both contain consumption as well as investment goods. $NX$ is based on data collected by the U.S. Customs.

[3] Conceptually, it should thus be the depreciation rate of capital goods that is key and that, at any frequency of sampling, determines whether it is to be regarded as capital or an intermediate good. In the 1990s, the national accounts started treating "software" as investment, as opposed to an intermediate good; the line between these two categories is sometimes quite thin.

[4] Hence, gross output should fall as the aggregation is broader and broader. The deduction of intermediates from within the industry are not necessarily made in practice.

[5] Some very small firms are excluded in the annual data but are included in censuses, conducted by the Census Bureau, every five years. Hence, the BEA needs to estimate the missing data in any year and the censuses then lead to ex-post GDP revisions. The Census also provides input-output data.

Price collection, which is crucial for the construction of real measures, is entirely based on surveys. They are both conducted by the BEA and by the Bureau of Labor Statistics (BLS), which produces a number of price indexes such as the CPI and the PPI. Here, the BLS trades off the costs of collecting more observations against the benefits of obtaining a more informative sample. Note that, as a result of how data is collected, nominal GDP is (almost) covering the universe of activity, whereas real GDP is survey-based—because prices are. Another aspect of real comparisons is that across space: There are also, for example, price indexes constructed for different metropolitan areas.

Consumption measures in NIPA are sometimes compared with numbers from the Consumer Expenditure Survey, CEX, conducted by the BLS. The CEX reports what a sample of individuals consume in great detail based on monthly interviews and diaries individuals are asked to fill in. The CEX is thus of great value for studying inequality in consumer well-being.

Measures of the capital stock are also provided by the BEA and based on the perpetual inventory method. That is, rather than measuring the value of all (or a sample of) capital in use at point in time, one bases the amount/value of each type of capital on how much investment of this type was made in the past, taking into account how capital depreciates. Depreciation rates differ greatly between, say, structures (with low depreciation rates) and equipment (with high depreciation rates). Note also that depreciation is endogenous to an important extent; for example, an old computer may still work but is retired because new computers are much more powerful. Residential structures is a special category: they are counted in the capital stock when produced; for rental units the income accruing directly to the owner is capital income and directly included in GDP, wheras for owner-occupied housing the contribution to GDP is imputed, i.e., estimated based on market rents for similar units.

#### The pandemic: what is going on?

A nice parallel to what occurred during the Great Depression—another **what's going on** question—is sitting right in front of us. During the writing of this book, a new need for data has emerged: high-frequency observations on various activities in our economies—day to day, or at least week to week. The reason, of course, is the pandemic, and a need to track not only economic variables but also social interactions. The internet has helped; it is now possible to gather massive amounts of transactions' data very quickly, and such data has already been extremely useful in identifying and comparing different proposed mechanisms regarding the interaction between the pandemic and the economy.[6] These are new developments that will not be discussed at length in this book.

### 1.1.2 Keeping track of long-run growth

By the mid-1950s, the economy was more or less back on its track and now there was a sufficient amount of aggregate data that it was possible to analyze its growth performance.

[6]Similarly, data on cell phone use by GPS location has become very valuable.

Robert Solow's 1956 paper—the basic neoclassical growth model—and his 1957 paper on growth accounting became the impetus for a burgeoning literature on economic growth that was both empirical and theoretical. The idea in Solow's growth-accounting paper is that one could use measures of output and inputs, along with the prices of inputs, and a basic theory of production to break down aggregate growth into the contributions of each input and, finally, a residual, which could be thought of as technical change: the *Solow residual*. It was thus due to the systematic measurement of quantities and prices that this kind of analysis could now be performed.

**Which factors, then, accounted for most of U.S. growth?** Solow found technical change to be of great, and direct, importance. However, in accordance with his 1956 theory, capital accumulation was an indirect result of technological change. This theory, thus, went beyond accounting and concluded that technological progress is the one (and only) fundamental reason why growth in output keeps going, and going, and going.

One reaction to the empirical finding was that whereas it was plausible that technological change is key to growth, its importance may have been overstated; after all, it was measured as a residual, and if input growth rates were underestimated, the role for technological change would not be as large. A particularly likely reason for this was that workers' skills were improving; there was an ongoing trend of increased schooling and work experience—"on-the-job learning"—was thought to contribute to worker productivity as well. In 1958, labor economist Jacob Mincer wrote his influential paper relating individual wages to years of schooling and experience and found very strong regularities in the data that have been imported and used in constructing better measures of labor input: "human capital." The so-called Mincer equation, which can be derived based on a simple opportunity cost-based theory, says in its estimated form that one more year of schooling adds a little below 10 percent to your wage. Another important measurement development in the growth-accounting literature was the notion of a firm's user cost of capital; firms buy capital and often own it until it gets scrapped, so how should one measure the "year-by-year price" of this input? Hall and Jorgenson (1967) developed an answer that was consistent with microeconomic theory and has been used ever since.

The growth-accounting literature developed virtually into an industry, where productivity performance was computed and accounted for on a disaggregated level. A natural accompanying project was to construct similar data series across countries and compare them. In 1978, Kravis, Summers, and Heston published a paper with comparable data series for 100 countries, an effort that was later continued and today takes the form of the Penn World Tables, a crucial data source for students of economic growth. In a related effort, Maddison (1995) used various sources of data to estimate GDP levels for a range of countries going back into the early nineteenth century. Harmonizing data across countries is challenging, and comparing real output too: should nominal outputs be compared in real terms by use of nominal exchange rates? Because the purchasing power of different monies vary by country, as it does within countries too, a so-called PPP adjustment gradually arose as a new standard; we discuss it and its implications in the growth chapter.

As a result of the multi-country data sets, new light could now be shed on the process of

growth; in particular, **what made some countries grow so fast and others stagnate?**. The endogenous-growth literature of the late 1980s and 1990s thus asked these questions, which were partly phrased as challenges to the Solow model. The endogenous nature of technological change—in particular how it is driven by incentives to innovate—as well as of human capital accumulation came in focus, first theoretically, and later—in order to distinguish and compare different hypotheses—on measurement. Today, for example, we have access to large patent data sets that are currently under the magnifying glasses of hordes of researchers. So questions such as **Who becomes an inventor?** now occupy many macroeconomists. Here, data sets on individuals and their characteristics enter in focus. We briefly comment on those below.

An aspect of economic growth is **structural change**. Structural change typically refers to how some sectors shrink over time and others grow; roughly speaking, a typical path is one where countries gradually build their income—and "develop"—starting with agriculture as a dominant activity, gradually then moving into manufacturing, and finally growing the service sector. Today, agriculture only employs a percent or so of the total workforce in the U.S., whereas in the poorest countries in the world the number is 80 percent; manufacturing has furthermore been overtaken by services. Today, many macroeconomists worry about another expression of structural change: how information technology (IT) changes the workplace. In particular, **what is the role of robotization?**, for macroeconomic performance, for inequality, and for competition across firms.

Another element of structural change is women's labor-market participation, which has risen steadily and significantly in the U.S. over the whole postwar period and stands at a very high level today (not quite as high as that for men, but close). In contrast, it is much lower in many other developed countries, though it is also high in a number of countries at a lower level of development. The government sector—its size and role—can also be depicted as part of a process of structural change, as international trade can be, both for a given country and for the global economy. An important ambition of macroeconomists studying medium- to long-run issues is thus to analyze the sources of these changes as well as their effects. An element they have in common is that they are slow-moving and never causes of immediate media attention, but nevertheless crucial for our economic welfare.

In sum, research on economic growth appears to have come in waves, where theory and measurement interact in very central and mutually reinforcing ways.

### 1.1.3 The 1970s: an oops, with stagflation, high unemployment, and more

The period up to the early 1970s was generally perceived as one of steady growth and increased prosperity. As for macroeconomic policy, the Keynesian recipes were adopted in most countries in the form of regular interventions to stabilize the economy; the term *fine tuning* was often used. Then came an "oops": the sharp recession in 1973, along with a number of severe macroeconomic problems that ended up being quite persistent. Although the cause-and-effect question is still debated, many interpret the events as a result of the oil-

price hike orchestrated by OPEC in October of 1973, and the challenging era that followed was not specific to the U.S. but shared by most of the western world. Two primary difficulties involved lackluster GDP performance—along with a slowdown in productivity that at the time appeared permanent to many (the "productivity slowdown" period)—and sharply rising inequality. The effects on inequality in the labor market had different expressions in different countries: in the U.S. and the U.K., wage inequality rose sharply, while in many other EU countries unemployment rose to very high levels and stayed high for many years.

Interestingly, a new development in theory, that would turn out to have heavy impact on macroeconomics, occurred in the early 1970s, before the drastic downturns in aggregate activity occurred: the development of search models in labor economics (McCall, 1970, and Mortensen, 1972). This was to become one of the cornerstones of a theory of unemployment that later was adopted in macroeconomics and that is a core chapter of this text. The measurement of the concept of being "unemployed" goes back in time much further, to the late 1930s; the new search theory in the 1970s thus benefited immediately from available data. However, the interest of macroeconomists in the topic rose sharply during the productivity slowdown period and generated further data needs. Today, a central part of our analyses of unemployment includes detailed data both on individuals and on firms, most of it in survey form. Moreover, keeping track of inequality trends, such as the average wage gap between skilled and unskilled labor that took off beginning in the second half of the 1970s, has become a central activity for macroeconomists.

**Labor-market data**

Labor-market data is virtually all survey-based. Total labor income is available from NIPA but how it breaks down into employment, hours worked, and wages/salaries for different workers is all based on how individuals answer questions in questionnaires. A key source is the Current Population Survey, conducted monthly by the BLS in collaboration with the Census Bureau. The CPS has a (limited) panel feature, i.e., it interviews the same people at more than one point in time. The CPS also measures unemployment, i.e., it asks people if they are not working and looking actively for a job, thus in line with search theory. An important source of data on individual labor-market outcomes is the Panel Study of Income Dynamics (PSID), conducted by the University of Michigan, started in 1968: it follows individuals over a significant amount of time and has data on a number of individual variables, all self-reported. The breakdown of labor earnings into hours worked and a wage per hour in these data sets is based on individual reporting on how many hours they work. Other panel data sets on individuals include the National Longitudinal Survey (NLS, conducted by the BLS) and Survey of Income and Program Participation (SIPP, conducted by the Census Bureau).[a] Firms are also surveyed and report work hours for their employees (e.g., in the Annual Survey of Manufactures from the Census Bureau), but then the same workers may have multiple jobs so measures of how much an individual works in total still rely on asking the individual. Around the turn of the millenium, and in part as a result of the new

developments in the field of macro labor, where search frictions for firms and workers are in focus, BLS also started to collect micro data, published from 2002 and on: the Job Openings and Labor Turnover Survey (JOLTS). This data set has since become invaluable in the evaluation of further developments of our theories.[b]

How individuals spend their total amount of time is also measured. Since 2003, the BLS has produced the American Time Use Study (ATUS), a survey documenting daily activities in great detail, including how much time is leisure versus various forms of "work at home." Using ATUS, it is also possible to find out not only whether people searched for jobs, but how much time they spent on this activity.

[a] A longitudinal survey is a panel, i.e., a study that follows the same individuals over time.

[b] Today, data from labor markets—e.g., recent trends, particular skills in demand—are also provided commercially, often with real-time information from the internet, and used by human resource departments and head-hunting agencies.

Governments attempted to stabilize the fall in GDP, in particular with expansionary monetary policy, i.e., by cutting interest rates. However, there was very limited success; rather, the 1970s was also a period of unusually high inflation (with annual rates in the 10–20 percent range in many countries, and even higher for some). The combination of stagnation and inflation was dubbed *stagflation*. The Keynesian paradigm came under increasing scrutiny and a number of economists focused on weaknesses in the Keynesian theory itself. A particularly powerful point was the "Lucas critique," which explained how reduced-form relationships between aggregates—a cornerstone in the applied Keynesian apparatus—could break down if policy changed. The Phillips curve—the negative relationship between the inflation rate and the unemployment rate—was a particular case in point: in a paper that actually predated the oil crisis and stagflation, Lucas (1972) advanced a theory showing how attempts to exploit this seeming trade-off with monetary policy would make the relationship itself break down. When the relationship did break down, Lucas's sharp critique gained added force and, in hindsight, marked a clear break in the development of macroeconomics.

Curiously, today—as of writing this text—we are experiencing events that are reminiscent of those during the 1970s: a sharp rise in inflation, in part due to higher energy prices (now due to the Ukraine war and other world events), and fears of a new stagflation period. Undoubtedly, macroeconomists are better equipped today in confronting this episode, but every challenge seems to have its unique properties and we are still far from business as usual.

### 1.1.4 Kydland and Prescott: a way forward

Lucas's critique was not just destructive, in pointing to weaknesses in the Keynesian theory approach, but he suggested, at least conceptually, how an alternative framework could be built up. The idea was to build explicitly on microeconomic theory, and with Kydland and Prescott's 1982 paper it became clear just how to do this in a way that also allowed systematic comparison with data: they offered *quantitative theory*. This paper suggested basing the microeconomics on empirical studies in applied fields, such as labor economics

and consumption studies, not just in terms of the structure but also by importing parameter values from the empirical microeconomic literatures. Kydland and Prescott's paper, which led to an explosion in macroeconomic studies, also added another important aspect of measurement: that often, data needs to be detrended, or "filtered," in order to be ready for analysis, in their case in examining the sources of business cycles.

**Filtering**

When macroeconomic models are built and compared to data, the data are almost always filtered first. To understand what filtering means, you need to see macroeconomic models as dynamic systems, i.e., as some form of vector difference equations, that contain random variables. Thus, macroeconomic models in fact define a *stochastic process*. Such a process could thus be simulated by the researcher and, in principle, compared to data. However, the idea is rarely that the theory is constructed to explain everything. Kydland and Prescott, for example, were interested in recessions and booms, which are movements upward and downward in macroeconomic aggregates around some overall trend, but this trend was not the subject of their study. In order to compare theory and data, most researchers therefore use filters to extract the aspect of the data they are interested in analyzing. To do this, theory comes in handy: stochastic processes can, quite generally, be thought of as sums of sub-processes, each one with a different frequency, i.e., periodicity. For business cycles, one thinks of periodicities of between 5 and 10 years perhaps, and so-called band-pass filters offer can be used to take any data series and remove any frequency outside a specified range. In contrast, studies of medium- to long-term movements in variables require removing high frequencies and retaining low to medium frequencies.[a] In financial economics, when day-to-day, or minute-to-minute changes in stock prices are analyzed, all but the very high frequencies are removed before the data can be analyzed.

[a] Kydland and Prescott (1982) used a specific filter: the so-called Hodrick-Prescott filter, which is an intuitive way of extracting data. It has very wide-spread use in macroeconomics.

Kydland and Prescott's theory of the business cycle was quite stylized and stripped down—among other things, it was phrased entirely in real terms and had no role for monetary policy—and the literature that followed enriched their framework in a multitude of directions. A key point here is that the first wave of models had perfectly working markets; later, a number of frictions were added and today, virtually no macroeconomic model that is used in practice is free of market imperfections. A key friction that was added was price stickiness: firms setting dollar prices of products face costs in doing so, and therefore only adjust prices infrequently. This makes monetary policy have direct effects on the economy, something it might otherwise not have. Thus, the "New Keynesian" framework was built up, where monetary policy is in focus. Again, the new theory led to measurement efforts. In particular, studies such as Bils and Klenow (2004) looked at the survey data available underlying the CPI and recorded the frequency of price adjustments; their work allowed researchers to

parameterize the microeconomic structure for adjustment costs assumed in the models.

### 1.1.5 Different waves of macroeconometrics

The comparison between models and data has also undergone waves. As with much of economics, it is challenging to discern causal relationships in the historical data given to us. As more and more data have become available, however, more and more thinking has been devoted to the development of different statistical methods for doing this. A central question has thus been the purely methodological one of how historical macroeconomic data can be used to make inferences.

Macroeconomic models are all simplifications of a highly complex system and therefore there is little point in "testing a model" by assessing whether it can be the true data generating process. As the saying goes, all models are wrong but some can be useful. A useful model allows us to answer an important question in a convincing way and the model must be consistent with the relevant data we can observe in order to be convincing. Of course, if the answer to our question can be directly observed then there is no point using a model. So a useful model allows us to bridge the gap from the data we observe to the questions we want to answer. Knowing which data are important to match in order for the answer to be convincing is often argued to be an art. But art, too, can be taught.

The empirical implementation of Keynesian theory involved estimating large systems of (usually linear) relationships, often with ad-hoc specifications of short-run dynamics, i.e., with lags of variables added so as to provide a better fit. Sometimes instrumental variables were used, but that was more uncommon. The critique that came in the 1970s forced macroeconomists back to the drawing board. One approach was to estimate the new, now microeconomics-based, structural models that rapidly developed using a classical statistical methods. A literature using maximum-likelihood and Bayesian techniques for estimation was developed; a related development involved use of the generalized method of moments, which could be applied to a subset of the model's equations.

Another theory-based path, labeled calibration, was the method favored by Kydland and Prescott in their work. The calibration approach is very common in current macroeconomic research and can be used to derive quantitative conclusions from a theoretical model. For example, in their work on business cycles, Kydland and Prescott wanted to know to what extent movements in technology could generate fluctuations in aggregates that resemble those in the data. The spirit of calibration is to select the model's parameter values based on other moments of the data than those in focus in the study. For example, Kydland and Prescott based their parameter choices on two kinds of data: (i) micro data, e.g., for people's attitudes toward risk and intertemporal substitution; and (ii) long-run facts (that is, low-frequency data not in focus for their high-frequency interests). Once the model parameters have been selected, one can then derive the model's predictions for the moments of interest and compare them to the moments observed in the data. For example, Kydland and Prescott calculated the variances and correlations of aggregate variables to assess whether the model could generate business cycles that resembled those observed in the data. The sentiment of "all models are wrong" may explain the wide use of calibration within macroeconomics. Cal-

ibration does offer discipline in the sense described above—parameters are not to be chosen to match the moments the researcher wants to explain—but, at the same time, does not lend itself to hypothesis testing.[7] Relatedly, as all models are wrong we are more interested in the broad patterns they predict: can the model at all account for the phenomenon under study, or are the magnitudes severely off? If the model can generate patterns similar to those in the data, it is typically judged "potentially useful" and elaborated on further, possibly adding detail and examining auxiliary implications. This is how "technology shocks" entered our vocabulary and are, still today, considered relevant for (but far from alone in) explaining business cycles.

A much less structural approach was proposed in Sims (1980): vector autoregressions (VARs). Sims's focus was much more on the identification of causal effects in aggregate data, and the core of his methods involved ways to observe plausibly exogenous shocks, such as an unexpected increase in the Fed funds rate, and then trace out the effects of the shocks on macroeconomic variables, including the effects on the subsequent movements in the rate. In its simplest form, a VAR is a linear system of variables including lags, thus containing both intra- and intertemporal relationships, and a shock to each variable at each point in time. A literature also evolved that took structural models and derived their linearized VAR approximations, which could then be compared to the estimated VARs as well as offer some structural interpretations of some the coefficients in the VAR. VAR analysis is a very common tool in macroeconomics.

Another approach to identifying causal effects is to use natural experiments. Increasingly, macroeconomists use information from natural experiments to identify the strength of causal relationships at the microeconomic level and use these moments as targets when calibrating a macroeconomic model. An important example is measuring the marginal propensity to consume out of wealth. In 2001, the U.S. government paid tax rebates to most households and randomly gave some households their payments sooner than others. Johnson, Parker, and Souleles (2005) used the random timing of the payments to measure how strongly consumption spending responds to additional income. This type of information is now an important calibration target for many macroeconomic studies as the marginal propensity to consume is important for understanding the effects of certain government policies. In this case, the natural experiment occurs at the level of the household as some households are paid earlier than others. Natural experiments at the level of an entire economy are more rare but there are some examples. One type of application is exemplified by Acemoglu, Johnson, Robinson (2001), who used excerpts of historical records from the colonial era to make causal statements about the effects of institutions on long-run economic growth and well-being. Another example is the "narrative" approach to the evaluation of monetary policy, where minutes from Federal Reserve meetings are analyzed to identify exogenous events affecting Federal Reserve policy (Romer and Romer, 1989). The Romer and Romer study is an example of another phenomenon, which is to use text analysis (e.g., words used in media) in

[7]The lack of hypothesis testing is shared with Bayesian analysis. Of course, in Bayesian analysis, the parameter selection is informed by the data under study; in calibration it is not: all the weight is on the prior.

macroeconomic contexts; the wave of big-data tools has thus also entered our field.

### 1.1.6 Models: intuition vs. quantitative use

Kydland and Prescott's work was a game changer also in how macroeconomists approach model building, which previously had been largely oriented toward building an intuitive understanding of mechanisms—such as Lucas's 1972 Phillips-curve paper. Of course, there are a huge number of different mechanisms at play, so how does a macroeconomist oriented toward giving policy advice choose which mechanism(s) to focus on? Kydland and Prescott's answer was to move away from building models aimed at intuition in favor of larger models that could be parameterized and calibrated to deliver quantitative output. This quantitative output would then, within a single model, aggregate across the many mechanisms inherent in the model. The analysis of larger, nonlinear, models is much harder and, with sufficient complexity, impossible to undertake with "pencil and paper." Thus a new sub-field of macroeconomics developed rapidly: that focusing on solving dynamic models with numerical techniques. Nowadays, the most common approach in applied macroeconomics is arguably to formulate rich models allowing several mechanisms believed to play a role, solve the models numerically, and then simulate them to study model output and compare different mechanisms quantitatively. This approach has not replaced the need to formulate much smaller models to build intuition, but the aim is to ultimately be equipped not only with an intuitive understanding of what goes into building the macroeconomic equivalent of a bridge but also with quantitative assessments that allow us to cross the bridge without fear.

### 1.1.7 Macroeconomics and inequality

As already mentioned, the late 1970s saw sharply rising wage inequality, a phenomenon that has continued, though with different intensity, and hit different groups differently, during different decades. **What explains these developments?** Technological change, increased exposure to trade, or changes in unionization? Many macroeconomists have turned their attention to this question. They have developed theory and examined how different theories match the data, thus making the overlap with labor economics a particularly vibrant one. Here, data on individuals has been a key input. Again, most of this data is taken from surveys, but it is increasingly common for researchers to make use of administrative data, i.e., data on the whole population.

**administrative data**

The basis for taxation of individuals and firms is reports on incomes and transfers (including bequests). Here, employers report wages and salaries paid out for all taxable individuals, and individuals complement these data. For example, the self-employed report their own earnings. Similarly, firms and government agencies report transfers made, such as social security payments, and these are based on earnings. Tax authorities

also have records of capital income for all individuals—dividends, interest, and capital gains—but the specific assets are not recorded. Relatedly, there is no administrative data on wealth in the U.S., since it is not taxed, but some countries do employ wealth taxes and, therefore, administrative data on wealth can, in principle, be accessed there. Thus, all the underlying data is registered and potentially a source for researchers to use. Access is restrictive, but can be granted. The Internal Revenue Service (IRS) and the Social Security Administration (SSA), for example, have been employed to provide detailed account of the distribution of incomes for the full population.

Of course, tax records and other administrative data is not freely available and, depending on the particular data set, may be more or less difficult to obtain permission to use. In all cases, the data the researcher is given is made anonymous: individuals' identities are never revealed. A particularly interesting possibility for researchers is to link different data sets (whether a administrative data set or not) but this is rarely allowed.[a]

[a]The reason is the risk of inadvertently compromising anonymity. There are exceptions. For example, in Sweden, local researchers can cross-link administrative data sets; applications to do this require careful descriptions of the purpose of the study, the methods employed, and involve various ethical considerations.

In one strand of the literature, the focus has been on data sets that contain detailed information both on firms and on their employees, hence shedding light on what kinds of firms "match" with what kinds of workers and how wages are then set and vary over time. Although the share of total income paid to labor has been remarkably stable during the postwar period in the U.S., it has had a **recent trend downward** over the last decades, a trend that can be observed also in many other countries. Hence, macroeconomic researchers are examining various hypotheses for this phenomenon and here, structural change and technological change, possibly along with changes in the degree of competition—have markups increased?—are being examined. To this end, access to data on firms is critical, and the overlap with another field, industrial organization, is evident.

**Firm data**

Various firm-level (and establishment-level) data sets are provided by both governmental institutions and commercial vendors.[a] In the U.S., the Census Bureau and the BLS provide data from administrative sources and from surveys. We can thus obtain information about (among other things) entry, exit, and employment dynamics of firms and establishments at annual and quarterly frequencies. Although micro-level data are often confidential, many useful summary statistics are publicly available through their websites.

Some of the data sets include information about inputs and outputs on the firm level. Among other things, these can be of use for trying to estimate firms' marginal costs, which are never directly observable; their movements over time are important

for understanding macroeconomic phenomena. Relatedly, innovative activity can be measured based on firms' R&D expenditures, and the patents generated can be accessed from the patent office, along with citations to patents (that allows measured of impact).

It is somewhat rare but for some countries and sectors, there are data sets that allow looking at firm-worker matches. These data sets are also all survey-based.

[a] Commonly used commercial vendors include Compustat, Orbis, and NETS.

Over the last two decades, we have also observed a sharp increase in macroeconomists' interests in **wealth inequality**. Just like during the Great Depression, a common perception has been that of increasing gaps between "rich" and "poor," along with various forms of polarization, but what does the data—to the extent we even have it—really say about wealth inequality? There are (at least) two reasons for macroeconomists to care about this question, and about the underlying trends driving wealth inequality to change over time. One is an intrinsic interest in inequality as a key aggregate phenomenon: the view that it is an undesirable feature in society and should be taken into account even if it is in conflict with other goals. An additional intrinsic reason is political stability: as expressed in Piketty's 2013 book, one may worry that democracy is threatened if inequalities rise above certain levels. As a final example, one dimension of inequality of relevance in macroeconomics is that between women and men and across racial and ethnic groups; we are now seeing an increasing number of contributions documenting and analyzing, in particular, how the relative wages and the relative hours worked across groups have evolved over time.

A second reason for macroeconomists to care about inequality is that it captures heterogeneity that is important to take into account when examining the workings of the macroeconomy. When, for example, a tax rebate is implemented with the purpose of stimulating consumer spending, we have reasons to think that cash-constrained, poorer households would spend a large fraction of the rebate whereas richer households will save most, if not all, of it. Thus, distributional data on wealth appears as a determinant of the efficacy of many policy interventions. The development of so-called heterogeneous-agent models, which began in the 1990s and has generated a very large literature, is a response to both these reasons to keep track of, and understand, wealth inequality.

**Measuring individual wealth**

In the U.S., since wealth is not taxed, there is no direct administrative data on it.[a] The Survey of Consumer Finances (SCF), also conducted by the Federal Reserve Board, is available every three years since 1983 and has data on individual assets; it is a key source of information about the wealth distribution. Unlike some of the surveys mentioned above (such as the ASM), this survey is voluntary, but efforts are made to make it representative. The IRS administrative data has capital income, so it is possible to estimate wealth by observing the annual income it generates—if one is willing to assume

a rate of return on the wealth. This is called the *capitalization* method.

[a] A small set of countries have taxed wealth over various periods in time and therefore have administrative data on it.

As it turns out, the various different sources used by different researchers do indicate a rather significant increase in wealth inequality in the U.S. beginning in the late 1970s. Similar trends have also been documented in a number of other countries. In sum, macroeconomics today, including at the level of policy making, is concerned with a much broader view of inequality than in the past. A comprehensive database on inequality in income and wealth across countries is the World Inequality Database (https://wid.world/).

### 1.1.8 Taxes and government activities

Many western countries, including the U.S., have experienced slow, long-run increases in the role of government, both when it comes to its total share of GDP and employment and in terms of transfers, such as social security and welfare systems more generally. Thus, the **nature, determinants, and effects of taxation** have become a central theme for macroeconomists. In the U.S., marginal tax rates were increasing and peaked shortly before Ronald Reagan took office and thereafter the degree of progressivity was lowered significantly, and has stayed at a historically low level since. The taxation of corporate profits has also changed over time. How have these changes affected hours worked, economic activity, and inequality? These questions preoccupy many macroeconomists.

Aside from these changes over time, an overall key question, aside from the degree of progressivity of the tax code, has centered around the choice between different tax bases: taxes on capital income, taxes on labor income, corporate taxes, indirect taxes (such as sales taxes), seignorage, property taxes, and so on. That is, macroeconomics and public finance intersect in important ways.

When it comes to the efficiency features of different tax rules, there is also an important overlap with economic theory. For example, it may be tempting to use tax and transfer schemes to fill in where private insurance markets appear to be missing—thus improving the economic situation of those experiencing unexpected adverse events. However, if the private markets are imperfect because of fundamental information asymmetries—leading to problems of moral hazard and adverse selection—it is important to think about potentially negative incentive consequences of government interventions. Such questions have been examined by theorists (e.g., Mirrlees, 1971) and a relatively recent subarea of macroeconomics has applied and further developed their contributions. There are also other impediments in the pursuit of efficient government policy. One, which has been a major issue in macroeconomic research at least since the 1970s, is the fundamental inability of governments to commit to its future policy choices. For example, basic public finance theory says that it is efficient to levy taxes on already installed capital—since it is not distorting any choices—but, if firms/investors know in advance that their capital will be taxed in the future, current investment is distorted. Thus, the government would like to say that it will not tax capital in the future, but, ex

post, change its mind. This kind of *time inconsistency* of policy choice has been held as a key explanation for how central banks, including the U.S. Fed, failed to stabilize the economy in the 1970s: they acted on short-run incentives to raise inflation, but to the extent market participants predicted these actions in advance they simply adjusted prices and wages upward, leading to stagflation: no improvement in output, and higher inflation. This reasoning was a key argument behind why many central banks were given significant independence in many countries around the world (including the U.S.): with statutes focusing more on long-run goals, there would be less room for time inconsistencies. The tensions inherent in differences between a government's optimal plans and their ex-post temptations to change them has generated another vibrant area of research with significant components of economic theory.

Related to the lack of commitment to future policy choices, there is another challenge. For one might imagine that, armed with well-researched insights about how policy decisions can best be made to achieve a given set of goals—be they about efficiency or redistribution—it is straightforward to approach policymakers with proposals, who would then turn around and implement them. Unfortunately, in practice, such a direct influence rarely materializes. One reason for this is what is usually labeled "political constraints"; i.e., for policy changes to be implementable in practice, political support is necessary. For this reason, the political economy of macroeconomic policymaking is another subject of a significant amount of research. One goal of this research is to simply try to understand **what policies tend to be chosen, and why**; a second, and ultimately more ambitious goal, is to take a **normative perspective on institutions** and evaluate how different features (ranging from basic rules for how governments are chosen to details in their decision-making processes) lead to different policy outcomes. On the positive side, one might note that research can have significant influence also in this arena. One example is the politico-economic reform in Sweden in the early 1990s that came as a result of the work of the so-called Lindbeck Commission. The commission was appointed after a deep recession and lackluster economic performance during a period of time. Its proposals to change key features of the political system—even involving the frequency of elections and key rules surrounding how the government could use its budget—in order to improve outcomes were actually implemented in full.

### 1.1.9 The Great Recession: another oops

In 1979, in an effort to end the high-inflation era, the new Fed chairman Paul Volcker announced, and implemented, a period of very tight monetary policy. Most macroeconomists attribute the ensuing recessions in the early 1980s to this change in monetary policy stance. Inflation did come down, and it did so also in most other western countries; the stagflation period had been a worldwide phenomenon. In Europe, steps were gradually taken toward tighter monetary policies as well and a currency union was eventually created: in 1999, ten country currencies ceased to exist and the euro took their place. This was a period of financial integration not only among European countries, but also between developed and emerging economies. Restrictions to the movement of financial assets, goods, and services were loosened, triggering a "globalization" process that generated great interdependence

among economies. From the mid-1980s and for over two decades, until 2007, the U.S. economy experienced recessions but they were minor and the overall aggregate performance was viewed to be very satisfactory. In 2002, macroeconometricians James Stock and Mark Watson dubbed this era the Great Moderation: a period of time when the macroeconomic aggregates displayed healthy growth and very low volatility. By some, the Great Moderation was attributed to the new and transparent policies followed by the independent Fed. Researchers also advanced other hypotheses that were more structural, such as changes in the nature of technological change. Furthermore, the stabilization frameworks used at many central banks now relied on the New-Keynesian macroeconomic model: a setting based on microfoundations and including a number of frictions, most importantly sticky prices and sticky wages. A prescription from these models was for the central bank to systematically counteract macroeconomic shocks so as to lower volatility and improve our welfare.

Whatever may have caused the Great Moderation, there is no doubt that few expected the events that surprised the world in 2007: a severe economic downturn that was worldwide as well. The recession was nowhere near as deep as the Great Depression, but it was nevertheless a very problematic period: unemployment rose sharply and only fell back very slowly in a manner that was uncharacteristic, compared at the very least to recent experience. The crisis immediately hit Europe as well, and in 2009 a multi-year debt crisis, sometimes labeled the eurozone crisis, was set off. A number of European countries thus suffered from high national debt levels and difficulties in rolling over their debt; this period was one of significant uncertainty. The uncertainty not the least involved what paths government policy would take, and there was ample speculation that some countries would leave the currency union. In the end, they did not, but the crisis was long and painful, as were the macroeconomic and political debates about whether debts should be forgiven or not. The crisis highlighted the potential problems of a globalized economy, triggering some countries to 'close down.' The exit of Great Britain from the European Union, labeled "Brexit," was the most notable example. An important role in combating the crises was played by central banks, but now with methods very different than those pursued during the previous decades.

A period of time of intense research followed. **What were the deep causes of the Great Recession?** How could it have been avoided? Were our main theories flawed? This research is still ongoing so it is hard to draw definite conclusions, but there is consensus that a combination of excessive risk-taking in housing markets—arguably rooted both in private and government decisions—and severe frictions in financial markets together slowly sowed the seeds of the downturn. Indeed, the term Global Financial Crisis is equally often used to refer to these developments.

As a result of the experiences during this period, massive research has gone into studying the workings of financial markets and financial institutions and how government regulation affects their performance. It reminded us how asset markets, debt buildups, and (excessive?) risk-taking can be intricately intertwined with the workings of the macroeconomy. The downward trend in the real interest rate is a related phenomenon. It is suggested to be connected to more severe asset-price fluctuations, including the formation of price bubbles; here the emergence of cryptocurrency is perhaps particularly noteworthy.

**Financial data**

One of the key questions in this part of macroeconomics is *financial stability*. In particular, if key financial actors have strong interdependencies in their asset portfolios and liability structures, then domino effects can occur, whereby relatively minor shocks can have severe aggregate consequences. As an example, the house-price decline and the resulting defaults on mortgages in the U.S. during the Great Recession were significant, but their quantitative magnitudes were quite small compared to many other asset-price movements throughout history that barely even generated recessions at all. The reason why the small shock had major consequences was to be found in how mortgage liabilities had been packaged and distributed among key financial actors, all of which was quite nontransparent not just to policymakers but to market participants as well. Hence, the need for data in this area is, and was, great, and a major challenge is that high-frequency data on portfolios is proprietary information and, when it is available, can be hard to interpret.

In its Flow of Funds section, the Federal Reserve Board produces the quarterly Financial Accounts of the United States, a comprehensive set of accounts that includes detail on the assets and liabilities of households, businesses, governments, and financial institutions. These are aggregate data, allowing us to track, among other things, trends in indebtedness—which was key in the analysis of the Great Recession, but also not quite sufficient for detecting interdependencies among financial institutions. There are also data on individual households and firms; the SCF is discussed above.[a]

Asset prices and returns for publicly traded firms are of course available from numerous sources. Assessments of values for non-traded firms are much harder to come by, and even many large firms are not publicly traded. In the past, a typical path from the birth of a firm to an established, large company involved a mix of individually provided funds, bank loans and possibly bond issues, with an eventual public offering and public trading. Today, the path toward public trading of the firm's equity often takes longer and goes via risk-capital and private-equity financing.

How private individuals make portfolio decisions is another important input into how the macroeconomy works, but raw data on this is much harder to come by; as discussed above, data on wealth management is typically not available except in limited surveys.

[a] An often-used financial database for firms is Compustat, a commercially provided service containing information about publicly traded firms (also outside the U.S.). An even larger database is Orbis, which also contains non-traded firms, including smaller businesses.

Though it seems clear by now that the basic macroeconomic framework was not abandoned as a result of the Great Recession, it is equally clear that it has been changed and enriched in the direction of including financial frictions that play a prominent role. Macroeconomists are, perhaps painfully so, aware that the next recession will rarely have the same characteristics as the most recent one, and as a result their theories grow richer and more

complex, rather than themselves undergoing cyclical fluctuations.[8]

### 1.1.10 Climate change

The intersection between macroeconomics and environmental economics was close to empty until it became clear toward the end of the 20th century that **climate change**, at least in important part caused by human emission of carbon dioxide into the environment (primarily by burning fossil fuels like oil, natural gas, and coal), was a potentially critical threat to our welfare. Climate science itself is rather young part of the natural sciences and although there was consensus about the mechanisms involved early on in the 20th century, it was not until at the end of the century that consensus was built that a significant amount of global warming—of a little under 1 degree Celsius—had occurred as a result of human economic activity. The Intergovernmental Panel on Climate Change (IPCC) formed in 1988 and has issued 6 comprehensive reports on the subject.

As a result of these developments, increasingly many economists have become engaged in climate research and contributed importantly to our understanding of how climate change interacts with economics. Here, clearly, collaboration with climate scientists has been important. Part of the economists' research is about assessing the effects of global warming on economic activity, on health, and more generally human welfare and it has engaged many subdisciplines of economics. Macroeconomists have helped in damage measurement by studying aggregates and how they react to weather as well as climate. But macroeconomists have also contributed by constructing global economic models aimed at *integrated assessment*: examining how different economic policies would influence the world's market economies and jointly determine climate and economic outcomes, hence providing advice for policymakers. This endeavor has led to an upsurge in quantitatively oriented modeling of global macroeconomic interactions.

The focus on fossil fuels also directs the attention of macroeconomists to natural-resource and energy economics, which attracted attention already in the 1970s as the oil shocks hit. The questions now are similar to what they used to be, and they have been underscored by recent events such as the war in Ukraine. In short: **what is the nature of energy supply, and what are the potentials for, and effects of, technological change in this area?** The climate-economy nexus is an area where research, along with communication of the research results to policymakers, can turn out to be of extremely high value, especially in the developing world which, by all estimates, are likely to be the most severely hit affected by further warming. Quite fundamentally, limiting the use of fossil fuels will amount to the use of government policy around the world to steer individuals and markets in a direction away from these fuels. How this is best done is a matter of understanding economic decisions and market interactions: it is a question for economists.

A broad interpretation of the environment includes health hazards caused by epidemics and pandemics. Here too, macroeconomists have contributed both empirically and theoretically; the theoretical contributions primarily amount to building assessment models quite like

[8] Voices were certainly raised suggesting that a return to classic Keynesian theory was called for.

those in the climate area, though now focusing on integrating epidemiology and economics.

### 1.1.11 Where do we stand?

Clearly, our economies are constantly evolving as a result of a number of societal changes, including technological developments and policy reforms. With these changes, we have indicated how macroeconomics has changed course, sometimes abruptly and sometimes merely by expanding on existing frameworks. We currently have a body of knowledge that allows for a more nuanced understanding of macroeconomic events and policies, and the macroeconomic models used in practice are accordingly much richer than in the past.

Do we lack sufficient self-criticism, however? One regularly hears arguments that macroeconomics never really admits that it is wrong, nor that it recognizes that it needs to change course. In concrete terms, can we really say that we are in a better position today to meet the next major macroeconomic challenge? Our perception, first, is that macroeconomists do admit mistakes, as indicated by the number changes in our thinking that have been described above. On some occasions, our theories simply have not incorporated all relevant features—this would be "our" mistakes—and as a result we have tried to build these features in as swiftly as possible. On other occasions, though, we have simply been surprised by events that did not have economic origins and yet necessarily generated economic downturns that, once we were informed of the "shock," we could understand with existing models.

Our own firm belief is in fact instead that, guided by research, macroeconomic policy has been increasingly successful over time. Not all recessions are alike, but they are not all distinct either, and lessons from one will typically be useful in the future too. In particular, we count the responses to the Great Recession and to the recent pandemic as having benefited in major ways from macroeconomic research. At the same time, of course, not all governments acted alike and there always remain differences in views on policy, especially since strong views on economic policy tend to be linked to strong political views.

Thus, as different shocks hit our economies, we do not cycle back and forth between models, each model emphasizing one type of shock and how to respond to it. Rather, we strive to combine insights as they come and try to isolate how they add to, rather than erase, our previous understanding of the economy. For illustration, consider Keynes's core insights: they were entirely new and crucial for building up an understanding of how stabilization policy could be usefully conducted, and although the 1970s saw a definite break in the foundational elements underlying macroeconomic models and a temporary return to very stylized models, the Keynesian insights have since been added back into the newer models. These models are simply more sophisticated today than before and are more clear on the circumstances under which Keynes's insights can be applied. The models keep developing; for example, the reliance on rational expectations may develop, as we obtain more data on how forecasts are actually made. Macroeconomists appear to be in agreement that frictions in financial markets can be critical and even central in explaining some recessions, but how to identify the key frictions and prevent them from playing out is still an open question; clearly, a prohibition of all loans would by definition have prevented the mortgage market from causing problems, but very clearly such restrictions are highly undesirable. Thus, research

on this issue today is trying to identify how we can reach a balance between stability and business-as-usual market efficiency. Many challenges remain in macroeconomic research but we are nevertheless optimistic that there will be fewer and fewer oopses in the future.

Finally, the construction of rich and complex models is of course not an end in itself and especially for communication and in teaching—even at the graduate level—it is important to simplify and abstract from many of the complexities. Therefore, IS-LM models can still be useful in building an understanding of how the macroeconomy works, as can RBC models and simple models of debt crises. But beliefs that either of these models is sufficient has been proven wrong many times over.

## 1.2 Looking ahead

This introductory chapter has sampled a number of important topics addressed by macroeconomists over the course of the last century. Many topics have been left largely without comments, such as trade liberalization and immigration. This is not to suggest that they are any less important: they remain very active research areas, some overlapping with further sub-disciplines of economics such as international trade—and the main focus here has instead been to illustrate the large variety of topics and methods.

The development of rich and complex models means that macroeconomics is not becoming easier. This textbook is a living proof of this statement: although it makes a major effort to mix data and historical episodes with theory, it does require new students of this area to make serious investments in methods. In particular, the switch toward microeconomics-based theory, with an accompanying aim to match the main historical facts—*quantitative theory*—motivates many of the early chapters. The belief on which this text is based is that this material is here to stay. One possibility is that the theory will be supplemented with elements of behavioral economics, but there is little consensus yet on which features are key. This regards both how we view the choices made by consumers and firms and how they form expectations. For now, full rational optimization and expectations are viewed to be a very reasonable starting point, and the insights from investments in these methods will also highly likely be relevant as behavioral elements may be added. Thus, we try to view the methods parts as fun, because they are, though they are never there for their own sake: their only aim really is to help us understand the macroeconomy. This is the nature of macroeconomics; it is challenging, but it is engaging.

The overall text has two basic parts. The first one focuses on methods, while introducing the methods—in particular, the framework used—with explicit reference to historical data. This part is core material and should be read in advance of the rest of the text. The second part has applications. The set of applications perhaps contains slightly more material than is covered in a typical first-year PhD course, but only slightly; the recommendation is nevertheless to read the chapters in order. In the first part, Chapter 2 is of particular importance since it is a lead-in to the rest of the text. In particular, it dives into the macroeconomic data and, bit by bit, introduces Solow's way to make sense of this data: macroeconomic aggregates are generated by the neoclassical growth framework. This framework is the core

setting used in macroeconomics. The framework is not an arbitrary one: as the chapter explains in some detail, it is instead precisely motivated by a need to square our theory with some striking, long-run facts that are very hard, if not impossible, to explain without this theory. The chapter also moves beyond Solow's treatment by arguing that some parameters he treated as exogenous constants—saving rates and hours worked—are better described as conscious choices of households in a market environment. The chapter concludes with a preview of the remainder of the text, emphasizing that virtually every chapter thereafter, dealing with the main, applied topics in macro (growth, business cycles, asset prices, labor markets, etc.), build directly on the market version of the neoclassical growth model.

# Part II

# Foundations

# Chapter 2

# A framework for macroeconomics

The purpose of this chapter is threefold. First, we go through the main macroeconomic time series, focusing primarily on their historical properties. We mostly use U.S. data and thus encourage the reader to examine the corresponding graphs for other countries.

The second, and key, purpose of the chapter is to gradually introduce the basic framework—the macroeconomic model—that will constitute the core tool in the textbook. Thus, each graph will be interpreted from the perspective of the proposed framework. An underlying assertion is that it is hard, if not impossible, to account for the data except with the kind of framework we use. This framework goes back to Solow's growth model and then builds in conscious choices, such as firms' choices of inputs, consumers' choices for saving and hours worked, later on the purposeful development of human capital and technology, and so on. The emphasis on explicit choices necessitates a microeconomic approach, not just in terms of theory but also in terms of the data we will look at; nowadays, much macroeconomic research directly studies cross-sectional data (for households, firms, etc.). There is no presumption that markets work perfectly. Instead, much of the analysis, especially when it comes to looking at macroeconomic policy, centers around pinpointing specific weaknesses in the functioning of markets. Finally, a key feature of the core framework is that it is *quantitative*: the aim is to formulate a model that can account for the magnitudes of macroeconomic phenomena and not just their qualitatively features.

The third purpose of the chapter is to be a stepping-stone into the rest of the text. Thus, this chapter will offer a brief description of the topics studied in later chapters.

## 2.1 The facts and interpretations: real aggregates

In this section we document some basic facts relevant for macroeconomic analysis. The facts are presented in a stylized manner; for example, the unemployment series will be described as "stationary" and this term should not be interpreted in a statistical sense but rather as a series that does not have a marked trend (toward, say, zero or one). Of course, the swings in the series will be pointed out, including rather persistent ones. The main facts we go over in this section, moreover, emphasize the longer run; short-run facts are discussed in more

detail later. The growth facts we will focus most on are from the United States, but we will show some data from other countries as well. They are, for the most part, referred to as the Kaldor facts, but there is no strict adherence here to the facts originally pointed to in Kaldor (1957).

### 2.1.1 Output grows steadily

One of the most remarkable facts in economics is the steady growth of output over the last centuries. The path for (the logarithm of) real U.S. output is shown in Figure 2.1.[1] The figure reveals almost constant growth over more than a century and the swings up and down seem minor from a bird's-eye perspective[2]. The notable exception is the Great Depression episode and the rebound after that, but after that hick-up the economy lands on "the same" growth path again. Thus, our regular business-cycle movements, including the most notable recent recession (the Great Recession, 2007–2009), are barely visible.

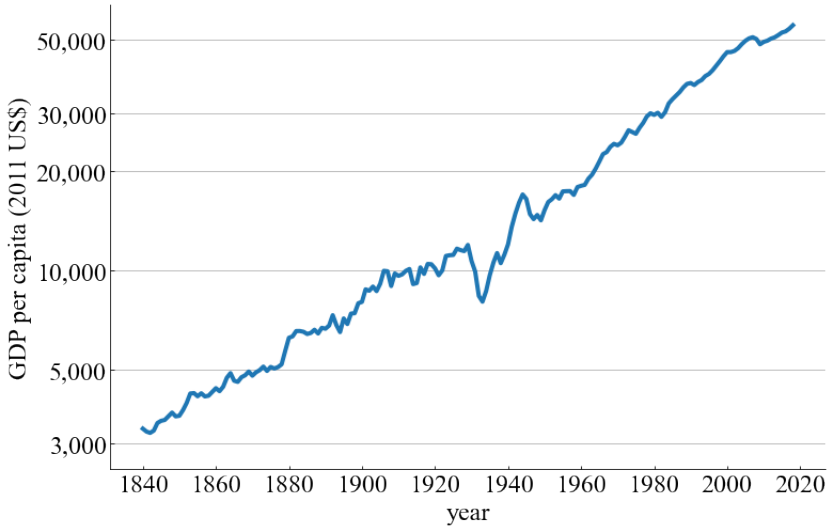



Figure 2.1: GDP per-capita in the U.S.

**Notes**: The figure plots GDP per-capita in 2011 prices in the U.S. 1840-2018. **Source:** Maddison project.

One of our key goals now is to try to "account" for output growth, i.e., to provide a theory that offers a deeper understanding of the remarkable fact in Figure 2.1. We do so by focusing on the production side, i.e., how the basic inputs into production have evolved over time. We also look at their prices.

[1] As discussed in the measurement section, systematic measurement did not begin until about a third into the twentieth century. The Maddison database goes much further back and then output estimates are based on available time series for, e.g., production, employment, and prices.

[2] A linear fit suggests that the series is well approximated by a annual growth rate of 1.86 percent. (Such a linear fit obtains an $R^2$ of more then 0.98.)

### 2.1.2 The basic resources behind output—and their prices

We begin by looking at capital.

#### Capital input

As we shall argue, the process of growth is driven, at least in part, by capital accumulation. Figure 2.2 shows the capital-output ratio in the U.S. since the late 1920s. Apart from a marked jaggedness early on—during the Great Depression especially—we also see clear stability at a value of around 3. The measure for capital here is the standard one: "accumulated investments, minus depreciation."

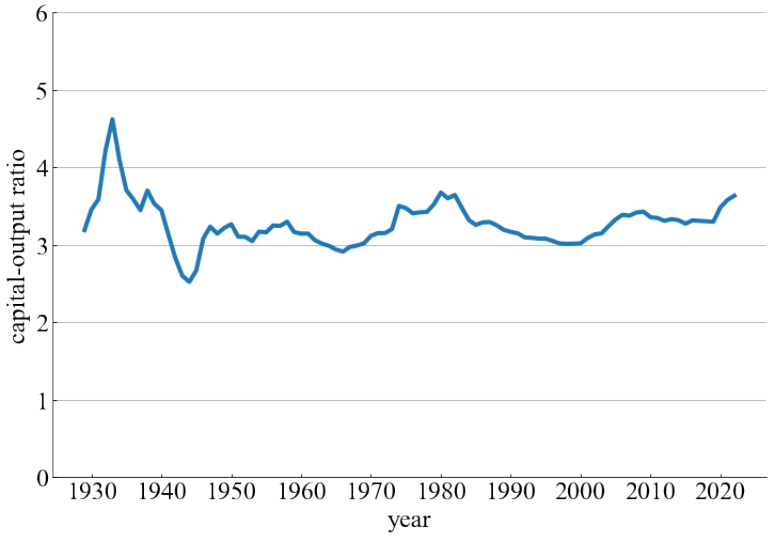



Figure 2.2: Capital-output ratio in the U.S., 1929-2022.

**Source:** FRED. Numerator: Current-Cost Net Stock of Fixed Assets and Consumer Durable Goods (K1WTOTL1ES000), Annual, Not Seasonally Adjusted, converted to billions of dollars. Denominator: Nominal GDP (GDPA), Annual, Not Seasonally Adjusted, reported in billions of dollars, The figure plots the ratio between fixed capital and consumer durables relative to the GDP.

The focus here is capital as an input into production. It is nevertheless interesting to note that a broader interpretation of capital is wealth, which would include the value of land, housing, and so on. In Figure 2.3, we show the wealth-output data as computed in Piketty (2014).[3] We see marked stability again, though large changes in the composition of the capital stock, toward manufacturing capital and, especially, housing, and a total that is 4–5 rather than 3.

[3]We use his data starting 1880.

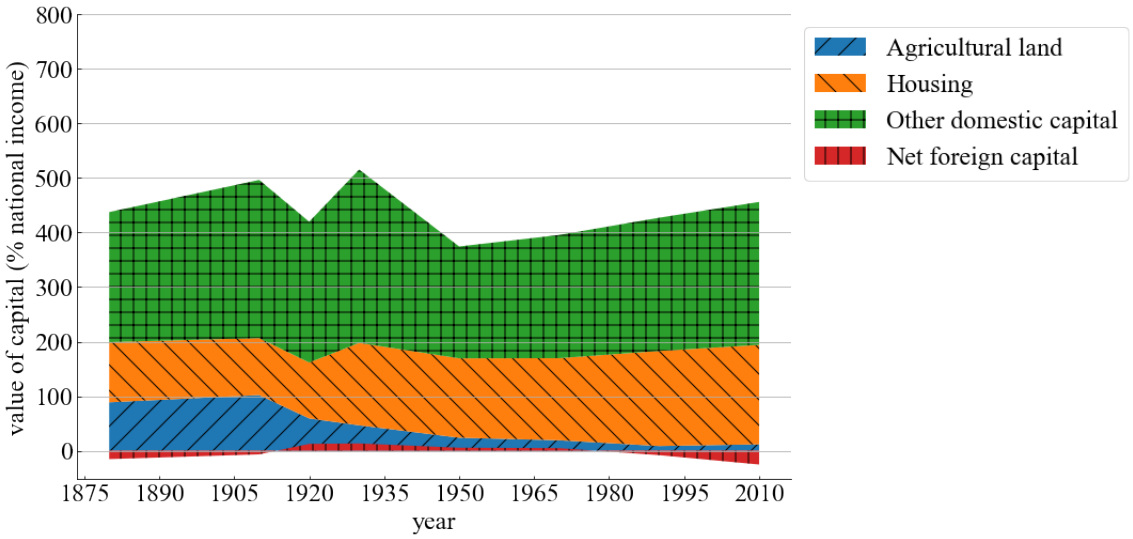



Figure 2.3: Wealth-output ratio in the U.S.

**Source:** Capital in the 21st century, Chapter 4, Figure 4.10. Data here.

### The price paid for using capital

How expensive has it been for producers to use capital over time? Most commonly, firms buy capital and use it until they scrap it, or sell it in market for used capital goods. It has become increasingly common for firms to instead rent capital (such as machinery or buildings), in which case the price paid for the use of capital is clearly the rent. But due to lacking systematic historical data on rents, measures of the cost of capital are instead constructed based on (a minimum of) theory. One way is thus to look at a measure of the returns on investments: on the margin, under competitive markets, this return should equal the marginal cost of the investment: the price we are looking for. Thus, we can look at stock-market returns as one measure of capital's price. Using binned data on stock-market returns—a return to investments in the capital of firms whose shares are publicly traded—from three long time-periods, Figure 2.4 shows no strong trends. If one were to look at shorter time periods, the variations in the stock market returns are of course very noticeable and large. These fluctuations are likely due to changes in asset valuations more than to changes in costs, which is why longer-run averages seem more appropriate.

Alternatively, one can attempt to measure the cost side, but also using theory. The "user cost of capital" is based on the foregone return to saving: by buying and owning capital, a firm is losing the return it would have received by merely saving the money. This measure also takes into account depreciation: a part of the capital is lost by using it, or needs maintenance to be kept in good shape. Moreover, it takes into account capital's change in value over time; computers, while not physically depreciating, lose value on the used market since new computers always out-compete old computers. Thus, the user cost is, roughly

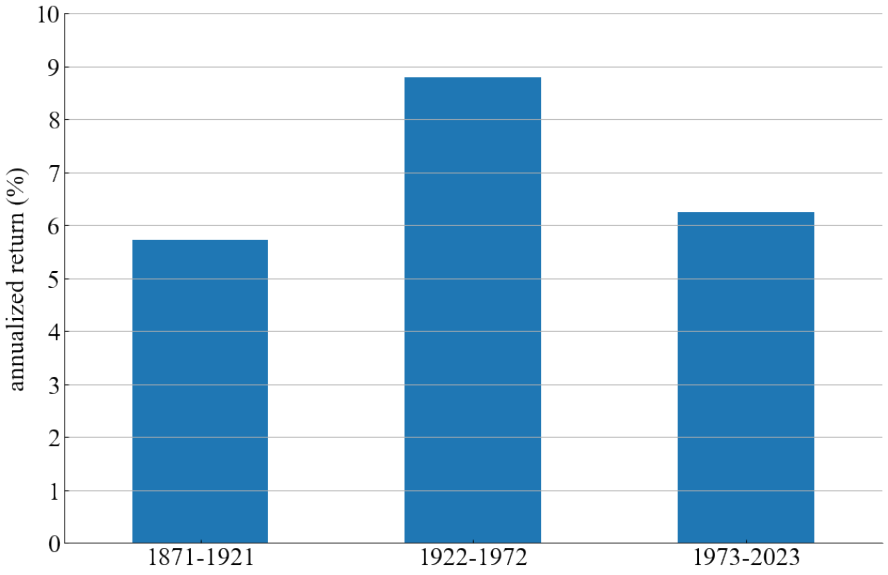



Figure 2.4: Return on capital.

**Source:** Online Data from Robert Shiller, "Irrational Exuberance." Annual, geometrically compounded returns to U.S. Stock Markets.

speaking, a market interest rate plus depreciation plus a the fall in value.[4] The user cost is also stationary, but of course also includes short-run swings, as Figure xyz (missing, need to find one) shows.

### Labor input

The second major input into production is labor. Employment is one measure of input, but it is often relevant to take into account how many hours each employee works given that hours worked vary widely and many people have more than one job. Hence, a common measure of labor input is hours worked (in the marketplace) per adult. Various measures are available and we will show two here. First, Figure 2.5 shows that, since the beginning of the last century, hours worked per week have fallen, from around 28 hours to around 23 hours. Looking more closely at the graph, we see that since the end of World War II, hours look rather stable, without a net downward trend. This is a fact that is often referred to—that U.S. hours are stationary—but actually only accurate over the postwar period. Second, we see very large departures from trend in the figure; during the Great Depression, extreme unemployment rates account for the low hours, with a subsequent war-related upswing. In the U.S., unemployment movements, which are large, account for a big share of the movements in hours.

[4]Other factors can appear in a user-cost formula, such as the role taxation plays for capital income an in deductions for depreciation.

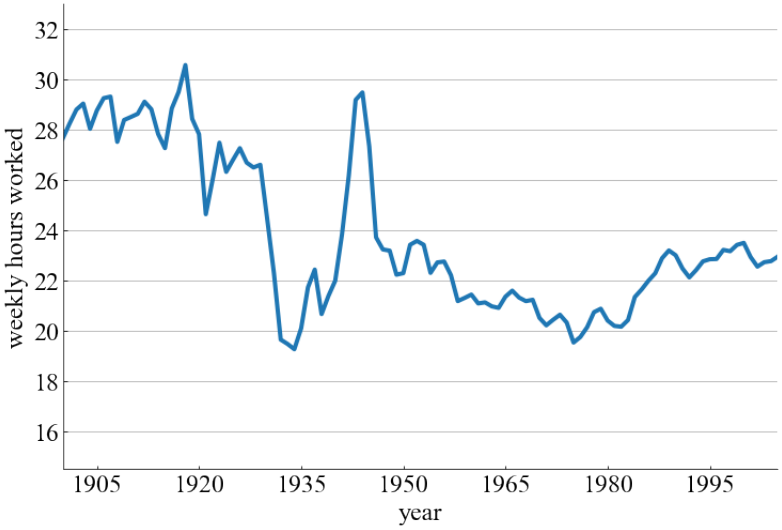



Figure 2.5: Average weekly hours worked in the U.S (age 14+)

**Source:** Francis and Ramey (2009). Data here.

Figure 2.6 shows hours worked over a longer time period, along with average real wages. Here, the downward trend in hours is even clearer (the graph depicts hours per employed, so the numbers are overall higher and do not account for changes in participation). The cumulative decline is almost 50%.

#### Wages

Figure 2.6 also shows real wages. Wages have risen at a remarkable average rate of around (or even above) the rate of output, with an exception in the most recent period. The secular increase is not surprising in some general sense: the standards of living have, slowly but surely, risen steadily for most Americans, and the major source of this rise is coming from higher and higher earnings. Given that hours are not showing an upward trend—in fact, the opposite—it must be that real wages have risen steadily. There are also some movements in how total output is divided up into capital and labor income; we will discuss these later, but the first-order aspect here is that the shares have been quite stable.

### 2.1.3 Taking stock: a "neoclassical" picture emerges

The data on capital and output had puzzled economists; in particular, the constancy of the capital-output ratio at a value around 3 suggested something quite stark—it suggested a rigid technology structure where labor played no role and capital and output were always in the same proportions—and this was hard to square with how we knew that production took place in practice.

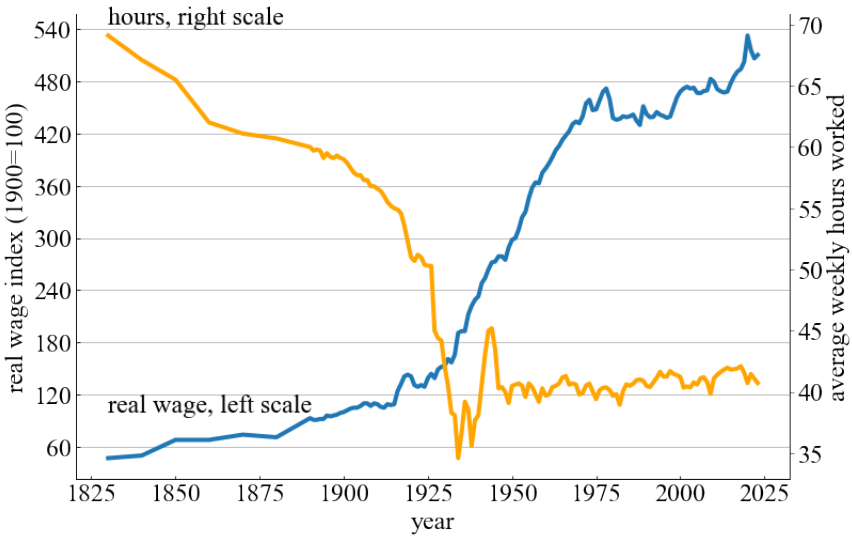



Figure 2.6: Average weekly hours worked in the U.S. (Manufacturing)

**Source:** Period 1830-1880: Whaples 1990, Table 2.1. Period 1890-1970: U.S. Census Bureau, "Bicentennial Edition: Historical Statistics of the United States, Colonial Times to 1970," Series 765 and 803. Period 1970-2023: FRED, monthly series AWHMAN, annualized. Wage data source: Period 1830-1888: Williamson 1995, Table A1.1, (1900=100). Period 1890-2023: FRED, quarterly series LES1252881600Q, annualized (1982-84 CPI Adjusted Dollars) **Note**: Converted wage index in 1982-84 dollars from FRED to 1900=100 index by multiplying the FRED series by the ratio of wage index in 1983 from Williamson with wage index in 1983 from the FRED.

Solow (1956), however, took a less immediate production perspective on these facts. He noted what you saw in the previous two sections: there has been a steady rise in output, an equally steady rise in capital with stationary, or even declining, hours worked. At the same time, the price of capital has been stationary while wages have had a significant upward trend. These facts did not, by themselves, appear mysterious. Solow in particular noted that it is natural that as the price of an input rises, the use of the input declines. More precisely, from the perspective of production theory, at a higher relative input price, a firm uses less of that input relative to other inputs. Labor has become more and more expensive relative to capital, and its use has fallen, again relative to capital: the capital-labor ratio has risen at the rate of output growth, or even slightly more.

The other side of the coin, which was important to address, is about accounting for these changes in relative input prices: what made labor more and more expensive relative to capital? One natural explanation would have its roots in technological change. In particular, suppose it is directed toward labor: labor becomes more productive over time per unit of hour worked. If capital and labor are complementary, this factor would lead to an increased value of capital on the margin. However, as we saw, the market return to capital has remained stationary—it has not risen. Solow noted, however, that the stationarity could be explained

by *neoclassical* forces. A production function that has decreasing marginal returns to each input is labeled neoclassical and with such a production function, as the relative amount of capital—capital used per worker—rises, the marginal value of capital would fall. If firms buy inputs in competitive input markets, moreover, this would be reflected directly in the market return to capital. Thus, a story emerges where technological change directed toward labor, along with neoclassical forces, can, at least potentially, account for the historical data on relative input quantities and relative input prices.

The neoclassical features led Solow to describe *aggregate* output as being generated by an *aggregate* production function, with *aggregate* capital and *aggregate* labor as inputs. Solow also posited, along the lines above, that the production function likely changed nature over time—that there was some technological progress—and built a model around these ideas; we will briefly describe this model momentarily but let us first focus on how, given Solow's perspective, one could take the next natural step: accounting for the sources of aggregate growth as coming from growth in inputs and growth in technology. This accounting procedure, along with the development of Solow's neoclassical growth model, would allow us to obtain a much less mysterious, and in fact quite natural and operational, account of how output grows over time. As a matter of microeconomic theory, the existence of an aggregate production function, i.e., a functional mapping from the total (economy-wide) quantities of inputs only into some measure of aggregate output, is not easy to establish in general and has also, as the box below discusses, been subject to some controversy.

**The existence of an aggregate production function**

There are assumptions under which a functional relationship can be established between input quantities and a price-independent output measure. However, these assumptions are extremely specific, indeed knife-edge, cases. Think of a static economy producing two goods, $x_1$ and $x_2$, both from capital and labor but with different production functions; also imagine that there are no restrictions on how inputs can be allocated across the two sectors. Then if a relative price between $x_1$ and $x_2$ is assumed—let us call it $p$—it is straightforward to see that a competitive equilibrium will generate a mapping between the total input quantities and the value of total output: perfectly competitive markets would allocate, or a planner could equivalently allocate, capital and labor so as to maximize output. As the total amounts of capital and labor would vary, total output would change. Thus, we would obtain a mapping from inputs to output. But this mapping would nontrivially involve $p$: it would not be a pure production-function relationship. Thus, one would need to add a demand side, thus endogenizing $p$, to obtain a pure mapping from input quantities to a measure of output. But then preferences (or whatever gives rise to demand) need to be described, and they will generally, as would $p$ in a direct sense, influence the mapping. Thus, a pure production mapping is hard to imagine.[a]

The existence and usefulness of an aggregate production function was hotly debated in the so-called Cambridge capital controversy during the 1950s. This controversy, which

had its two head quarters in the two well-known Cambridges (the University of Cambridge, U.K., chiefly represented by Joan Robinson and Piero Sraffa, vs. Paul Samuelson and Robert Solow at MIT, Cambridge, Mass., U.S.), also involved the notion of "aggregate capital," but in essence it focused on the aggregate production function. The controversy on the existence conditions can be summarized as having been won by Cambridge, England, whereas the usefulness controversy was arguably won by Cambridge, Mass. With the tools of modern macroeconomics, one can solve large models and see to what extent departures from aggregation play an important role quantitatively. Some such endeavors have been undertaken and point to limited departures, but no fully systematic analyses have been conducted yet.

[a] If you assume that the two production functions have identical isoquants—one is a scalar multiplication of the other—then it is possible to construct the mapping, as a relative price $p$ is implied from the production technologies alone. Try to show this as an exercise!

### 2.1.4 Growth accounting

Solow (1957) introduced "growth accounting" as a way to implement the notion that one could break down aggregate output growth into sub-components. With the use of a rather limited amount of theory, Solow was thus able to quantify the relative importance of different sources of U.S. growth.

Solow's growth accounting made the following assumptions: (i) aggregate output $Y_t$ is generated from an aggregate production function $F_t(K_t, L_t)$, where $K_t$ is aggregate capital input, $L_t$ aggregate hours worked, and the subscript $t$ denotes that technological change may move the production function upwards over time; (ii) $F$ has constant returns to scale (CRS) and neoclassical properties; (iii) there is perfect competition for inputs and, hence, firms maximize profits. The CRS assumption, we know from microeconomics, leads to zero pure profits in a perfectly-competitive equilibrium, which Solow thought was a good approximation and, besides, being able to replicate production would at least ensure that if all inputs double, output would double, so less than constant returns to scale was not thought of as reasonable.

Throughout the text (when not otherwise noted), we will use the convention that lower-case letters are in per-capita real terms. Thus, in this chapter we use $y_t$ to denote per-capita output, i.e., $Y_t$ divided by the size of the population; denoting population by $N_t$, we have $y_t = Y_t/N_t$. Note that, since $F_t$ is CRS, we can write $y_t = F_t(k_t, \ell_t)$, where $k$ and $\ell$ are thus capital and hours worked per-capita terms. Most of the time in this chapter, beginning here and now, we will also abstract from population growth and simply consider a population of constant size ($N_t = N$). It is sometimes convenient to normalize the population size to 1 so that we have $y_t = Y_t$.

Given differentiability, we can write the total differential of output (with time subscripts suppressed) as

$$dy = \frac{\partial F}{\partial t}dt + \frac{\partial F}{\partial k}dk + \frac{\partial F}{\partial \ell}d\ell = \frac{\partial F}{\partial t}dt + \frac{\partial F}{\partial k}k\frac{dk}{k} + \frac{\partial F}{\partial \ell}\ell\frac{d\ell}{\ell},$$

where all the partials are evaluated at $(t, k_t, \ell_t)$; for the purpose of this derivation, we also treat time as continuous. Dividing by output, using $r$ and $w$ to denote capital's rental rate and the wage (the prices, as well as the quantities, also depend on time but this dependence is omitted for notational convenience), and then using firm profit maximization, we obtain

$$\frac{dy}{y} = \frac{\partial F}{\partial t}\frac{dt}{y} + \frac{rk}{y}\frac{dk}{k} + \frac{w\ell}{y}\frac{d\ell}{\ell}.$$

Here, $r$ and $w$ replaced the two marginal products: this is what follows from taking first-order conditions of the profit maximization problem

$$\max_{k,\ell} F_t(k, \ell) - rk - w\ell.$$

We finally let $1 - \alpha$ denote labor's share of income, which we know from NIPA; $\alpha$ here of course may depend on time. Then the CRS assumption gives us the capital share as $\alpha$: total input costs equal output, thus delivering zero profits. This equation allows us to obtain an account of the sources of growth:

$$\frac{dy}{y} = \frac{\partial F}{\partial t}\frac{dt}{F} + \alpha\frac{dk}{k} + (1 - \alpha)\frac{d\ell}{\ell}. \tag{2.1}$$

The term $\frac{\partial F}{\partial t}\frac{dt}{F}$ is labeled the *Solow residual*, because it can be calculated as a residual of the growth in output that cannot be accounted for by the growth in capital and labor (the second and the third term of the right-hand side). We see that output growth is a weighted sum (over small intervals in time, as we have used derivatives) of capital input growth and labor input growth, where the weights are their respective income, or cost, shares, plus the Solow residual, which is a measure of how the production possibility frontier has moved out over time.

The Solow residual in (2.1) expresses the direct effect of time on production (in percentage terms). In general, this effect depends on at which input pair $(k, \ell)$ $F_t$ is evaluated. If one makes further assumptions on how technology shifts the production function, it is however possible to derive specific series for technological change that are independent of the economy's current capital-labor mix. We will discuss two such assumptions because they are commonly used; they are by no means the only ones imaginable, but especially the second one will play a key role later.

One assumption is to let $F_t(K_t, L_t)$ takes the form $z_t F(K_t, L_t)$, where $F(K_t, L_t)$ is a time-independent function. In this case, equation (2.1) can be expressed as

$$\frac{dy}{y} = \frac{dz}{z} + \alpha\frac{dk}{k} + (1 - \alpha)\frac{d\ell}{\ell}. \tag{2.2}$$

Here, $z$ is TFP: total factor productivity. It can equivalently be thought of as a common, "Hicks-neutral" factor multiplying both inputs.[5]

[5]This is true since we have assumed that $F$ is homogeneous of degree 1 in $(k, \ell)$.

Alternatively, we can make the assumption that technology is *labor-augmenting*: we define $F_t(K_t, L_t) = F(K_t, z_t L_t)$, again with a time-independent function $F$. Now $z$ has a different meaning. In this case, since $\frac{\partial F}{\partial t} = \frac{\partial F}{\partial \ell}\ell$, we can write the original growth-accounting equation (2.1) as

$$\frac{dy}{y} = (1-\alpha)\frac{dz}{z} + \alpha\frac{dk}{k} + (1-\alpha)\frac{d\ell}{\ell}, \tag{2.3}$$

after having replaced $\frac{\partial F}{\partial \ell}$ by $w/z$ and recognized that $w\ell = (1-\alpha)F$.[6]

We now show some results from growth accounting, implemented the way Solow came up with it. In practice, it is important to take into account how the quality of the inputs change over time; in particular, one may want to adjust labor input to account for human capital accumulation, or else part of the Solow residual will reflect the increasing quality of this input.

We first look at the traditional measure of productivity: labor productivity. Figure 2.7 shows an average growth rate of a little below two percent per year, with significant ups and downs; currently, we are in a down period.

Figure 2.8 shows a full time series (along with those for some other countries; perhaps remove here) over a longer time period. These series are smoothed.

We thus see stable, positive labor productivity growth, hovering around two and a half percent per year. How about the growth in total factor productivity? Figure 2.9 gives the answer from two sources, now in un-smoothed form. Here too, we see growth at a little below two percent per year, with significant movements up and down that we will return to later.

Figure 2.10 shows smoothed data for TFP growth over a longer time period. The patterns are similar.

### 2.1.5 The dynamic system

We are now equipped with measures of output and input aggregates, as well as with a measure of aggregate technology, in the form of TFP. Solow's next step was to use his framework to probe further into the mystery of the all but constant capital-output ratio.

Step one in this endeavor was to explicitly link time periods (years) by noting that tomorrow's capital aggregate stock is today's stock, plus new investment minus the part of capital that has depreciated.[7] With a constant rate of capital depreciation $\delta$, this yields

$$k_{t+1} = (1-\delta)k_t + i_t. \tag{2.4}$$

We see an equation that is consistent with the long-run data, if all the variables appearing in it grow at a common rate. We can rewrite the above equation as

$$\frac{k_{t+1}}{y_{t+1}}\frac{y_{t+1}}{y_t} = (1-\delta)\frac{k_t}{y_t} + \frac{i_t}{y_t}$$

[6] Similarly, one could define a capital-augmenting technology series by letting the $z$ multiply capital.

[7] A similar accumulation equation can be formulated for human capital. We delay this discussion until our chapter on growth.

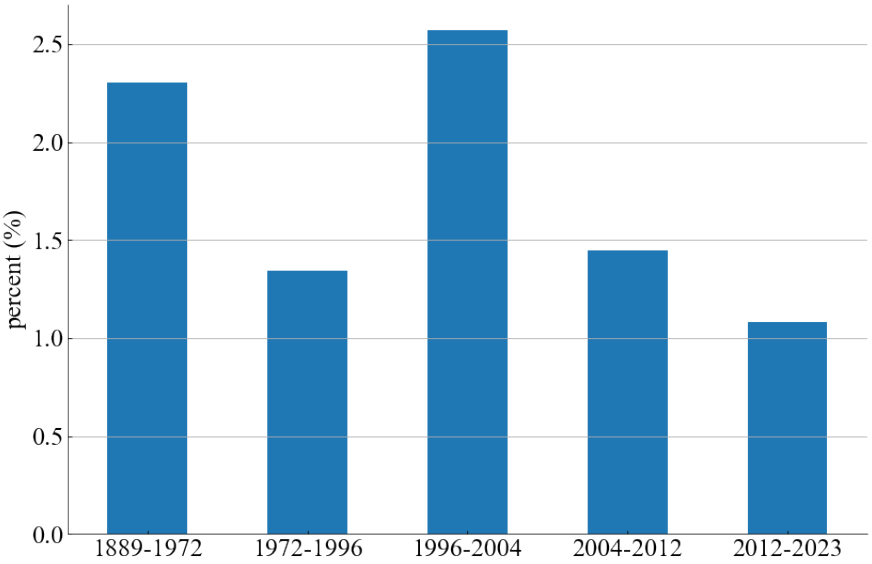



Figure 2.7: Labor productivity in the U.S., sub-periods

**Source:** Period 1889-1948: U.S. Census Bureau, Historical Statistics of the United States, Colonial Times to 1970, Part II, Series W 1, pp. 948, see here.here. Period 1948-2023: U.S. Bureau of Labor Statistics and FRED Quarterly total economy hours worked (in billions of hours) series are from BLS "Hours Worked in Total U.S. Economy and Subsectors". Data here. Quarterly real GDP (in billions of 2017 dollars) is from FRED Real gross domestic product (GDPC1).
**Note 1**: Data constructed following Robert Gordon (2012). Gordon defines "labor productivity as real GDP divided by an unpublished quarterly BLS series on hours for the total economy, including the private economy, government, and institutions." Percentage logarithmic growth rates are calculated between the first quarter of each of the listed years, e.g., 1948:Q1 to 1972:Q1. To extend the series back from 1948 to 1891, annual NIPA data on real GDP prior to 1929 are ratio-linked to the real GDP data of Balke-Gordon (1989), and the BLS hours data prior to 1948 are ratio-linked to the man-hours data of Kendrick (1961, pp. 330-32)." **Note 2**: Unlike Gordon (2012), to extend the series back from 1948, we used Census Bureau Hisitorical Statistics. Then the series for both periods are then re-indexed to 1948=100.

and denoting the (net) growth rate of output as $\gamma_t$, this equation can be expressed as

$$\frac{k_{t+1}}{y_{t+1}}(1+\gamma_t) = (1-\delta)\frac{k_t}{y_t} + \frac{i_t}{y_t}.$$

Clearly, if capital, investment, and output all grow at a same constant rate $\gamma$ and $i_t/y_t$ is equal to a constant value $s$, then this equation is consistent with capital-output ratio that does not change over time:

$$\frac{k_t}{y_t} = \frac{s}{\gamma+\delta}$$

for all $t$. The investment-output ratio is not constant over time; in particular, it fluctuates

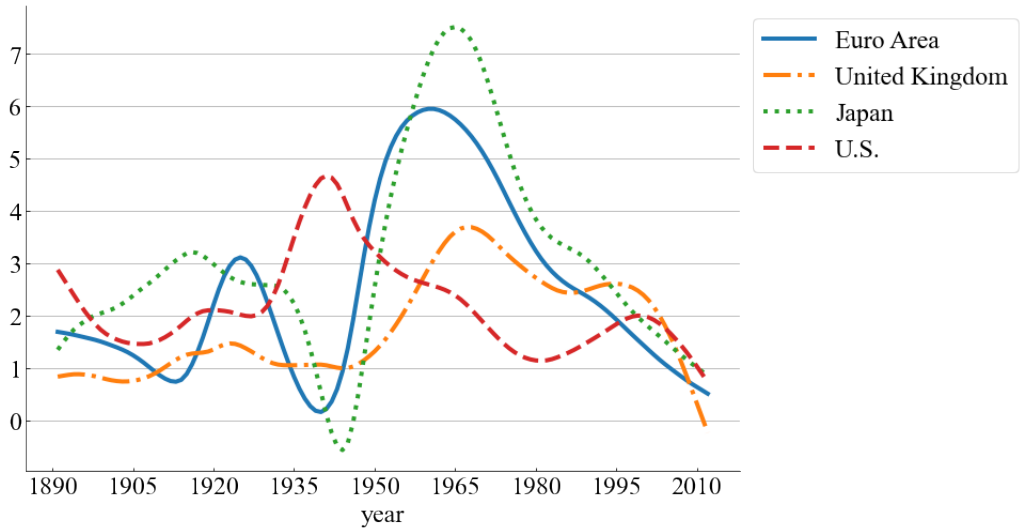



Figure 2.8: Labor productivity for a selection of countries

**Source:** Bergeaud, A., Cette, G. and Lecat, R. (2016) "Productivity Trends in Advanced Countries between 1890 and 2012," Review of Income and Wealth (see here). Hodrick-Prescott-filtered annual growth of labor productivity per hours worked. **Note**: Bergeaud et al. (2016) states: "In order to establish the stylized facts of productivity growth, we smooth the annual productivity growth rate over the period using the Hodrick–Prescott filtration (HP). Considering the very high volatility of our data, the choice of the lambda coefficient, which sets the length of the cycle we capture, is of paramount importance. Setting too high a value for lambda would tend to absorb smaller cycles, while setting too low a value would result in major cyclical effects being considered to be trends, especially around WWII. We decided to focus on 30-year cycles, which implies a value of 500 for lambda, according to the HP filter transfer function."

significantly. The consumption-output ratio for the U.S. is plotted in Figure 2.11; here, consumption is defined as private plus government. One minus this measure is close to the ratio of investment (again, private plus government) to output, since net exports are near zero in the U.S. as a fraction of GDP.[8] We see significant movements in $s$ early in the 20th century and thereafter small movements, possibly with a slight downward trend. But the overall assumption of a constant investment, or saving, rate appears to be a good one.

The above discussion allows a mechanical account of how one can interpret the capital-labor ratio: it is a simple function of the rate of saving, the economy's growth rate, and the rate of depreciation of capital, along what has been labeled the *balanced growth path*. The numerical values can be squared, too: with a depreciation rate of around 0.08 and a net growth rate of around 0.02, a saving rate of 0.30 delivers a capital-output ratio of 3.

However, this account still does not give an answer to the question why: why is the economy always (almost) at this value? That is, it explains if the capital-output ratio is 3 at

[8]This comes from the national accounting identity $Y = C + I + NX$, where $C$ and $I$ both contain private as well as government expenditures.

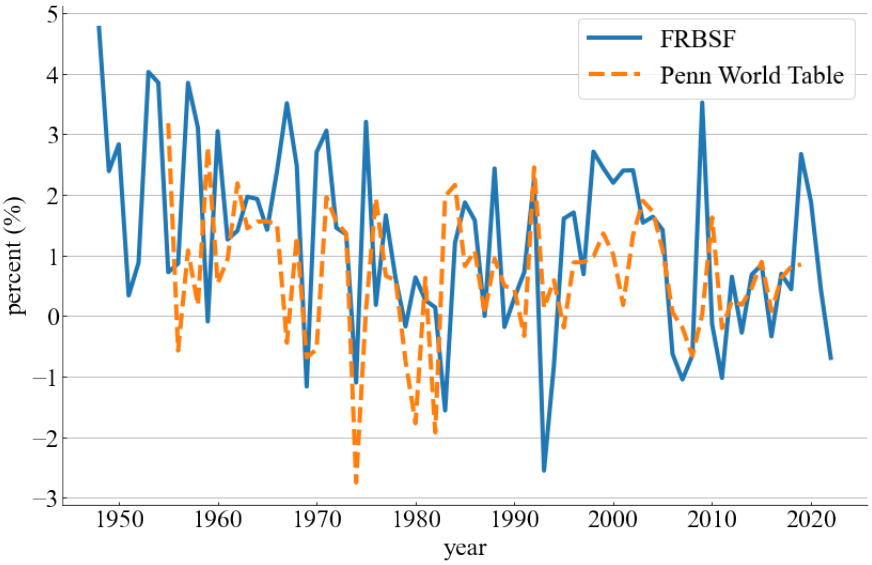



Figure 2.9: TFP in the U.S., two measures

**Sources:** Series 1: FRED, "Total Factor Productivity at Constant National Prices for United States (RTFPNAUSA632NRUG)," reported by Penn World Tables. Available for 1955-2019 Series 2: Utilization-adjusted quarterly TFP series for the US Business Sector, 1948-2022 John G. Fernald, "A Quarterly, Utilization-Adjusted Series on Total Factor Productivity." FRBSF Working Paper 2012-19 (updated March 2014).

some point in time, it will remain 3. But why is it 3 to start with? Solow found an answer.

Solow considered the following dynamic system, which is the logical implication of the above reasoning:

$$k_{t+1} = sF(k_t, (1+\gamma)^t \ell) + (1-\delta)k_t$$

for all $t$. This system is almost exactly what we have looked at before. First, we have replaced $i$ by $sy$, since we take saving to be a constant fraction of output, in consistency with the above evidence. Second, there are two additional assumptions: we have set hours worked to a constant, $\ell$, and we have made technology growth appear only in a *labor-augmenting* form, i.e., technical change is equivalent to raising labor input, or the quality of labor input. This was based on the hunch above: if labor units become more and more productive due to technological change, the return to capital does not have to fall due to a higher ratio of capital to hours worked. It turns out that this assumption is actually crucial. Namely, Uzawa (1961) proved that, for a production function of two inputs where one is limited—labor hours, in this case, are constant—to admit exact balanced growth in a model of this kind, technological change has to take this form.[9] So we can obtain balanced growth if and only if the above assumptions are met; whether labor input is constant or declining does not matter.

[9]The proof of Uzawa's important theorem can be found in Appendix 3.A, which discusses the Solow model in detail.

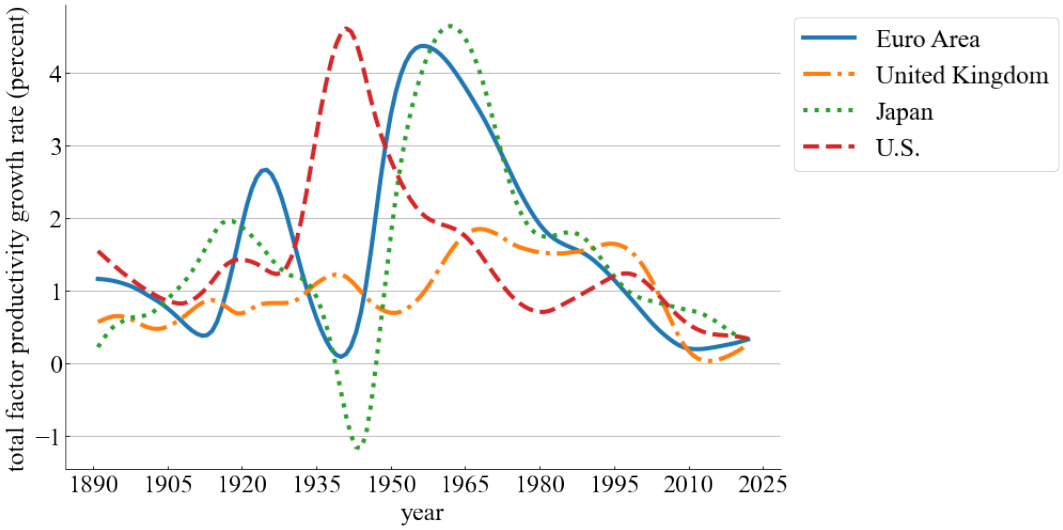



Figure 2.10: Historical TFP for a broader set of countries

**Source:** Bergeaud, A., Cette, G. and Lecat, R. (2016) "Productivity Trends in Advanced Countries between 1890 and 2012," Review of Income and Wealth (see here). Hodrick-Prescott-filtered annual growth of total factor productivity (TFP) per hours worked. **Note**: Bergeaud et al. (2016) states: "In order to establish the stylized facts of productivity growth, we smooth the annual productivity growth rate over the period using the Hodrick–Prescott filtration (HP). Considering the very high volatility of our data, the choice of the lambda coefficient, which sets the length of the cycle we capture, is of paramount importance. Setting too high a value for lambda would tend to absorb smaller cycles, while setting too low a value would result in major cyclical effects being considered to be trends, especially around WWII. We decided to focus on 30-year cycles, which implies a value of 500 for lambda, according to the HP filter transfer function." TFP growth rate for years after 2012 is from OECD data. OECD (2023), Multifactor productivity (indicator). doi: 10.1787/a40c5025-en (Accessed on 27 December 2023). Aggregate TFP growth rate for EU19 countries is calculated by taking a weighted average of the growth rates of each country where weights are the share of each country in the total GDP of the EU19 in each year. OECD (2023), Gross domestic product (GDP) (indicator). doi: 10.1787/dc2f7aec-en (Accessed on 27 December 2023).

The dynamic system can be rewritten

$$(1+\gamma)\tilde{k}_{t+1} = sF(\tilde{k}_t, \ell) + (1-\delta)\tilde{k}_t;$$

the equation is obtained by means of a simple variable transformation—$\tilde{k}_t \equiv k_t/(1+\gamma)^t$, a stationary variable if $k$ grows at the net rate $\gamma$—and division by $(1+\gamma)^t$.

Now note, first, that there is a constant solution to this system: $\bar{\tilde{k}}$. It is the unique value that (under some minimal conditions) solves

$$(1+\gamma)\bar{\tilde{k}} = sF(\bar{\tilde{k}}, \ell) + (1-\delta)\bar{\tilde{k}}.$$

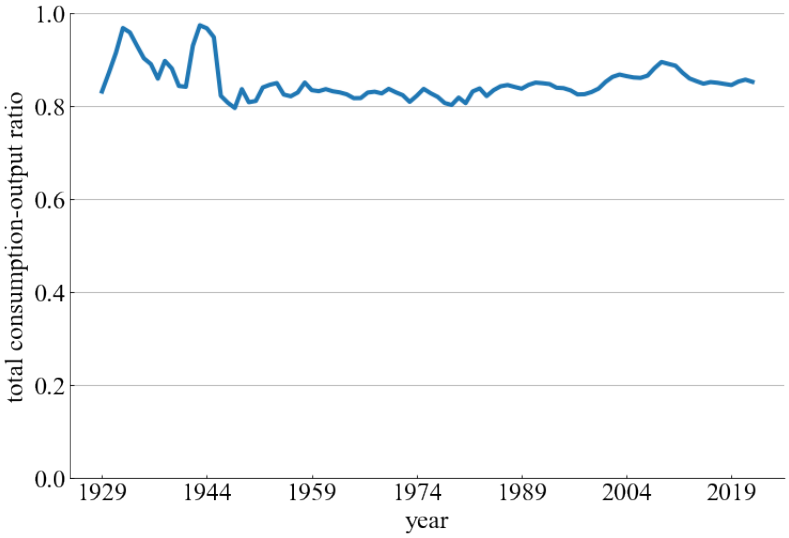



Figure 2.11: Ratio of total consumption to output

**Source:** BEA, NIPA Table 1.1.5 Calculated as the ratio of two series: Numerator: Sum of Personal consumption expenditures (DPCERC) and Government consumption expenditures and gross investment (A822RC), Annual, Millions of dollars Denominator: Gross domestic product (A191RC), Annual, Millions of dollars The figure plots the ratio between total consumption expenditures (sum of all goods and services) relative to the GDP.

So in a situation where $\tilde{k}_0 = \bar{\tilde{k}}$, we obtain $\tilde{k}_1 = \tilde{k}_2 = \cdots = \bar{\tilde{k}}$. That is, capital grows at a constant rate $\gamma$, since $k_t = \tilde{k}_t(1+\gamma)^t$. The same is then true for investment and output, since $y_t = F(k_t, (1+\gamma)^t h) = (1+\gamma)^t F(\bar{\tilde{k}}, h)$ follows when $F$ is CRS.

We have established that if the initial capital stock has a particular value, this economy will grow at a constant rate. However, the second, and most remarkable, thing to note about this system is a convergence property. Solow showed, under very weak conditions, that for any given initial condition on capital, the ensuing capital sequence, and hence the economy's aggregate variables will *converge* to the balanced growth path. We will explain this in detail in the next chapter. Intuitively, it is the neoclassical production function—the very feature that was used to make sense of why a higher input price is consistent with lower input use, and at the same time, under competition, a higher marginal productivity—that explains convergence: when capital is comparatively low, growth is comparatively fast, because its marginal productivity is high, delivering higher capital accumulation per unit of capital. Conversely, when capital is high, its growth rate is low, so there is movement back toward the balanced growth rate no matter where the initial capital stock is.

The neoclassical convergence mechanism de-mystifies why the value of the total capital stock on average is worth roughly 3 times annual output: it doesn't have to be exactly 3 at all times, but deviations bring it back toward 3. It also produces a dynamic framework that, as we shall see in the rest of the textbook, has become the workhorse model for macroeco-

nomics, much because it offers a coherent account of how the macroeconomic aggregates have evolved historically. The analysis of business cycles, for example, builds on the neoclassical framework with various stochastic shocks added to the dynamic system. These shocks could be "supply shocks" or "demand shocks," but they have in common that their *propagation* through the economy—how the macroeconomic variables respond in the short run, which can differ greatly across different kinds of shocks—eventually takes us back toward the balanced growth path. This is ensured by the convergence property of the system.

The final, unresolved, issue is that some of our assumptions above are mere mechanical descriptions of the data: investment is a constant rate $s$ of output, and hours worked are constant at $\ell$. In the real economy, these two features should be the results of conscious choices made by households. Consumption choices are made continuously, and people can influence how much they work. Moreover, a theory that adds consumer choice will also allow us to make welfare statements, which the analysis so far does not. We will turn to households' choices momentarily: we will "rationalize" the $s$ and the $\ell$ based on microeconomic theory—utility maximization. However, let us first briefly revisit the movements of input shares.

### 2.1.6 Input shares

On a balanced growth path, the shares of income paid to capital and to workers, respectively, are constant. This is because in the theory Solow proposed, $rk$ grows at the rate of output ($r$ is constant but $k$ grows at the rate of output) and $w\ell$ does as well ($w$ grows at the rate of output and $\ell$ is constant, at least over the postwar period). Thus, as shares of output, they should be constant. Have they been? We see in Figure 2.12 that there is rather remarkable constancy over time.

In the most recent years, a downward trend of the labor share can been noticed, however, as is evident if the time interval is restricted. Figure 2.13 illustrates, for a few developed countries, that the labor share has been declining. Figure 2.14 shows the same fact as a global average. The downward trend has been subject to much scrutiny and research recently, but for now we will maintain the stylized fact as "labor share is close to constant over time."

Virtually all applied macroeconomic studies employ an aggregate production function that is of the Cobb-Douglas variety, i.e., we have $F(k,\ell) = Ak^{\alpha}\ell^{1-\alpha}$, where $\alpha$ is constant over time. This function has the property that $F_k(k,\ell)k/F(k,\ell) = \alpha$, i.e., the income shares under perfect competition are independent of the values of $k$ and $\ell$. We see in the data that although the shares are not literally constant, their movements are relatively minor, even as economies go through recessions and booms. The Cobb-Douglas function thus conveniently generates a decent approximation to the data, which is why it is so often used.

### 2.1.7 Summing up

We summarize the main stylized (notice the repeated appearance of the word "roughly" in the descriptions below) facts:

1. output per capita has grown at a roughly constant rate

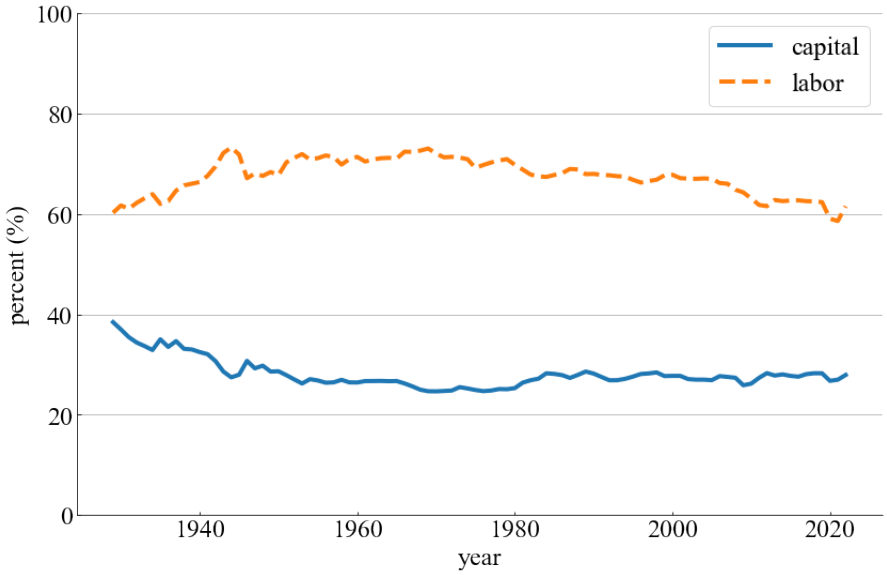



Figure 2.12: U.S. factor shares over time

**Source:** NIPA Table 2.1 Factor shares are calculated as compensation of employees and capital income divided by personal income. Capital income is calculated as proprietors income with inventory valuation and capital consumption adjustments plus rental income of persons with capital consumption adjustment plus personal income receipts on assets.

2. the capital-output ratio (where capital is measured using the perpetual inventory method based on past consumption foregone) has remained roughly constant
3. consumption as a fraction of output has been roughly constant
4. the wage rate has grown at a roughly constant rate equal to the growth rate of output
5. the real interest rate has been roughly constant, seen over a longer period of time
6. labor income as a share of output has remained roughly constant
7. hours worked per capita have been roughly constant over the recent half a century.

These facts are consistent with aggregates obeying a neoclassical structure whose core is a CRS production function with labor-augmenting technical change and decreasing marginal products in each input, constant labor supply, a constant rate of capital depreciation, and a constant investment-output (saving) ratio.

### 2.1.8 Rationalizing saving and labor-supply choices

We now discuss how macroeconomists theorize further to make sense of households' observed choices for saving and labor supply. The aim is to add a richer microeconomic structure that

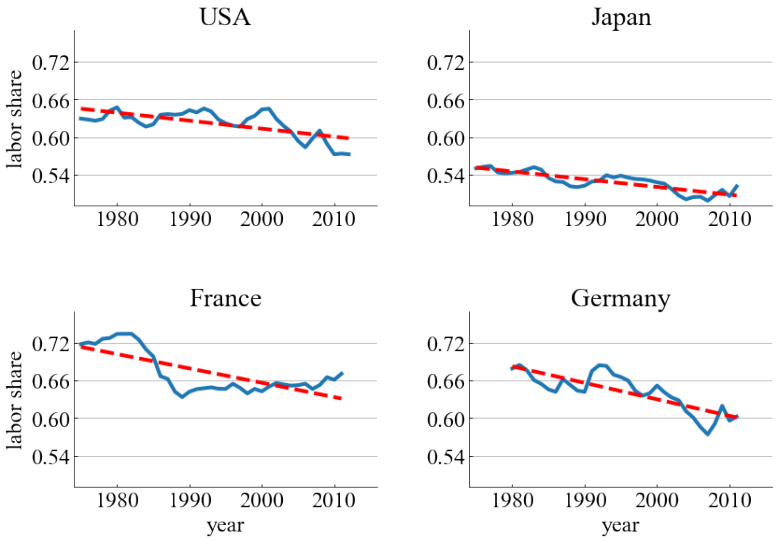



Figure 2.13: Labor share

**Source**: Karabarbounis, Loukas, and Brent Neiman. 2014. "The Global Decline of the Labor Share." Quarterly Journal of Economics, 129(1), 61-103. **Note**: No corporate labor share data is available for Japan, thus total labor share is plotted instead. Moreover, the plotted series for Japan looks different from the one in the paper.

is qualitatively, and quantitatively, consistent with the data. The central object here will be utility functions and a key question is whether there are utility functions consistent with the observed choices. A follow-up question is whether, given such utility functions, the main dynamic properties—in particular convergence—will be maintained. We briefly address the first of these questions in the present chapter and postpone the second until a later chapter.

Before we introduce utility functions, we will briefly discuss the population structure, along with our general approach.

### Time and people

We begin with time and then describe our people.

**Time** The dynamic system above was described in discrete time, i.e., time periods are integers. It is also possible to describe the system in continuous time (here, we would use $t$ as an argument of our functions and not as a subscript, e.g., $y(t)$ vs. $y_t$). The main workhorse model developed in this book is using discrete time, mostly because we consider it somewhat easier to teach.[10] There is, however, no substantive difference between the two approaches and both are common in practice.

[10]The reason is that it is easier to make concrete; for example, dynamic optimization can be more straightforwardly connected to basic optimization theory in one or more (a finite number of) variables; when there are stochastic shocks, continuous time requires more investment still, and becomes a bit more abstract as

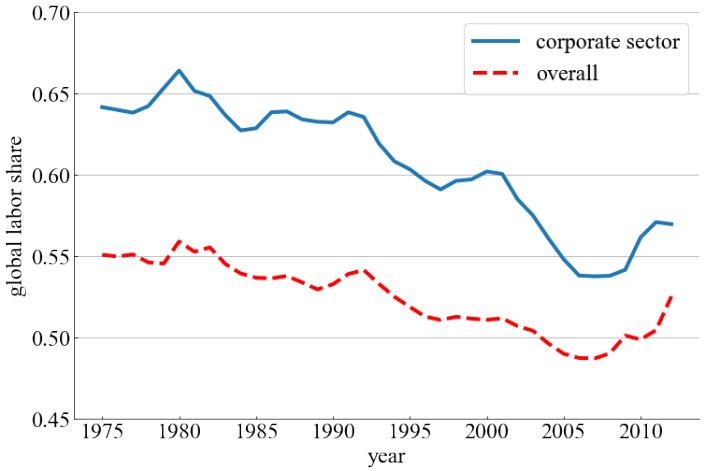



Figure 2.14: Global labor share

**Source**: Karabarbounis, Loukas, and Brent Neiman. 2014. "The Global Decline of the Labor Share." Quarterly Journal of Economics, 129(1), 61-103.
Note: Global labor share of the corporate sector is the average of country-level labor shares weighted by corporate gross value added measured in US dollars at market exchange rates. Global labor share of the overall economy is the average of country-level labor shares weighted by GDP measured in US dollars at market exchange rates.

Second, we will almost always assume that time is infinite; the main exception is when we illustrate mechanisms in, say, two-period models. The reason for adopting infinite time is mostly practical but to some extent also conceptual. The same economic structures as we study under the infinite-horizon assumption can also be analyzed under the assumption that time is finite. However, then there is a built-in non-stationarity: time itself obtains importance—because it captures how far we are from "the end" and, certainly, if the end would be near, many decisions in the economy would change radically. So long as decisions and outcomes today are not more than marginally affected by the exact end date, it is simply more convenient to consider time to be infinite; then, time is not important in itself. Most macroeconomic models with an end date $T$ have the feature that $T$ is virtually irrelevant for decisions far from $T$ (i.e., decisions at $t \ll T$); when the models we look at do not have this property, we will point it out.

**People** At any point in time, the economy is inhabited by individuals, or households, and our maintained assumption will be that they are utility maximizers. By assuming utility maximizers we mean we "rationalize" the choices we observe in the data by treating these choices as optimal given well-defined utility maximization problems. Thus, we back out how consumers must value consumption (and leisure) based on their own behavior. This approach is powerful, since it allows us to make statements that directly reflect the welfare

well.

of the population. In the analysis of economic policy, being able to make welfare statements is crucial, since it allows us to compare policies on normative grounds.[11]

The most common macroeconomic model has a very stark, and highly stylized, description of the population: there are many identical individuals alive at any one point in time, and these individuals, moreover, live forever. "Identical" here means that they have the same utility functions and constraints and, therefore, face identical maximization problems. "Many" is important in that it allows us to make sense of price-taking behavior, but the number itself does not matter. "Live forever" sounds much sillier than it is: the notion here is that people are representatives in a dynasty, i.e., a consumer at time $t$ values not only her own consumption but also that of her children, grandchildren, and so on. Given how people appear to care about their children and grand-children—given the resources they spend on them while alive and also in the form of bequests—it does seem a very natural starting point. It should be pointed out that the assumption here is not just that people care about their offspring: they are truly altruistic about their offspring. In other words, their preferences are aligned. We will touch on the possibility that they are not later on in the text.

Of course, macroeconomic models with more explicit and realistic structures are also common. First, building in a life-cycle pattern—since, at the very least, as a person ages, needs and abilities change—is common.[12] Then, the addition of children and altruism toward them would still deliver a model that in its macroeconomic features is quite similar to the simpler model. If life-cycle models do not have altruism toward children, such as in the simple "overlapping-generations model," they can (but do not necessarily) behave differently than than dynastic model. We will discuss this in some detail too.

Second, nowadays a "heterogeneous-agent" framework has been developed and become a very commonly used workhorse for macroeconomic analysis. In it, households are different in a variety of ways (income, wealth, preferences, etc.). Such models share many properties with the simplest dynastic setting, but of course also add richness. One aspect of these models is that they allow us to jointly study macroeconomics and inequality. Another one is that they behave quite differently (and, arguably, in ways closer aligned with the data) in some respects, in particular in terms of how the economy responds to various shocks and to policy changes. From our perspective here, however, the key observation is that heterogeneous-agent models can be viewed as extensions of the dynastic representative-agent setting rather than as fundamentally different. Third, some macroeconomic models also have richer models of the household structure, explicitly incorporating couples and children. This is, so far, a smaller literature, however.

[11] There is a strand of macroeconomic literature that builds in behavioral elements; after all, the literal interpretation of the data that consumers make perfect decisions at all time is of course very strong. One interpretation of this research is as a robustness check: if minor departures from rationality create major changes in the results, we should perhaps worry. This research is ongoing and we have no systematic treatment of it in this text.

[12] Death can then be modeled as deterministic or stochastic.

### Preferences

Households will be assumed to have utility functions that are time-additive, with consumption in periods $t$ and $t+1$ evaluated as $u(c_t) + \beta u(c_{t+1})$, where $u$ is strictly increasing and strictly concave. Thus, the same function $u$ is used for consumption in both periods, but there is a weight $\beta$ on $t+1$ consumption. The fact that $u$ is the same for both consumption goods implies that consumption in both periods are normal goods: with more income, consumers would like to consume more of both goods, which seems very reasonable in this application.[13] Another aspect of this setting is an element of *consumption smoothing*: there is decreasing marginal utility to consumption in both periods so spending all of an income increase in one period will in general never be optimal. Furthermore, $\beta < 1$ captures impatience, or a probability of death—or any other reason for down-weighting future utility—and will be a typical assumption.

### Choice

An important part of the text will explain intertemporal choice from first principles: different methods for solving intertemporal problems, with and without uncertainty, along with a number of important macroeconomic applications. Here, the purpose is to very briefly explain the key steps, heuristically, in order to account for the growth facts.

Conceptually, the way consumers make decisions—if able to choose when to consume their income—is according to basic microeconomic principles: so as to set their marginal rate of substitution equal to the relative price. We will now go through the two key choices using these principles. Before looking at the specific choice examples, let us note that by rationality, in the present context, we include the notion of *perfect foresight*: given that no shocks are occurring, consumers know what prices prevail not only today but also in the future. If there are shocks—and we will study shocks later in the text—rationality is interpreted as *rational expectations*, i.e., knowing the probability distribution for variables in the future.

**Consumption vs. saving** The relative price between consumption at $t$ and $t+1$ is the *real interest rate*: it is the amount of goods at $t+1$ that a consumer can buy for one unit of the good at $t$. We will denote the gross real interest rate between $t$ and $t+1$ $R_{t+1}$ here. The marginal rate of substitution between the goods can be obtained by defining an indifference curve relating to these two goods. Thus, write $u(c_t) + \beta u(c_{t+1}) = \bar{u}$, take total differentials, i.e., $u'(c_t)dc_t + \beta u'(c_{t+1})dc_{t+1} = 0$, and then solve for $-dc_{t+1}/dc_t$. Setting the resulting expression equal to the gross real interest rate, we obtain

$$\frac{u'(c_t)}{\beta u'(c_{t+1})} = R_{t+1}.$$

[13] Try to verify this by showing that, if the income allocated to the two goods is increased and consumers can choose between the goods freely, given a fixed relative price, both consumption levels will rise.

This equation, which equivalently can be written

$$u'(c_t) = \beta u'(c_{t+1})R_{t+1},$$

is commonly referred to as the *Euler equation* and it is a central element of macroeconomic theory. It says that an optimizing consumer sets the marginal utility loss of saving one consumption unit for tomorrow (the left-hand side) equal to the gain tomorrow in consumption terms (the right-hand side), that is, $R_{t+1}$ (the return on the savings) times the marginal utility of each unit tomorrow, $\beta u'(c_{t+1})$.

We argued above that a constant saving rate will imply convergence toward a constant level of capital relative to technology, i.e., $\hat{k}_t$ becomes constant—this is implied by Solow's analysis, which we will elaborate more on later. In particular, a constant saving rate is associated with aggregate consumption growing at a constant rate. We also saw that balanced growth requires a constant real interest rate. So the question now is whether individuals, when faced with a constant interest rate, would choose a consumption path that grows at a constant rate, despite the desires to smooth consumption over time. The question boils down to whether the Euler equation could hold for constantly growing consumption, and we now address this question.

Let us begin with the answer: there is a sharp characterization saying that the utility function $u$ is consistent with exact balanced growth if and only if it is a power function. It is easy to verify the "if" part: $u'(c_t)/u'((1+\gamma)c_t)$ becomes constant if $u(c)$ is a power function, and that constant contains the growth rate. Hence, under a power utility function a constant interest rate will lead the consumer to choose a constant consumption growth rate. What that growth rate is precisely depends on the interest rate, on $\beta$, and on the curvature of $u$, but not on how wealthy the consumer is. The precise class of functions is captured by

$$u(c) = \frac{c^{1-\sigma} - 1}{1 - \sigma} \tag{2.5}$$

with $\sigma > 0$ and $\sigma \neq 1$; the case where $\sigma$ approaches 1 yields $u(c) = \log c$.[14] The "only if" result is harder to prove, but within reach; the proof is in the appendix to Chapter 4.

The above discussion has been carried out heuristically and in terms of simply selecting two adjacent time periods, without reference to how many periods the consumer lives in total. This discussion also means that the arguments above hold whether in dynastic or overlapping-generation economies, or some combination of these, so that the restrictions placed on preferences in order to be consistent with balanced growth hold rather generally.

Note that the utility-function characterization came from a requirement of exact balanced growth. One could imagine growth paths that are asymptotically balanced and still match our historical data. In the limit, however, the underlying utility functions would then look like our $u$ defined in equation (2.5).[15]

[14] To understand the $\sigma = 1$ case, take the limit as $\sigma$ goes to 1 but use l'Hôpital's rule. One obtains $\lim_{\sigma\to 1} \frac{d(c^{1-\sigma}-1)/d\sigma}{d(1-\sigma)/d\sigma} = \lim_{\sigma\to 1} \frac{-c^{1-\sigma}\log c}{-1} = \log c$.

[15] An example is $u(c + 0.01)$, where $u$ satisfies (2.5).

Let us summarize: in macroeconomic modeling, where consumption-saving choices are viewed to come from optimizing consumers, the utility function employed in almost all applications is a power function. This choice is made because we want our frameworks to account for the basic historical facts. For utility functions that are not in this class, we would therefore, for example, see very different saving rates 100 years ago than today; typically, saving rates would go to zero or to one over time as the economy grows, and balanced growth as observed in the data would not be possible.

**Labor vs. leisure** Turning to labor supply, the idea is to allow households to choose how much to work. But do people choose how much they work? In many countries, work hours are regulated, and your own specific employer may not offer you much choice. From the historical, macroeconomic perspective, however, there is no doubt that labor supply is a choice. First, we have seen that in the longer run, work hours per adult have fallen appreciably. Second, looking across countries, a similar pattern emerges (one we will revisit later): households in rich countries work significantly less than do households in poor countries. These patterns reflect differences (over time and across space) that involve both the intensive margin, i.e., how many hours each individual works when she works, and the extensive margin, i.e., whether a given individual works at all and the fraction of her lifetime the individual works positive hours. We thus see differences across time and space as reflecting choice and, for example, labor-market regulations stipulating a 40-hour workweek should be seen as an outcome reflecting people's choice at the time these regulations were decided upon.[16] In sum, and especially with our long-run perspective, we consider the choice of how much labor to supply as a very natural one.

In more concrete terms, and focusing on the intensive margin only, the period utility function has to allow the agent to explicitly value leisure. Then, we can choose to have leisure as an argument: $u$ would depend on $(c, l)$, where $l$ is leisure.

We thus need to insist on balanced growth in consumption jointly with a constant labor supply in the long run. We therefore need the condition $\frac{u_l(c_t,l_t)}{u_c(c_t,l_t)} = w_t$ to be met at all points in time on a balanced growth path where, as shown above, $c$ and $w$ grow at the same rate.

Let us begin with the requirement that work hours are constant along a balanced path, as motivated by the postwar U.S. data. Then, just like in the above case, there is a sharp characterization of what preferences are consistent with an exact balanced-growth path. In particular, they are consistent with an exact balanced-growth path if and only if the utility function is of the form $u(c) = \frac{(cg(l))^{1-\sigma}-1}{1-\sigma}$, where $g$ is strictly increasing and such that $cg(l)$ is strictly quasiconcave. It is again straightforward to show the "if" part—it is a matter of looking at the Euler equation and the first-order condition for the hours choice jointly and verifying that balanced growth is consistent with the derived equations—but, as before, more demanding to show the "only if" part.

A second possibility is that we would like our theory to be consistent also with the longer-run data and across countries with very different standards of living. This is actually possible,

[16] Of course, in countries that are not democratic, one could imagine that work hours are dictated. But even then people's preferences would likely play a role.

since it also seems—though it was perhaps not clear from the earlier graphs—that as output has grown at a roughly constant rate, hours worked have declined at a roughly constant rate. The rate of hours decline is only, say, a third of a percent per year, but over time these small changes accumulate and become visible. It turns out, moreover, that an outcome with exact balanced growth where hours fall at a constant rate is possible as an outcome within our framework if and only if the utility function satisfies $u(c,l) = \frac{c^{1-\sigma}v(c^{\frac{\nu}{1-\nu}}l)-1}{1-\sigma}$, $\nu > 0$. With technology and output growth proceeding at a net rate $\gamma$, we then have hours grow at $(1+\gamma)^{-\nu}$ and consumption at $(1+\gamma)^{1-\nu}$. Wages now grow slightly faster than output, and the labor share remains a constant share of output.[17]

The case just mentioned collapses to the one described earlier with $\nu = 0$. With $\nu = 0$, as wages grow along a balanced path, the substitution effect (making it more beneficial to work at higher wages) exactly cancels the income effect (making it more attractive to choose more leisure as income rises); if $\nu > 0$, the income effect is slightly stronger. Thus, the interpretation is that at lower levels of income, people work more because consumption is more important to them.

In conclusion, we have now arrived at a utility-function specification that is (the only one) consistent with choosing a constant saving rate and constant (or constantly declining, at a low rate) labor supply in the long run. The precise population structure and the length of households' lives can, however, satisfy a variety of assumptions and our main two applications below will be the representative-agent dynasty and the simplest overlapping-generations model.

## 2.2 The rest of the text

In the next five chapters, we will go over the main macroeconomic tools: (i) the Solow growth model (Chapter 3); (ii) dynamic optimization (Chapter 4); (iii) dynamic equilibrium theory (Chapter 5); (iv) welfare (Chapter 6); and (v) uncertainty (Chapter 7). The methods presented in these chapters are core material that will then be used and applied over and over. The methods will also make the material discussed in the present chapter more precise; for example, convergence in the Solow model will be discussed in detail, the maximization problems of consumers will be fully described and a convergence theorem under optimal saving will also be provided, a dynamic competitive equilibrium will be defined and characterized in its application to the growth model, the welfare properties of such equilibria will be discussed in some detail, and it will be made clear under what assumptions the methods and results extend straightforwardly to the case of uncertainty.

The last methodological chapter, Chapter 8, goes over some materials on the empirical methods. The chapter also discusses the issues of how we quantitatively evaluate the theoretical models in macro context.

[17] Constantly declining, or growing, hours worked do not pose a problem from the perspective of the production side.

We now very briefly discuss some of the key issues and contents in the applied chapters that then follow in the second part of our text.

### The applied issues: Chapters 9–22

**Chapter 9: consumption** This chapter looks more in detail at how key consumer choices are made on the individual level. It will be made clear that the very simplest consumption-saving model delivers predictions for marginal propensities to consume—these are a key component of macroeconomic propagation mechanisms—that are hard to square with micro data: they are too low. Imperfect insurance against individual shocks to labor-market outcomes, an a priori plausible addition to the basic model, will change this and bring the model closer in line with the data. At this point, heterogeneity will become an important element of the analysis and the connection between macroeconomics and inequality will appear for the first time.

**Chapter 10: labor supply** As just discussed in the present chapter, the choice of labor supply can be seen as an extension to basic consumption theory. Allowing for both an extensive and an intensive margin of choice is of particular relevance, the former referring to whether or not to work at all and the latter to how many hours to work, conditional on working. Given this foundation, it is of significant interest to understand how labor choice varies with wages: the wage elasticity of labor supply. There are different notions of elasticities and they are useful in different contexts, with a particular distinction between how hours vary over time in response to wage changes and how hours vary across countries. Another key distinction that will be discussed is individual vs. aggregate labor supply.

**Chapter 11: growth** Now the wide disparity of incomes across countries will be addressed. What explains why many countries remain poor, and more generally how do the distributions of GDP, consumption, and other key variables across all countries in the world evolve over time? What is the role of technological progress, and what determines it? Human capital accumulation and its role in the growth process will also be discussed. Relative to the material in the earlier chapters, this chapter will introduce some additional features and, at the same time, flesh out quantitative predictions and compare them with data. At the end of the chapter, there will be an attempt at providing quantitative answers to the core growth questions.

**Chapter 12: real business cycles** Already in Chapters 3 and 7 there are some glimpses into how the neoclassical model fares when the economy is hit by random shocks. The so-called *real business cycle* (RBC) model, as laid out by Kydland and Prescott in their 1982 paper, proposed that macroeconomic fluctuations are indeed to an important extent a result of unforeseen changes (mostly increases in) technology. In addition, and even more importantly, their paper introduced quantitative, microeconomics-based macroeconomic theory as a tool. Their idea was to formulate a stochastic, dynamic general-equilibrium (DSGE)

model, solve it numerically for parameter values that were plausible given micro data and long-run facts, and then use it to address macroeconomic phenomena such as fluctuations and the effects of policy changes. Virtually all analyses of macroeconomic fluctuations since have adopted their approach, though many of the frameworks that came later departed in various ways from Kydland and Prescott's basic RBC model. In particular, many other shocks were proposed, money was introduced, and numerous frictions to how markets operate were included. This chapter thus begins by describing the key facts: how macroeconomic aggregate fluctuate and correlate. It then describes the core model, if nothing else because many of the later models treat it as a benchmark. What is the jury's verdict on the role of technology shocks in accounting for business-cycle fluctuations? This question will be addressed in the later chapters. Finally, the chapter also discusses how to filter the raw data in order to focus on the business-cycle frequencies.

**Chapter 13: government** So far, very little has been said about the government. The U.S. government is sizable, and many other governments are much larger still as a share of total expenditures or employment. Governments spend resources and make transfers aimed at redistribution, e.g., from rich to poor and across age groups. The chapter will thus first describe and discuss the key facts pertaining to government variables. Second, the chapter will use our basic theory to examine, given some specified objectives, how different government financing schemes compare, both positively and normatively? Taxation, for example, tends to involve distortions, so some effort will be devoted to understanding its effects. Is it important that the government runs a balanced budget, or does debt management even matter? As a part of this effort, the chapter will give us an introduction to how one can formulate optimal policy problems aiming to maximize consumer welfare while being restricted to the use of distortionary taxes.

**Chapter 14: asset prices** The previous chapters will have studied many intertemporal issues, including borrowing and lending, but the analysis of asset markets, and the determination of asset prices in particular, is important in its own right. The chapter will thus describe the key facts—for example, asset prices fluctuate "wildly" and risky assets pay a much higher return on average than do riskless assets such as U.S. Treasury bonds—and then proceed to analyze these facts through the lens of our basic theories. A core framework is the so-called consumption Capital Asset Pricing Model (CAPM), which derives asset prices and their stochastic features in relation to explicit household choices over its stochastic consumption path. An asset of specific relevance is housing, which will also be discussed in some detail both empirically and using quantitative theory.

**Chapter 15: money** A very particular asset is fiat money. "Fiat" refers to the fact that money is intrinsically without value and, nowadays, the value of money is in no sense backed by any real objects (as it was historically in many economies: its value was backed by gold). This raises the question of why it has value at all. From this perspective, inflation means that money loses value. The chapter will go over basic data on inflation and basic theories

of how the value of money can be determined. It will briefly touch on the determination of exchange rates too; although the value of money in terms of real goods in any given economy does not necessarily fluctuate much, exchange rates do and the chapter will briefly address this volatility. A related concept is an asset's liquidity, and we will very briefly describe the Diamond-Dybvig model of banking, which offers a way of thinking of banks as transforming illiquid into liquid assets, along with an associated potential susceptibility to "bank runs." The chapter, finally, will provide a bridge into the next chapter by explaining how money is introduced in the so-called New-Keynesian model of business cycles.

**Chapter 16: nominal frictions and business cycles** Prices and wages appear to move sluggishly on in the micro data. The chapter will begin by documenting some key facts on this and also review some evidence suggesting that monetary policy can have real effects because prices and wages are "sticky." The New-Keynesian model will then be introduced here. This model has become a workhorse for central banks around the world. It builds on the RBC model but adds nominal frictions: costs associated with changing prices and wages. The extension to the RBC model involves introducing long-lived firms with market power: these firms set prices knowing that prices will be costly to change in the future. The framework also has a description of how the central bank behaves; in particular it introduces a notion of monetary "policy shocks," as an additional source of macroeconomic fluctuations. The chapter will also discuss the evidence on the role of monetary policy in accounting for aggregate fluctuations.

**Chapter 17: frictional credit markets** By many economists, the Great Depression is viewed to have in part been caused by frictions in the credit market, i.e., impediments to borrowing for firms. Similarly, the 2007–2009 Great Recession is also considered to have had its roots in financial-market malfunctioning. The chapter will begin by documenting some correlations that suggest that financial frictions might be important. It will then show how such frictions can be introduced into the core framework and how macroeconomic propagation sometimes, but not always, changes nature in the presence of such frictions. Policy measures to address the frictions will also briefly be touched upon.

**Chapter 18: frictional labor markets** Often, the rate of unemployement is even used to define the business cycle: it is highly countercyclical—rises in recessions and falls in booms. The chapter begins by reviewing not only the key facts on aggregate unemployment but also on individuals' movements in and out of jobs over time. It then introduces the most common framework for analyzing unemployment: the search and matching model. It begins by looking at worker search and then introduces a full general-equilibrium model with matching frictions as in Pissarides (1985). The resulting model is then confronted with data and the so-called Shimer puzzle is introduced and discussed. Finally, the chapter shows how the Pissarides model can be extended so as to incorporate capital accumulation and, thus, as such can be seen as an important extension of our basic macroeconomic framework.

**Chapter 19: heterogeneous consumers** In this chapter, we discuss inequality between households. The first part of the chapter views inequality as interesting in its own right and, hence, reviews both data and theory. The focus is broad, thus covering labor-market inequalities (wages, earnings, and hours) as well as inequality in consumption and in wealth, and for each variable of interest it surveys the main theories. The discussion, again, aims to be quantitative, i.e., the theories are evaluated based on how much of the observed inequality they can plausibly account for. The second part of the chapter then turns around and asks how inequality might matter for macroeconomic aggregates. We have already touched on one way in which it could: to the extent a model with significant inequality generates a marginal propensity to consume that is higher on average, it will alter many of the model's predictions. This is an active research area in macroeconomics; the HANK model—a Heterogeneous-Agent New-Keynesian setting—in particular has already had significant impact in applied monetary policy contexts.

**Chapter 20: heterogeneous firms** The introduction of the workhorse model is based on an aggregate production function. Clearly, this is an abstraction, at the same time as it hopefully offers a good approximation to the properties of a more realistic framework with a multitude of firms. This chapter examines this issue, both by looking at data on firms and by constructing models with firm heterogeneity. Like the chapter on household heterogeneity, it discusses how firm heterogeneity suggests new mechanisms and thus add insight into how the macroeconomy works. Two channels are studied in particular. One involving misallocation of input factors across firms when there are frictions. The other makes specific assumptions about firm size and discusses granularity: a notion of extreme firm inequality where some very large firms can be relevant to the whole economy. The chapter also briefly touches on markups and the degree of competition.

**Chapter 21: international macro** Many readers of this textbook will perhaps not primarily feel at home in the "large, closed economy" version of our macroeconomic theories. There are, in fact, even strong arguments to suggest that the U.S. economy of today is much more dependent on the rest of the global economy than it used to be and, therefore, issues of trade, exchange rates, and international borrowing and lending ought to take a more central place than it does in many textbooks. The present chapter thus tries to make amends. In particular, it builds toward an up-to-date international business cycle model with monetary and other frictions. The focus is on conceptual issues and mechanisms rather than on a full quantitative model.

**Chapter 22: emerging markets** Yet many other readers may feel that the focus of our text is on the highly developed, world-leading economies such as the U.S. while their interest at least in part is in macroeconomic issues in emerging markets: countries that have opened up to trade and hope to grow rapidly to begin catching up with the leading countries. These countries are argued to have specific vulnerabilities, for example in their ability to borrow in times of severe recessions; indeed sovereign default has been commonly observed, i.e.,

episodes where debts are not paid back and capital flight occurs. This chapter discusses these issues, again as an extension of our workhorse model: what features need to be added, or changed, to deliver a framework that can be used to study macroeconomic fluctuations in emerging markets?

**Chapter 23: sustainability** The final chapter should concern all readers, independent of country of origin, as it deals with global topics that have risen to the top of the political agenda virtually everywhere: areas where human economic activity causes environmental problems, such as climate change. The specific question of climate change concerns macroeconomics, as macroeconomic activity, at least historically, is closely tied to carbon dioxide emissions—as a byproduct of using fossil fuels for energy generation. The chapter mostly focuses on climate change and thus goes through the necessary basic natural-science background and the way in which climate and economics interact. A simple "integrated assessment model" is developed and used to study how policy can be used to address the issue. A brief discussion of natural resource use is also included.

# Chapter 3

# The Solow model

One of the long-run economic facts presented in Chapter 2 is the stability of the capital-output ratio $K_t/Y_t$ (where $Y_t$ represents aggregate output at period $t$ and $K_t$ represents aggregate capital at period $t$) over time: capital is roughly three times annual GDP. Earlier contributions of growth theory, such as the so-called Harrod-Domar model, considered this stability as representing a technological property and incorporated it as one of the assumptions in the model. The capital stock, traditionally divided into structures and equipment but nowadays also containing some intangible components (e.g., software), is one of the important production inputs, but it is of course not the only one. It is possible to produce products in a very capital-intensive way, but clearly there is a choice and using labor—different people's time, skills, and effort—is the most obvious alternative, or complement. Given the many possibilities in which production process can be set up, it is therefore not obvious why, at the macro level, $K_t/Y_t$ is almost constant over time. As argued in Chapter 2, this was Solow's starting point and he managed to resolve the tension between the stable aggregate ratio and the intuitive notion that capital and labor are quite substitutable by a sequence of insights that led both to the construction of a framework for studying macroeconomic dynamics and for measurement of technological change. The purpose of the present chapter is thus to detail the Solow model: the basic assumptions underlying it and their implications.

A central element in the Solow model is the aggregate production function. In the aggregate production function, there are three economic variables that can affect the growth of GDP: technology $A_t$, capital $K_t$, and labor $L_t$. The Solow model focuses on the endogenous accumulation of capital $K_t$. We will see that $K_t$ not only reacts to the saving rate but also to $A_t$ and $L_t$. After solving the model, which will deliver a stable capital-output ratio, we will focus on two main takeaways: (i) the fundamental source of long-run growth in per capita income is the growth in $A_t$; (ii) if all parameter values are common, different economies converge to the same (both in terms of level and growth rate) income per capita in the long run.

## 3.1 The basic model

We start our exposition using the simplest version of the model, where there is neither technological progress nor any growth in the size of the population or the skills of workers. The centerpiece of the Solow model is the aggregate production function

$$Y_t = F(K_t, L_t).$$

Note that we can interpret $Y_t$ as the GDP in this economy. We make the following assumptions for the function $F(K, L)$:

1. $F(K, L)$ is strictly increasing in both $K$ and $L$.
2. $F(K, L)$ is strictly quasiconcave in $(K, L)$ (it has strictly convex isoquants).
3. $F(K, L)$ exhibits constant returns to scale in $(K, L)$: when $K$ and $L$ change to $cK$ and $cL$, with any $c > 0$, $F(K, L)$ becomes $cF(K, L)$.
4. $F(0, L) = 0$.
5. $\lim_{K\to 0} F_1(K, L) = \infty$, where $F_1(K, L) \equiv \partial F(K, L)/\partial K$.
6. $\lim_{K\to\infty} F_1(K, L) = 0$.

Assumptions 5 and 6 are often called Inada conditions and are stronger than we need but these assumptions simplify the exposition.[1]

In the basic model, we assume that the population is constant and that hours worked per worker is constant, so that $L_t$ is constant. We normalize both population and hours per capita to 1; therefore, the only variable input for production is $K_t$, and because of this normalization, we can write $F(K_t, 1) = F(k_t, 1)$, where we remind the reader that lower-case letters are per-capita measures. Let us use $f(k_t)$ to denote $F(k_t, 1)$. Then the production function can be expressed

$$y_t = f(k_t). \tag{3.1}$$

The second important piece of the Solow model is the equation that describes the evolution of the capital stock:

$$k_{t+1} = i_t + (1 - \delta)k_t, \tag{3.2}$$

where $i_t$ is investment in period $t$. The existing capital stock loses value, $\delta k_t$, while being used, where $\delta \in (0, 1)$ is capital's depreciation rate.

These two centerpieces are connected through individual behavior. First, the goods supply $y_t$ is equal to the demand for goods, $c_t + i_t$:

$$y_t = c_t + i_t. \tag{3.3}$$

[1] We also assume that $F$ is twice continuously differentiable. This means that first-order conditions to maximization problems involving $F$ can be differentiated and then generate continuous functions.

Here, $c_t$ is consumption and $i_t$ is investment (both per capita). Note that here we implicitly assume (as in the large part of the following chapters) that goods are homogeneous and can be used for both consumption and investment. Because output $y_t$ is also the total income for consumers, it is either consumed or saved. Therefore, we know that total saving has to equal total investment. In an open economy—where there is trade—this does not necessarily hold. Finally, in this chapter section, we can interpret both $c$ and $i$ as including government consumption and investment, respectively.

In the Solow model, instead of explicitly modeling the consumption-saving decisions of consumers, the consumers are assumed to mechanically save a constant fraction of their income, so that investment is given by

$$i_t = sy_t, \tag{3.4}$$

where $s \in (0, 1)$ is the constant saving rate. This behavioral assumption is relaxed and replaced by consumers' optimizing consumption-saving behavior in Chapter 4.

Inserting equation (3.4) into equation (3.2) and using equation (3.1) yields

$$k_{t+1} = (1 - \delta)k_t + sf(k_t). \tag{3.5}$$

This difference equation expresses the dynamics of the capital stock $k_t$ over time. This is *the fundamental equation of the Solow model*. Note that the only endogenous variable on the right-hand side of the fundamental equation is $k_t$. Therefore, the next period's stock of capital $k_{t+1}$ can be determined only with the knowledge of the current capital stock $k_t$, given values of exogenous objects: the scalars $\delta$ and $s$ and the function $f$. Note also that, starting from a given $k_0$, once we obtain the series of $\{k_{t+1}\}_{t=0}^{\infty}$ from the fundamental equation (3.5), the time series $y_t = f(k_t)$, $c_t = (1 - s)y_t$, and $i_t = sy_t$ can readily be obtained.

### 3.1.1 Steady state and dynamics

To analyze the difference equation (3.5), we first consider a special situation where $k_t$ is constant over time. Call this situation the *steady state* and denote it with an upper bar: $k_t = \bar{k}$ for all $t$. From the fundamental equation (3.5), the steady-state capital stock can be determined by the solution of the equation

$$\bar{k} = (1 - \delta)\bar{k} + sf(\bar{k}).$$

This equation implies $\delta\bar{k} = sf(\bar{k})$. It is straightforward to verify that, under the assumptions for the aggregate production function in the previous section, a strictly positive value of $\bar{k}$ that solves this equation always exists and is unique. Graphically, plot the left- and right-hand sides of the equation $\delta\bar{k} = sf(\bar{k})$; the left-hand side is a straight line through the origin with a positive slope and the right-hand side, which also starts in the origin, is strictly increasing and strictly concave, with a slope of infinity at 0 and one that approaches 0 as $\bar{k} \to \infty$. Clearly, we see that an intersection exists and is unique. The Inada conditions are used here to guarantee the existence of $\bar{k}$. In the context of the basic model and as pointed

out above, the Inada conditions are stronger than necessary: they can be replaced by weaker versions $\lim_{k\to 0} F_1(k,1) > \delta/s$ and $\lim_{k\to\infty} F_1(k,1) < \delta/s$.

A particularly useful production function is the Cobb-Douglas production function: $F(K,L) = K^\alpha L^{1-\alpha}$, so that $f(k) = k^\alpha$, where $\alpha \in (0,1)$. This production function satisfies all assumptions we need, including the Inada conditions. With Cobb-Douglas production, $\bar{k}$ can be solved for analytically:

$$\bar{k} = \left(\frac{s}{\delta}\right)^{\frac{1}{1-\alpha}}. \tag{3.6}$$

From this expression, we can see that the capital stock in the steady state is increasing in the saving rate $s$ and decreasing in the depreciation rate $\delta$. Because aggregate output (GDP) is $\bar{y} = f(\bar{k})$, $\bar{y}$ is also increasing in $s$ and decreasing in $\delta$.

Now, let us use a diagram to analyze the dynamics of $k_t$ when $k_0 > 0$ is not at the steady-state level. Figure 3.1 plots the equation (3.5) with the 45-degree line (that is, representing $k_{t+1} = k_t$). In the figure, the intersection of (3.5) and the 45-degree line represents the steady-state $\bar{k}$.

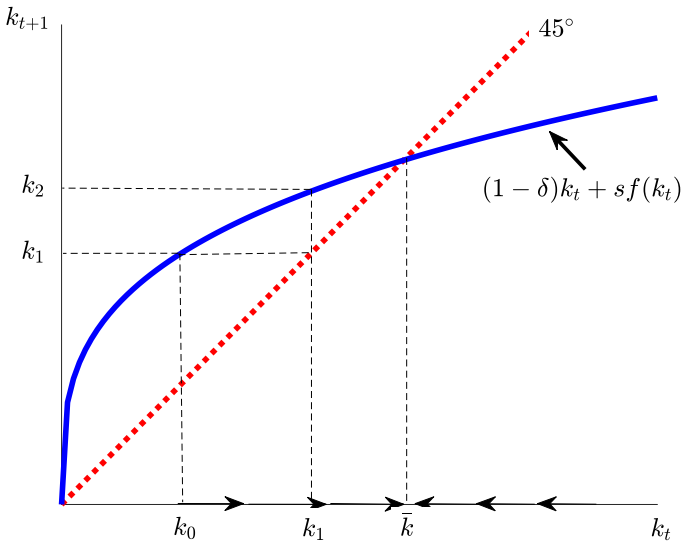



Figure 3.1: Dynamics in the Solow model

In the figure, when we start from a given $k_0$ on the horizontal axis, we can obtain $k_1$ on the vertical axis by using the (3.5) curve. By placing this $k_1$ back on the horizontal axis and using the (3.5) curve again, we obtain $k_2$, and so on. This procedure yields the full time series of the capital stock: $\{k_{t+1}\}_{t=0}^{\infty}$. One can easily verify the dynamics of $k_t$ exhibits a global and monotonic convergence to $\bar{k}$, regardless of the initial value $k_0$. That is, whatever the starting point is, the time series of $k_t$ gradually approaches $\bar{k}$ over time.

The figure works very well as a graphical argument, but how would a mathematical proof be put together? Suppose that $k_t < \bar{k}$. It is then straightforward to show that (i) $k_{t+1} > k_t$

(because $sf(k_t) > \delta k_t$ when $k_t < \bar{k}$) and also that (ii) $k_{t+1} < \bar{k}$ (because $k_t < \bar{k}$). Repeating this procedure, we can see the sequence $\{k_t, k_{t+1}, k_{t+2}, ...\}$ is monotone and bounded by $[k_t, \bar{k}]$. From the Monotone Convergence Theorem, the sequence has a limit. The limit has to be $\bar{k}$, as the limit is unique under the conditions given.

Intuitively, the convergence occurs because the aggregate production function $f(k_t)$ has decreasing returns to capital (Assumption 2 above). Equation (3.2), rewritten in terms of the change in the capital stock,

$$k_{t+1} - k_t = i_t - \delta k_t,$$

reveals two forces that go in opposite directions: (gross) investment and depreciation. When the total capital stock is small, output per unit of capital is large, and the constant saving rate then implies that a large (gross) investment is made relative to the existing capital stock. This process enables the aggregate capital stock to increase. As $k_t$ increases, output per unit of capital becomes smaller due to the decreasing returns property, and when $k_t$ is very large enough, the gross investment cannot cover total deprecation, $\delta k_t$. Thus, the investment force is stronger when $k_t$ is small and the depreciation force is stronger when $k_t$ is large. This relationship is perhaps even clearer if we write the above equation in terms of the growth rate:

$$\frac{k_{t+1} - k_t}{k_t} = \frac{sf(k_t)}{k_t} - \delta,$$

where we have replaced $i_t = sf(k_t)$. The assumptions $f''(k_t) < 0$ and $f(0) = 0$ imply that $f(k_t)/k_t$ is decreasing in $k_t$, generating the negative relationship between the investment force of pushing up the capital stock and the level of the capital stock. In fact, when saving behavior (as captured by $s$ here) is modeled explicitly as a choice, $s$ can counteract this force toward convergence, but it turns out not to be strong enough to overturn the convergence result. This issue will be discussed in detail in the next chapter.

Let us go back to our motivating fact: the stability of $k_t/y_t$ over time. In the basic model here, $k_t/y_t$ is of course constant in steady state, as are all the variables. Below, however, we will see that, even in a growing economy where $k_t$ and $y_t$ keep increasing over time, the economy settles to a situation where $k_t/y_t$ is constant over time.

In the Cobb-Douglas case above, the steady-state $k/y$ ratio can be solved out as

$$\frac{\bar{k}}{\bar{y}} = \frac{s}{\delta}.$$

Clearly, in the long run $k_t/y_t$ is larger when $s$ is larger and when $\delta$ is smaller.

### Other kinds of dynamics

The growth model can, in principle, generate very rich (and complex!) dynamics if its neoclassical feature is not present, i.e., if the production function is not strictly concave in capital. There are applications in the economics literature that, in reduced form, have such non-neoclassical features, and we now briefly illustrate how they can work.

**Endogenous growth** Consider the situation where $F_1(k, 1)$ is uniformly above $\delta/s$. Figure 3.2 below draws such an example. In this case, the steady state with $\bar{k} > 0$ does not exist, and $k_t$ keeps growing larger over time. That is, there is unbounded growth "by itself": growth in *endogenous*. This concept will be discussed more in Chapter 11 below but in a richer model where other production inputs can also be accumulated. Here, given that one expects decreasing returns to each input—such as capital—it is hard to take this case very seriously.

In the special (and illustrative) case where $F_1$ is a constant—as illustrated in the graph—we can think of output as linear in capital: $y_t = Ak_t$, with no role for labor (make $\alpha = 1$ in the Cobb-Douglas setting).[2] Given that labor commands about two thirds of the income from production, this setup does not seem empirically plausible. In a setup with endogenous growth such as this, two identical countries starting out with different capital stocks will be forever different; the gap between them, in percentage terms, will stay constant.

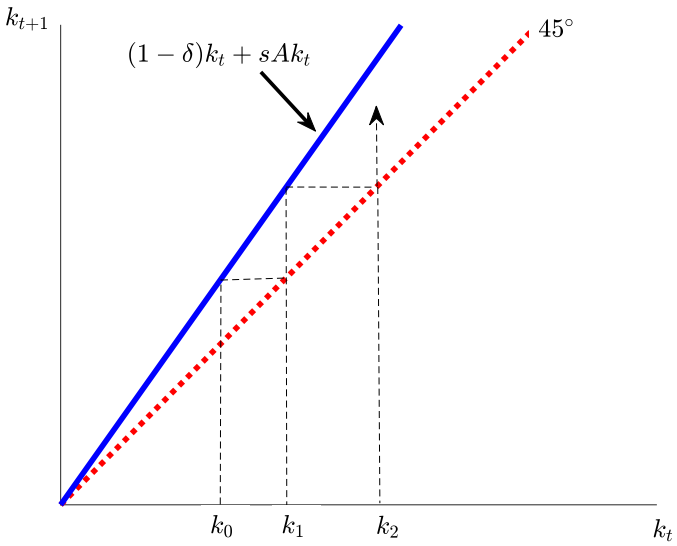



Figure 3.2: Endogenous growth in the Solow model

**Poverty traps** Suppose that the production function is not globally concave in $k$: it has a middle section that is convex. This could be true if there are some regions of $k$ with increasing returns, say, as a result of large infrastructure investments—the building of transportation networks. In such a case, the right-hand side of (3.5) will not be concave, and it may cross the 45-degree line multiple times, as illustrated in Figure 3.3 below.

Clearly, in this case there are multiple steady states and at least one of the steady states will then not be "stable": $k_t$ will not converge to that steady state even when $k_0$ starts very

[2]One can imagine a role for labor if $Y_t = AK_t + BL$, which is CRS, but here labor would not matter asymptotically if $A > \delta/s$.

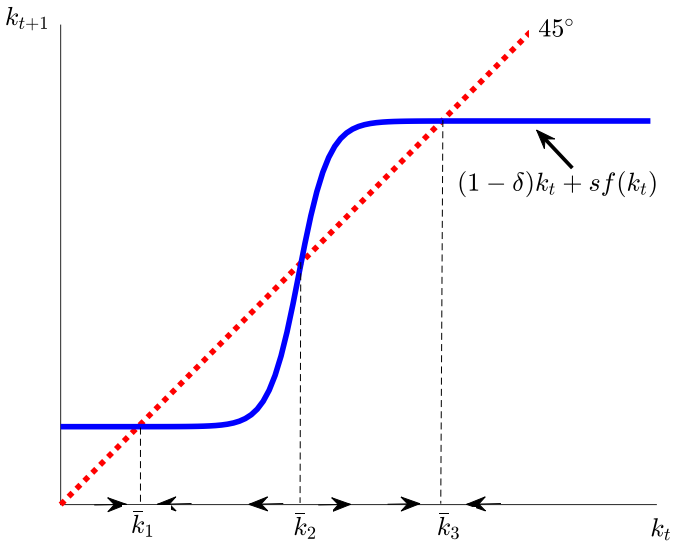



Figure 3.3: Poverty traps in the Solow model

close to it (a small perturbation away from the steady state leads further away from it). When there are multiple steady states, an economy can get stuck in the steady state with a low $\bar{k}$ (and thus a low GDP) when it starts from a low $k_0$. This situation is often called the *poverty trap*. The gap between a poor country with a small $k_0$ and a rich country with a large $k_0$ may never close in this setting. In Figure 3.3, there are three steady states, $\bar{k}_1$, $\bar{k}_2$, and $\bar{k}_3$. Of these three, $\bar{k}_1$ and $\bar{k}_3$ are stable steady states. When the economy starts from a very low $k_0$, the economy converges to $\bar{k}_1$ and gets trapped in it. To escape from the trap, the capital stock would have to be pushed up to a larger level than $\bar{k}_2$, from which it would converge to $\bar{k}_3$. One way to achieve this movement is to (temporarily) encourage very high saving. If the saving rate is raised permanently so that the $(1-\delta)k_t + sf(k_t)$ curve moves up sufficiently, the steady states $\bar{k}_1$ and $\bar{k}_2$ will disappear and the economy converges globally to $\bar{k}_3$. The growth/development literature does not appear to have identified sufficiently large increasing returns leading to results of the kind described here, but it is an interesting possibility.

**Non-monotonic dynamics and chaos** An even more radical departure from the neoclassical setting is if $f(k)$ declines in $k$ at high levels of $k$. Conceptually, if a bakery has no ovens, ovens have high marginal productivity, and as more ovens are added, the marginal productivity declines, and it will become negative once there are so many ovens in the bakery that there is neither space for bakers nor for the dough. This possibility is more esoteric in a macroeconomic context but let us nevertheless study it briefly. So when $f(k)$ decreases steeply enough, the right-hand side of (3.5) will become decreasing in $k_t$. Illustrating this graphically, we will see that convergence, if convergence is at all possible, will not be mono-

tonic.[3] In fact, $k_t$ can exhibit forever oscillating (or even chaotic) dynamics.[4] An example of is drawn in Figure 3.4.

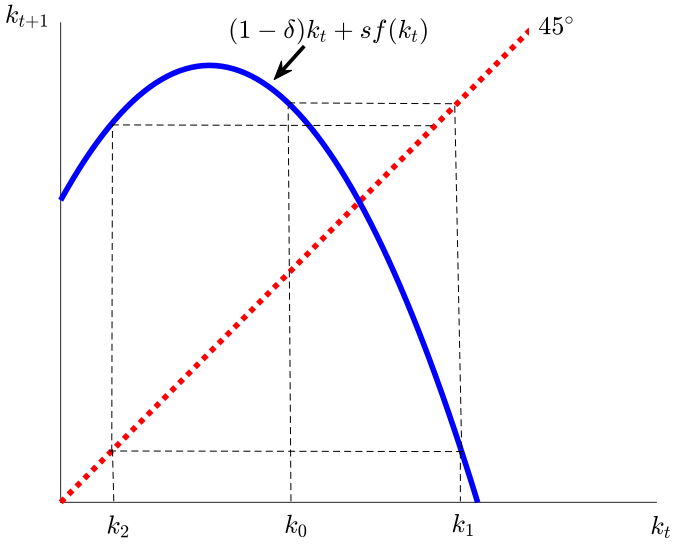



Figure 3.4: Complex dynamics in the Solow model

## 3.2 The growing economy

Now we extend the model to the situation where $A_t$ and $L_t$ grow over time. In the previous section, the long-run outcome was the steady state where there is no growth. This extension is necessary for addressing the facts related to the growth issues described in Chapter 2.

We assume that the aggregate production function takes the form of

$$Y_t = F(K_t, A_t L_t).$$

There are two changes from the basic model: first, we allow the labor input (population times hours worked per person) $L_t$ to grow over time. Second, and more importantly, we allow for technological progress. In this production function, the variable representing the technology level, $A_t$, is multiplied by labor input $L_t$. Technological progress thus takes a form of improving the labor input, and $A_t L_t$ is often referred to as the total number of *efficiency units of labor* (or *effective labor*). This form of technological progress is labor-augmenting; it was introduced in the previous chapter.[5] As was also asserted there, Uzawa

[3] Locally stable dynamics will occur if the slope at the steady state is less than 1 in absolute value.

[4] Chaos is a mathematical term; it involves great sensitivity to initial conditions and forever non-monotone behavior that never settles down to a repeated pattern (a repeated pattern could be a two-cycle): it looks "random."

[5] This form of technical progress is sometimes also called Harrod-neutral.

(1961) proved that labor-augmenting technical change is the only form of technical progress that is consistent with exact balanced growth, that is, the growth path where aggregate variables such as output and capital grow at a constant rate. Uzawa's theorem is formally stated and proved in Appendix 3.A.

We assume that the (net) growth rate of $A_t$ is $\gamma$ and that the growth rate of $L_t$ is $n$.[6] The same manipulations of equations as in the basic model yield

$$K_{t+1} = (1-\delta)K_t + sF(K_t, A_t L_t).$$

Dividing both sides by $A_t L_t$, we obtain

$$\frac{K_{t+1}}{A_t L_t} = (1-\delta)\frac{K_t}{A_t L_t} + s\frac{F(K_t, A_t L_t)}{A_t L_t}.$$

Let us define a new variable $\tilde{k}_t$ by

$$\tilde{k}_t \equiv \frac{K_t}{A_t L_t}.$$

Then, because the above equation can be rewritten as

$$\frac{A_{t+1}}{A_t}\frac{L_{t+1}}{L_t}\frac{K_{t+1}}{A_{t+1}L_{t+1}} = (1-\delta)\frac{K_t}{A_t L_t} + sF\left(\frac{K_t}{A_t L_t}, 1\right),$$

we obtain

$$(1+\gamma)(1+n)\tilde{k}_{t+1} = (1-\delta)\tilde{k}_t + sf(\tilde{k}_t). \tag{3.7}$$

Here, $f(\tilde{k}_t) \equiv F(\tilde{k}_t, 1)$ as in the basic model. After the capital stock $K_t$ is normalized by $A_t L_t$, we thus obtain a very similar difference equation as the fundamental equation (3.5) in the basic model. Once we characterize the dynamics of $\tilde{k}_t$, we can "untransform" it into the core macro variables $Y_t$, $K_t$, and $C_t$.

### 3.2.1 Balanced growth and dynamics

The characterization of the fundamental equation (3.7) follows similar steps as for the basic model. The concept corresponding to the steady state in the basic model is the *balanced growth path* (some researchers still prefer to use the name "steady state" for the balanced growth path, because the normalized variables are "steady" also in this case). Along the balanced growth path, the normalized capital stock $\tilde{k}_t$ is constant, and typical economic variables, such as $Y_t$, $K_t$, and $C_t$, grow at a constant rate. Once again, we use a notation with upper bar: $\tilde{k}_{t+1} = \tilde{k}_t = \bar{\tilde{k}}$. The value of $\bar{\tilde{k}}$ along the balanced-growth path solves

$$(1+\gamma)(1+n)\bar{\tilde{k}} = (1-\delta)\bar{\tilde{k}} + sf(\bar{\tilde{k}}).$$

[6] $n$ has two origins: a growing population and changes in hours worked per person, which we saw from Chapter 2 is best characterized by a decline in the longer run.

With a Cobb-Douglas production function we can obtain a closed-form solution:

$$\bar{\tilde{k}} = \left( \frac{s}{(1+\gamma)(1+n)+\delta-1} \right)^{1-\alpha}. \tag{3.8}$$

The dynamic property of the model can be analyzed similarly to the basic model. As in the basic model, starting from any $\tilde{k}_0$, $\{\tilde{k}_{t+1}\}_{t=0}^{\infty}$ converges monotonically to $\bar{\tilde{k}}$. In (3.8), the rate of technological progress $\gamma$ and the population growth rate $n$ work similarly to depreciation: maintaining a level of $k_t = K_t/(A_t L_t)$ is harder as $A$ and $L$ grow faster; i.e., each unit of untransformed capital needs to grow faster, as if making up for depreciation.

In this framework, we can analyze how various economic variables grow over time. For example, suppose $L_t$ is simply population size (assuming that all citizens work one unit) and then consider income per capita, defined as $y_t \equiv Y_t/L_t$. Because $Y_t/(A_t L_t) = f(\tilde{k}_t)$, in the long run, $Y_t/(A_t L_t)$ converges to $f(\bar{\tilde{k}})$. Therefore, in the long run, income per capita converges to

$$y_t = f(\bar{\tilde{k}}) A_t$$

and the growth rate of $y_t$ in the long run is

$$\frac{y_{t+1}-y_t}{y_t} = \frac{f(\bar{\tilde{k}})A_{t+1} - f(\bar{\tilde{k}})A_t}{f(\bar{\tilde{k}})A_t} = \frac{A_{t+1}-A_t}{A_t} = \gamma.$$

The growth in technology $A_t$ is essential in sustaining long-run growth in per capita income. Surprisingly, no other parameters affect the long-run growth of per capita income. For example, encouraging saving (an increase in $s$) does not affect the long-run growth rate of per capita income in the economy. Note that this result does not mean that the change in $s$ does not have any effect on economic outcome: it has an effect on the *level* of per capita income, rather than the growth rate. It also has an effect on the growth rate in the short run (when the economy is not yet on the balanced-growth path).

In the short run, $\tilde{k}_t$ changes over time and its movement has an effect on the economic outcome. For example, the growth rate per capita income is now

$$\frac{y_{t+1}-y_t}{y_t} = \frac{f(\tilde{k}_{t+1})A_{t+1} - f(\tilde{k}_t)A_t}{f(\tilde{k}_t)A_t} = \frac{f(\tilde{k}_{t+1})}{f(\tilde{k}_t)}(1+\gamma) - 1.$$

From the fundamental equation (3.7), we know that when $\tilde{k}_t < \bar{\tilde{k}}$, $\tilde{k}_t$ increases over time, that is, $\tilde{k}_{t+1} > \tilde{k}_t$. Therefore, in this case, $f(\tilde{k}_{t+1})/f(\tilde{k}_t) > 1$ and the growth rate of $y_t$ in the short run is larger than $\gamma$. Similarly, when $\tilde{k}_t > \bar{\tilde{k}}$, the growth rate of $y_t$ is smaller than $\gamma$. In other words, the Solow model predicts that income per capita of a poor country grows faster than at rate $\gamma$ and that income per capita of a rich country grows slower than at rate $\gamma$ in the short run. This difference in growth rate is another representation of the convergence prediction of the Solow model.

## 3.3 Stylized facts and the Solow model

The model with growth, presented above, can match various stylized facts of economic growth. First, going back to our motivating fact on $K_t/Y_t$, because $Y_t/(A_tL_t) = f(\tilde{k}_t)$ and $K_t/(A_tL_t) = \tilde{k}_t$ are both constant along the balanced growth path, $K_t/Y_t = \tilde{k}_t f(\tilde{k}_t)$ is also constant in the long run. Once again, the Solow model can replicate the constant $K_t/Y_t$ in the data.

The first fact in Chapter 2 was the steady growth of the GDP per capita. As we have seen above, the GDP per capita grows at the rate $\gamma$ along the balanced growth path (towards which the economy converges from any starting point). This fact, therefore, is consistent with the Solow model with technological progress.

Another stylized fact is that the return to physical capital has been nearly constant. Here, we need to first compute the return to physical capital in the model. Suppose that firms maximize profit under competitive markets:

$$\max_{K_t, A_tL_t} F(K_t, A_tL_t) - r_tK_t - w_tA_tL_t. \tag{3.9}$$

Here, output is taken as the numéraire and the price is set at one. Therefore, $F(K_t, A_tL_t)$ is the revenue and $r_tK_t + w_tA_tL_t$ is the cost. Let us assume, for simplicity, that the capital stock is owned by the consumers and rented to the firms with the rental rate $r_t$. Thus, $r_t$ represents the return to capital. The other component of the cost is the wage payment: $w_t$ is the wage per efficiency unit of labor. From the first-order condition for $K_t$, the return to physical capital is equal to the marginal product of capital:

$$r_t = F_1(K_t, A_tL_t).$$

Differentiating both sides of the equation $f(K_t/A_tL_t) = F(K_t, A_tL_t)/(A_tL_t)$ with respect to $K_t$, we obtain that

$$r_t = f'(\tilde{k}_t).$$

Along the balanced growth path, therefore, $r_t$ is constant because the right-hand side is constant at $f'(\tilde{\bar{k}})$.

Another prominent fact is the stability of the labor share and the capital share. The capital share is equal to $r_tK_t/Y_t$, and it can readily be seen that it is constant, because we have already seen that both $r_t$ and $K_t/Y_t$ are constant along the balanced-growth path. The labor share is $w_tA_tL_t/Y_t$. From the first-order condition of (3.9), the wage is equal to the marginal product of labor:

$$w_t = F_2(K_t, A_tL_t).$$

When the production function has constant returns to scale, it is homogeneous of degree one.[7] Then it follows that production becomes

$$Y_t = K_tF_1(K_t, A_tL_t) + A_tL_tF_2(K_t, A_tL_t).$$

[7]Recall that a function $f(x)$ is homogeneous of degree $r$ (is $H(r)$) when $f(sx) = s^r f(x)$ for all $(s, x)$; here $x$ is a vector and $r$ and $s$ are scalars. If $r = 1$, it then follows, using differentiation with respect to $s$ and each element of $x$, that $f(x) = \sum_i (\partial f/\partial x_i)x_i$ for all $x$.

Dividing both sides by $Y_t$, we obtain that the labor share is one minus the capital share. Therefore, the labor share is also constant when the capital share is constant.

We can also confirm the stability of the labor share by direct calculation. As for the case of $r_t$, it can be shown, by differentiating $f(K_t/A_tL_t) = F(K_t, A_tL_t)/(A_tL_t)$ with respect to $A_tL_t$, that

$$w_t = f(\tilde{k}_t) - \tilde{k}_t f'(\tilde{k}_t).$$

It can readily be seen that $w_t$ is constant when $\tilde{k}_t$ is constant at $\bar{\tilde{k}}$. Because $A_tL_t/Y_t = 1/f(\tilde{k}_t)$, it is also constant along the balanced growth path. Therefore, $w_tA_tL_t/Y_t$ is also constant. Note that, for a Cobb-Douglas production function $Y_t = K_t^{\alpha}(A_tL_t)^{1-\alpha}$, the capital share is $\alpha$ and the labor share is $1-\alpha$ regardless of the values of $K_t$ and $A_tL_t$, and therefore the factor shares are constant even outside the balanced-growth path.

## 3.4 Convergence

We have already seen that, in the Solow model, the economy monotonically converges to the steady state (or balanced growth path). Here, we look at this convergence property more in detail and take a quick look at the data.

### 3.4.1 Local properties: the speed of convergence

In the basic model, where we again set $L_t = 1$ and use the per-capita notation $k_t$, the fundamental equation (3.5) can be approximated around the steady-state as

$$\Delta k_{t+1} = \left(1 - \delta + sf'(\bar{k})\right)\Delta k_t, \tag{3.10}$$

where $\Delta k_t$ represents the deviation of $k_t$ from its steady-state value, that is, $k_t - \bar{k}$, when the deviation is small.[8] When the production function is in the Cobb-Douglas form, using the steady-state solution (3.6),

$$\Delta k_{t+1} = (1 - \delta(1-\alpha))\Delta k_t$$

holds. Replacing $\Delta k_t$ by $k_t - \bar{k}$ and dividing both sides by $\bar{k}$, (3.10) can be expressed as

$$\frac{k_{t+1} - \bar{k}}{\bar{k}} = (1-\lambda)\frac{k_t - \bar{k}}{\bar{k}},$$

where $\lambda \equiv \delta - sf'(\bar{k})$ represents the *convergence speed*. A large value of $\lambda$ implies that $\Delta k_{t+1}$ becomes smaller (in absolute value) more quickly, implying a faster convergence. This is illustrated in Figure 3.5 below: a higher $\lambda$ represents a flatter slope at the steady state and "more steps until you reach steady state."

In the Cobb-Douglas case, using (3.6),

$$\lambda = \delta(1-\alpha) \tag{3.11}$$

[8] This is obtained from a first-order Taylor approximation of the expression around $\bar{k}$.

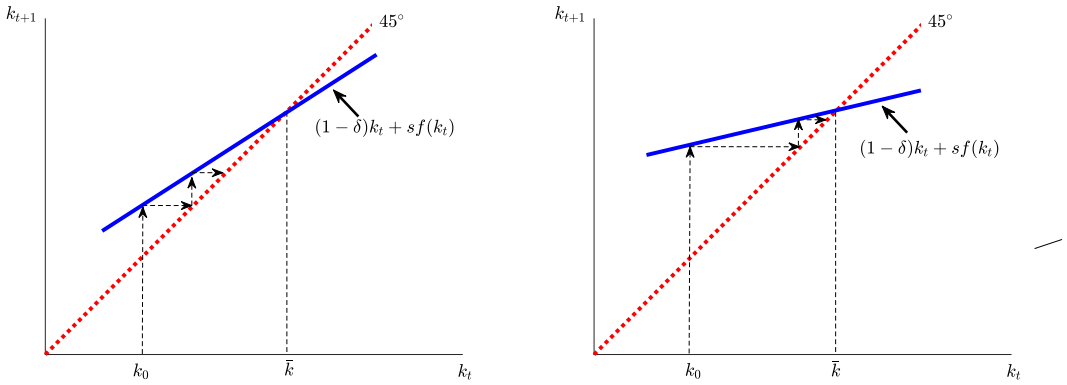



Figure 3.5: Slow and fast convergence

holds, and thus the convergence is faster when $\alpha$ is small and $\delta$ is large. Note that $s$ does not affect the convergence speed in this case. The convergence speed, in general, is affected by how $k_t$ affects $k_{t+1}$. In an extreme case, if $k_t$ has no effect on $k_{t+1}$ (a flat line), convergence is immediate. The parameter $s$ has two opposing forces to this mechanism. For a given $k_t$, a large $s$ implies a larger impact of $k_t$ on $k_{t+1}$. However, the steady-state value of capital is larger when $s$ is larger, and thus the marginal product of capital at the steady state, $f'(\bar{k})$, is smaller when $s$ is larger, implying a smaller impact of $k_t$ on $k_{t+1}$. These two opposing forces exactly offset each other when the production function is in the Cobb-Douglas form.[9]

In the case with growth, the fundamental equation (3.7) can be approximated by

$$(1+\gamma)(1+n)\Delta\tilde{k}_{t+1} = \left(1-\delta+sf'(\bar{\tilde{k}})\right)\Delta\tilde{k}_t. \tag{3.12}$$

When the production function is in the Cobb-Douglas form, using the steady-state solution (3.8),

$$\Delta\tilde{k}_{t+1} = \left(\alpha + \frac{(1-\alpha)(1-\delta)}{(1+\gamma)(1+n)}\right)\Delta\tilde{k}_t$$

holds. The equation (3.12) can be rewritten as

$$\frac{\tilde{k}_{t+1}-\bar{\tilde{k}}}{\bar{\tilde{k}}} = (1-\lambda)\frac{\tilde{k}_t-\bar{\tilde{k}}}{\bar{\tilde{k}}},$$

where the convergence speed is now given by $\lambda = 1-(1-\delta+sf'(\bar{\tilde{k}}))/(1+\gamma)(1+n)$. In the Cobb-Douglas case we obtain

$$\lambda = (1-\alpha)\left(1-\frac{1-\delta}{(1+\gamma)(1+n)}\right). \tag{3.13}$$

[9]The Cobb-Douglas function is very convenient because it often simplifies the algebra and leads to simple expressions. This simplicity, however, can be deceiving as we see here: the functional form often makes fundamental forces going in opposite direction cancel. That is, under the surface there may be very strong forces but, as if by magic, the Cobb-Douglas form makes them invisible.

### 3.4.2 Cross-country data

Is convergence observed in the data? Recall that the convergence prediction implies that a country that starts with a smaller per-capita GDP experiences faster subsequent growth. Figure 3.6 plots this relationship across countries. The data is taken from the Penn World Table 10.0 (https://www.rug.nl/ggdc/productivity/pwt/). The horizontal axis is per-capita real GDP in 1960, which we take as the starting point. The vertical axis is the subsequent growth rate (annualized using geometric averages) in per-capita real GDP from 1960 to 2019.

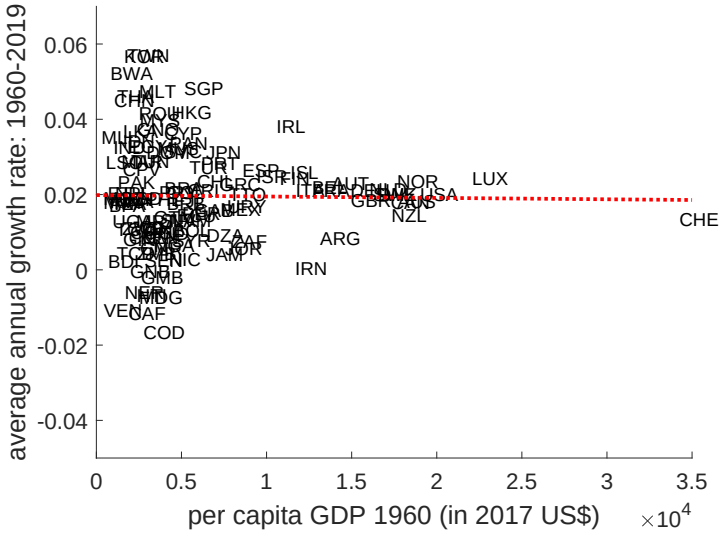



Figure 3.6: All countries, 1960–2019

**Source**: Penn World Tables 10.0. The GDP variable used is RGDPNA.

One can immediately see that there is no systematic tendency for initially poor countries to grow faster. The fact that there is no tendency for countries to converge, however, does not imply a rejection of the Solow model. In fact, the Solow model does not predict that different countries will always converge to the same balanced growth path (a phenomenon called "unconditional convergence" or "absolute convergence"). Rather, it predicts that countries converge if they share the same parameter values ("conditional convergence"). We know, for example, that saving rates differ widely across countries and, at least over shorter time horizons, it is reasonable to think that the growth rates of $A_t$ also differ.

To examine conditional convergence, a useful exercise is to look at the same kind of graph restricted to a smaller, and more similar, set of countries. Figure 3.7 thus plots the same data as Figure 3.6, but only for the original members of the Organisation for Economic Co-operation and Development (OECD). OECD was formed by high-income countries that share relatively similar economic and political institutions, and we can expect the underlying parameters in the Solow model to be relatively similar among these countries.

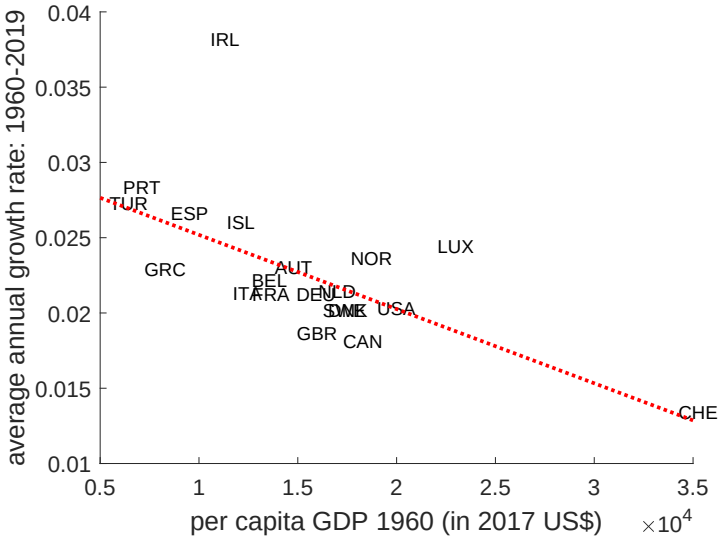



Figure 3.7: OECD countries, 1960–2019

**Source**: Penn World Tables 10.0. The GDP variable used is RGDPNA.

For this set of countries, we observe a clear tendency for convergence: poor countries in 1960 on average experience faster subsequent growth. Barro and Sala-i-Martin (2004, Chapter 12) conduct a similar exercise across regions within countries, treating each region as a different "country." For U.S. states, Japanese prefectures, and European regions, they find a clear tendency of convergence, supporting the prediction of the Solow model.

In a recent paper, Kremer et al. (2020) show that in recent years, the data actually show a tendency for unconditional convergence. Figure 3.8 below repeats the same exercise as in Figure 3.6 for the same set of countries, but setting the initial date to 2000. We can see that some convergence (negative correlation) is observed in the recent years. The authors argue that this tendency arose because some of the underlying factors that affect growth (the factors that likely affect the growth rate of $A_t$), such as policies, institutions, and human capital have become more similar across countries in recent years. In Chapter 11, we will discuss this and many related issues in greater detail.

### 3.4.3 Quantitative use of the Solow model

What are the *quantitative* predictions of the model for convergence? To answer this question, we need to assign functional forms and specific parameter values. This procedure will give us a numerical value for $\lambda$ in (3.13). Once $\lambda$ is computed, one can of course also conduct counterfactual experiments by simulating a hypothetical situation using the quantitative model.

If we are interested in the model's quantitative predictions for the speed of convergence,

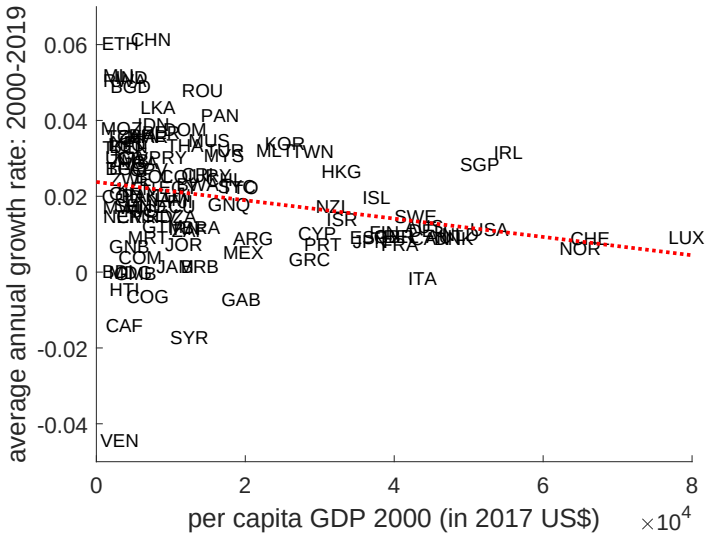



Figure 3.8: All countries, 2000-2019

**Source**: Penn World Tables 10.0. The GDP variable used is RGDPNA.

one way to proceed would be to simply see if it is possible to choose parameters so as to hit the "observed $\lambda$" (e.g., as measured by the slope in Figure 3.7). More generally, one could specify a full stochastic model, say, with explicit shocks to variables (such as $A_t$) and estimate the resulting structure against the data we just looked at. Clearly, we could generate a good fit in this case if we are free to choose parameters; for example, given any production function, we could match the $\lambda$ by an appropriate choice of $\delta$. This choice, however, may not be consistent with what we know about depreciation rates from microeconomic data. More generally, we would like our model's different components (functional forms and parameter values) to be selected to be in line with microeconomic studies (and perhaps aggregate data too). This way, the quantitative evaluation is disciplined. As briefly discussed in Chapter 1 above, a way forward is *calibration*. The procedure consists of two distinct steps, each guided by data. First, we assign a parameterized functional forms to the unknown functions in the model. In our case, the production function $F$ corresponds to the unknown function. Second, we assign specific values to the parameters. Because we use various moments (such as means and variances) in the data to assign parameter values, this second step bears close resemblance to estimation by the method of moments.

Before starting the calibration, we have to decide on the length of the time period. Here, we are chiefly interested in movements in aggregates that occur over the medium run and thus set one period to be one year. For the first step, we choose the Cobb-Douglas form, used above, for the production function: $F(K_t, A_t L_t) = K_t^{\alpha}(A_t L_t)^{1-\alpha}$. As we have seen, this functional form yields constant factor shares, which is consistent with the rather striking data on an absence of major movements in the shares. A Cobb-Douglas function is special,

however, in that it has the property that the substitution elasticity between inputs (capital and labor) is equal to one, so we need to make sure that it is consistent with studies of production functions, at least at a high level of aggregation.[10] These studies rarely suggest major departures from 1.

In the second step, we need to assign values to the parameters $\alpha$, $\delta$, $\gamma$, $n$, and $s$. Here, we have implicitly assumed that the production function has constant returns to scale, i.e., that the sum of the exponents on capital and labor is one, so that it suffices to choose the $\alpha$. Before discussing the parameter selections, note that for the convergence speed $\lambda$ in (3.13), the information on $s$ is not necessary. Below we will therefore skip assigning a value to $s$.

Calibration usually draws on multiple data sources. Overall, there are two methods of assigning the parameter values based on data. First, if particular parameters have been considered as important objects for investigation elsewhere and we know plausible parameter values from past studies, it is convenient to directly assign the parameter values accordingly. Second, we can choose data moments that involve several parameters and assign parameter values such that these moments, when generated by our theory, line up with observed moments. The first method can be considered a special case of the second method, because the parameter values from past studies have to come from certain data moments used in these studies. From this perspective, an alternative interpretation of the calibration procedure is an implementation of the method of moments with multiple data sets.

We set the value of $\alpha$ from national income accounting. As we have seen from the previous section, $\alpha$ corresponds to the capital share. From Figure 2.12 in Chapter 2, $\alpha = 1/3$ is a good approximation. We can set the value of $\gamma$ from the long-run growth rate of per capita income in advanced countries. The population growth rate $n$ can be measured directly from the data. Barro and Sala-i-Martin (2004, p.58) use $\gamma = 0.02$ and $n = 0.01$ as a benchmark at the annual frequency. The fundamental equation (3.7) can be rewritten as

$$(1+\gamma)(1+n) = 1 - \delta + \frac{I_t}{K_t},$$

where we used $sf(\tilde{k}_t)/\tilde{k}_t = I_t/K_t$. The investment-capital ratio in the U.S. economy is about 0.076 (Cooley and Prescott, 1995) at an annual frequency. Given the above values of $\gamma$ and $n$, this equation implies $\delta = 0.046$.[11] Clearly, here, one could alternatively have used depreciation rates directly from data on depreciation (by capital type, or from aggregate data on depreciation rates) and then this equation would have implied a value for the average size of $I/K$ on a balanced growth path. Given the accounting practices and the fact that capital remains at roughly three times annual GDP as time passes, on average investment precisely will have to make up for depreciation (taking population and technology growth into account), so either way the number for $\delta$ ends up around 0.05.

[10] The substitution elasticity is given by the percentage change in the ratio of the inputs when the ratio of their prices change by one percent, i.e., $-d\log(K/L)/d\log(r/w)$. Using the firm's first-order conditions, we can see that this expression must equal 1 for a Cobb-Douglas production function.

[11] Given that $I/K = (I/Y)(Y/K)$, we could alternatively have measured $I/Y$—the saving rate—and $Y/K$, which we know is about 1/3. Conversely, we can now obtain the implied $I/Y$ as 0.076/0.33, which approximately equals 0.23.

With these values, we find that $\lambda = 0.049$. The empirical counterpart of $\lambda$ is about 0.015 to 0.03 (Barro and Sala-i-Martin 2004, p,59), and therefore the Solow model over-predicts the convergence speed. The model value of $\lambda$ is about 0.02 if $\alpha$ is raised to 0.73. An argument for a larger value of $\alpha$ is that a part of labor income is the return to human capital, which can be accumulated in a similar manner as physical capital, and thus some part of labor income should be included in the capital share. Relatedly, one can view $A_t$ as an accumulable factor—after all, technological development is often part of conscious investments into R&D, and hence it too is a capital stock. These issues are returned to in Chapter 11.

Once the model has been assigned specific functional forms and parameter values, we can also conduct quantitative experiments. Suppose, for example, the saving rate $s$ equals 0.1. How would increasing $s$ to 0.2 affect the normalized level of output along the balanced-growth path, $f(\bar{\tilde{k}})$? With a Cobb-Douglas production function, we have already obtained the solution for $\bar{\tilde{k}}$ in (3.8). Inserting our calibrated parameter values, along with $s = 0.1$, we obtain $\bar{\tilde{k}} = 1.20$ and $f(\bar{\tilde{k}}) = \bar{\tilde{k}}^{\alpha} = 1.06$. When $s$ goes up to 0.2, $\bar{\tilde{k}}$ rises to 1.90 and $f(\bar{\tilde{k}}) = 1.24$. Therefore, doubling the saving rate from 10% to 20% increases the normalized level of output by 17%, because $1.24/1.06 = 1.17$. In addition, we could compute the quantitative predictions for how fast output would rise to eventually reach a 17% higher value.

## 3.5 Business cycles

The usefulness of the Solow model goes much beyond the study of economic growth; due to its ability to account for the broad features of the macroeconomic data over our modern economic history, it constitutes the core of macroeconomic modeling. A prominent illustration of this is the fact that virtually all modern theories of business cycles build on a version of, or elaboration on, the Solow model. Although business cycle theories are detailed later in Chapters 12 and 16, we now briefly review how these theories are linked to the Solow model and exhibit some of the key associated tools, such as impulse response diagrams.

### 3.5.1 Various theories of business cycles

The studies of business cycles are primarily the analysis of the arrival of shocks to the economy and how the economy reacts to these shocks. The way the economy reacts to the shocks is usually called the "propagation mechanism." Once we introduce uncertainty in Chapter 7, we can treat these "shocks" more precisely. Here, we consider a general (exogenous and deterministic) movement in certain variables as the source of business cycle fluctuations.

Below, first, we extend the basic model in several directions. In particular, we outline how the Solow model can be modified and accommodate various shocks in three different business cycle models. Throughout we conduct the analysis in per-capita terms and, hence, use lower-case letters.

- The first business cycle model is the so-called real business cycle (RBC) model. As will be explained in Chapter 12, the RBC model provides a simple mechanism whereby macroeconomic variables comove, as is clear in the data. The prototypical RBC model considers the following aggregate production function

$$y_t = A_t F(k_t, \ell_t),$$

where $\ell_t$ is variable labor input, and considers a shock to $A_t$ (often called the "neutral technology shock"). That is, it assumes that $A_t$ changes over time and the movement of $A_t$ is the source of the business cycle. For example, consider a model where $A_t$ switches around between two values, $A_H$ and $A_L$, where $A_H > A_L$. When $A_t$ moves to $A_H$, the economy starts moving towards the corresponding steady state. Then $A_t$ switches to $A_L$, and the economy now moves towards a different steady-state value. We can interpret these movements as business cycles: movements around some average. Augmented with steady growth we would have movements around the balanced path.

Note that the production function in this section is different from the Harrod-neutral form $F(k_t, A_t\ell_t)$ earlier. Note that the distinction between Harrod-neutral technological progress and the Hicks-neutral technological progress is not essential when the production function is in the Cobb-Douglas form: the Harrod-neutral production function $k_t^\alpha (A_t\ell_t)^{1-\alpha}$ can be rewritten as $A_t^{1-\alpha} k_t^\alpha \ell_t^{1-\alpha}$, and by defining $\tilde{A}_t \equiv A_t^{1-\alpha}$, the same production can be interpreted as the Hicks-neutral production function $\tilde{A}_t k_t^\alpha \ell_t^{1-\alpha}$.

If we maintain the assumptions of the basic model, we have $\ell_t$ constant and $i_t/y_t = s$ (and $c_t/y_t = 1 - s$) constant even with the shocks to $A_t$. These features are at odds with the business cycle data. To accommodate the business cycle facts that (i) $\ell_t$ comoves positively with the business cycle, (ii) $i_t$ is more volatile than $y_t$, and (iii) $c_t$ is less volatile than $y_t$, the basic equation would have to allow for the saving rate and $\ell_t$ to react to $A_t$ (and possibly to $k_t$). The economy evolves, therefore, following the difference equation

$$k_{t+1} = (1 - \delta)k_t + s(k_t, A_t)A_t F(k_t, \ell(k_t, A_t)).$$

This is a modified form of the fundamental equation (3.5) of the basic Solow model.

- Next, we consider a different kind of shock. Suppose that the final goods market clearing condition (3.3) is modified to

$$y_t = c_t + i_t/\nu_t,$$

where $\nu_t$ moves over time (and often called the "investment-specific technological progress"): when it is high, it is cheaper to produce investment goods. One can think of this equation as reflecting the two-sector structure of the economy: $y_t$ and $c_t$ are measured in consumption goods, and investment goods have a production process that can create $i_t$ units of investment goods by using $i_t/\nu_t$ units of consumption goods. The fundamental equation (3.5) can now be modified to

$$k_{t+1} = s\nu_t F(k_t, \ell) + (1 - \delta)k_t;$$

this equation can also be extended to include endogenous $s$ and $\ell$ as in the case of the neutral technology shock.

- In the third example, we consider a very different model structure: one with "demand shocks." First, assume that $c_t$ is exogenous and that it moves around over time. The movement of $c_t$ serves as the (demand) shock. Suppose, further, that we maintain the assumption that $i_t/c_t = s/(1-s)$. Because of this assumption, since $s$ is constant, $i_t$ is proportional to $c_t$ and moves along with it. Therefore, the total demand for goods is a function of $c_t$:

$$c_t + i_t = \frac{1}{1-s} c_t.$$

When $c_t$ is not sufficiently large, $c_t + i_t$ would be less than the full capacity output $F(k_t, \ell)$ (here we assume again that $\ell$ is given by labor-force participation: those who want to work). Assume, then, that when there are demand shortages, total output $y_t$ is determined by the demand side and so that a fraction $u_t$ of the labor force become unemployed (therefore, $u_t$ is the unemployment rate):

$$y_t = \frac{1}{1-s} c_t = F(k_t, \ell(1-u_t)).$$

From the second equality, $u_t$ can be represented as the function of $c_t$ and $k_t$: $u(c_t, k_t)$. The fundamental equation (3.5) can therefore be modified as

$$k_{t+1} = sF(k_t, \ell(1-u(c_t, k_t))) + (1-\delta)k_t.$$

This framework is very Keynesian in spirit but clearly begs the question of how output can end up below full capacity, and hence be driven by demand. In a well-functioning market, this phenomenon could not occur. This book contains several chapters on frictions that could lead to something like the setting just described, and it then becomes central for policymakers to understand the exact nature of these frictions.

In all three cases above, we can represent $k_{t+1}$ as a function of $k_t$ and a shock ($A_t$, $\nu_t$, or $c_t$). This representation allows us to characterize the dynamics of $k_t$ (and other macroeconomic variables, such as $y_t$, $c_t$, and $i_t$) in response to these shocks.

### 3.5.2 Impulse responses

One method of describing how the economic variables respond to shocks is to draw an impulse-response function. Consider the RBC example above, with a Cobb-Douglas production function, an exogenous saving rate, and fixed labor supply. Suppose that before period 0, the value of $A_t$ is constant at $\bar{A}$. After a sufficiently long time, the value of $k_t$ settles close to the corresponding steady-state value $\bar{k}$. Then suppose that at time 0, $A_0$ is $(\varepsilon \times 100)\%$ higher, that is, $A_0 = (1+\varepsilon)\bar{A}$. For $t = 1, 2, 3, \ldots$, the value of $A_t$ is $A_t = (1+\rho^t\varepsilon)\bar{A}$, where $\rho \in (0, 1)$.

We can then generate the resulting time-path of $k_t$, starting from $k_0 = \bar{k}$, with

$$k_{t+1} = sA_t k_t^{\alpha} \ell^{1-\alpha} + (1-\delta)k_t, \tag{3.14}$$

for $t = 0, 1, 2, \ldots$. This time path is called the impulse-response function. The time-path of $A_t$ is drawn in Figure 3.9—the impulse—along with the response of $k_t$. For the impulse-response function for $k_t$ we use $s = 0.2$, $\delta = 0.046$, and $\alpha = 1/3$. The starting value of $A$ and the value of $\ell$ are normalized to 1. The initial value of the shock to $A$, $\varepsilon$, is 1%, and the persistence $\rho = 0.9$.

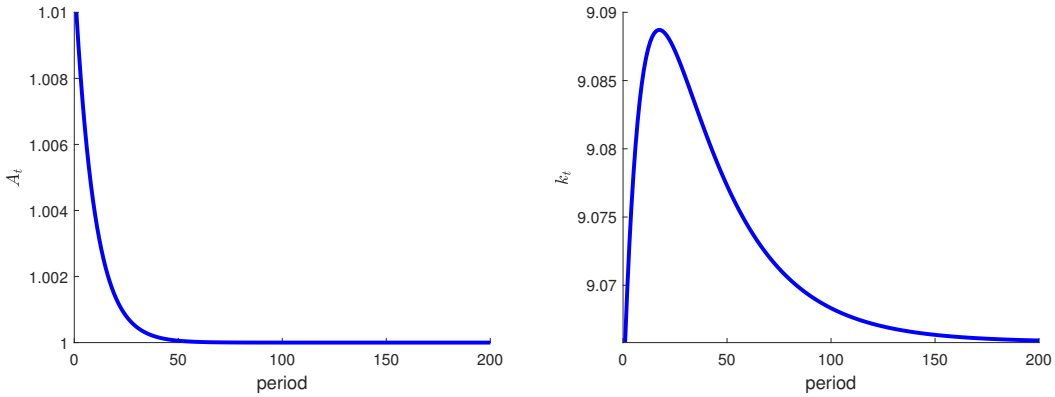



Figure 3.9: Impulse response: how $A_t$ (left) affects $k_t$ (right)

In the figure, $k_t$ increases from its steady-state value of 9.066 to a maximum value of 9.089 and then gradually goes back to the steady-sate value: a hump shape. Three properties of the graph are worthwhile emphasizing. First, the movement of $k_t$ is much slower than the movement of $A_t$. The deviation of $A_t$ becomes very close to zero around period 50, whereas the response of $k_t$ is more persistent. Second, the magnitude of the response of $k_t$ is significantly smaller than the impulse: the maximum deviation as a fraction to the steady-state level is $9.089/9.066 - 1 = 0.0025$, that is, 0.25%. This magnitude is much smaller than the initial $A_t$ deviation of 1%. Third, $k_t$ eventually comes back to the steady-state value. The convergence force is always at work when we consider the response to recurrent shocks, as in the business-cycle examples.

### 3.5.3 Log-linearized impulse responses

Often it is more convenient to approximate the dynamics around the steady state by a log-linarized system. Log-linearization expresses the system in terms of deviation of the logarithms of variables from their corresponding steady-state values. An arbitrary variable $x_t$ can be expressed as

$$x_t = \bar{x} e^{\hat{x}_t}, \tag{3.15}$$

where

$$\hat{x}_t \equiv \log\left(\frac{x_t}{\bar{x}}\right),$$

is the log deviation from the steady-state value $\bar{x}$. Note that because

$$\hat{x}_t \equiv \log\left(\frac{x_t}{\bar{x}}\right) \approx \frac{x_t - \bar{x}}{\bar{x}},$$

where the approximation is the first-order Taylor expansion around $\bar{x}$, $\hat{x}_t$ can be interpreted as the percent deviation from the steady state. Also note that

$$\bar{x}e^{\hat{x}_t} \approx \bar{x}(1 + \hat{x}_t) \tag{3.16}$$

from the first-order Taylor approximation of the expression.

Now, consider the impulse-response experiment of the system (3.14). Using the transformation (3.15),

$$\bar{k}e^{\hat{k}_{t+1}} = s\bar{A}e^{\hat{A}_t}(\bar{k}e^{\hat{k}_t})^{\alpha}\ell^{1-\alpha} + (1-\delta)\bar{k}e^{\hat{k}_t}$$

holds. Using (3.16) and the fact that

$$\bar{k} = s\bar{A}(\bar{k})^{\alpha}\ell^{1-\alpha} + (1-\delta)\bar{k} \tag{3.17}$$

holds from the definition of $\bar{k}$, we obtain

$$\bar{k}\hat{k}_{t+1} = s\bar{A}(\bar{k})^{\alpha}\ell^{1-\alpha}(\hat{A}_t + \alpha\hat{k}_t) + (1-\delta)\bar{k}\hat{k}_t.$$

Because equation (3.17) implies $\delta\bar{k} = s\bar{A}(\bar{k})^{\alpha}\ell^{1-\alpha}$, this equation can be rewritten as

$$\hat{k}_{t+1} = (1 - \delta(1-\alpha))\hat{k}_t + \delta\hat{A}_t. \tag{3.18}$$

Let us again use $\lambda \equiv \delta(1-\alpha)$, which is the notation for the convergence speed in (3.11). In fact, the log-linearization procedure here is essentially the same as the procedure for obtaining the percentage deviation in the convergence section, in the sense that both are applying the first-order Taylor approximation to (3.14).

The above impulse-response experiment corresponds to setting $\hat{A}_0 = \varepsilon$ and $\hat{A}_t = \rho^t\varepsilon$ for $t = 1, 2, 3, \ldots$. Solving (3.18) with this specification yields

$$\hat{k}_{t+1} = \sum_{\tau=0}^{t}\rho^{t-\tau}(1-\lambda)^{\tau}\delta\varepsilon = \rho^t\frac{1 - \left(\frac{1-\lambda}{\rho}\right)^{t+1}}{1 - \frac{1-\lambda}{\rho}}\delta\varepsilon.$$

Quantitatively, the log-linear approximation performs well in our calibrated model. The maximum error (in the units of $K_t$) of the approximation is about 0.00001. If we were to plot the log-linear solution and the nonlinear solution in the same figure, the difference would not be visible.

Applying the above log-linearization procedure to the production function $y_t = A_t k_t^{\alpha}\ell^{1-\alpha}$, the log-deviation of output is

$$\hat{y}_t = \hat{A}_t + \alpha\hat{k}_t = \rho^t\frac{\alpha\delta}{\rho}\frac{1 - \left(\frac{1-\lambda}{\rho}\right)^{t}}{1 - \frac{1-\lambda}{\rho}}\delta\varepsilon$$

for $t = 1, 2, \ldots$ (for $t = 0$, $\hat{y}_0 = \hat{A}_0 = \varepsilon$). This equation explicitly describes how $y_t$ moves over the cycle, in response to the realization of the shock $\varepsilon$. Log-linearization allows us to derive this explicit expression.

An alternative to the log-linear approximation is a linear approximation in *levels*. Let the production function be $A_t F(k_t, \ell_t)$. Assuming that $\ell_t = 1$ for any $t$ and defining $f(k_t) \equiv F(k_t, 1)$, the fundamental equation (3.5) (with the modification in the production function) can be written as

$$k_{t+1} = g(k_t, A_t),$$

where $g(k_t, A_t) \equiv (1 - \delta)k_t + sA_t f(k_t)$. Using the notation we employed earlier,

$$\Delta k_{t+1} = g_k \Delta k_t + g_A \Delta A_t,$$

where $g_k$ and $g_A$ are the partial derivatives of $g(k, A)$ with respect to $k$ and $A$, respectively (evaluated at the steady state), and $\Delta k_t = k_t - \bar{k}$ and $\Delta A_t = A_t - \bar{A}$. When $f(k_t) = k_t^\alpha$, we obtain $g_k = 1 - \delta(1 - \alpha)$ and $g_A = s^{\frac{1}{1-\alpha}} \delta^{-\frac{\alpha}{1-\alpha}} A^{\frac{\alpha}{1-\alpha}}$.

# Chapter 4

# Dynamic optimization

In the Solow model presented in Chapter 3, the evolution of the capital stock, along with technological change, is the driving force of output growth. In the model, because agents are assumed to save a constant proportion $s$ of income, saving and, therefore, investment decisions are exogenously given. Cass (1965) and Koopmans (1965) developed the first optimizing models of growth by adding *microeconomic foundations*, i.e., by describing how saving comes about as a result of (rational) choice. These foundations were based on the inter-temporal trade-off between consumption today and consumption in the future, also known as the consumption-saving model. Augmenting the Solow model to incorporate endogenous investment decisions by individuals gave rise to the "optimizing neoclassical growth model," which is nowadays at the core of modern macroeconomic theory. This model delivers the same long-run implications as the original Solow model, but because individual utility is explicit, the model can also be used for analyzing normative issues (individual welfare outcomes). In addition, because the decisions are forward-looking, we can obtain richer policy implications even on the positive side. With the introduction of methods dealing with stochastic model elements ("shocks"), it also serves as the foundation of the real business cycle theory, which we will discuss in more detail in Chapter 12.

In this chapter, we introduce a simple dynamic optimization model and discuss the main characteristics of this class of models. A key objective is to describe how to characterize solutions to dynamic optimization problems, and to introduce discrete dynamic programming methods. We work with a representative agent who faces a dynamic trade-off, meaning that sacrifices today yield gains in the future. For example, in the standard consumption-saving model, the agent can save today (reducing current consumption) in exchange for increasing consumption possibilities in the future. In such a model, the saving rate $s$ is endogenously determined. When extending this model to incorporate production, saving and investment are optimally determined. Because saving depends on income in the neoclassical growth model, the level of capital and the productivity level will now affect the saving's rate.

A key underlying assumption, first introduced by Milton Friedman in 1957, is that individuals are *forward-looking*. This means that they do not make decisions based just on their current income (as in the textbook Keynesian model) but also considering their expected future income. In other words, consumption and saving decisions depend on expectations.

Subsequent research on consumption also incorporated the idea of *rational expectations*, which states that individuals use all the information available to them at each point in time to make the best possible forecast about the future. While this assumption may seem extreme, it is arguably a good first approximation to the average behavior of individuals for important decisions in their lives. Deviations from rational expectations are studied in behavioral economics, some of which will be discussed in later chapters of the book. In the present chapter, we assume that agents are fully rational and forward-looking. Because their future income is taken into account when making dynamic decisions, it is important to determine the time horizon of their decision-making process. There are two common approaches followed in the literature: (i) agents live a finite number of periods, or (ii) agents live forever. The latter is interpreted as a *dynastic* structure in which individuals alive today care about the welfare of their descendants, as discussed later in this chapter. Because the infinite-horizon models require more mathematical sophistication, we start with a finite-horizon model. We also discuss two alternative ways of solving dynamic optimization problems: using sequential methods and using recursive methods. Sequential methods involve maximizing over sequences. Recursive methods, also labeled dynamic programming methods, involve functional equations. We begin with sequential methods and then move to recursive methods.

## 4.1 A dynamic optimization problem

Economic decisions are made by agents. These could be: (i) individuals (or households) deciding how much of a good or service to consume, (ii) firms choosing how much to produce, or (iii) a government deciding on policy. In this chapter, we focus mainly on the decisions of individuals. The problem of the firm is described in detail in Chapter 5, whereas the government policy decisions are deferred to Chapter 13.

Agents live for $T$ periods, and this time horizon can, in principle, be finite or infinite. With reference to the next chapter, which deals with market economies inhabited by many consumers, we assume that consumers are all identical, so we now study a **representative agent**. This agent makes decisions over sequences of **allocations** in order to maximize a lifetime objective. Allocations are quantities of goods or services, such as consumption, hours worked, investment, etc. The typical dynamic optimization problem in sequential form studied in macroeconomics takes the following form

$$\begin{aligned} \max_{\{y_t, x_{t+1}\}_{t=0}^{T}} \quad & \sum_{t=0}^{T} \beta^t \hat{\mathcal{F}}(y_t) \\ \text{s.t.} \quad & x_{t+1} = h(x_t, y_t) \\ & x_{t+1} \in \Gamma(x_t). \end{aligned} \tag{P1}$$

In this problem, $\sum_{t=0}^{T} \beta^t \hat{\mathcal{F}}(y_t)$ represents the **objective function**. The summation between 0 and $T$ indicates that decisions must be made for each and all of those time periods. The function $\hat{\mathcal{F}}$ is the instantaneous objective, representing, for example, a per-period utility

function (for individuals) or profit function (for firms). The constant $\beta$ is referred to as the stationary discounting weights: it is our discount factor; $\frac{1}{\beta} - 1$ is the discount rate. They are called stationary because the ratio between the weights of any two different dates $t = i$ and $t = j > i$ only depends on the number of periods elapsed between $i$ and $j$, and not on the values of $i$ or $j$. In other words, $\frac{\beta^{t+k}}{\beta^t} = \beta^k$.

The sequence $\{y_t\}_{t=0}^T$ represents the choice variables, sometimes referred to as **control variables**. Examples of these are consumption, leisure, saving, and investment levels at each point in time. The sequence $\{x_t\}_{t=0}^T$ represents the **state variables**. Examples are the stocks of capital, debt, or housing. States and controls are related through the constraint $x_{t+1} = h(x_t, y_t)$. For example, $h$ can reflect a budget constraint or a production technology. The value of the initial state $x_0$ is exogenously given. To differentiate states from controls, notice that if an agent chooses $y_0$ in period 0, the value of the state next period, $x_1$, is automatically determined from $x_1 = h(x_0, y_0)$ because $x_0$ is exogenous.[1] In period 1, the choice of the control $y_1$ determines the state in the following period, $x_2$, and so on. In other words, choosing the control variable optimally at $t$ determines the value of the state variable at $t + 1$. Finally, $\Gamma(x_t)$ represents the feasible set which, given the value of the current state $x_t$, restricts the values that $x_{t+1}$ can take; we will be specific on standard forms $\Gamma$ can take later.

For dynamic optimization problems to be well defined (i.e., for solutions to exist), we need to make assumptions about the primitives of the model. Quite generally, we know from basic math—the Weierstrass theorem—that a continuous function attains both a maximum and a minimum when evaluated over a non-empty and compact set. A compact set (of values for a finite vector) means that the set is closed and bounded.[2] Sufficient conditions for this theorem are that (i) $\hat{\mathcal{F}}(y_t)$ is continuous for all $y_t$, (ii) $h(x_t, y_t)$ is continuous and, in its second argument, strictly monotone for all $(x_t, y_t)$, and (iii) $\Gamma(x_t)$ is non-empty, closed, and bounded for all $x_t$. Assumption (ii) ensures that we can express $y_t$ as a (continuous) function of $(x_t, x_{t+1})$, which allows us to write the period objective as a continuous function of this vector. Hence, we have an overall continuous function (of the sequence $\{x_1, x_2, \ldots, x_T, x_{T+1}\}$) to be maximized over a non-empty, compact set.

The two most widely used models in macroeconomics are the consumption-saving model and the neoclassical growth model (NGM). The agent's objective in both models is to maximize lifetime utility choosing the optimal path of consumption. They differ in the production structure and assets available to consumers. In the consumption-saving model, the agent has a time-varying endowment, and can save or borrow at market prices. The NGM, instead, considers the production structure from the Solow model but extends it by endogenizing consumption and investment decisions. In order to fix ideas, it is useful to discuss the main assumptions underlying these models and to map them into our generic formulation.

[1] Often, the problem explicitly says "with $x_0$ given," to emphasize that it is not a choice variable. However, the fact that it is not a choice variable is already clear since it is not listed among them (under the "max").

[2] By drawing simple graphical examples with non-continuous functions, open sets, and unbounded sets, you can illustrate what can go wrong and why a supremum or infimum may exist but neither a maximum nor a minimum exist.

### 4.1.1 The consumption-saving model

In the consumption-saving model, there is a representative agent who lives for $T$ periods and must choose the optimal stream of consumption $\{c_t\}_{t=0}^T$, where $c_t$ denotes consumption at time $t$. We can think of consumption at $t$ as a different good from consumption at $t+1$. Preferences are represented by a utility function $U\left(\{c_t\}_{t=0}^T\right)$. A standard assumption is that this function exhibits "additive separability":

$$U\left(\{c_t\}_{t=0}^T\right) = \sum_{t=0}^{T} \beta^t u\left(c_t\right).$$

Additive separability implies that the marginal utility of consumption at $t$ does not depend on consumption at other times. Notice that the per-period (or instantaneous) utility index $u(\cdot)$ does not depend on time either. The stationary discounting weights satisfy $0 < \beta < 1$, which is consistent with the observation that individuals seem to deem consumption at an early time more valuable than consumption further off in the future. Formally, if consumption were constant over time, i.e., $c_t = c$ for all $t$, the marginal utility of $c_t$ decreases in $t$ because $u'(c)$ is multiplied by $\beta^t$. Of course, consumption in the future can be more valued on the margin if there is less of it, i.e., if it is sufficiently low relative to consumption today.

We assume that the instantaneous utility function $u(c)$ satisfies the following properties.

1. $u(c)$ is strictly increasing.
2. $u(c)$ is strictly concave.
3. $\lim_{c \to 0} u'(c) = \infty$.

The first property states that individuals have positive marginal utility, $u'(c) > 0$. We also assume that marginal utility of consumption is diminishing, or $u''(c) < 0$. The last property is an Inada condition, stating that agents have infinite marginal utility of consumption as $c$ approaches zero.

We abstract from preferences over leisure and, hence, the determination of the number of hours worked (e.g., labor). These will be studied briefly in the next chapter, and more in depth in Chapter 9. For now, we simply assume that individuals have an endowment of one unit of time, and supply it inelastically to production. They are paid a wage $w_t$, which may vary over time. In addition, they have access to borrowing and lending. We denote their level of *assets* with $a_t$, with the understanding that $a_t < 0$ indicates that the agent has *debt*. For example, $a_t = 10$ means that another individual owes him or her 10 units of the consumption good, so the agent is a lender, whereas if $a_t = -10$ then the agent owes someone else that amount, so he or she is a borrower. The initial amount of assets is given by $a_0$. The net interest rate earned on savings is $r_t$. The budget constraint of the agent can be written as

$$c_t = w_t + (1 + r_t)a_t - a_{t+1}.$$

Borrowing allows the agent to consume more today (in the amount $-a_{t+1}$), but lowers consumption in the future because he or she would need to pay the principal and interest on the loan, $(1+r_t)a_t$. A standard assumption is that agents cannot borrow more than a certain amount, which is incorporated as the borrowing limit $a_{t+1} \geq \underline{a}$, with $\underline{a}$ being an exogenous constant. In addition, we impose that $a_{T+1} \geq 0$ in order to eliminate the possibility that the agent ends with a positive level of debt in the last period. Absent this constraint, the agent would choose to borrow up to the limit and never pay back, which clearly could not occur in equilibrium (since no lender would be willing to lend the funds in the last period). We will further discuss the issue of the borrowing limit and the terminal condition later in the chapter.

The dynamic optimization problem of the agent can now be written as

$$\begin{aligned} \max_{\{c_t, a_{t+1}\}_{t=0}^{T}} \quad & \sum_{t=0}^{T} \beta^t u\left(c_t\right) \\ \text{s.t.} \quad & c_t = w_t + (1+r_t)a_t - a_{t+1}, \forall t \in \{0, ..., T\} \\ & c_t \geq 0, \forall t \in \{0, ..., T\} \\ & a_{t+1} \geq \underline{a}, \forall t \in \{0, ..., T\} \\ & a_{T+1} \geq 0. \end{aligned} \tag{P2}$$

The key sequences to be optimally determined are consumption and asset holdings $\{c_t, a_{t+1}\}_{t=0}^{T}$, to maximize the lifetime utility of the agent. We assume that agents know the sequences of prices (wages and interest rates) when making decisions, and take them as given. We defer the discussion of how these are determined in equilibrium to Chapter 5. The main trade-off faced by the agent, then, is whether to consume today or to save (borrow) a unit of consumption in exchange for $1 + r_{t+1}$ additional (less) units in the future.

This model can be mapped to the generic specification as follows: the per-period objective $\hat{\mathcal{F}}\left(\cdot\right)$ corresponds to the instantaneous utility function $u\left(\cdot\right)$, the control variables are given by the consumption sequences $\{c_t\}_{t=0}^{T}$, and the states correspond to the level of assets $x_t = a_t$. The equation relating controls to states is the resource constraint, $h(a_t, c_t) = w_t + (1+r_t)a_t - c_t$. The feasible set is given by $\Gamma(a_t) = [\underline{a}, w_t + (1 + r_t)a_t]$. The lower bound ensures that $a_{t+1} \geq \underline{a}$, whereas the upper bound ensures that savings do not exceed current income so that $c_t \geq 0$ for all $t$.

### 4.1.2 The neoclassical growth model

The utility function in the NGM is the same as in the consumption-saving model: an agent wants to maximize his or her discounted lifetime welfare. The assumptions underlying the instantaneous utility function $u$ are the same as in the previous section: the marginal utility of consumption is positive but diminishing, and the Inada condition must hold.

What changes relative to the consumption-saving model is that this is not an endowment economy, but a production economy instead. Rather than focusing on assets delivering an exogenous return, we consider capital accumulation, where returns are determined by the productivity of capital. The production structure is identical to the one in the Solow model.

Following what we learned in Chapter 3, we assume that technology can be represented by the production function $y_t = f(k)$, where we have used the fact that aggregate labor $L = 1$. From the properties of $F(K, L)$, it is easy to show that $f$ is a strictly increasing and strictly concave function in $k$. In some applications, we relax the assumption to weak concavity in order to accommodate linear production functions. Capital evolves as in the Solow model, i.e.,

$$k_{t+1} = (1 - \delta)k_t + i_t \tag{4.1}$$

with $k_0$ given. We abstract from population growth and technological progress to simplify the exposition. The resource constraint is $c_t + i_t \leq y_t$, assuming that this is a closed economy where savings are equal to investment. We can combine it with equation (4.1), substituting away investment, to write the resource constraint in terms of consumption and capital:

$$c_t + k_{t+1} \leq f(k_t) + (1 - \delta)k_t. \tag{4.2}$$

There are alternative ways to specify how the markets for labor, consumption, and capital are organized. For example, we could consider firms who own the capital stock and accumulate it over time. They produce output and sell it to consumers, who work in firms. Alternatively, we could assume that agents own capital, make investment decisions, and rent it to firms every period in exchange for a rental rate. These decentralized production structures will be discussed at length in Chapter 5, where we also explain how prices are determined. In Chapter 6, we show that as long as markets are perfectly competitive, the allocations (i.e., quantities of consumption, capital, and investment) in a competitive equilibrium are Pareto optimal and solve the following *planning problem*.

$$\begin{aligned} \max_{\{c_t, k_{t+1}\}_{t=0}^{T}} \quad & \sum_{t=0}^{T} \beta^t u\left(c_t\right) \\ \text{s.t.} \quad & c_t + k_{t+1} \leq f\left(k_t\right) + \left(1 - \delta\right) k_t, \forall t \in \{0, ..., T\} \\ & c_t, k_{t+1} \geq 0, \forall t \in \{0, ..., T\}. \end{aligned} \tag{4.3}$$

The planner chooses allocations directly (note that there are no prices in the equations above) to maximize the lifetime welfare of all agents in the economy. Because all agents are identical, this objective is equivalent to maximizing the welfare of the representative agent. The key sequences to be optimally determined are consumption and capital $\{c_t, k_{t+1}\}_{t=0}^{T}$. The main trade-off faced by consumers is whether to consume today or in the future. Any amount of production that is not consumed can be saved, and therefore invested in new capital, yielding future consumption through the additional amount of goods produced, $f'\left(k_t\right) + 1 - \delta$. The rationale is analogous to that in the consumption-saving model. Because we assume strictly positive marginal utility of consumption, the resource constraint always holds with equality.

We can map this model to the generic specification as follows: the per-period objective $\hat{\mathcal{F}}\left(\cdot\right)$ corresponds to the instantaneous utility function $u\left(\cdot\right)$, the control variables are given by the consumption sequences $\{c_t\}_{t=0}^{T}$, and capital represents the states $x_t = k_t$. The equation relating controls to states is the resource constraint, $h(k_t, c_t) = f\left(k_t\right) + \left(1 - \delta\right) k_t - c_t$. Finally, the feasible set is given by $\Gamma(k_t) = [0, f\left(k_t\right) + \left(1 - \delta\right) k_t]$. The lower bound ensures

that $k_{t+1} \geq 0$, whereas the upper bound ensures that $c_t \geq 0$. The next step is to discuss how to solve for the optimal sequences.

## 4.2 Sequential methods: finite horizon

When the horizon is finite, $T < \infty$, it is possible to use the Kuhn-Tucker Theorem to solve a sequential maximization problem (4.3). The associated Kuhn-Tucker conditions are necessary and sufficient for an optimum if the objective function is strictly concave in the choice vector and the constraint set is closed, bounded, and convex. Let's start with the simplest case, in a two-period economy.

### 4.2.1 A two-period consumption-saving model

Consider the consumption-saving model when $T = 2$, that the agent is borrowing constrained $\underline{a} = 0$, and that initial assets holdings are zero, $a_0 = 0$. Moreover, let us consider a specific example where, for illustration, the endowment profile is decreasing over time, $w_0 \geq w_1$ and the interest rate is time invariant, $r_t = r$. The budget constraints can be written as

$$c_0 = w_0 - a_1 \quad \text{and} \quad c_1 = w_1 + (1+r)a_1,$$

where we already used the fact that the lower bound on assets holdings binds in the last period, $a_2 = 0$. Intuitively, the representative agent will never choose to save in period 1 if the economy ends in period 1. These constraints can be combined by replacing $a_1$ from the second period constraint into the first period constraint, and re-organizing terms, as follows

$$c_0 + \frac{c_1}{1+r} = w_0 + \frac{w_1}{1+r}.$$

This equation is known as the lifetime budget constraint. It states that the discounted value of lifetime consumption (left-hand side) must be equal to the lifetime value of income (right-hand side). Future periods are discounted by $1+r$, the relative price of consumption between periods 0 and 1. In addition, we have that $a_1 \geq 0$, which can alternatively be written as $w_0 - c_0 \geq 0$. The Lagrangian can be written as

$$\begin{aligned} \mathcal{L} &= u(c_0) + \beta u(c_1) + \mu \left\{ w_0 + \frac{w_1}{1+r} - c_0 - \frac{c_1}{1+r} \right\} \\ &+ \lambda\{w_0 - c_0\}, \end{aligned}$$

where we introduced the Lagrange multiplier $\mu$ on the lifetime budget constraint and $\lambda$ on the non-negativity constraint. We do not need to consider the non-negativity constraints on consumption because the Inada condition $\lim_{c_t \to 0} u'(c_t) = \infty$ ensures that the agent always

chooses $c_t > 0$.[3] The first-order conditions with respect to $c_0$ and $c_1$ are

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial c_0} &: u'(c_0) - \mu - \lambda = 0, \\ \frac{\partial \mathcal{L}}{\partial c_1} &: \beta u'(c_1) - \mu \frac{1}{1+r} = 0, \end{aligned}$$

and the additional Kuhn-Tucker conditions read

$$\lambda[w_0 - c_0] = 0 \tag{4.4}$$

$$w_0 - c_0 \geq 0 \quad \text{and} \quad \lambda \geq 0. \tag{4.5}$$

Consider an interior solution with $w_0 - c_0 > 0$, implying $\lambda = 0$. The assumption of a decreasing wage path $w_0 > w_1$, strictly decreasing marginal utility of consumption, and a mild assumption on $\beta(1+r)$ guarantee that the solution will indeed be interior. Combining these conditions, we obtain the first-order condition known as the **Euler equation**

$$u'(c_0) = \beta(1+r)u'(c_1).$$

The left-hand side captures the marginal cost of saving an additional unit (which causes a decrease in consumption in the initial period). The right-hand side captures the marginal benefit of saving, which is given by the discounted value of consumption in period $t = 1$ obtained through the returns to savings, $1 + r$. Note that when $\beta(1 + r) = 1$, the agent chooses to consume a constant proportion of lifetime earnings every period

$$c_0 = c_1 = \frac{w_0 + \beta w_1}{1 + \beta}.$$

Using the budget constraints, this in turn implies that $w_0 - c_0 = a_1 = \frac{w_0 - w_1}{2+r} > 0$.[4]

### 4.2.2 Generic $T$-period model

When the budget constraint is non-linear in the state variables, it is not possible to construct the lifetime budget constraint as we did in the example above. However, it is still possible to use the Kuhn-Tucker theorem to solve the finite-horizon model, as the next result shows.

**Result 1** *Consider a finite-horizon problem* (P1)*, with* $\Gamma(x_t) = [\underline{x}, \gamma(x_t)]$*. Now use the constraint* $x_{t+1} = h(x_t, y_t)$ *to solve for* $y_t$ *as* $y_t = \hat{h}(x_t, x_{t+1})$*, replace it in the instantaneous objective function, with* $\mathcal{F}(x_t, x_{t+1}) \equiv \hat{\mathcal{F}}\left(\hat{h}(x_t, x_{t+1})\right)$*, so that we obtain*

$$\max_{\{x_{t+1} \in \Gamma(x_t)\}_{t=0}^{T}} \sum_{t=0}^{T} \beta^t \mathcal{F}(x_t, x_{t+1}).$$

[3] The proof of this statement is straightforward and omitted for brevity.

[4] In the general case (where $\beta(1+r) = 1$ or $w_0 > w_1$ may not be met), we can proceed by first finding an interior *candidate* solution and then verifying whether or not $a_1$ satisfies its constraint (in this case that it be non-negative); if this works, it is a solution—given the strictly concave utility function and convex constraint set. If $a_1$ violates its constraint, set it equal to its boundary value and solve for $\lambda$, in order to verify that it is non-negative.

*Suppose that $\mathcal{F}(x_t, x_{t+1})$ is increasing in its first argument, decreasing in its second argument, continuously differentiable, and jointly concave in $(x_t, x_{t+1})$. If*

*(i)* $\mathcal{F}_2\left(x_t^*,\ x_{t+1}^*\right) + \beta\mathcal{F}_1\left(x_{t+1}^*,\ x_{t+2}^*\right) = 0$, $\forall t < T$ *with* $\left\{x_{t+1}^*\right\}_{t=0}^{T-1} \in int\Gamma(x_t)$.

*(ii)* $x_{T+1}^* = \underline{x}$.

*Then the sequence* $\left\{x_{t+1}^*\right\}_{t=0}^{T}$ *maximizes the objective.*

We omit the proof of this result here. In our section on infinite-horizon optimization below, we state the corresponding result for that setting and, in our appendix, prove it from first principles. That proof can easily be adapted to the present, finite-horizon case.

The result states that it is optimal to choose the lower bound of the feasible set in the last period, $\underline{x}$, and that an interior solution for $t < T$ satisfies the first-order condition (i), which represents the Euler equation for the general model. This equation is sometimes referred to as a "variational" condition (as part of "calculus of variation"): given two boundary conditions $x_t$ and $x_{t+2}$, we vary the intermediate value $x_{t+1}$ so as to achieve the best outcome. Combining these variational conditions, we notice that there is a total of $T+2$ equations and $T+2$ unknowns, namely the sequence of states, plus the initial and terminal conditions. This is called a *difference equation* in the sequence of state variables. It is a *second-order* difference equation because there are two lags of $x$ in the equation. Since the number of unknowns is equal to the number of equations, the difference equation system will typically have a solution.

### 4.2.3 The finite-horizon neoclassical growth model

In the NGM, we have that $\underline{x} = 0$. Solving for consumption from the resource constraint and inserting it into the objective, the problem can be written more compactly as

$$\max_{\{k_{t+1}\}_{t=0}^{T}} \sum_{t=0}^{T} \beta^t u\left(f(k_t) + (1-\delta)\,k_t - k_{t+1}\right) \quad \text{s.t.} \quad k_{t+1} \geq 0 \quad \forall t \leq T. \tag{P3}$$

Our assumptions on the utility and production functions ensure concavity of the objective function and a closed, bounded, and convex constraint set, so that the assumptions guaranteeing that there is a unique maximizer are fulfilled.[5] As in the two-period consumption-saving model, we are omitting the constraint $c_t \geq 0$, given the Inada condition in the utility function.

The Lagrangian associated to our maximization problem (P3) is

$$\mathcal{L} = \sum_{t=0}^{T} \beta^t \left\{u(f(k_t) + (1-\delta)\,k_t - k_{t+1}) + \mu_t k_{t+1}\right\},$$

[5]To show that the period objective is concave in $(k_t, k_{t+1})$, it is necessary to go beyond concavity in $k_t$ and $k_{t+1}$ separately: one needs to check that the Hessian is globally negative-definite.

where we introduced the Lagrange/Kuhn-Tucker multipliers $\beta^t \mu_t$ corresponding to the non-negativity constraints on capital for each period $t$. The first-order conditions are as follows.

$$\frac{\partial \mathcal{L}}{\partial k_{t+1}} : -u'(c_t) + \beta u'(c_{t+1})[f'(k_{t+1}) + 1 - \delta] + \mu_t = 0, \quad t \in \{0, ..., T-1\}. \tag{4.6}$$

$$\frac{\partial \mathcal{L}}{\partial k_{T+1}} : -u'(c_T) + \mu_T = 0. \tag{4.7}$$

The first-order condition in the last period is different from the previous ones because the economy ends at that time. Finally, the Kuhn-Tucker conditions also include

$$\begin{aligned} \mu_t k_{t+1} &= 0, \ t \in \{0, ..., T\} \\ k_{t+1} &\geq 0, \ t \in \{0, ..., T\} \\ \mu_t &\geq 0, \ t \in \{0, ..., T\}. \end{aligned} \tag{4.8}$$

Equation (4.8) is usually referred to as the **complementary slackness condition**; it parallels equation (4.4) above in the two-period consumption-savings problem. Because we assumed that $u(c_t)$ is increasing in its first argument, equation (4.7) implies that $\mu_T > 0$. From (4.8) evaluated at $t = T$, we find that $k_{T+1} = 0$. This result establishes the terminal condition for our maximization problem: consumers leave no capital for production after the last period, since they receive no utility from that capital and would rather use it for consumption during their lifetime. The insight is trivial here but will be relevant also when we consider an infinite-horizon economy.

Given the Inada conditions, in particular $\lim_{k \to 0} f'(k) = \infty$ and $\lim_{c \to 0} u'(c) = \infty$, it is optimal to set $k_{t+1} > 0$ for all $t < T$. Hence, the non-negativity constraint on capital will never be binding, implying that $\mu_t = 0$ for $t < T$. Replacing this in equation (4.6) delivers the Euler equation,

$$u'\left[f\left(k_t\right) + (1-\delta)k_t - k_{t+1}\right] = \beta u'\left[f\left(k_{t+1}\right) + (1-\delta)k_{t+1} - k_{t+2}\right]\left[f'\left(k_{t+1}\right) + 1 - \delta\right], \tag{4.9}$$

which holds for all $t \in \{0, ..., T-1\}$. This, together with the initial condition $k_0$ and the terminal condition derived before, $k_{T+1} = 0$, determines the capital sequence $\{k_{t+1}\}_{t=0}^{T}$ (e.g., $T+2$ equations and $T+2$ unknowns). Because the first-order conditions are sufficient in the example, there is a unique solution to the difference equation (4.9) describing the evolution of capital over time.

To interpret the key equation for optimization, it is useful to break the Euler equation down in three components:

$$\underbrace{u'\left(c_t\right)}_{\substack{\text{marginal cost} \\ \text{of investment}}} = \underbrace{\beta u'\left(c_{t+1}\right)}_{\substack{\text{utility increase} \\ \text{per unit invested}}} \cdot \underbrace{\left[f'\left(k_{t+1}\right) + 1 - \delta\right]}_{\text{return on investment}}.$$

The left-hand side represents the marginal cost of investing one unit of consumption today, generating a disutility loss of $u'(c_t)$. The right-hand side represents the marginal benefit of this investment. Higher investment increases capital next period, which produces additional output (plus un-depreciated capital) $f'(k_{t+1})+1-\delta$. This increases consumption next period, with a discounted utility gain of $\beta u'(c_{t+1})$.

### 4.2.4 Solving a finite-horizon model

We have seen how to derive the set of equations that will determine the solution to the maximization problem. How does one solve these equations, however? Several methods are available. The first one uses "backward induction": starting at the final period, $T$, we can use the resource constraint and the Euler equation iteratively moving backwards. This method is illustrated in the following example, which is designed to yield closed-form solutions at each stage (it is the only known example with strictly concave utility and production functions that can be solved analytically).

**Example 4.1** *Consider a $T$-period economy where utility is logarithmic, $u(c) = \log c$, the production function is Cobb-Douglas $f(k) = Ak^{\alpha}$, and there is full depreciation $\delta = 1$. The Euler equation (4.9) for $t < T$ becomes*

$$\frac{1}{Ak_t^{\alpha} - k_{t+1}} = \beta \frac{1}{Ak_{t+1}^{\alpha} - k_{t+2}} \cdot \alpha A k_{t+1}^{\alpha-1}. \tag{4.10}$$

*The last term is the marginal product of capital $f'(k) = \alpha A k^{\alpha-1}$. Evaluating equation (4.10) at period $t = T-1$, replacing the terminal condition $k_{T+1} = 0$ in its right-hand side, and simplifying delivers*

$$k_T = \frac{\alpha\beta}{1+\alpha\beta} A k_{T-1}^{\alpha}.$$

*We can now use this result, together with equation (4.10) evaluated at period $T-2$, to obtain*

$$k_{T-1} = \frac{\alpha\beta(1+\alpha\beta)}{1+\alpha\beta+(\alpha\beta)^2} A k_{T-2}^{\alpha}.$$

*Going backwards in time, it is possible to see that*

$$k_{T-t} = \frac{\alpha\beta(1+\alpha\beta+...+(\alpha\beta)^t)}{1+\alpha\beta+....+(\alpha\beta)^{t+1}} A k_{T-t+1}^{\alpha}.$$

*Using the fact that $\alpha\beta < 1$, we can use the properties of geometric series in the numerator and the denominator to simplify this expression. That, together with a change of variables, allows us to obtain a formula that describes the evolution of capital in closed form:*

$$k_{t+1} = \alpha\beta \frac{1-(\alpha\beta)^{T-t}}{1-(\alpha\beta)^{T-t+1}} A k_t^{\alpha}.$$

*Consumption becomes*

$$c_t = \frac{1-\alpha\beta}{1-(\alpha\beta)^{T-t+1}} A k_t^{\alpha}.$$

There are a few characteristics of the solution that we would like to highlight. First, we obtained an expression that, given the initial condition $k_0$, fully describes the evolution of capital, output, and consumption for the whole time horizon: the outcomes depend explicitly

on parameters and on $k_0$. Second, we see that it is optimal to save (and invest) a proportion $s_t = \alpha\beta\frac{1-(\alpha\beta)^{T-t}}{1-(\alpha\beta)^{T-t+1}}$ of output every period. In contrast to the assumption of the Solow model, the optimal saving rate depends on time (the time left to the final date $T$). But, like in the Solow model, the saving rate does not depend on the level of the capital stock. Two key parameters determine the saving rate: the discount factor and the degree of concavity of the production function. Third, although the utility function is strictly concave, consumption is not fully smoothed: it will, in general, vary over time. The reason for the lack of full smoothing is that the level of the capital stock influences the marginal return on saving when the production function is neoclassical: the higher the capital stock, the lower is this return. The Euler equation tells us that the higher the marginal return on saving, the higher the consumption growth should be, implying that for capital stocks below (above) steady state, consumption rises (falls) over time.[6]

It is possible to obtain analytical solutions in a few other cases; one is where the production function is linear and preferences are represented by a $u$ as given in (2.5); this function is often called the Constant Relative Risk Aversion (CRRA) utility function. Thus, $u(c) = \dfrac{c^{1-\sigma}-1}{1-\sigma}$, where $\sigma \geq 0$ and $\sigma \neq 1$.[7] The CRRA function is one of the most commonly used additively separable utility functions in macroeconomics. It has, as special cases,

$$\begin{array}{ll} \sigma = 0 & \text{linear utility,} \\ \sigma > 0 & \text{strictly concave utility,} \\ \sigma \to 1 & \text{logarithmic utility.} \end{array}$$

The limit case where $\sigma \to \infty$ is the zero function. However, the relevant limit should be seen as that obtained by raising $\sum_{t=0}^{\infty}\beta^t c_t^{1-\sigma}$ to a power $1/(1-\sigma)$, in which case the limit becomes a "Leontief function": $\min_t \{c_t\}_{t=0}^{\infty}$.[8]

The elasticity of intertemporal substitution (EIS) is defined as the percentage change in consumption between periods $t$ and $t+s$ in response to a percentage change in the returns to investment between the same two periods:

$$EIS \equiv \frac{\frac{\partial\left(\frac{c_{t+s}}{c_t}\right)}{\frac{c_{t+s}}{c_t}}}{\frac{\partial R_{t,t+s}}{R_{t,t+s}}}.$$

In the next example, we show that the CRRA function has a constant intertemporal elasticity of substitution, which is equal to $\frac{1}{\sigma}$. For this reason, a CRRA utility function is also sometimes referred to as a Constant Elasticity of Intertemporal Substitution (CEIS) utility function.

[6] The special case $\alpha = 1$ means that the marginal return on saving is always $A$. Then, unless $\beta A = 1$, the economy will grow or shrink over time at a constant rate, as will consumption. Consumption, along with capital, will thus only be constant if $\beta A = 1$.

[7] We do not consider uncertainty in this chapter, so the concept of the "relative risk aversion" is not directly relevant. Here this term serves only as a label for a class of utility functions.

[8] The stated utility function here is a monotone transformation of the original function and we know that the behavior given by a utility function is preserved under monotone transformations.

**Example 4.2** *Consider a $T$-period economy where utility is CRRA, $u(c) = \dfrac{c^{1-\sigma}-1}{1-\sigma}$ and the production function is linear $f_t(k_t) = R_t k_t$. In this case, the Euler equation:*

$$u'(c_t) = \beta u'(c_{t+1}) R_{t+1}.$$

*Replacing repeatedly, we have*

$$\begin{aligned} u'(c_t) &= \beta^k u'(c_{t+s}) \underbrace{R_{t+1}R_{t+2}...R_{t+s}}_{\equiv R_{t,t+s}} \\ u'(c) &= c^{-\sigma} \Rightarrow c_t^{-\sigma} = \beta^k c_{t+s}^{-\sigma} R_{t,t+s} \\ \frac{c_{t+s}}{c_t} &= \left(\beta^k\right)^{\frac{1}{\sigma}} \left(R_{t,t+s}\right)^{\frac{1}{\sigma}}. \end{aligned}$$

*This means that the EIS becomes*

$$\frac{\frac{\partial\left(\frac{c_{t+s}}{c_t}\right)}{\frac{c_{t+s}}{c_t}}}{\frac{\partial R_{t,t+s}}{R_{t,t+s}}} = \frac{\partial \log \frac{c_{t+s}}{c_t}}{\partial \log R_{t,t+s}} = \frac{1}{\sigma}.$$

*When $\sigma \to 1$, the relative expenditure shares $c_t/(c_{t+s}/R_{t,t+s})$ do not change: this corresponds to the logarithmic case. When $\sigma > 1$, an increase in $R_{t,t+s}$ would lead $c_t$ to increase and investment to decrease: the income effect, leading to smoothing across all goods, is larger than the substitution effect. Finally, when $\sigma < 1$, the substitution effect is stronger than the income effect: investment rises whenever $R_{t,t+s}$ increases. When $\sigma = 0$, the elasticity is infinite and investment responds discontinuously to $R_{t,t+s}$.*

Another, and absolutely central, reason that the CRRA utility function plays an important role for us is that it is the only one that is consistent with balanced growth, as described in Section 4.3.3.

## 4.3 Sequential methods: infinite horizon

In this section, we extend our model to infinite periods. The main advantage of an infinite horizon is that the household problem becomes stationary: the maximization problem at date $t$ is exactly the same as in period $t+1$ (for a given starting level of capital). This property is in contrast to that in the previous section, where decisions were significantly affected by how many periods the individual had left (see Example 4.1). A large number of macroeconomic applications, particularly those studying the long-run evolution of aggregate economic variables, use infinite-lived agents as their main building block.

For the typical models that macroeconomists use, the infinite-horizon version behaves very similarly to the finite-horizon version when the latter's remaining time horizon is long

enough.[9] However, let us also discuss whether an infinite time horizon is a sensible assumption: after all, people do not live forever. However, to the extent that individuals are altruistic, they care about their descendants. Let $u(c_t)$ denote the utility flow to generation $t$. We can then interpret $\beta^t$ as the weight individuals attach to the utility enjoyed by their descendants, $t$ generations down the family tree. Their total welfare is given by $\sum_{t=0}^{\infty} \beta^t u(c_t)$. As long as $\beta < 1$, agents care more about themselves than about their offspring.[10]

### 4.3.1 Mathematical considerations

Because agents are now choosing infinite sequences of consumption and investment, models with an infinite time horizon demand more advanced mathematical tools.

A basic question is whether the solution to the planner's problem exists once choices are no longer elements of the Euclidean space $\Re^T$. In more general notation, suppose we are seeking to maximize a function $U(x)$, where $x \in S$ and $S$ is a set that includes infinite sequences. If $U$ is continuous, we can invoke the Weierstrass theorem, provided that the set $S$ is nonempty and compact. For finite sequences, continuity and compactness are defined in standard ways. But for infinite-dimensional sequences, several issues arise. How do we define continuity in this setup? What is an open set? What does compactness mean? Answering these questions in detail is beyond the scope of this book; we refer you, for example, to Stokey and Lucas (1989). For illustration, we will, however, provide some specific examples where the maximization problem may be ill-defined (i.e., have no solution) unless a set of necessary conditions holds.

**Unbounded utility** Continuity of the objective requires boundedness. A necessary condition for the lifetime utility $U$ to be bounded is that consumption streams do not yield "infinite" utility. If two consumption streams do so, they cannot be compared and the maximization problem is ill-defined. For example, consider a plan specifying equal amounts of consumption each period, $\{c_t\}_{t=0}^{\infty} = \{\overline{c}\}_{t=0}^{\infty}$, delivering $U = \sum_{t=0}^{\infty} \beta^t u(\overline{c})$. Clearly, this function is unbounded (e.g., does not have a finite limit) unless $\beta < 1$.

This may not be sufficient, however. Suppose that the constraints allow for a constantly increasing consumption stream $\{c_t\}_{t=0}^{\infty} = \left\{c_0 (1+\gamma)^t\right\}_{t=0}^{\infty}$. The lifetime utility is now $U = \sum_{t=0}^{\infty} \beta^t u\left(c_0 (1+\gamma)^t\right)$. Even if $\beta < 1$ (so $\beta^t$ is decreasing to 0), the argument inside the utility function is growing at rate $\gamma$. The shape of the utility function, hence, is key to determining whether the maximization problem is well defined. In the case of a CRRA utility function $u(c) = (c^{1-\sigma} - 1)/(1-\sigma)$, we obtain a level of utility that is a geometric sum. Hence, boundedness requires $\beta(1+\gamma)^{1-\sigma} < 1$. If $\sigma < 1$, so that there is less than logarithmic curvature, this requirement involves an upper bound on $\gamma$ for utility to be a

[9] Game theory is a sharp contrast here: we know that an infinite horizon can then open up to the existence of many equilibria, e.g., the trigger-strategy outcomes in repeated games.

[10] In this simple example we attach only one consumption level to each generation. The example could be extended to the case where people of generation $t$ consume in multiple periods; though more complicated to formulate, this extension would be straightforward and the main insights would carry over.

positive, finite number. If $\sigma > 1$, positive growth ($\gamma > 0$) implies that the boundedness condition is met; with negative growth at a sufficiently high rate, it will not be.[11]

**Constraint sets that are "too large"** The problem can be ill-defined if the constraint sets are "too large". In the consumption-saving model discussed at the outset of the chapter, we imposed the constraint that borrowing could not exceed the exogenous amount $\underline{a}$. This condition ensures that the constraint set is bounded, and hence that the problem is well defined.

In some applications, however, imposing the constraint $a_{t+1} \geq \underline{a}$ for each $t$ can be "too restrictive"; it will rule out feasible borrowing. Instead, for example, we could only impose that $a_{T+1} \geq 0$, i.e., that the agent cannot borrow in the very last period. This terminal constraint is important: if this constraint was not imposed, any individual would have incentives to accumulate an infinite amount of debt going into the final period. Such a restriction also makes sense: knowing that the consumer would engage in such a scheme, no lender would be willing to lend at any positive rate, as the loan would be defaulted with probability 1.

With an infinite horizon, a final-period constraint loses meaning in a literal sense. Instead, we impose a constraint that is its appropriate infinite-period extension, known as the **no Ponzi game** (nPg) condition. In words, this condition rules out "borrowing at infinity, measured in present value." This requirement represents a restriction on the agent's constraint set, preventing it from being so large as to allow the agent to attain arbitrarily high utility. To see how the condition comes about, let us look at a simple example.

Suppose we endow a consumer with a given initial amount of net assets, $a_0$, representing claims against other agents. Additionally, suppose that the agent has no other sources of income, so the budget constraint is

$$c_t + a_{t+1} = (1 + r)a_t, \forall t \geq 0,$$

where we assume that $r > 0$.[12] Consider a candidate solution to consumer's maximization problem $\{c_t^*\}_{t=0}^{\infty}$. Absent further constraints, the agent could improve on $\{c_t^*\}_{t=0}^{\infty}$ as follows:

1. Let $\tilde{c}_0 = c_0^* + \epsilon$, with $\epsilon > 0$, thus making $\tilde{a}_1 = a_1^* - \epsilon$.
2. For every $t \geq 1$ leave $\tilde{c}_t = c_t^*$ by setting $\tilde{a}_{t+1} = a_{t+1}^* - \epsilon(1 + r)^t$.

Given a strictly increasing utility function, the agent is clearly better off under this alternative consumption allocation, which satisfies the budget constraint period-by-period. Because this sort of improvement is possible for *any* candidate solution, there cannot be a maximum for lifetime utility.[13] Note that with this alternative allocation, the agent's debt is growing

[11] Negative growth, i.e., $\gamma < 0$, does not occur in the context of our standard balanced-growth model, but it can hypothetically occur if natural resources are assumed to be finite and essential for production. In this case, utility could hence become unboundedly negative—recall that $u$ is negative in this case—and there may even be no allocation of resources with finite utility. For more, see Chapter 23.

[12] Cases with $r \leq 0$ can sometimes be relevant but we do not consider them here.

[13] One could imagine a maximum if an arbitrary upper bound is placed on consumption at each date, but such ad-hoc assumptions are undesirable.

without bound at rate $1+r$, and it is never repaid. This type of scheme, borrowing $(1+r)a_t$ every period $t$ to keep rolling the debt, is often called the "Ponzi scheme" or the "Ponzi game". It is crucial in the infinite-horizon model to impose a constraint that rules out the Ponzi scheme. A condition that works is the nPg condition in the following form:

$$\lim_{t\to\infty} \frac{a_{t+1}}{(1+r)^t} \geq 0. \tag{4.11}$$

Intuitively, the agent cannot engage in borrowing and lending so that their "terminal asset holdings" (in present-value terms) are negative, because this means that they would borrow and not pay back.

We can use the nPg condition to simplify, or *consolidate*, the sequence of budget constraints. To do this, solve for $a_1$ from the budget constraint in period 1 and substitute the expression into the budget constraint in period 0: we obtain a budget containing $c_0$, $c_1$, $a_0$, and $a_2$. Next, replace $a_2$ in this expression by solving for it in the next budget constraint, and proceed this way forward. After $T$ substitutions, we obtain

$$\sum_{t=0}^{T} c_t \frac{1}{(1+r)^t} = a_0(1+r) - \frac{a_{T+1}}{(1+r)^T}.$$

Taking limits, we arrive at

$$\sum_{t=0}^{\infty} c_t \frac{1}{(1+r)^t} = a_0(1+r) - \lim_{T\to\infty} \frac{a_{T+1}}{(1+r)^T} \leq a_0(1+r), \tag{4.12}$$

where the inequality comes from the nPg condition. This is the lifetime budget constraint in the infinite horizon model.

We motivated ruling out Ponzi schemes as a natural infinite-horizon extension of the constraint $a_{T+1} \geq 0$ in a finite-horizon model. There is also a constraint often labeled the "natural borrowing limit." It can be used both in the finite-horizon and the infinite-horizon cases and it captures, precisely, the notion of a loosest possible constraint on borrowing: the only restriction is that you are able to pay back, in a present-value sense, if you set consumption to zero at all future times. In the particular model here, the natural borrowing limit is zero at all times, but in general the constraint depends on the future stream of non-asset income. This case is analyzed in Appendix 4.A.1, which shows that imposing the natural borrowing limit, imposing the lifetime budget constraint, and imposing the nPg condition are all equivalent.

Note that the lifetime budget constraint is often written with equality, i.e.,

$$\sum_{t=0}^{\infty} c_t \frac{1}{(1+r)^t} = a_0(1+r), \tag{4.13}$$

because, as we shall see below, $\lim_{T\to\infty} a_{T+1}/(1+r)^T$ will never be chosen to be positive, so long as the utility function is strictly increasing. The reason is that assets do not themselves

contribute to utility, so it is always better to increase consumption as long as it is positive. The inequality in the opposite direction of (4.11),

$$\lim_{T\to\infty}\frac{a_{T+1}}{(1+r)^T}\leq 0, \tag{4.14}$$

therefore, is a part of the optimality condition. The condition (4.14) will be referred to as the **transversality condition** (TVC). The equality (4.13) is the combination of the inequalities (4.12) (which comes from the nPg condition (4.11)) and the TVC (4.14). The nPg and the TVC are different in nature: the former is a restriction on what paths the consumer is allowed to choose whereas the latter is a self-imposed condition—it is chosen by the consumer.

Let us now work out a full solution to the problem above.

**Example 4.3** *Consider the infinite-horizon version of the consumption-saving model (without borrowing constraints) given a logarithmic utility function* $u(c)=\log c$*. The optimization problem is:*

$$\begin{array}{rl}\max\limits_{\{c_t,a_{t+1}\}_{t=0}^{\infty}} & \sum\limits_{t=0}^{\infty}\beta^t\log c_t\\ s.t. & c_t+a_{t+1}=a_t(1+r),\forall t\geq 0\\ & \textit{nPg condition.}\end{array}$$

*To solve this problem, replace the period budget constraints with a consolidated one evaluated at equality (hence, already using the TVC, i.e., the fact that it is not optimal to save a positive amount, in present-value terms, at infinity):*

$$\sum_{t=0}^{\infty}c_t\left(\frac{1}{1+r}\right)^t=a_0(1+r).$$

*With this simplification, the first-order conditions are*

$$\beta^t\frac{1}{c_t}=\lambda\left(\frac{1}{1+r}\right)^t,\forall t\geq 0,$$

*where* $\lambda$ *is the Lagrange multiplier associated with the consolidated budget constraint. From the first-order conditions it follows that*

$$c_t=[\beta(1+r)]^t c_0,\forall t\geq 1.$$

*Substituting this expression into the consolidated budget constraint, we obtain*

$$\begin{aligned}\sum_{t=0}^{\infty}\beta^t(1+r)^t\frac{1}{(1+r)^t}c_0 &= a_0(1+r)\\ c_0\sum_{t=0}^{\infty}\beta^t &= a_0(1+r).\end{aligned}$$

*From here,* $c_0=a_0(1-\beta)(1+r)$*, and consumption in the periods* $t\geq 1$ *can be recovered from* $c_t=[\beta(1+r)]^t c_0$*.*

**Sufficient conditions: the transversality condition (TVC)** In general, the infinite-horizon maximization problems involve the same mathematical techniques as the finite-horizon ones. In particular, we make use of (Kuhn-Tucker) first-order conditions. In the neoclassical growth model, these lead to a second-order difference equation, the Euler equation, defining a path for the state variable given the initial condition $k_0$. But unlike in the finite horizon case, where it was optimal to set $k_{T+1} = 0$, there is no final condition that allows us to pin down the sequence of capital. Therefore, the difference equation that characterizes the first-order condition may have an infinite number of solutions. To determine the solution, we need to make use of an additional optimality condition that we already briefly discussed: the *transversality condition*. This condition, which we will now discuss in more general terms, captures the principle that it cannot be optimal for an agent to choose a sequence of capital involving, in present-value utility terms, a positive shadow value as $t \to \infty$. In the consumption-saving problem such behavior is clearly sub-optimal: lower savings would be feasible and yield higher lifetime utility.

We will not prove the necessity of the TVC here. We will, however, provide a sufficiency condition for a generic optimization problem. We will, moreover, offer a proof strategy that also allows us to derive what form the TVC must take (it is not always obvious what its precise form should be). The message of the following proposition is that if we have a convex maximization problem (utility is concave and the constraint set convex) and a sequence $\{x_{t+1}\}_{t=0}^{\infty}$ that satisfies the Kuhn-Tucker first-order conditions and the transversality condition, then indeed we have a maximum. Formally, we have the following.

**Proposition 4.4** *Consider the infinite-horizon version of the maximization problem* (P1)*, where we use* $x_{t+1} = h(x_t, y_t)$ *to replace* $y_t$ *in the objective:*

$$\max_{\{x_{t+1}\in\Gamma(x_t)\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t \mathcal{F}(x_t, x_{t+1}).$$

*If* $x_{t+1}^* \in$ *int* $\Gamma(x_t)$ *for all* $t$*,*

*(i) Euler equation:* $\mathcal{F}_2\left(x_t^*,\ x_{t+1}^*\right) + \beta\mathcal{F}_1\left(x_{t+1}^*,\ x_{t+2}^*\right) = 0\ \forall t$

*(ii) TVC:* $\lim_{t\to\infty} \beta^t \mathcal{F}_1\left(x_t^*,\ x_{t+1}^*\right) x_t^* = 0,$

$\mathcal{F}\left(x_t, x_{t+1}\right)$ *is jointly concave in* $\left(x_t, x_{t+1}\right)$ *and increasing in its first argument, and* $\Gamma(x)$ *is a convex set for all* $x$*, then* $\left\{x_{t+1}^*\right\}_{t=0}^{\infty}$ *maximizes the objective.*

**Proof.** See Appendix 4.A.3 ■

The proof is relatively mechanical: it repeatedly uses the Euler equation, (i), and a definition of concavity, namely, that a concave function is always globally below its tangent hyperplane. It thus shows that the stated sequence gives a higher value to the objective than any other feasible sequence so long as a remaining inequality is met. That remaining inequality is met when the TVC is satisfied: condition (ii) of the theorem.

In the neoclassical growth model, the partial derivative $\mathcal{F}_1(x_t, x_{t+1})$ becomes $u'(c_t)[f'(k_t)+1-\delta]$, which corresponds to the marginal utility of increasing capital in period $t$. The TVC thus reads

$$\lim_{t\to\infty} \beta^t u'(c_t)[f'(k_t)+1-\delta]k_t = 0.$$

It states that the discounted present-value of each additional unit of capital times the stock of capital has to be zero in the limit. If this requirement is not met, it would be beneficial to modify the path of capital in order to increase consumption, generating higher lifetime utility, without violating feasibility.

In cases where a proposed chosen sequence becomes stationary in the limit—such as in the neoclassical growth model without growth—the TVC is satisfied "automatically" so long as $\beta < 1$. When consumption and capital grow, i.e., due to technical change, the condition is less trivial. Under the form of utility that allows balanced growth, $u(c) = \frac{c^{1-\sigma}-1}{1-\sigma}$, $\beta^t u'(c_t)k_t$ will grow at the gross rate $\beta\gamma^{1-\sigma}$, since both $c_t$ and $k_t$ grow at gross rate $\gamma$ on a balanced path. The transversality condition thus requires $\beta\gamma^{1-\sigma} < 1$, which we recall is also the condition that makes utility bounded.

It is yet again worth emphasizing that the transversality condition and the no-Ponzi game conditions are conceptually distinct. The TVC (jointly with the Euler equation being met at all times) is a sufficient condition for optimization. It states that the value of assets cannot be positive in the limit, because the agent could otherwise be made better off by reducing them. Hence, it is eliminating sequences that cannot be optimal by ruling out over-accumulation of wealth.[14] The nPg condition, on the other hand, is an institutional constraint ensuring that the agent cannot have positive debt (in present value terms) in the limit. An agent would want to choose to increase debt unboundedly if he or she were allowed to violate it. The latter could not reasonably be thought to occur in a market economy and therefore the constraint is imposed. In sum, the TVC is a self-imposed constraint ruling out over-accumulation of assets, whereas the nPg condition is an externally imposed constraint ruling out of extreme under-accumulation (exploding debt).

### 4.3.2 Solving the infinite-horizon neoclassical growth model

The infinite-horizon maximization problem can be solved taking the limit of the solution we found in the finite horizon case, and making sure that the transversality condition is met. A question, to which we will return later, is whether this model, like the Solow model, will imply global convergence to a steady state. A preliminary inquiry involves finding the set of steady states. This is straightforward: the Euler equation (4.9) implies that for any constant positive level of consumption—which means that $u'(c_t) = u'(c_{t+1})$ can be eliminated from the equation—we obtain

$$1 = \beta(f'(\overline{k}) + 1 - \delta). \tag{4.15}$$

Given that $f$ is strictly concave and satisfies Inada conditions, this implies a unique solution.

[14] The nature of this argument is that TVC is a necessary condition for optimization, which it is under some conditions; we do not prove this here.

We will return to the general formulation later, once we have covered recursive methods, and then establish global convergence to $\overline{k}$ from any positive starting level of capital. For now, we will look at an example where we can solve for dynamics explicitly.

**Example 4.5** *Consider an infinite-horizon economy where utility is logarithmic, $u(c) = \log c$, the production function is Cobb-Douglas $f(k) = Ak^{\alpha}$, and there is full depreciation: $\delta = 1$. Truncating the economy at $t = T$ delivers the same optimality conditions and analytical expression for capital next period that we found in Example 4.1. Denote the solution to the truncated problem with $k_{t+1}^{T}$. Taking limits, we can find a candidate solution to the infinite horizon problem $k_{t+1}$ as*

$$\begin{aligned} \lim_{T\to\infty} k_{t+1}^{T} &= \lim_{T\to\infty} \alpha\beta \frac{1-(\alpha\beta)^{T-t}}{1-(\alpha\beta)^{T-t+1}} Ak_t^{\alpha} \\ k_{t+1} &= \alpha\beta Ak_t^{\alpha} \end{aligned} \tag{4.16}$$

*To check that the TVC holds, as pointed out above, we only need to look at a limit and not at other aspects of the sequence for capital. So, (i), (4.16) implies that the capital sequence converges to a limit (as will, then, consumption) and, (ii) therefore the limit behavior of the expression $\beta^t \mathcal{F}_1(k_t,\ k_{t+1})\, k_t$ will boil down to a constant times $\beta^t$. Thus, the TVC will be met. To see the concrete expressions, first note that $\mathcal{F}_1(k_t,\ k_{t+1}) = u'(c)f'(k_t)$, since consumption is simply $c = f(k_t) - k_{t+1}$. The TVC can be written as*

$$\lim_{t\to\infty} \beta^t u'(c_t) f'(k_t) k_t = \lim_{t\to\infty} \beta^t \frac{A\alpha k_t^{\alpha-1}}{Ak_t^{\alpha} - k_{t+1}} k_t.$$

*Using equation (4.16) and simplifying delivers $\lim_{t\to\infty} \beta^t \alpha/(1-\alpha\beta) = 0$, since $\beta < 1$.*

Another method typically used to solve infinite horizon models analytically is called "guess and verify." The strategy consists of guessing a generic functional form for $k_{t+1}$ as a function of $k_t$ and using the Euler equation to verify that the guess is correct. We illustrate this method in the following example.

**Example 4.6** *Consider the economy from Example 4.5. Guess that $k_{t+1} = sAk_t^{\alpha}$, where $s$ is an unknown parameter (the saving rate). Insert this guess into the Euler equation (4.9) to obtain*

$$\frac{1}{Ak_t^{\alpha} - sAk_t^{\alpha}} = \beta\alpha A \frac{k_{t+1}^{\alpha-1}}{Ak_{t+1}^{\alpha} - sAk_{t+1}^{\alpha}}.$$

*After some manipulations, we can verify that $s = \alpha\beta$ satisfies the equation above no matter what $k_t$ is.*

An obvious complication with this approach is of course how to come up with an insightful initial guess. A standard procedure is to use a combination of the truncation-at-$T$ method and the guess and verify method. In other words, it is possible to solve the problem backwards

a couple of periods to get a sense of the possible functional form, and then verify whether this guess is correct using the Euler equation.

It is worth noticing that, in the infinite-horizon economy of Example 4.5, the stock of capital evolves as assumed by the Solow model: agents invest a *constant* proportion, $s = \alpha\beta$, of their income every period. Moreover, using what we have learned in Chapter 2, we can show that this economy converges to a steady state $\overline{k}$. To compute it, simply evaluate our solution at the steady state, $k_t = k_{t+1} = \overline{k}$,

$$\overline{k} = \beta\alpha A\overline{k}^{\alpha} \Rightarrow \overline{k} = (\beta\alpha A)^{\frac{1}{1-\alpha}},$$

which is identical to what we found in Chapter 3, given $\delta = 1$.

In general, when closed-form solutions are not possible to solve for, we need to resort to numerical methods. It is possible to solve for sequences numerically, but seeking numerical solutions using dynamic programming is also possible. Dynamic programming also delivers conceptual insights. The following section turns to this method.

In the case of a relatively simple model like the NGM, yet another method of characterizing the dynamics is to use *phase diagrams*. The solution to a general NGM can be written as a set of two difference equations for two variables $(k_t, c_t)$:

$$k_{t+1} - k_t = f(k_t) - \delta k_t - c_t$$

and

$$u'(c_t) = \beta u'(c_{t+1})(f'(k_{t+1}) + 1 - \delta).$$

The first is the resource constraint for the social planner, and the second is the Euler equation. We can look for the dynamic path of $(k_t, c_t)$ that satisfies these two difference equations, together with the nonnegativity constraints for both variables and the TVC. The graphical analysis using the phase diagram is explained in Appendix 4.A.4.

### 4.3.3 Balanced growth in the neoclassical growth model

The economies described in the previous sections eventually converge to a steady state, exhibiting no growth in the long run. It is possible to study a version of the NGM in which there is balanced growth, with a production structure similar to the one studied in Chapter 3.2. There, we considered $Y_t = F(K_t, A_t L_t)$, where $A_t L_t$ denotes "efficiency units of labor." An important assumption that allows us to obtain balanced growth is that $L_t$ and $A_t$ grow at the (constant) rates $n$ and $\gamma$, respectively. Because of population growth, we now need to take a stand on the welfare criterion by giving weight to the current population and people born in the future. In what follows, we assume that the social planner is "utilitarian" and maximize the sum of utility in the entire economy:

$$\sum_{t=0}^{\infty} \beta^t L_t u\left(c_t\right),$$

where $c_t = C_t/L_t$ is the per capita consumption. An alternative formulation is to assume the social planner cares about the "per capita" utility in each period: $\sum_{t=0}^{\infty} \beta^t u(c_t)$. As will become clear with the procedure below, this different assumption amounts to a different discount factor by the social planner. In our benchmark formulation, with positive population growth, we give more weight to future utility than in the alternative formulation. Quantitatively, the population growth rates tend to be small in advanced economies, and the impact of the alternative assumptions on the outcome is unlikely to be significant.

The resource constraint with growth becomes

$$C_t = F(K_t, A_t L_t) + (1-\delta)K_t - K_{t+1}.$$

Defining variables in "per efficiency units of labor", $\tilde{x}_t = X_t/(A_t L_t)$, we can follow similar steps as those in Section 3.2 to write the resource constraint as

$$\tilde{c}_t = f(\tilde{k}_t) + (1-\delta)\tilde{k}_t - (1+\gamma)(1+n)\tilde{k}_{t+1}. \tag{4.17}$$

We further assume that the instantaneous utility is CRRA, $u(c_t) = c_t^{1-\sigma}/(1-\sigma)$, $\sigma > 0$ and $\sigma \neq 1$. Appendix 4.A.2 shows that this function (and the logarithmic function, which should be seen as the case $\sigma \to 1$) is the only form that is consistent with balanced growth. Normalizing $A_0 = 1$ and $L_0 = 1$, we can rewrite the objective function as

$$\begin{aligned}\sum_{t=0}^{\infty} \beta^t L_t \frac{c_t^{1-\sigma}}{1-\sigma} &= \sum_{t=0}^{\infty} \beta^t (1+n)^t \frac{((1+\gamma)^t \tilde{c}_t)^{1-\sigma}}{1-\sigma} \\ &= \sum_{t=0}^{\infty} \tilde{\beta}^t \frac{\tilde{c}_t^{1-\sigma}}{1-\sigma},\end{aligned} \tag{4.18}$$

where now we use the adjusted discount factor $\tilde{\beta}$:

$$\tilde{\beta} \equiv \beta(1+n)(1+\gamma)^{1-\sigma}. \tag{4.19}$$

The maximization problem for the social planner can then be written as maximizing (4.18) subject to the resource constraint (4.17) and the non-negativity constraints $\tilde{c}_t \geq 0$ and $k_{t+1} \geq 0$. The problem can be solved using standard procedures, and the Euler equation can be derived as:

$$(1+n)(1+\gamma)(\tilde{c}_t)^{-\sigma} = \tilde{\beta}(\tilde{c}_{t+1})^{-\sigma}\left[f'(\tilde{k}_{t+1}) + (1-\delta)\right].$$

Note that, given the definition of $\tilde{\beta}$ in (4.19), the term $(1+n)$ drops out when the Euler equation is written with the original discount factor $\beta$. Therefore, the social planner's intertemporal allocation is not affected by the population growth under our utilitarian assumption.

Along the balanced growth path, each variable grows at a constant rate. From our normalized model above, the balanced growth path can be found by imposing steady-state conditions in the $(\tilde{c}_t, \tilde{k}_{t+1})$ variables. This procedure reduces the Euler equation to

$$(1+n)(1+\gamma) = \tilde{\beta}\left[f'(\bar{\tilde{k}}) + (1-\delta)\right].$$

By setting $n = \gamma = 0$, we can check that the above equations coincide with the ones derived under no growth. If we assume a Cobb-Douglas production function, replacing $f'(\bar{\bar{k}}) = \alpha \bar{\bar{k}}^{\alpha-1}$ in the equation above, we can deliver $\bar{\bar{k}}$ as a function of the relevant parameters of the model.

The question of whether or not the economy will converge to its balanced growth path from arbitrary initial conditions will be dealt with in the next section.

## 4.4 Recursive methods

In the previous section, we solved the maximization problem searching for a sequence of real numbers $\left\{x_{t+1}^*\right\}_{t=0}^{\infty}$ that achieves the highest value of the objective function. This involved finding a solution to an infinite sequence of equations (e.g., a difference equation). It is conceptually useful to break down this high-dimensional problem into a sequence of similar, but smaller problems, which all are tied to each other: "recursive" refers to this tie. This principle is at the core of the recursive method known as *dynamic programming*, introduced by Richard E. Bellman in the 1950s. A key difference with sequential methods is that the solution to the optimization problem will now be a *function* rather than a sequence of numbers. In what follows, we present the idea behind recursive methods and how they can be used; we discuss the precise theoretical links between the sequential approach and the functional approach briefly in Section 4.4.3 below.

### 4.4.1 Dynamic programming and the Bellman equation

An implicit assumption in the sequential formulation was that the whole path of $x_{t+1}$ was chosen in the initial period. The key to dynamic programming is to think of dynamic decisions as being made not once-and-for-all but period by period instead. In other words, the value of $x_{t+1}$ is decided in period $t$ rather than at date 0. A key question is whether the two formulations are identical. They will be, as long as the problem at hand is **stationary**. This is the case whenever the structure of the maximization problem that a decision maker faces is identical in nature at every point in time. To make ideas more concrete, let's revisit the infinite-horizon version of the problem (P1).

$$V(x_0) \equiv \max_{\{x_{t+1} \in \Gamma(x_t)\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t \mathcal{F}(x_t, x_{t+1}) \tag{4.20}$$

where $\Gamma(x_t)$ represents the feasible choice set for $x_{t+1}$ given $x_t$. In the expression, the "value function" $V(x_0)$ represents the value of the objective function at the optimum given the initial condition $x_0$. This high-dimensional problem can be broken down and re-written as

$$
\begin{aligned}
V(x_0) &= \max_{\{x_{t+1}\in\Gamma(x_t)\}_{t=0}^{\infty}} \left\{ \mathcal{F}(x_0, x_1) + \beta \sum_{t=1}^{\infty} \beta^{t-1} \mathcal{F}(x_t, x_{t+1}) \right\} \\
&= \max_{x_1\in\Gamma(x_0)} \left\{ \mathcal{F}(x_0, x_1) + \beta \left[ \max_{\{x_{t+1}\in\Gamma(x_t)\}_{t=1}^{\infty}} \sum_{t=1}^{\infty} \beta^{t-1} \mathcal{F}(x_t, x_{t+1}) \right] \right\} \\
&= \max_{x_1\in\Gamma(x_0)} \left\{ \mathcal{F}(x_0, x_1) + \beta \underbrace{\left[ \max_{\{x_{t+2}\in\Gamma(x_{t+1})\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^{t} \mathcal{F}(x_{t+1}, x_{t+2}) \right]}_{\equiv V(x_1)} \right\}
\end{aligned}
$$

The simple mathematical idea that $\max_{x,y} f(x,y) = \max_y \{\max_x f(x,y)\}$, provided that each of the max operators is well-defined, allows us to maximize "in steps." To do so, first notice that the problem in squared brackets (in the last row) has the same structure as equation (4.20), but with a different initial condition. In other words, it represents the value of the objective function attained by an agent that chooses the optimal sequence of $x_{t+1}$ given the initial condition $x_1$. Because the time horizon is the same, and neither the instantaneous objective function nor the feasible set $\Gamma$ change over time, then it must be the case that the problem in squared brackets is equal to $V(x_1)$. This implies that we can write

$$
V(x_0) = \max_{\{x_1\in\Gamma(x_0)\}} \left\{ \mathcal{F}(x_0, x_1) + \beta V(x_1) \right\}.
$$

The two representations are identical as long as the problem is stationary. Intuitively, a dynamic problem is stationary if the problem at $t$ and at $t+1$ look the same, so one can capture all relevant information for the decision maker in a way that does not involve time. Not all problems are stationary. For example, consider the finite-horizon version of the neoclassical growth model. In it, agents care about how many periods are left when choosing investment. The decision problem changes as the terminal period approaches. One can still describe the problem recursively: break it up into a current choice between consumption and investment and all future choices as implicitly captured by a value function. The difference is that the value function will depend on time, $V_t$.

With infinitely many periods, the remaining horizon is the same at every $t$. For different values of the initial capital stock, the choices differ, but do not depend on the time period in which the choice is made. This means that a decision at any point in time does not depend on anything but the level of capital at that point in time.

In general, the predetermined, payoff-relevant information for the consumer is—as was mentioned earlier in this chapter—called a *state variable*. The state variable for the planner in the neoclassical growth model is the current stock of capital $k_t$. In our generic example, it is simply given by $x_t$. When transforming a sequential problem into a recursive one, it is important to choose the state variables appropriately so that the resulting problem is indeed recursive.

When the problem is stationary, decisions take a stationary form:

$$x_{t+1} = g\left(x_t\right).$$

In particular, the function determining how future capital depends on the current state does not vary with time. The function $g\left(\cdot\right)$ is known as the **decision rule** or **policy function**.

Given an arbitrary initial condition $x_t$, we can write the recursive problem as

$$V(x_t) = \max_{x_{t+1} \in \Gamma(x_t)} \{\mathcal{F}(x_t, x_{t+1}) + \beta V(x_{t+1})\}.$$

This is the dynamic-programming formulation. The derivation was completed for a given value of $x_t$ on the left-hand side of the equation. On the right-hand side, however, we need to know $V$ evaluated at any value for $x_{t+1}$ in order to be able to perform the maximization. Going forward, we now change our notation and use $x$ to denote current values and, adding a prime, $x'$, to denote next period's values. This change of notation is conceptually important: in a stationary dynamic program, "$t$" actually has no place as time is no longer of essence. Above, it was used to connect the sequence problem to the recursive one; the only remaining need is to distinguish today's $x$ from tomorrow's $x'$. Thus, we write our dynamic-programming equation as the quest for a $V(x)$ satisfying

$$V(x) = \max_{x' \in \Gamma(x)} \{\mathcal{F}(x, x') + \beta V(x')\} \tag{4.21}$$

for all values of $x$. This equation is called the "Bellman equation." It is a *functional equation*: the unknown is a function $V(x)$; i.e., it needs to satisfy eq. (4.21) for all values of the argument of the function, $x$. It can be intuitively interpreted as follows: the discounted lifetime value of our representative agent's objective function is equal to the instantaneous value $\mathcal{F}(x, x')$ received today, plus the discounted value enjoyed from tomorrow into the future, $\beta V(x')$, all assuming optimal choices.

We use the function $g$ alluded to above to denote the arg max in the functional equation:

$$g(x) = \arg\max_{x' \in \Gamma(x)} \{\mathcal{F}(x, x') + \beta V(x')\}$$

for all $x$, or the decision rule for $x'$: $x' = g(x)$. This notation presumes that a maximum exists and it is unique. Otherwise, $g$ would either not exist or not be a function and, rather, a correspondence. Note that the following must hold for all $x$ by definition.

$$V(x) = \mathcal{F}(x, g(x)) + \beta V(g(x)).$$

### 4.4.2 Writing a problem recursively

Sometimes we are presented with the sequential formulation of a problem, and it is necessary to 'translate it' into a recursive formulation. We can do this for several reasons. Perhaps the most important one is conceptual: finding a recursive formulation offers an added understanding of the problem at hand. That is because when a problem has been formulated

recursively, we know what behavior will depend on: the state variable(s) of the recursive formulation, and nothing else. In practice, this often means that we can obtain insights into how a problem can be solved.

Consider the infinite-horizon version of the neoclassical growth model, which corresponds to (P3) with $T = \infty$. We first need to make sure that the problem is indeed stationary. That is, that the decision-maker faces the same type of problem every period given the same value of the state variables. That brings us to the question of what the relevant state variables ought to be.

**The standard NGM:** In the NGM, the state variable is given by the stock of capital at the outset of the period, $k$. In other words, $x = k$ in this model. The value function depends on $k$, implying that we can write $V(k)$. The control variable is $c = f(k) + (1-\delta)k - k'$, where we already used the notation that next period variables are denoted with primes (e.g. $k_{t+1}$ is written as $k'$). The recursive formulation of the neoclassical growth model is

$$V(k) = \max_{k' \in \Gamma(k)} \{u(f(k) + (1-\delta)k - k') + \beta V(k')\}, \tag{P4}$$

with $\Gamma(k) = [0, f(k) + (1-\delta)k]$, for all $k$. While this problem is relatively straightforward to write in recursive form, this is not always the case, as we see with the next two examples.

**The periodic NGM:** Consider a small deviation from the problem above, where the level of technology oscillates deterministically between two values $A_h$ and $A_l$, where $A_h > A_l$. In particular, period $t$ output equals $A_h f(k_t)$ if $t$ is even and $A_l f(k_t)$ if $t$ is odd. The resource constraint is $c_t + k_{t+1} = A_t f(k_t) + (1-\delta)k_t$, as before. At first sight, this problem does not look stationary because the level of productivity is changing over time. However, it is changing deterministically and predictably. We can write the problem recursively exploiting the periodicity in the evolution of TFP. To do so, we need to incorporate an additional state variable: whether we are in an even or odd period, which in turn corresponds to an $h$ or $l$ value for $A_t$. We can therefore write

$$V_h(k) = \max_{k' \in \Gamma_h(k)} \{u(A_h f(k) + (1-\delta)k - k') + \beta V_l(k')\}$$

$$V_l(k) = \max_{k' \in \Gamma_l(k)} \{u(A_l f(k) + (1-\delta)k - k') + \beta V_h(k')\}$$

with $\Gamma_i(k) = [0, A_i f(k) + (1-\delta)k - k']$ for $i \in \{h, l\}$, for all $k$. Clearly, given the same values of these two states, the decision maker would always select the same $k'$. So we were able to write the problem recursively with the appropriate choice of state variables.

**The delayed depreciation NGM:** The objective and resource constraints are standard,

$$\begin{array}{ll} \max\limits_{\{c_t\}_{t=0}^{\infty}} & \sum\limits_{t=0}^{\infty} \beta^t u\left(c_t\right) \\ \text{s.t.} & c_t + i_t = F\left(k_t\right), \end{array}$$

but capital depreciates fully in two periods, and does not depreciate at all before that, so that the law of motion for capital given a sequence of investments $\{i_t\}_{t=0}^{\infty}$ is given by

$$k_t = i_{t-1} + i_{t-2}.$$

In the first period, $k_0 = i_{-1} + i_{-2}$, with two initial conditions $i_{-1}$ and $i_{-2}$. The recursive formulation for this problem (with "primes" denoting consecutive periods) becomes

$$\begin{aligned} V(i', i) = &\max_{c, i''} \{u(c) + V(i'', i')\} \\ \text{s.t.} \quad &c = f(i' + i) - i''. \end{aligned}$$

Notice that there are two state variables in this problem. That is unavoidable here; there is no way of summarizing what one needs to know at a point in time with only one state variable. Both $i_{-1}$ and $i_{-2}$ are natural state variables: they are predetermined, they affect utility and decision making, and neither is redundant.

### 4.4.3 Properties of the value function

We have argued that there are two different ways of solving a dynamic maximization problem: (i) the sequential method, where the task is to find a sequence (that maximizes the objective function), and (ii) dynamic programming, where the task is to find a function solving the recursively stated problem, i.e., solving a functional equation. A formal mathematical statement that these two methods are equivalent is beyond the scope of this text; Stokey and Lucas (1989) has all the details. For each of the two approaches, it is important to specify a space within which one looks for a solution—a space of "allowable" sequences and a space of allowable functions, respectively. For dynamic programming, one most conveniently restrict attention to bounded, continuous functions $V$. For this choice to work out, one needs assumptions on primitives ($\mathcal{F}$, $\Gamma$, and $\beta$ in our generic formulation); minimum conditions involve (i) continuity of $\mathcal{F}$, (ii) non-emptiness, continuity, and compactness of $\Gamma$, and (iii) $\beta < 1$.[15] Under these assumptions, the functional equation defined by the dynamic program has a unique solution (it is the value function what is unique; the policy function can be a correspondence, unless we place further restrictions on $\mathcal{F}$ and $\Gamma$.) Moreover, for any given value of the state, it allows us to construct a sequence of choices corresponding to solutions to the sequentially stated problem. Conversely, solutions in the sequence space can be used to construct a value function solving the dynamic programming problem.

We now state some key properties of the dynamic programming problem and briefly comment on them. For proofs, see Stokey and Lucas (1989). You may have studied the necessary ingredients into the proofs in preparatory math, in particular the fact that the dynamic program defines a *contraction mapping*.

1. It is possible to find $V(x)$ by an iterative process. The procedure is as follows.

[15] Some maximization problems involve discrete choice; we could, for example, restrict the choice of capital to belong to a finite set of values. The conditions on primitives then become correspondingly weaker: $\mathcal{F}$ has to be bounded and $\Gamma$ nonempty.

i. Select any initial $V_0(x)$ function. One example is $V_0(x) = 0 \ \forall x$.

ii. Define a sequence of functions as follows: for all $x$,

$$V_{n+1}(x) = \max_{x' \in \Gamma(k)} \mathcal{F}(x, x') + \beta V_n(x')$$

for $n = 0, 1, 2, \ldots$.

Then the following is true:

(i) the resulting sequence $\{V_j(x)\}_{j=0}^{\infty}$ converges to $V(x)$, i.e., to the function that solves the dynamic program;

(ii) the distance to the solution $V$ gets smaller and smaller at a constant rate: $\|V_{n+1} - V\| \leq \beta \|V_n - V\|$, where $\|f\|$ denotes a distance between functions.[16]

Notice that the particular initial guess $V_0 = 0$ delivers an interpretation of $V_n$: it represents present-value utility for an economy consisting of $n$ periods. Clearly, $V_n$ is an addition of "utils" over time which involves optimal choice of $x'$ in all periods, including the last one, $V_1$.

Feature (ii) is rather remarkable: it holds no matter what the initial guess is and it is therefore very useful in practical applications: for finding a solution numerically. In particular, if one can bound the initial error by some value $\bar{\epsilon} \equiv \|V_0 - V\|$ and then generate the sequence of functions, one knows that after $n$ iterations the error is at most $\beta^n \bar{\epsilon}$. Finally, (ii) is extremely useful in allowing us to prove properties of the value function, such as those we state next.

2. Assume that $\mathcal{F}$ is strictly increasing in its first argument and that $\Gamma$ is monotone: if $x \leq \tilde{x}$, then $\Gamma(x) \subseteq \Gamma(\tilde{x})$. Then $V$ is strictly increasing.

3. Assume that $\mathcal{F}$ is strictly concave (in its two arguments jointly) and that $\Gamma(x)$ is convex in the following sense: if $x' \in \Gamma(x)$ and $\tilde{x}' \in \Gamma(\tilde{x})$, then $\theta x' + (1-\theta)\tilde{x}' \in \Gamma(\theta x + (1-\theta)\tilde{x})$. Then $V$ is strictly concave and the policy unique (i.e., $g$ is a well-defined function).

4. Assume that $\mathcal{F}$ and $\Gamma$ satisfy the properties in the previous statement and, in addition, that $\mathcal{F}$ is continuously differentiable. Then for any $x$ where the choice is interior, $V$ is differentiable, i.e., $V'(x)$ exists.

5. Assume that $\mathcal{F}$ and $\Gamma$ satisfy all the conditions above. Then, the policy function $g(x)$ is strictly increasing.

Very briefly, property 1 above follows from the Bellman equation being a contraction mapping.[17] Properties 2 and 3 are proved using property 1 as follows: take a $V_0$ with the

[16] The distance is a norm; in practice the sup-norm is used, i.e., the largest difference across all values of the argument of the function, i.e., $\|f\| \equiv \sup_x f(x)$.

[17] A contraction mapping obtains for the Bellman equation so long as $\beta < 1$.

desired characteristic (e.g., concavity), show that if $V_n$ has the desired characteristic then so does $V_{n+1}$, and finally argue that the limit $V^*$ inherits the characteristic at least weakly (e.g., $V^*$ is concave). Strictness of the characteristic (e.g., $V^*$ is strictly concave) follows by applying the argument again, with the limit $V^*$ inserted on the right-hand side of the Bellman equation. Property 4 is also possible to prove relatively straightforwardly but note that it applies only in the case where $V$ is concave. Property 5 follows from the first-order condition

$$-\mathcal{F}_2(x, x') = \beta V'(x').$$

The left-hand side of this equality is clearly increasing in $x'$, since $\mathcal{F}(x, x')$ is strictly concave in its second argument, and the right-hand side is strictly decreasing in $x'$, since $V(x)$ is strictly concave under the stated assumptions. Furthermore, since the right-hand side is independent of $x$ but the left-hand side is decreasing in $x$, the optimal choice of $x'$ is increasing in $x$.

### 4.4.4 Solving for the value function

As for the sequentially formulated maximization problem, analytical solutions are only available in very special cases. When they do exist, there are again a few different methods available to find them. One is a "guess and verify method," also known as the "method of undetermined coefficients." This method, however, requires an insightful initial guess for $V(x)$. When the initial guess is unavailable, we can use the iterative process described above, starting from $V_0(x) = 0$, quite similarly to how we proceeded to solve the finite-horizon model backwards.

We again illustrate these solution methods using the neoclassical growth model, (P4).

**Example 4.7** *Consider the economy from Example 4.5. The corresponding Bellman equation is*

$$V(k) = \max_{k' \geq 0} \left\{ \log\left(Ak^\alpha - k'\right) + \beta V(k') \right\}.$$

*Let the initial guess of the value function be* $V_0(k) = 0$. *Then,*

$$V_1(k) = \max_{k' \geq 0} \left\{ \log\left[Ak^\alpha - k'\right] \right\}. \tag{4.22}$$

*The right-hand side is maximized by taking* $k' = 0$, *yielding* $V_1(k) = \log A + \alpha \log k$. *Using this in the next iteration, we obtain*

$$V_2(k) = \max_{k' \geq 0} \left\{ \log\left[Ak^\alpha - k'\right] + \beta\left[\log A + \alpha \log k'\right] \right\}.$$

*The first-order condition delivers*

$$\frac{1}{Ak^\alpha - k'} = \frac{\beta\alpha}{k'} \Rightarrow k' = \frac{\alpha\beta Ak^\alpha}{1 + \alpha\beta}.$$

*We can interpret the resulting expression for $k'$ as the rule that determines how much it would be optimal to save if we were at period $T-1$ in the finite horizon model. We can substitute this into $V_2(k)$ to yield*

$$\begin{aligned} V_2(k) &= \log\left[Ak^\alpha - \frac{\alpha\beta Ak^\alpha}{1+\alpha\beta}\right] + \beta\left[\log A + \alpha\log\frac{\alpha\beta Ak^\alpha}{1+\alpha\beta}\right] \\ &= \underbrace{\log\left(A - \frac{\alpha\beta A}{1+\alpha\beta}\right) + \beta\log A + \alpha\beta\log\frac{\alpha\beta A}{1+\alpha\beta}}_{=a_2} + \underbrace{\left(\alpha+\alpha^2\beta\right)}_{=b_2}\log k. \end{aligned}$$

*The same procedure can be used to obtain a $V_3(k)$, and so on. This procedure would make the sequence of value functions converge to $V(k)$.*

*We can also, at this point, guess and verify a functional form for $V$. Grouping terms in $V_2$, we can see that it takes the form $V_n(k) = a_n + b_n\log k$ for all $n$. Therefore, we may already guess that the function to which this sequence is converging has to be of the form:*

$$V(k) = a + b\log k.$$

*In order to determine the corresponding parameters $a$, $b$, we take first-order conditions and find that $k' = \dfrac{\beta b}{1+\beta b}Ak^\alpha$. Inserting this into the right-hand side of the Bellman equation*

$$RHS \equiv \max_{k'\geq 0}\left\{\log\left(Ak^\alpha - k'\right) + \beta\left(a + b\log k'\right)\right\}.$$

*and equating the resulting expression to our guess delivers a system of two equations in two unknowns, $a$ and $b$. The solutions will be*

$$b = \frac{\alpha}{1-\alpha\beta} \quad \textit{and} \quad a = \frac{1}{1-\beta}\frac{1}{1-\alpha\beta}\left[\log A + \log\left(1-\alpha\beta\right)^{1-\alpha\beta} + \log\left(\alpha\beta\right)^{\alpha\beta}\right].$$

*The resulting decision rule is exactly the same policy rule obtained using sequential methods,*

$$k' = \alpha\beta Ak^\alpha.$$

### 4.4.5 The functional Euler equation

In the sequentially formulated maximization problem, the Euler equation turned out to be a crucial part of characterizing the solution. With the recursive strategy, an Euler equation can be derived as well. Consider the Bellman equation for a general instantaneous objective function $\mathcal{F}(x, x')$:

$$V(x) = \max_{x'\in\Gamma(x)}\left\{\mathcal{F}(x,x') + \beta V(x')\right\}.$$

Under suitable assumptions, this problem delivers the policy function $x' = g(x)$. Hence,

$$V(x) = \mathcal{F}(x,\ g(x)) + \beta V(g(x)). \tag{4.23}$$

Assuming an interior solution, $g(x)$ satisfies the first-order condition

$$\mathcal{F}_2(x,\ x') + \beta V'(x') = 0.$$

Evaluated at the optimum, i.e., at $x' = g(x)$,

$$\mathcal{F}_2(x,\ g(x)) + \beta V'(g(x)) = 0.$$

In contrast to how we derived the sequential formulation, we now need to know the derivative of the value function $V'(\cdot)$ in order to solve for the policy rule. Even though it is not possible, in general, to write $V(x)$ in terms of primitives, we can find its derivative. Using the equation (4.23) above, one can differentiate both sides with respect to $x$ (recall that the equation holds for all $x$ and, again under some assumptions stated earlier, is differentiable). We obtain

$$V'(x) = \mathcal{F}_1(x,\ g(x)) + \underbrace{g'(x)\Big[\mathcal{F}_2(x,\ g(x)) + \beta V'(g(x))\Big]}_{\text{indirect effect through optimal choice of } x'},$$

where "$g'(x)$" represents the derivative of the policy function. The differentiability of $g$ was not established above, and it is harder to prove, but it is actually not required in order to derive the result that just will follow. Namely, from the first-order condition, the argument in brackets in the equation just stated is zero and hence

$$V'(x) = \mathcal{F}_1(x,\ g(x)),$$

which again holds for all values of $x$. The indirect effect thus disappears: this is an application of a general result known as the **Envelope Theorem**.

Updating, we know that $V'(g(x)) = \mathcal{F}_1\Big(g(x),\ g(g(x))\Big)$ also has to hold. The first-order condition can now be rewritten as follows:

$$\mathcal{F}_2(x,\ g(x)) + \beta\mathcal{F}_1\Big(g(x),\ g(g(x))\Big) = 0\ \forall x. \tag{4.24}$$

This is the Euler equation stated as a functional equation: it does not contain the unknowns $x_t$, $x_{t+1}$, and $x_{t+2}$ but, rather, has to hold for all $x$. Recall our previous Euler equation formulation

$$\mathcal{F}_2(x_t,\ x_{t+1}) + \beta\mathcal{F}_1(x_{t+1},\ x_{t+2}) = 0, \forall t,$$

where the unknown was the sequence $\{x_t\}_{t=1}^{\infty}$. Now instead, the unknown is the function $g$. That is, under the recursive formulation, the Euler equation turned into a functional equation: the **functional Euler equation**.

**Solving directly for policy** The previous discussion suggests that a third way of searching for a solution to the dynamic problem is to consider the functional Euler equation, and solve it for the function $g$. We have previously seen that we can (i) look for sequences solving a nonlinear difference equation plus a transversality condition; or (ii) we can solve a Bellman (functional) equation for a value function.

The functional Euler equation offers another way to solve directly for behavior, thus bypassing the value function. Its key feature is that it expressed an intertemporal tradeoff. Here, the recursive approach provides some extra structure relative to the sequential Euler equation: it tells us that the optimal sequence of capital stocks needs to be connected using a stationary function. Mathematically, unlike the Bellman equation, the functional Euler equation is not a contraction mapping. It is also only a necessary condition on the optimal policy function $g$ and, viewed in isolation, can allow multiple solutions. Only one of these solutions, however, is the policy function that solves the right-hand side of the Bellman equation. It will become clear below how, at least locally, multiple solutions are possible.

The functional Euler equation approach is often used in practice in solving dynamic problems numerically. We now show an example for which an analytical solution exists.

**Example 4.8** *Consider the model used in Example 4.7. With full depreciation* $\mathcal{F}(k,k') = u(f(k)-k')$. *Then, the respective derivatives are:*

$$\begin{aligned}\mathcal{F}_1(k,k') &= u'(f(k)-k')f'(k)\\ \mathcal{F}_2(k,k') &= -u'(f(k)-k').\end{aligned}$$

*In the particular parametric example, and replacing* $k' = g(k)$, *equation (4.24) becomes:*

$$\frac{1}{Ak^\alpha - g(k)} - \frac{\beta\alpha A\left(g(k)\right)^{\alpha-1}}{A\left(g(k)\right)^\alpha - g\left(g(k)\right)} = 0, \forall k.$$

*This is a functional equation in* $g(k)$. *Guess that* $g(k) = sAk^\alpha$, *i.e., saving is a constant fraction of output. Substituting this guess into functional Euler equation delivers*

$$\frac{1}{(1-s)Ak^\alpha} = \frac{\alpha\beta A\left(sAk^\alpha\right)^{\alpha-1}}{A\left(sAk^\alpha\right)^\alpha - sA\left(sAk^\alpha\right)^\alpha}.$$

*As can be seen,* $k$ *cancels out, and the remaining equation can be solved for* $s$. *Collecting terms and factoring out* $s$, *we obtain*

$$s = \alpha\beta.$$

*Thus,* $\alpha\beta Ak^\alpha$ *satisfies the functional Euler equation for all values of* $k$. *It is, moreover, the same answer that we arrived at in Example 4.7.*

### 4.4.6 Dynamics in the optimizing neoclassical growth model

We now apply recursive methods to characterize the solution to the planning problem involving the standard neoclassical growth model. We could also apply our methods to the case

of consumer saving under a constant wage and interest rate, but we leave that application for the reader. It may be useful to first note that there is a unique maximum attainable level of the capital stock, $k_u$. That is, if the capital stock starts out below (or at) $k_u$, it cannot take a higher value than $k_u$ in the future. This result follows because for any $k$, $f(k) + (1 - \delta)k$ is the highest feasible saving $k'$, because this implies zero consumption. Hence, $f(k_u) + (1 - \delta)k_u = k_u$ defines $k_u$ and we can restrict attention to a search for value and policy functions over the closed and bounded set $[0, k_u]$. There are two solutions to this equation: a strictly positive $k_u$ and 0; we discard the latter.

We know from the functional Euler equation (4.24) that, for all $k \in [0, k_u]$, $g(k)$ satisfies

$$u'\left(f(k) + (1 - \delta)k - g(k)\right) = \beta u'\left(f(g(k)) + (1 - \delta)g(k) - g(g(k))\right)\left(f'(g(k)) + 1 - \delta\right) \tag{4.25}$$

In particular, at a steady state $\bar{k}$, $g(\bar{k}) = \bar{k}$, which—given that the argument of $u'$ is non-zero—allows us to write the steady-state condition as

$$1 = \beta(f'(\bar{k}) + 1 - \delta), \tag{4.26}$$

which is identical to (4.9) above. Now we can proceed to show global convergence to the steady state.

We use the property that under the standard assumptions the value function $V(k)$ is strictly concave.[18] That implies that we can write

$$\left[V'(k) - V'(g(k))\right]\left[k - g(k)\right] \leq 0 \quad \forall k \in [0, k_u].$$

To see that the inequality holds, note that whenever $g(k) > k$, the expression contained in the left-most bracket is positive, whereas the right-most bracket contains a negative quantity, and vice versa.

Using the envelope theorem, we know that

$$V'(k) = u'\left(f(k) + (1 - \delta)k - g(k)\right)\left(f'(k) + 1 - \delta\right).$$

From the first-order condition in the Bellman equation, we obtain

$$V'(g(k)) = u'\left(f(k) + (1 - \delta)k - g(k)\right)\frac{1}{\beta}.$$

Inserting these two expressions into the left-most bracket of the inequality and factorizing $u'$, we obtain

$$u'(f(k) - g(k))\left[f'(k) + 1 - \delta - \frac{1}{\beta}\right]\left[k - g(k)\right] \leq 0 \quad \forall k \in [0, k_u]. \tag{4.27}$$

When $k = \overline{k}$, from (4.26), $f'(k) + 1 - \delta = \frac{1}{\beta}$ and $g(k) = k$ holds. When $k > \overline{k}$, then $f'(k) + 1 - \delta < \frac{1}{\beta}$, and thus (4.27) implies $g(k) < k$, and capital decreases over time. When $k < \overline{k}$, capital increases. This result implies, given that we also know that $g$ is increasing (property 5 above), that the system is globally stable and converges monotonically to $\overline{k}$.

[18] This requires us to show that $u(f(k) + (1 - \delta)k - k')$ is concave in $(k, k')$. This can be accomplished by computing the Hessian and showing that it is negative definite.

**Approximating the policy function** Most often, it is not possible to solve for the policy function in analytical form. In such cases, we can approximate the solution locally with a linear function,

$$g(k) \sim a_0 + a_1 k$$

where $a_0$ and $a_1$ are the coefficients that we need to solve for, around the steady-state value of capital. We need two equations to solve for our two unknowns, but since $g(k) = k = \overline{k}$ in steady state we obtain one condition as

$$a_0 = \overline{k}(1 - a_1),$$

where $\overline{k}$ solves (4.9). To obtain $a_1$, we use a procedure that parallels linearization techniques discussed in Chapter 3, but is less cumbersome because we exploit the fact that the functional Euler equation must hold for all $k$. Thus, we can simply differentiate equation (4.25) with respect to $k$, i.e., take the derivative of the left-hand side, LHS, and set it equal to the derivative of the right-hand side, RHS. We obtain, with the obvious notation, $LHS_k(k) = RHS_k(k)$, which must hold for all $k$. Evaluating this expression at steady state allows us to obtain an equation that provides a second condition relating $a_1$ to $a_0$. Hence,

$$LHS_k(\overline{k}) = u''(\overline{c})[f'(\overline{k}) + 1 - \delta - \underbrace{g'(\overline{k})}_{=a_1}]$$

$$RHS_k(\overline{k}) = \beta u''(\overline{c})\Big[(f'(\overline{k}) + 1 - \delta)a_1 - \underbrace{g'(\overline{k})g'(\overline{k})}_{=a_1^2}\Big]\Big[f'(\overline{k}) + 1 - \delta\Big] + \beta u'(\overline{c})f''(\overline{k})a_1.$$

where we used the fact that, under our guess, $g(g(k)) = a_0 + a_1[a_0 + a_1 k]$; furthermore, $\bar{c}$ is defined as $f(\overline{k}) - \delta\overline{k}$. Setting $LHS_k(\overline{k}) = RHS_k(\overline{k})$ and simplifying delivers

$$u''(\overline{c})\left[1 - \beta a_1\right] = u''(\overline{c})\left[a_1 - \beta a_1^2\right] + \beta^2 u'(\overline{c})f''(\overline{k})a_1, \tag{4.28}$$

which can be used to solve for $a_1$ given $\overline{k}$. Clearly, $a_1$ satisfies a second-order polynomial equation. It is straightforward to show that this equation has two real solutions, of which one is strictly between zero and one, and the other is strictly greater than one. Thus, locally, $g(k)$ can be approximated by two functions. This was alluded to above and it is not a surprise, given that the functional Euler equation is only a necessary condition. A function $g$ which solves this equation also has to attain the maximum on the right-hand side of the Bellman equation. Of the two functions obtained here, only one has that property: the one with an $a_1$ between zero and one. It is the one corresponding to a slope less than one in $(k, k')$ space, thus giving monotone convergence to the steady state. Concretely, by denoting $\Theta \equiv \beta^2 u' f''/u'' > 0$, the solution of (4.28) is

$$a_1 = \frac{1 + \beta + \Theta - \sqrt{(1 + \beta + \Theta)^2 - 4\beta}}{2\beta}. \tag{4.29}$$

It can easily be checked that $a_1 \in (0, 1]$, $a_1$ is decreasing in $\Theta$, and $a_1 = 1$ when $\Theta = 0$ and $a_1 \to 0$ as $\Theta \to \infty$.

Using this procedure, it is possible to also obtain higher-order approximations to the policy function $g$ around the steady state. To obtain the second-order Taylor approximation, simply differentiate the functional equation once more with respect to $k$, which delivers $g''$ at the steady state: it can be solved for as a function of $g$ and $g'$, which were previously solved-out. Further differentiations will, successively, give us any desired higher-order approximations.[19]

**Calibration** We discussed the concept of calibration in Chapter 3: the adoption of specific functional forms and parameter values with the purpose of generating quantitative predictions. We applied it in that chapter to gauge the Solow model's quantitative predictions for the speed of convergence to a steady state. The NGM developed in this chapter, instead, differs in its convergence properties: the saving rate is not constant away from steady state but is instead endogenous and, as we have seen, a function of the current level of the capital stock. How strong the dependence on capital is depends on the utility function. More precisely, equation (4.29) determines $a_1$, the slope of the saving function as it crosses the steady-state line: a slope of zero implies convergence in one period (infinite speed) and a slope of 1 implies no (infinitely slow) convergence. The remaining model parameters can be selected as in Chapter 3 so our only question here is: what is $a_1$, or rather, what is the value of $\Theta$ at the steady state? Recall that $\Theta = \beta^2 u' f''/u''$, $\bar{c} = f(\bar{k}) - \delta\bar{k}$, and $f'(\bar{k}) = 1/\beta - (1-\delta)$. Furthermore, suppose that $u(c) = \frac{c^{1-\sigma}-1}{1-\sigma}$. Then it can be shown that

$$\Theta = -\frac{\beta^2 f''(\bar{k})(f(\bar{k}) - \delta\bar{k})}{\sigma}.$$

Because $f''(\bar{k})$ is negative, $\Theta$ is positive and decreasing in $\sigma$. When $\sigma = 0$, the utility function is linear, and in this case $\Theta \to \infty$. The above result implies $a_1 \to 0$ in this case, implying $k'$ does not depend on $k$. Convergence, in this case, is immediate. The opposite extreme is when $\sigma \to \infty$, $\Theta \to 0$ and $a_1 \to 1$. In this case, the utility function is "infinitely curved" (Leontief utility). When $\sigma$ is very large, consumers are extremely unwilling to change consumption over time and convergence is very slow.

Therefore, $\sigma$ is an important parameter in determining the speed of convergence. What do empirical studies of consumption suggest as an appropriate value of $\sigma$? Using aggregate consumption data, Hall (1982) estimates that $1/\sigma = 0.1$. Attanasio (1993, 1995) uses micro data instead and finds that $1/\sigma \in [0.3, 0.8]$. In many applied macroeconomic studies using the CRRA function, however, there are also stochastic components, and under uncertainty, $\sigma$ is also equal to the coefficient of relative risk aversion (uncertainty is discussed in Chapter 7 below) and this coefficient is often estimated to be above 1. Perhaps for this reason, much of the applied literature focuses on $\sigma = 1$ (log utility) or $\sigma = 2$, though rarely much higher than one.

Further assume that the production function $f(k)$ is a power function: $f(k) = k^\alpha$. This formulation implies the aggregate production function is of the Cobb-Douglas form. In this

[19] Approximations based on Taylor approximations, as those carried out here, of course require differentiability of the function. Thus, we approximate $g$, which is endogenous, with that proviso.

case, $\Theta$ can be solved as

$$\Theta = \frac{\beta^2(1-\alpha)}{\sigma}\left(\frac{1}{\beta} - 1 + \delta\right)\left[\frac{1}{\alpha}\left(\frac{1}{\beta} - 1 + \delta\right) - \delta\right]$$

which is decreasing in $\alpha$ and increasing in $\delta$. Convergence is slower when the production function is closer to linear and the depreciation rate is small. This property is qualitatively similar to the Solow model.

## 4.5 Concluding remarks

In this chapter, we showed a set of tools that are useful for solving standard dynamic optimization models used in macroeconomics. We discussed the sequential formulation, where the aim is to choose the best sequence of allocations (quantities) that maximize lifetime utility, starting with a finite horizon economy and then moving to an infinite horizon economy, highlighting the mathematical complications that arise. We then showed how to use dynamic programming methods to write the optimization problem recursively, where the key is to solve for policy functions determining optimal allocations. In practice, all these methods are used. We spent less time on motivating specific functional forms and, throughout, only used two examples: a pure consumption-saving problem under price-taking and an optimizing neoclassical growth model. Neither of these cases allowed valued leisure or other, richer optimization problems. However, the methods we introduce here are straightforwardly extended to other, richer contexts and will indeed be used over and over in the rest of the text.

# Chapter 5

# Dynamic competitive equilibrium

The introductory chapter looking at long-run data, Chapter 2, argued that a certain view on production and technical change was important for understanding our macroeconomic history. It also argued, informally at least, that a market economy with certain features was another important component of the overall picture: firms and private households making decisions in their self interest. The most commonly used framework in economics is a market setting with perfect competition—indeed, it is the setting the earlier chapter alluded to—and in the present chapter we will develop such theory carefully in the context of a dynamic economy.

Like many of our assumptions, perfect competition should be viewed as a useful approximation rather than an exact description of our economy. Indeed, most firms have a degree of market power and in some industries market power is critical for understanding how the industry works. However, a macroeconomic approach naturally starts from a benchmark case that is (i) not too far from a description of "average" behavior and (ii) allows tractable analysis, while still capturing the key features of markets. These features involve how private incentives—of consumers and firms—jointly steer production and consumption through a price mechanism. Market scarcity—expressed as a demand that exceeds supply—will lead to higher prices, lowering demand and/or raising supply. For example, the level of investment, a key macroeconomic variable, is affected by the interest rate, which is a relative price between goods today and goods in the future: consumers are willing to forsake goods today by lending more to investors the higher is the interest rate; and investors thus invest until the interest rate they have to pay on a loan no longer balances the productive returns on the investment. This market mechanism, moreover, is at play not just for traditional consumption and investment goods and services but for the development of new ideas and technologies that are key to long-run development.

How do agents decide what to buy or sell? They solve an optimization problem that is—in a macroeconomic context—often dynamic, as we have seen in Chapter 4. The solution to the optimization problem provides the quantities of the objects that the agent would like to buy and/or sell at given prices; for example, a household could sell labor services and get paid a wage (the price for labor). Therefore, an important component of a competitive equilibrium is the characterization of individual demands and supplies for goods, services and

assets, *given* the corresponding prices.[1] Since in the economy there are many agents who buy and sell the traded objects, the derivation of economy-wide demands and supplies requires the aggregation of demands and supplies over all agents. The core framework described in this chapter makes this procedure simpler by assuming that all consumers are identical and that all firms are identical—we use the "representative consumer" and the "representative firm" as our key constructs. A large part of the current macroeconomic research literature of course focuses on consumer and firm heterogeneity, which will be amply discussed later in the text.

In summary, a perfectly competitive equilibrium occurs when two conditions are met: agents' choices are optimal taking the prices as given and markets clear, i.e., demand equals supply, for all goods and services. In this chapter, we will also see that the market allocation delivers a Pareto-optimal outcome. With a representative consumer, this amounts to verifying that, in the perfectly competitive equilibrium, the consumer's utility is maximized given all technological constraints. When we depart from this setting, and when monopoly elements and other "frictions" are introduced, markets need no longer deliver optimal outcomes. We look at such instances in some detail already in Chapter 6 but they will then appear again and again in the applied chapters later.

## 5.1 Different equilibrium concepts

When considering our macroeconomic models, it is very helpful to be precise in formulating equilibrium concepts. A model economy, first, has some descriptive "deep" features, such as a set of agents (consumers, firms, a government, etc.) along with their objectives (such as utility functions) and relevant constraints (time available, technologies, etc.). Second, what the set of traded goods and services is—along with their prices—must be specified. The equilibrium concept then lists a set of conditions that need to be fulfilled: (i) agents maximizing their objectives subject to their constraints (e.g., consumers maximizing utility subject to their budget constraints and firms maximizing profits taking their technological possibilities into account) and (ii) resource feasibility: that what is consumed is also produced, often expressed as "demand equals supply," for all traded goods and services. The definition of an equilibrium thus organizes this information in the form of a set of conditions that need to be fulfilled. In this chapter, we will formulate equilibrium definitions in a rather mathematical way, i.e., with a minimum of words. Thus, we will try to refrain from conditions such as "taking prices as given" in a maximization problem; rather, if a maximization problem is specified, it must list the choice variables, and it then follows that any variables that are not listed as choices (such as prices in a perfectly competitive equilibrium) must, then, be taken as given.

Formulating equilibria in dynamic models, moreover, involves a choice regarding how time is dealt with. We will, in particular, consider three different definitions of competitive

[1] The notion that agents take prices as given allows us to analyze the economy without game-theoretic elements: in a given agent's decision problem, the behavior of others appears in the form of (endogenous) constants.

equilibria in turn:

1. Arrow-Debreu equilibrium in which trades occur at date 0,

2. sequential equilibrium in which trades occur period by period, and

3. recursive equilibrium in which prices and decisions are expressed as functions of the economy's state variables.

In the first two of these, an equilibrium consists of sequences (of quantities and prices) indexed by time, i.e., a specific values for each variable at each point in time. The third concept, in contrast, formulates the equilibrium as a set of functions of state variables (such as a capital stock or level of asset holdings). For many economies, one can define the competitive equilibrium in any of these three ways and they all give rise to the same allocations. Which one is chosen in a given application depends on the purpose, but, broadly speaking, the Arrow-Debreu equilibrium is the version that aligns most closely with microeconomic theory while being rather abstract as a concept. Sequential equilibria are defined more in accordance with how we think events actually play out in reality. A recursive equilibrium has the sequential trading structure but is expressed in terms of functions rather than sequences because it uses dynamic programming methods. Whereas equilibria based on sequences typically are defined given specific values of the state variables (such as the initial capital stock), recursive equilibria are not, i.e., they apply to all values of the state variables.

Throughout, we will focus on dynamic models with an infinite time horizon. Finite-horizon economies can of course be studied too; they should be seen as special cases of what we study here. As pointed out above, infinite-horizon economies are useful because there, time plays a less central role: the remaining time horizon is always the same. Occasionally (later in the text), we focus on one-period (where the "time horizon" plays no role) or two-period economies, but later in the text. In Sections 5.2–5.4 we study infinitely-lived dynasties and go through the different equilibrium definitions. In the final part of the chapter, Section 5.5, we study overlapping-generations (OG) economies, where time goes on forever but all individuals live a finite number of periods.

## 5.2 Arrow-Debreu equilibrium

In an Arrow-Debreu equilibrium, all trades take place at time zero. One way to think about this assumption is that every agent signs a contract at time zero with other agents. The contract specifies the quantities of the traded objects that the agent will deliver or receive at any future time $t$ from other agents at the prices specified. This market structure is called an Arrow-Debreu or date-0 market. It differs from the sequential market structure where trades for goods take place at all times $t$, not only at time zero. This different market structure will be considered in the other two equilibrium concepts.

It is perhaps worth emphasizing that signed contracts are fully enforceable, that is, promises made in a contract are always fulfilled. If some of the promises are not enforceable,

we are in an environment in which markets are incomplete. We will consider economies with incomplete markets in later chapters.

The best way to illustrate the concept of an Arrow-Debreu equilibrium is through the application to specific economies. We start with an endowment economy.

### 5.2.1 An endowment economy

The first example is an economy in which production is exogenously determined. There is a continuum of consumers, indexed by $i \in [0, 1]$. In every period $t$ each consumer produces $y_{i,t} \in \mathbb{R}_+$ of a single consumption good that cannot be stored for future consumption. The model does not specify how production takes place (for example with the input of labor) and the exogenous production is often referred to as endowment. An economy without the specification of a production technology is sometimes called an *exchange economy*, since the only economic activity that agents undertake, besides consumption, is the trade of the endowments. But, effectively, the exchange economy can be seen as a special case of a production economy. Finally, the consumer's utility from any given consumption path $\{c_{i,t}\}_{t=0}^{\infty}$ is

$$\sum_{t=0}^{\infty} \beta^t u\left(c_{i,t}\right). \tag{5.1}$$

Consumers' preferences are all the same (they have the same $u$ and $\beta$).

What is traded here is goods at different points in time: buying one unit of good at time $t$ means buying a contract that gives ownership of a unit of good at that point in time. Its price is denoted by $p_t$. Thus, $p_t/p_0$ represents the units of consumption goods delivered at time 0 that are needed to buy 1 unit of the consumption good delivered at time $t$. It is customary to normalize the price at time zero to 1 so that the good at time 0 becomes the numéraire. Then $p_t$ is the price of a time-$t$ good in terms of time-0 consumption goods.

Given the price $p_t$ for $t = 0, 1, ...$, the total value of the consumer's endowments is given by $\sum_{t=0}^{\infty} p_t y_{i,t}$ and the value of the consumer's total expenditures is $\sum_{t=0}^{\infty} p_t c_{i,t}$. The budget constraint then requires that the value of expenditures not be larger than the value of the endowments. In practice, since we always use strictly increasing utility functions, consumers will always use up all the resources and we will therefore impose this constraint with equality. Thus, we have

$$\sum_{t=0}^{\infty} p_t c_{i,t} = \sum_{t=0}^{\infty} p_t y_{i,t}. \tag{5.2}$$

We will define the equilibrium focusing on the mathematical conditions that must be satisfied, as opposed to their economic interpretation. The equilibrium definition provides a roadmap for solving for an equilibrium—we convert the equilibrium conditions into a set of equations that allow us to solve the model.[2]

[2]Here and in what follows, we omit any requirements that quantities and prices be non-negative. After we solve for an equilibrium we can verify that any such requirements are satisfied. Also, we do not specify the mathematical spaces to which the sequences must belong; for infinite sequences, this involves advanced concepts, which is why we omit them.

**Definition 1** *An **Arrow-Debreu competitive equilibrium** is a set of sequences* $\left\{c_{i,t}^{\star}\right\}_{t=0}^{\infty}$, *for each* $i \in [0,1]$, *and* $\{p_t\}_{t=0}^{\infty}$ *such that*

1. *for each* $i$, $\left\{c_{i,t}^{\star}\right\}_{t=0}^{\infty}$ *solves*

$$\max_{\{c_t\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t u(c_t) \quad s.t. \quad \sum_{t=0}^{\infty} p_t c_t = \sum_{t=0}^{\infty} p_t y_{i,t};$$

2. $\int_0^1 c_{i,t}^{\star} di = \int_0^1 y_{i,t} di$.

## Equilibrium characterization

To characterize the optimal decision of a consumer (point 1 in the above definition), we write the Langragian for the consumer problem. We then take the first-order conditions as shown in Chapter 4. This gives us

$$\beta^t u'(c_{i,t}) = \lambda_i p_t,$$

where $\lambda_i$ denotes the Lagrange multiplier for the (lifetime) budget constraint, equation (5.2), and the prime denotes the derivative of the utility function. The multiplier is the shadow value of lifetime wealth which could differ across agents if their lifetime endowments differ. However, the multiplier does not depend on time. If the utility function is strictly concave, its derivative is strictly decreasing in consumption. Thus, the above condition determines, uniquely, the optimal consumption for agent $i$.

If we eliminate the multiplier using the optimality conditions at time zero and at a generic time $t$, for the same consumer, we obtain the price

$$p_t = \beta^t \frac{u'(c_{i,t})}{u'(c_{i,0})}. \tag{5.3}$$

Remember that $p_0$ has been normalized to 1, which explains why it does not appear in this expression. Since the price $p_t$ is the same for all agents, this condition tells us that the ratio of marginal utilities at different times are the same across consumers. This equation states that the relative price of consumption at time $t$ in terms of time zero consumption has to equal the marginal rate of substitution between these two goods. This is the ratio between the present value of the marginal utility of consumption at time $t$ and the marginal utility of consumption at time zero.

To gain further intuition, we now consider the special case in which the utility function takes the logarithmic form, that is, $u(\cdot) = \log(\cdot)$. With this special utility, the above condition becomes

$$p_t = \beta^t \frac{c_{i,0}}{c_{i,t}}.$$

Since this condition must hold for all agents, we can use this condition to find equilibrium prices. Multiply both sides of the equation by $c_{i,t}$ and sum across all $i$. We obtain $p_t \int_0^1 c_{i,t} di =$

$\beta^t \int_0^1 c_{i,0} di$. Using market clearing, i.e., $\int_0^1 c_{i,t} di = \int_0^1 y_{i,t} di = Y_t$, we obtain

$$p_t = \beta^t \frac{Y_0}{Y_t}.$$

Thus, we see that prices decline over time because of discounting but also that periods with lower total endowments have higher prices. This is because marginal utility is strictly decreasing: lower total resources make each unit of consumption more valuable on the margin.

From the Euler equation, we see that in this economy all consumers experience the same consumption growth. More specifically, if we take any two consumers, $i$ and $i'$, we have

$$\frac{c_{i,t+1}}{c_{i,t}} = \frac{c_{i',t+1}}{c_{i',t}} = \frac{Y_{t+1}}{Y_t} = \beta \frac{p_t}{p_{t+1}}.$$

This condition is valid for any pattern of the individual endowments. Clearly, the growth rate of individual consumption is only determined by the aggregate endowment $Y_t$. Thus, individual consumption could be less volatile than individual income (consumption smoothing). To see this more clearly, suppose that that individual endowments change over time. However, the aggregation of the individual endowments is a constant, that is, $\int_0^1 y_{i,t} di = \bar{Y}$. In equilibrium it must be that $\int_0^1 c_{i,t} di = \bar{Y}$. In other words, aggregate consumption is constant. Since the consumption growth of all agents is the same, this implies that individual consumption must also be constant for all $t = 0, 1, 2, \ldots$. Therefore, in this special case with a constant aggregate endowment, but not necessarily constant individual endowments, the model features perfect consumption smoothing.

Lastly, the level of consumption for each consumer can be derived from the individual budget constraint. By combining the Euler equations for $t = 0, 1, 2, ...$ and $p_0 = 1$ we arrive at $p_t c_{i,t} = \beta^t c_{i,0}$. The budget constraint then becomes

$$c_{i,0} \sum_{t=0}^{\infty} \beta^t = Y_0 \sum_{t=0}^{\infty} \beta^t \frac{y_{i,t}}{Y_t} \quad \Rightarrow \quad c_{i,0} = (1-\beta) Y_0 \sum_{t=0}^{\infty} \beta^t \frac{y_{i,t}}{Y_t}.$$

Clearly, the level of consumption of individual $i$, as measured by its level at time 0, is a fraction $1-\beta$ of the individual's present-value income. This income is a function of endowments where (i) endowments further into the future obtain lower weights due to discounting and (ii) endowments in periods where aggregate income is high obtain a lower weight. The latter is true since resources in periods with high aggregate income are given lower value by consumers: their marginal utilities are lower then than in other periods. Thus, two consumers with the same average endowments can have different total wealth: the wealthy consumer is the one whose endowments are high when others' endowments are low.

Clearly, since consumption growth is identical for all consumers, we see that individual $i$'s share of aggregate consumption, and aggregate resources, is constant over time:

$$c_{i,t} = \theta_i C_t = \theta_i Y_t, \quad \text{with} \quad \theta_i = (1-\beta) \sum_{t=0}^{\infty} \beta^t \frac{y_{i,t}}{Y_t}.$$

In the special case in which agents have exactly the same endowment ($y_{i,t} = Y_t$), the share is 1 for all agents which effectively means that they consume their own endowment. We are then in the case of a *representative consumer*.

If the utility function is not logarithmic but maintains the balanced-growth form $(c^{1-\sigma} - 1)/(1-\sigma)$ (with $\sigma > 0$ and $\sigma \neq 1$; recall that $\sigma \to 1$ should be interpreted as $\log c$), then it is still possible to solve the model. It is easy to verify, using the consumer's Euler equation, that individuals' consumption growth rates will all be identical: the gross rates will equal $(\beta \frac{p_t}{p_{t+1}})^{\frac{1}{\sigma}}$. From this it follows that $p_t = \beta^t (\frac{Y_0}{Y_t})^{\sigma}$: again, resources are more valuable in times with lower endowments, and the more so the higher is the curvature of the utility function, as measured by $\sigma$.

The endowment framework studied here is useful in many contexts. One of these is asset pricing, covered in Chapter 14. The idea there is that any asset can be thought of as a stream of payments, "dividends," such as the endowment sequences described here, so the price of the asset is then the total market value of these endowments. Prices of endowments at each time period are straightforward to compute in endowment economies, including in cases with uncertainty; we will look at uncertainty in Chapter 7.

Finally, in many macroeconomic applications, the assumption of a representative agent is used. It makes sense as a special case of the above, when $u$ is strictly concave: then, given equal endowments for all agents, the consumption choices must all be the same. Such an equilibrium is often defined more compactly as

**Definition 2** *An **Arrow-Debreu competitive equilibrium** is a set of sequences* $\{c_t^\star\}_{t=0}^\infty$ *and* $\{p_t\}_{t=0}^\infty$ *such that*

1. $\{c_t^\star\}_{t=0}^\infty$ *solves*
$$\max_{\{c_t\}_{t=0}^\infty} \sum_{t=0}^\infty \beta^t u(c_t) \quad s.t. \quad \sum_{t=0}^\infty p_t c_t = \sum_{t=0}^\infty p_t y_t;$$
2. $c_t^\star = y_t$.

This equilibrium definition is mathematically precise but the economic context—that of many consumers, making the same choices—is only written "between the lines."

### 5.2.2 A production economy with labor

We now consider an economy where there is production but with only one factor input: labor. The economy is populated by a continuum of households, each supplying working hours $\ell_{i,t}$ to the market. However, working is costly for the household as it reduces utility. Given the consumption path $\{c_{i,t}\}_{t=0}^\infty$ and working hours $\{\ell_{i,t}\}_{t=0}^\infty$, the household's utility is

$$\sum_{t=0}^\infty \beta^t \Big[ u\left(c_{i,t}\right) - v(\ell_{i,t}) \Big]. \tag{5.4}$$

The function $v(\ell_{i,t})$ is the disutility from working. It is increasing and convex in $\ell_{i,t}$. We allow households to differ in their efficiency units of labor which we denote by $e_i$. What

this means is that, if household $i$ works $\ell_{i,t}$ hours, its contribution to the economy's input of labor (in efficiency units) is $e_i\ell_{i,t}$. We can think of $e_i$ as labor skills.

In this economy there is also a representative firm that produces consumption goods with the production function

$$Y_t = A_t L_t, \tag{5.5}$$

where $A_t$ could be time varying but not stochastic, and $L_t$ is the effective input of labor used in production.

In a competitive economy, the presumption is also that there are many firms. However, if these firms all have the same objective (profit maximization) and the same technology available, then they will face identical maximization problems. For this reason, we will use the concept of a *representative firm*. It thus hires labor from households paying the wage rate $w_t$ per effective unit of labor in terms of the consumption good at that time. Therefore, if the household supplies one efficiency unit of labor at time $t$, it receives $w_t$ units of consumption goods at that time, which translates to $p_t w_t$ units in terms of time-0 good.

The representative firm's objective is to maximize profits:

$$\max_{\{L_t\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \Big\{ p_t A_t L_t - p_t w_t L_t \Big\}. \tag{5.6}$$

Clearly, $L_t$ only appears in the time-$t$ term, so this problem reduces to an infinite sequence of static problems $\max_{L_t}(A_t L_t - w_t L_t)$; the within-period relative price $w_t$ is the price that will clear the time-$t$ market. Unlike for the concept of the representative consumer, whose optimal choice is unique, the optimal firm choice is—under constant returns to scale and price-taking—not unique in equilibrium: the firms is indifferent as to what scale to operate at.[3]

Given prices $p_t$ and $p_t w_t$ and labor supplies $\ell_{i,t}$, for $t = 0, 1, ...$, the value of lifetime income for household $i$ is $\sum_{t=0}^{\infty} p_t w_t e_i \ell_{i,t}$ and the value of its expenditures is $\sum_{t=0}^{\infty} p_t c_{i,t}$. The budget constraint requires that the value of expenditures not be larger than the value of incomes, that is,

$$\sum_{t=0}^{\infty} p_t c_{i,t} = \sum_{t=0}^{\infty} p_t w_t e_i \ell_{i,t}. \tag{5.7}$$

The following is then the compact definition of our equilibrium.

**Definition 3** *An **Arrow-Debreu competitive equilibrium** is a set of sequences $\left\{c^{\star}_{i,t}\right\}_{t=0}^{\infty}$ and $\left\{\ell^{\star}_{i,t}\right\}_{t=0}^{\infty}$, for each $i \in [0,1]$, $\{L^{\star}_t\}_{t=0}^{\infty}$, $\{p_t\}_{t=0}^{\infty}$ and $\{w_t\}_{t=0}^{\infty}$ such that*

1. *for each $i$, $\left\{c^{\star}_{i,t}\right\}_{t=0}^{\infty}$ and $\left\{\ell^{\star}_{i,t}\right\}_{t=0}^{\infty}$ solve*

$$\max_{\{c_t,\ell_t\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t u(c_t) \quad s.t. \quad \sum_{t=0}^{\infty} p_t c_t = \sum_{t=0}^{\infty} p_t w_t e_i \ell_t;$$

[3] In this case, an equilibrium price for labor, $w_t$, has to equal $A_t$, in which case the firm is indifferent as to its choice of $L_t$; if $A_t > w_t$, there is no optimal scale (profits rise without bound with scale), and if $A_t < w_t$, the firm's unique maximizer is $L_t = 0$.

2. $\{L_t^\star\}_{t=0}^\infty$ *solves*

$$\max_{\{L_t\}_{t=0}^\infty} \sum_{t=0}^{\infty} \Big\{ p_t A_t L_t - p_t w_t L_t \Big\};$$

3. *for each* $t$, $\int_0^1 e_i \ell_{i,t}^\star di = L_t^\star$ *and* $\int_0^1 c_{i,t}^\star di = A_t L_t^\star$ .

The first two points say that the allocation is what agents (households and firms) choose optimally to maximize their respective objective functions.[4] The third item defines the market clearing condition for labor (the aggregate supply of labor from households must be equal to the aggregate demand from firms) and for the goods markets (the aggregate demand of goods from households must be equal to the aggregate supply, which corresponds to production).

## Equilibrium characterization

The optimality conditions for the consumer's problem (again derived from the Lagrangian as shown in Chapter 4) are

$$\beta^t u'(c_{i,t}) = \lambda_i p_t, \tag{5.8}$$

and

$$v'(\ell_{i,t}) = e_i w_t u'(c_t), \tag{5.9}$$

where $\lambda_i$ is the Lagrange multiplier for the (lifetime) budget constraint, equation (5.7).

We can eliminate the multiplier using the optimality conditions for an agent $i$, at time $t$ and at time $t+1$, and obtain

$$\frac{p_{t+1}}{p_t} = \frac{\beta u'(c_{i,t+1})}{u'(c_{i,t})}. \tag{5.10}$$

As in the case of the endowment economy we just studied, the relative price of tomorrow's consumption in terms of today's consumption—the inverse of the (gross) real interest rate—equals the marginal rate of substitution between consumption at $t$ and $t+1$. With the normalization $p_0 = 1$, the equilibrium price at any time $t$ is

$$p_t = \frac{\beta^t u'(c_{i,t})}{u'(c_{i,0})}.$$

The firm solves (5.6) and the first-order condition requires that, for an interior solution—where the firm is indifferent between using more or less labor—we need to have

$$A_t = w_t.$$

[4] For the firm, we have written it so that the firm chooses labor and production at all times. Since the problem is not fundamentally dynamic—$L_t$ can be chosen independently of labor chosen at times other than $t$—we could alternatively had firms operating production at single dates only.

If $A_t$ were to exceed $w_t$, the firm's maximization problem would have no solution: the higher is $L_t$, the better. If $w_t$ were to exceed $A_t$ the firm would shut down (choose zero labor). Thus, for an equilibrium to exist, the relative price $w_t$ needs to adjust to be equal to $A_t$. When it has, the firm's size is indeterminate: all $L_t \geq 0$ deliver zero profits. This is a standard result under constant returns to scale and will apply also when the firm has more inputs (such as labor an capital, as in the neoclassical growth model). Our notation may suggest that there is one firm that produces the economy's entire output but, really, we think of a large number of firms solving the same problem under perfect competition, and the equilibrium will then determine total production but not which firm produces how much.

We now again consider the special case in which the utility function takes the logarithmic form, that is, $u(\cdot) = \log(\cdot)$. With this utility equation (5.10) can be written as

$$\frac{c_{i,t+1}}{c_{i,t}} = \frac{\beta p_t}{p_{t+1}}.$$

Since prices are the same for all agents, this implies that households experience the same growth in consumption. Thus, also in this case we have that individual consumption is a constant share of aggregate output, that is, $c_{i,t} = \theta_i Y_t$. However, output is not exogenous but it depends on the endogenous input of labor $L_t$. This is determined by the aggregation of individual labor supplies which are determined by the first-order condition (5.9). Using the log specification of the utility function and the fact that the wage rate is given by $w_t = A_t$, the condition can be rewritten as

$$\upsilon'(\ell_{i,t}) c_{i,t} = e_i A_t.$$

Next we use the property that individual consumption is a fixed share of aggregate output $c_{i,t} = \theta_i Y_t$. Substituting in the first-order condition we obtain

$$\upsilon'(\ell_{i,t}) \theta_i Y_t = e_i A_t.$$

Let us normalize skills so that $\int_0^1 e_i di = 1$. Given that skills are constant over time, different consumers have different total resources to spend in exact proportion to $e_i$, aside from differences in how much they work. However, we see that if we conjecture $\theta_i = e_i$, then the above condition implies that all households will supply the same labor $\ell_{i,t} = L_t$, confirming that total resources available to spend are simply proportional to $e_i$.

To see that the lifetime budget (5.7) is satisfied, use $c_{i,t} = \theta_i Y_t$, $w_t = A_t$, and $\ell_{i,t} = L_t$ in the budget constraint to obtain

$$\theta_i \sum_{t=0}^{\infty} p_t A_t L_t = e_i \sum_{t=0}^{\infty} p_t A_t L_t.$$

This is satisfied if $\theta_i = e_i$, confirming our guess.

Because high-skilled households earn higher incomes, they will enjoy higher consumption. High-skilled households, however, supply the same amount of labor as do low-skilled

households. This is a consequence of income and substitution effects offsetting each other. To see this, note that an extra unit of work effort earns $e_i w_t$, which increases with the wage and the skill of the worker. This generates a substitution effect in the direction of working more. However, since—without working more—the household earns higher consumption the higher is the wage, it experiences a lower marginal utility of consumption, which leads it to wish to work less (choose lower effort/higher leisure). With the utility function here, the two effects exactly cancel. These features play out even more clearly in a static optimization problem where the consumer has labor income only and spends it on consumption, $c = w\ell$, with the utility function $\log c - \upsilon(\ell)$: the choice of $\ell$ will not depend on $w$. This feature was alluded to in Section 2, where we argued that balanced growth with constant labor supply restricts us to a specific class of utility functions, to which the current example belongs.

Finally, imagine that individuals, in addition to being able to work, had independent ("asset") income. In particular, individual $i$ would be endowed with $a_{i,0}$ at time zero, with $\int_0^1 a_{i,0} di = 0$. This assumption implies that, if some agents have positive asset holdings, other agents must have negative holdings; thus, the resource constraints remain unchanged. Now, agents with different asset holdings will work different amounts because assets generate an income effect but not a substitution effect: higher asset holdings, by making the individual richer, will induce lower labor effort, as discussed above. Consider the simple static case just described: if the budget constraint reads $c_i = e_i A \ell_i + a_i$, individual labor supply will, using the first-order condition, be given by $\upsilon'(\ell_i) = a_i A/(e_i A \ell_i + a_i)$. Here, $\ell_i$ will depend nonlinearly on $a_i$ (so long as $\upsilon'$ is not a constant). Hence, total labor supply, along with total consumption, will depend on the distribution of assets: they will depend on wealth inequality. Consumption growth would still be equalized across agents, but total production will also depend on inequality since total consumption does.

### 5.2.3 The neoclassical growth economy

We extend the economy considered in the previous subsection by adding capital to production. This is essentially the neoclassical growth model with endogenous supply of labor. The accumulation of capital makes the problem more complex because what is produced in the future depends on the capital that is accumulated today. In most cases, a full analytical solution will not be available. Nevertheless, we can define the conditions that an equilibrium must satisfy.

The production function now takes the form

$$Y_t = A_t K_t^{\alpha} L_t^{1-\alpha}, \tag{5.11}$$

where $A_t$ is productivity, $K_t$ is the input of capital and $L_t$ is the input of labor.

We assume that capital is accumulated by households who then rent it to firms, similarly to labor. Therefore, in addition to the price for goods and labor, we now have a price for rental capital. The representative firm solves the profit maximization problem

$$\max_{\{K_t, L_t\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \Big\{ p_t A_t K_t^{\alpha} L_t^{1-\alpha} - p_t r_t K_t - p_t w_t L_t \Big\}, \tag{5.12}$$

where $r_t$ denotes the rental price of capital paid to households expressed in time-$t$ consumption. As in the previous sub-section, the optimal choice of the firm for period $t$ does not affect the optimal choices in other periods; we can equivalently state the problem as a sequence of sub-problems: for each $t$, $\max_{K_t,L_t} A_t K_t^{\alpha} L_t^{1-\alpha} - r_t K_t - w_t L_t$.

In this economy, production is used for consumption, $C_t$, and for investment, $I_t$:

$$Y_t = C_t + I_t.$$

Given the initial capital held by a household, $k_{i,0}$, and individual $i$'s investment $\iota_{i,t}$, for $t = 0, 1, ...$, the individual stock of capital evolves according to

$$k_{i,t+1} = (1-\delta)k_{i,t} + \iota_{i,t}. \tag{5.13}$$

Given $p_t$, $r_t$, $w_t$, investment $\iota_{i,t}$, and labor supply $\ell_{i,t}$, for $t = 0, 1, ..$, the value of lifetime income is $\sum_{t=0}^{\infty} p_t(r_t k_{i,t} + w_t e_i \ell_{i,t})$ and the value of expenditures for consumption and investment is $\sum_{t=0}^{\infty} p_t(c_{i,t} + \iota_{i,t})$. The budget constraint requires that the value of expenditures be equal to the value of incomes, that is,

$$\sum_{t=0}^{\infty} p_t(c_{i,t} + \iota_{i,t}) = \sum_{t=0}^{\infty} p_t(r_t k_{i,t} + w_t e_i \ell_{i,t}). \tag{5.14}$$

The compact equilibrium definition reads as follows.

**Definition 4** *An **Arrow-Debreu competitive equilibrium** is a set of sequences* $\left\{c^{\star}_{i,t}\right\}_{t=0}^{\infty}$, $\left\{\iota^{\star}_{i,t}\right\}_{t=0}^{\infty}$, *and* $\left\{\ell^{\star}_{i,t}\right\}_{t=0}^{\infty}$, *for each* $i \in [0,1]$, $\left\{L^{\star}_t\right\}_{t=0}^{\infty}$, $\left\{K^{\star}_{t+1}\right\}_{t=0}^{\infty}$, $\left\{p_t\right\}_{t=0}^{\infty}$, $\left\{p_t w_t\right\}_{t=0}^{\infty}$ *and* $\left\{p_t r_t\right\}_{t=0}^{\infty}$ *such that*

1. *for each* $i$, $\left\{c^{\star}_{i,t}\right\}_{t=0}^{\infty}$, $\left\{\iota^{\star}_{i,t}\right\}_{t=0}^{\infty}$, *and* $\left\{\ell^{\star}_{i,t}\right\}_{t=0}^{\infty}$ *solve*

$$\max_{\{c_t,\iota_t,\ell_t\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t u(c_t) \quad s.t. \quad \sum_{t=0}^{\infty} p_t(c_t + \iota_t) = \sum_{t=0}^{\infty} p_t(r_t k_t + w_t e_i \ell_t)$$

$$\textit{where, } \forall t, \quad k_{t+1} = (1-\delta)k_t + \iota_t \textit{ with } k_0 = k_{i,0},$$

2. $\left\{L^{\star}_t\right\}_{t=0}^{\infty}$ *and* $\left\{K^{\star}_t\right\}_{t=0}^{\infty}$, *where* $K^{\star}_0 = \int_0^1 k^{\star}_{i,0} di$, *with* $k^{\star}_{i,0} = k_{i,0}$, *solve*

$$\max_{\{L_t,K_t\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} p_t\Big\{A_t K_t^{\alpha} L_t^{1-\alpha} - r_t K_t - w_t L_t\Big\},$$

3. $\int_0^1 e_i \ell^{\star}_{i,t} di = L^{\star}_t$, $\int_0^1 k^{\star}_{i,t} di = K^{\star}_t$, *and* $\int_0^1 (c^{\star}_{i,t} + \iota^{\star}_{i,t}) di = A_t (K^{\star}_t)^{\alpha} (L^{\star}_t)^{1-\alpha}$ *for all* $t$.

Note that the $k_{i,0}$s are exogenous and that $K_0$ therefore is not included as an equilibrium object: we only include as equilibrium objects those that are endogenously determined. Also note, again, that since firms rent capital from consumers, their profit maximization problem is not dynamic (and we could equivalently have one firm per date). It is also possible to define an equilibrium where firms buy and own capital. A firm would then purchase a unit of the good at $t$ and use it as an investment good and then use it, together with labor at $t+1$. That way, the firm would have a dynamic problem, making profits in period $t+1$ and on that, in equilibrium, would be just large enough to offset the costs of investment at $t$ because of constant returns to scale.

## Equilibrium characterization

The optimality conditions for consumption, labor, and investment are derived by formulating the Lagrangian and taking derivatives. For consumption and labor we have, as in the previous economy,

$$\beta^t u'(c_{i,t}) = \lambda_i p_t, \tag{5.15}$$

$$v'(\ell_{i,t}) = e_i w_t u'(c_{i,t}), \tag{5.16}$$

where $\lambda_i$ is the Lagrange multiplier for the (lifetime) budget constraint, equation (5.14).

Since the optimality condition for consumption is the same as in the previous economy, we can again show that all households experience the same consumption growth independently of their skills $e_i$ and their initial wealth $k_{i,0}$, so long as $u(c)$ is a power function, which we will assume. This also implies that individual consumption is a share $\theta_i$ of aggregate consumption, that is, $c_{i,t} = \theta_i C_t$. What determines $\theta_i$ are, as before, the skills $e_i$ and the initial wealth $k_{i,0}$, along with how prices develop over time.

The optimality condition for investment is new and it takes the form

$$p_t = p_{t+1} r_{t+1} + (1-\delta) p_{t+1}. \tag{5.17}$$

The condition has a simple interpretation. If we buy one unit of capital today, it will cost us $p_t$. Next period, however, we can rent it to the representative firm earning the rental rate $p_{t+1} r_{t+1}$. In addition, we still have the non-depreciated capital $1-\delta$, which is worth $(1-\delta)p_{t+1}$. Thus, the left-hand-side is the cost of investing and the right-hand-side is the gross return, both expressed in terms of time-0 consumption. The equality simply says that the cost of investing must be equal to its return.[5] Note that if we divide both sides of equation (5.17) by $p_{t+1}$ we find $p_t/p_{t+1} = 1 + r_{t+1} - \delta$ so the price of consuming in date $t$ in terms of foregone consumption in $t+1$ is given by the gross return on capital.

The firm's problem now has two optimality conditions,

$$\begin{aligned} r_t &= \alpha A_t K_t^{\alpha-1} L_t^{1-\alpha}, \\ w_t &= (1-\alpha) A_t K_t^{\alpha} L_t^{-\alpha}. \end{aligned}$$

[5] From the consumer's perspective, one could conceive of prices and rental rates such that the stated equality is an inequality. This would either mean that a maximum does not exist or a corner solution; in equilibrium, neither are possible, so prices have to adjust to ensure equality.

The rental rate of capital and the wage rate are, respectively, the marginal productivities of capital and labor. Again, for general pairs $(r_t, w_t)$ these two conditions will not be met at the same time (in which a firm would make infinite profits or shut down): these prices will, however, adjust so that they are, and the firm is then indifferent as to the scale of the operation. However, the capital-labor ratio is pinned down uniquely.

Let us now specialize the utility function further: $u(c) = \log c$. Then, using the first-order condition for consumption, equation (5.15), at $t$ and $t+1$, we obtain

$$c_{i,t+1} = \beta \frac{p_t}{p_{t+1}} c_{i,t}.$$

Since $c_{i,t} = \theta_i C_t$ for all $t$, we obtain (equivalently, sum across all $i$)

$$C_{t+1} = \beta \frac{p_t}{p_{t+1}} C_t,$$

i.e., the Euler equation can also be expressed in terms of aggregate consumption. However, we know that from the discussion at the end of Section 5.2.2 that aggregate labor supply will depend on the asset distribution—as given by the distribution of $k_{i,0}$s here—so let us make a simplifying assumption here:

$$k_{i,0} = e_i K_0.$$

This assumption implies if you are rich in assets, you are also rich in productivity. Here, $e_i$ can then be interpreted also as the share of total initial capital $K_0$ held by agent $i$. We can also see that, conditional on working the same amounts, agents with different $e_i$s will have total wealth (including the asset holding) levels proportional to $e_i$. To see this, use $k_{i,t+1} = (1-\delta)k_{i,t} + \iota_{i,t}$ and (5.17) for all $t$ in the individual budget to obtain

$$\sum_{t=0}^{\infty} p_t c_{i,t} = \sum_{t=0}^{\infty} p_t w_t e_i \ell_{i,t} + (1 - \delta + r_0) k_{i,0}.$$

Here we see that, if we set $\ell_{i,t} = \ell_t$ independent of $i$—a guess we will confirm to be correct momentarily—then individual $i$'s total resources (the right-hand side) would be proportional to $e_i$, since $k_{i,0}$ also is. Since the Euler equations yield $p_{t+1}c_{i,t+1} = \beta p_t c_{i,t}$ for all $t$ and $i$, we can then insert, simplify, and rewrite the budget as

$$c_{i,t} = e_i \frac{\beta^t (1-\beta)}{p_t} \left( \sum_{t=0}^{\infty} p_t w_t \ell_t + (1 - \delta + r_0) K_0 \right).$$

We see that $\theta_i$, the individual's share of aggregate consumption, equals $e_i$.[6] Going back to equation (5.16), we see that the right-hand side becomes independent of $i$, and hence $\ell_{i,t}$ will be independent of $i$: it will equal $L_t$.

[6] To see this, note that $c_{i,t}$ has the structure $e_i X_t$ where $X_t$ is independent of $i$. Integrating across $i$ and using $\int e_i di = 1$ we find that $C_t \equiv \int c_{i,t} di = X_t$.

Given our assumptions, we can now think of there being a "representative consumer," even though consumers differ in endowments, since individuals' decisions all scale with $e_i$ (this representative is, rather, the aggregate of all consumers). We have not solved the model fully yet, however: aggregates remain to be determined. We can collect them as follows:

$$\frac{C_{t+1}}{C_t} = \beta\left(1 - \delta + \alpha A_{t+1} K_{t+1}^{\alpha-1} L_{t+1}^{1-\alpha}\right), \tag{5.18}$$

where we have used the aggregate Euler equation, (5.17), and the firm's first-order condition for capital;

$$\upsilon'(L_t) = \frac{(1-\alpha) A_t K_t^{\alpha} L_t^{-\alpha}}{C_t}, \tag{5.19}$$

where we have used the consumer's intratemporal first-order condition and the firm's first-order condition for labor; and

$$C_t = A_t K_t^{\alpha} L_t^{1-\alpha} + (1-\delta) K_t - K_{t+1} \tag{5.20}$$

from the resource constraint. Consumption can be eliminated and the system can be written as a difference equation (second-order in $K$, given $L$). It does not, in general, have a closed-form solution.[7]

How does the competitive equilibrium allocation compare to the solution to a social planner's problem of the kind we studied in Chapter 4? To explore this, let us suppose that all individuals have the same labor efficiency so $e_i = 1$ for all $i$. The social planner then wishes to maximize the utility of the representative household

$$\sum_{t=0}^{\infty} \beta^t \left[\log\left(C_t\right) - v\left(L_t\right)\right]$$

subject to the aggregate resource constraint (5.20). If we substitute the constraint into the objective function and take the first order condition for the planner's problem with respect to $K_{t+1}$ we obtain the exact same expression as (5.18). Similarly the first order condition with respect to $L_t$ is exactly the same as (5.19). The equations that characterize the solution to the planner's problem are therefore exactly the same as the equations that the competitive equilibrium must satisfy. As a result, the competitive equilibrium coincides with the choice of the planner. This is an important result that we will return to in Chapter 6.[8]

[7] $\delta = 1$ would deliver a constant saving rate equal to $\alpha\beta$, and $L_t$ would then be given by $\upsilon'(L_t)L_t = \frac{1-\alpha}{1-\alpha\beta}$ and not depend on time.

[8] We assumed here that all the individuals have the same labor productivity to simplify the exposition, but this assumption is not crucial. With heterogeneous skill levels we could show that the competitive equilibrium coincides with the solution to a different social planner's problem in which the social planner maximizes a weighted average of the utilities of the individuals with weights given by $e_i$.

## 5.3 Sequential equilibrium

In the previous sections we defined equilibria in a manner following Arrow-Debreu: all trades are decided on at time zero. In this section, instead, we assume that trades are decided on sequentially over time and deliveries take place either in the same period $t$ or in future periods. Future deliveries are especially relevant for financial contracts. For example, a debt contract is signed at time $t$, when borrowing occurs, but the repayment arises in one or more future periods.

Since trades arise sequentially, agents face a budget constraint in every period. This, however, does not imply that agents need to consume the whole income earned in the period. They can save by holding assets. In the rest of this section we assume that there is only one asset, denoted by $a_t$, that pays interest at a net rate $r_t$. We will define $q_t a_{t+1}$ as the amount saved at $t$ and $a_{t+1}$ as the amount delivered at $t+1$. This means that the real interest rate between $t$ and $t+1$ can be defined from $1 + r_{t+1} = 1/q_t$.[9]

The period resources that the agent does not use for consumption will be used to purchase assets. If instead consumption exceeds the resources available in the period, the agent will borrow. Borrowing means that the value of $a_t$ is negative, in which case the agent will pay an interest (the interests on borrowing and saving are the same). This will become clear in the applications.

### 5.3.1 The endowment economy

Returning to our endowment economy from Section 5.2.1, let us now assume that agents trade in every period. Agents thus trade an asset that pays the interest rate $r_t$: agent $i$ enters period $t$ with $a_{i,t}$ units of the asset and receives the endowment $y_{i,t}$. Therefore, the total resources available in the period are $a_{i,t} + y_{i,t}$. These resources are then used in part for consumption, $c_{i,t}$, and in part to purchase new units of the asset, $q_t a_{i,t+1}$. The budget constraint in period $t$ is thus

$$c_{i,t} + q_t a_{i,t+1} = y_{i,t} + a_{i,t}.$$

The agent thus maximizes lifetime utility (5.1) subject to the sequence of budget constraints, one for every period, and the no Ponzi game (nPg) condition introduced in Section 4.3.1. The initial asset holdings, which are exogenous, sum to zero: $\int_0^1 a_{i,0} di = 0$. We have the following.

**Definition 5** *A **sequential competitive equilibrium** is a set of sequences* $\left\{c^\star_{i,t}\right\}_{t=0}^{\infty}$ *and* $\left\{a^\star_{i,t+1}\right\}_{t=0}^{\infty}$*, for each* $i \in [0,1]$*, and* $\{q_t\}_{t=0}^{\infty}$ *such that*

[9] This convention is common; in an endowment economy, if alternatively $a_{t+1}$ is saving in consumption units at $t$ and $(1 + r_{t+1})a_{t+1}$ is the total return from this saving at $t+1$, then $r_0$ cannot be determined in equilibrium (or, equivalently, it can be set to any value). However, $q_0$ is determined.

1. *for each* $i$, $\left\{c_{i,t}^{\star}\right\}_{t=0}^{\infty}$ *and* $\left\{a_{i,t+1}^{\star}\right\}_{t=0}^{\infty}$ *solve*

$$\max_{\{c_t, a_{t+1}\}_{t=0}^{\infty}} \quad \sum_{t=0}^{\infty} \beta^t u(c_t)$$
$$s.t.$$
$$c_t + q_t a_{t+1} = a_t + y_{i,t} \quad \forall\, t, \ \textit{with } a_0 = a_{i,0},$$
$$\lim_{t\to\infty} \left(\Pi_{s=0}^{t} q_s\right) a_{t+1} \geq 0 \qquad \textit{(nPg condition)}$$

2. *for all* $t$, $\int_0^1 c_{i,t}^{\star} di = \int_0^1 y_{i,t} di$ *and* $\int_0^1 a_{i,t+1}^{\star} di = 0$.

The last two conditions are the market clearing conditions in goods and financial markets. The first says that the aggregate consumption must be equal to the aggregate quantity of goods available in every period. The second says that aggregate asset holdings must be zero in every period. In this economy, agents borrow and lend to one another. Each transaction therefore represents an increase in one person's assets and an offsetting decrease in another's. When we sum across all the agents, these trades net out to zero.

The equilibrium definition imposes two market clearing conditions, but in fact only one of them is needed due to Walras's Law. If we integrate the household budget constraints across $i$ and impose asset market clearing we arrive at the goods market clearing condition. Hence goods market clearing is insured by asset market clearing. It is also the case that goods market clearing implies the asset market clears.[10]

The optimality conditions can be derived by writing the Lagrangian and taking first-order conditions, which give

$$\beta^t u'(c_{i,t}) = \lambda_{i,t}, \tag{5.21}$$
$$q_t \lambda_{i,t} = \lambda_{i,t+1}, \tag{5.22}$$

where $\lambda_{i,t}$ is the Lagrange multiplier associated with the budget constraint. Differently from the first-order conditions in the setup with time-zero trading, the multiplier depends on time $t$. This is because we replaced the lifetime budget constraint (which is one constraint) with the sequence of period budget constraints. Therefore, we have one multiplier associated with each period budget constraint.

We now show that, despite this difference, we obtain the same optimality conditions. Using equation (5.21) at time $t$ and $t+1$ to eliminate the multipliers in equation (5.22) be obtain

$$q_t = \beta \frac{u'(c_{i,t+1})}{u'(c_{i,t})}.$$

[10] If we integrate the budget constraints at time 0 and impose $\int_0^1 a_{i,0} di = 0$ and goods market clearing we find $\int_0^1 a_{i,1} di = 0$. We can then proceed by induction to show that goods market clearing at date 1 implies asset market clearing at date 1 and so on.

This condition must be satisfied at any time $t$. If we use this condition from time 0 through time $t-1$, we obtain

$$q_0 \times q_1 \times \cdots \times q_{t-1} = \beta \frac{u'(c_{i,1})}{u'(c_{i,0})} \times .. \times \beta \frac{u'(c_{i,t})}{u'(c_{i,t-1})},$$

which can be rewritten more compactly as

$$\Pi_{s=0}^{t-1} q_s = \beta^t \frac{u'(c_{i,t})}{u'(c_{i,0})}.$$

The left-hand-side term corresponds to $p_t$ in the Arrow-Debreu equilibrium. Therefore, we obtained the optimality condition (5.3) we derived in the time-zero trade equilibrium. This illustrates that the Arrow-Debreu price $p_t$ is the present value at time zero of one unit of consumption at time $t$, discounted by the sequence of interest rates up to time $t$. It also shows that $p_{t-1}/p_t = 1/q_{t-1} = 1 + r_t$, that is, the gross interest rate between $t-1$ and $t$ is the ratio between the corresponding Arrow-Debreu prices: the relative price of consumption goods at $t-1$ in terms of goods at $t$.

The sequential equilibrium thus delivers the same equations determining quantities and, once translated back into Arrow-Debreu terms, the same prices as well. The literature uses both, guided by what is convenient in different contexts.

It is instructive to connect the budget constraints in the two setups: at first appearance they may not look equivalent, but they are. To see this, we start with the budget constraints at $t = 0$ and $t = 1$:

$$\begin{aligned} c_{i,0} + q_0 a_{i,1} &= a_{i,0} + y_{i,0}, \\ c_{i,1} + q_1 a_{i,2} &= a_{i,1} + y_{i,1}. \end{aligned}$$

Using the first equation to eliminate $a_{i,1}$ in the second equation (or viceversa) we obtain

$$c_{i,0} + q_0 c_{i,1} + q_0 a_{i,2} = a_{i,0} + y_{i,0} + q_1 y_{i,1}.$$

We use next the budget constraint at $t = 2$ to eliminate $a_{i,2}$, then the budget constraint at $t = 3$ to eliminate $a_{i,3}$, and so on. After $T$ substitutions we obtain

$$\sum_{t=0}^{T} \left(\Pi_{s=0}^{t-1} q_s\right) c_{i,t} + \left(\Pi_{s=0}^{T-1} q_s\right) a_{i,T+1} = a_{i,0} + \sum_{t=0}^{T} \left(\Pi_{s=0}^{t-1} q_s\right) y_{i,t}.$$

We have already shown that $\Pi_{s=0}^{t-1} q_s = p_t$. Furthermore, as $T$ converges to infinity, the second term on the left-hand side of the equation converges to something non-negative. Therefore, taking the limit $T \to \infty$ we obtain the lifetime budget constraint

$$\sum_{t=0}^{\infty} p_t c_{i,t} \leq a_{i,0} + \sum_{t=0}^{\infty} p_t y_{i,t}.$$

The budget constraint will be chosen to hold with equality given a strictly increasing utility function; that is, the limit of $\left(\Pi_{s=0}^{T-1} q_s\right) a_{i,T+1}$ as $t$ approaches infinity will be zero (a strictly positive amount would violate the TVC discussed in Section 4).

It is straightforward to apply a sequential equilibrium concept to other economies; to save space, we will only look at the neoclassical growth model, this time without valued leisure.

### 5.3.2 The neoclassical growth economy

We can similarly define an equilibrium of the neoclassical growth model in which agents trade period by period. We simply state the compact equilibrium definition here and leave it up to the reader to verify that it is equivalent to that described as an Arrow-Debreu equilibrium: that the allocations coincide and that prices, appropriately defined in comparable terms, do too. We use the case where labor is set exogenously to $e_i$ for agent $i$, with $\int_0^1 e_i di = 1$ so that aggregate labor supply equals 1.

**Definition 6** *A **sequential competitive equilibrium** is a set of sequences $\left\{c^\star_{i,t}\right\}_{t=0}^\infty$ and $\left\{k^\star_{i,t+1}\right\}_{t=0}^\infty$, for each $i \in [0,1]$, and $\{r_t\}_{t=0}^\infty$ and $\{w_t\}_{t=0}^\infty$ such that*

1. *for each $i$, $\left\{c^\star_{i,t}\right\}_{t=0}^\infty$ and $\left\{k^\star_{i,t+1}\right\}_{t=0}^\infty$ solve*

$$\begin{aligned}\max_{\{c_t,k_{t+1}\}_{t=0}^\infty} \quad & \sum_{t=0}^\infty \beta^t u(c_t) \\ & \text{s.t.} \\ & c_t + k_{t+1} = (1-\delta+r_t)k_t + w_t e_i \quad \forall\, t, \text{ with } k_0 = k_{i,0}, \\ & \lim_{t\to\infty} \frac{1}{\Pi_{s=0}^t (1-\delta+r_{s+1})} k_{t+1} \geq 0 \qquad \textit{(nPg condition)};\end{aligned}$$

2. *for each $t$, $(K_t, 1)$ solves $\max_{K,L} A_t K^\alpha L^{1-\alpha} - r_t K - w_t L$, where $K_t = \int_0^1 k^\star_{i,t} di$ (and $k^\star_{i,0} = k_{i,0}$); and*

3. *for each $t$,*

$$C_t + K_{t+1} = A_t K_t^\alpha + (1-\delta)K_t,$$

*where $C_t = \int_0^1 c^\star_{i,t} di$.*

Let us make two remarks. First, the use of Walras's Law here makes the last requirement redundant: add consumers' budgets at each point in time and use the firm's problem, which has constant returns to scale, to see this.[11] Second, we formulate the firm's problem as static here; it is simpler. In it, note that whereas $K$ and $L$ are mere choice variables and need not have time subscripts, the solution does need to be dated.

## 5.4 Recursive equilibrium

Just like it is possible to study dynamic optimization problems with recursive methods—dynamic programming—it is also possible to define equilibria that way. We saw in the context of optimization that recursive methods make the object of study different: we look for functions, not sequences. For a maximization problem, having solved a dynamic program

[11] Walras's Law was applicable in the Arrow-Debreu formulation too, but then only once—not at each $t$.

means that we have a function that we can apply to any value of its argument: the state variable (which can be a vector). A recursive equilibrium will also focus on functions.

To move slowly toward a general definition, let us first study how recursive methods can be used to define steady-state equilibria, i.e., equilibria where (aggregate) variables are constant over time. In the case of equilibria defined as sequences, we did not separately define steady-state equilibria: steady-state equilibria simply satisfy the definition of equilibria and have the additional property that all variables (at least aggregates) are constant over time. For the endowment economy, the case where all individuals' endowments are constant over time, all equilibria are steady-state equilibria. For the neoclassical economy, the economy's aggregate capital stock needs to have a certain starting value; otherwise, the equilibrium will not be in steady state from time zero. Steady-state equilibra defined using recursive methods are also just a subset of all equilibria. However, as we shall see, it is useful to define steady-state equilibria separately, as a first step.

### 5.4.1 Steady state

A steady-state equilibrium is one where the aggregate economy is at a rest point. This means, in particular, that prices are constant. Before proceeding, note that the present discussion also applies to the case where there is exact balanced growth, in which case the economy can be transformed into a stationary one whose steady state is then studied (in this case, wages will be growing along the balanced path but will be constant in the transformed economy).

Conceptually, now, for the definition of a steady-state equilibrium we will need to specify the constant prices and aggregates and to provide functions making clear that, at steady state, agents optimize: they consider behavior that is non-constant over time but choose to remain constant.

As before, we will go through definitions for different economies, beginning with a case without production.

#### The endowment economy

Consider, again, the economy with a continuum of consumers indexed by $i$, each now with and endowment that is constant over time, $y_i$. An equilibrium is defined as follows.

**Definition 7** *A **recursive steady-state competitive equilibrium** is a $q$ and a set of functions, $V_i^\star(a)$ and $g_i^\star(a)$, and asset values $a_i^\star$, for for each $i \in [0,1]$ such that*

1. *for each $i$, $V_i^\star(a)$ solves*

$$V_i(a) = \max_{a'} u(a + y_i - qa') + \beta V_i(a') \quad \forall a,$$

   *with $g_i^\star(a)$ attaining the maximum on the right-hand side for all $a$;*

2. *for each $i$, $g_i^\star(a_i^\star) = a_i^\star$; and*

3. $\int_0^1 g_i^\star(a_i^\star)di = 0$.

Here, the $V^\star$s and $g^\star$s capture individual optimization, given prices (a single $q$ in this case). The second condition expresses stationarity of individual asset holdings; at the same time, it does put a sharp restriction on the function. We are, thus, allowing different individuals to start with—and maintain—different holdings of assets. Market clearing is captured in the third condition. Our equilibrium definition also specifies a distribution of asset holdings, which is required to be constant over time. A special case is that of a representative agent, where the second condition would read $g^\star(0) = 0$ and the third condition would be superfluous.

**Equilibrium characterization** The functional Euler equation, which we arrive at after taking the first-order condition in the dynamic program and then using the envelope theorem, reads, for all $i$,

$$qu'\left(a + y_i - qg_i(a)\right) = \beta u'\left(g_i(a) + y_i - qg_i(g_i(a))\right).$$

Using stationarity at $a = a_i$, we obtain

$$qu'\left(a_i(1-q) + y_i\right) = \beta u'\left(a_i(1-q) + y_i\right).$$

Hence, $q = \beta$. We can now solve for $g_i$ from the functional Euler equation: it implies $a - qg_i(a) = g_i(a) - qg_i(g_i(a))$ and it is easy to see here that $g_i^\star(a) = a$ solves this functional equation for all $i$. No matter what asset level an agent has, they choose to keep it and just consume its accrued interest. This is the permanent-income result we have seen before, now in recursive form.

Given the simple form of the policy function, we immediately obtain the implied $V_i^\star$s. They satisfy $V_i^\star(a) = u(a(1-q) + y_i)/(1-\beta)$.

Having found all the equilibrium objects, we note, first, that the asset distribution can be chosen freely subject to market clearing and that all consumers have non-negative consumption. That is, this model has "no predictions" for long-run wealth distributions: any relative distribution of wealth is sustained over time. Since real-world wealth distribution have certain distinct features in common—regarding their overall shapes, across time and countries—this model therefore is not satisfactory. Richer models of wealth distribution that relate to data are therefore developed and discussed in Chapter 19. Second, we see that the shape of the utility function, $u$, does not play a role in any of the characterizations of the steady state. This is not true if there is balanced growth (in endowments); then, $u$ needs to have the usual balanced-growth shape.

### The neoclassical growth economy

Turning to the neoclassical growth model with optimal saving, let us proceed immediately to the definition of steady-state equilibrium, of course with the assumption that TFP, $A_t$, is

constant over time.[12] We will again consider a continuum of consumers. For simplicity, we will assume that labor supply is exogenous and equal to $e_i$ for agent $i$, with $\int_0^1 e_i di = 1$.

**Definition 8** *A **recursive steady-state competitive equilibrium** consists of scalars $r$ and $w$ and a set of functions, $V_i^\star(k)$ and $g_i^\star(k)$, and capital holdings $k_i^\star$, for for each $i \in [0,1]$ such that*

1. *for each $i$, $V_i^\star(k)$ solves*

$$V_i(k) = \max_{k'} u((1-\delta+r)k + we_i - k') + \beta V_i(k') \quad \forall k,$$

   *with $g_i^\star(k)$ attaining the maximum on the right-hand side for all $k$;*

2. *$(K^\star, 1)$ solves $\max_{K,L} F(K,L) - rK - wL$, where $\int_0^1 k_i^\star di = K^\star$; and*

3. *for each $i$, $g_i^\star(k_i^\star) = k_i^\star$.*

We see a close similarity between the agent's problem here and that in the endowment economy: there exogenous labor income and a constant interest rate. There are two prices, $r$ and $w$, but as we shall see they are closely related. Finally, we see that the market-clearing condition is different: assets sum up to the aggregate capital stock.

**Equilibrium characterization** For reasons identical to those used in the case of the endowment economy, we obtain $\beta = 1 - \delta + r$. This equation determines $r$. Then we know from firm maximization that $r = F_K(K^\star, 1)$. This equation determines $K^\star$. We also know that $w = F_L(K^\star, 1)$, which gives us $w$. The distribution of capital is indeterminate, subject to adding up to $K^\star$. The policy and value functions are given by $g_i^\star(k) = k$ and $V_i^\star(k) = u(k(r-\delta) + we_i)/(1-\beta)$, respectively, for all $k$ and $i$.

### 5.4.2 Dynamics

Turning to a full equilibrium, we now need to find a way to express, using functions, how prices and aggregates move over time. We will consider the neoclassical growth model only.[13] We will assume that TFP is constant, so as to isolate how these price and quantity movements are endogenous. We will, again, consider a continuum of households, but now assume as a benchmark that their asset holdings and labor productivities are the same. This means that the consumer we look at can be thought of as a representative agent from the outset, i.e., one among a [0,1] continuum of identical agents. We will revisit the question of actual heterogeneity at the end. We also begin with the assumption that labor supply is exogenous and equal to 1.

[12] Alternatively, it is growing at a constant rate, in which case the analysis requires a transformation of variables.

[13] Endowment economies can be considered as well but do not involve the same core issues as the neoclassical model allows us to illustrate.

Recall that recursive methods involve expressing outcomes as functions of *state variables*. A state variable has to be both relevant and predetermined. So, in our economy, what determines prices and aggregates over time? It is instructive to consider period 0: what determines prices at that point in time? In the neoclassical model $r_0$ and $w_0$ are determined by the capital/labor ratio, through the marginal products of capital and labor. Hence, at the very least these two prices will depend on the period-0 capital stock in the economy: besides being relevant, the aggregate capital stock is also predetermined. Is there another variable that qualifies as a state, determining prices? No. Hence, we conclude that two functions $r(K)$ and $w(K)$ need to be part of our equilibrium.[14]

Agents will thus take as given $r(K)$ and $w(K)$ when making their decisions. But, given that prices will move over time, how do they know what prices will prevail in the future? For this they need the equivalent of the planner's decision rule for capital: in the context of a recursive equilibrium, we will label it a law of motion, $K' = G(K)$. Thus, agents take $G$, $r$, and $w$ as given functions and then solve their dynamic programming problems. What will these problems look like? As in the case of the steady-state equilibrium we need to allow the agent's choice of capital to deviate from that in equilibrium. Therefore the dynamic program must read

$$V(k, K) = \max_{k'} u((1 - \delta + r(K))k + w(K) - k') + \beta V(k', G(K)) \quad \forall (k, K), \tag{5.23}$$

where $k$ is an individual agent's capital and $K$ is the aggregate capital stock, which they take as given.

We now define equilibrium:

**Definition 9** *A **recursive competitive equilibrium** consists of functions $r(K)$, $w(K)$, $G^\star(K)$, $V^\star(k, K)$, and $g^\star(k, K)$ such that*

1. *$V^\star(k, K)$ solves (5.23), for $G = G^\star$ and $g^\star(k, K)$ attains the maximum in this problem;*
2. *for all $K$, $r(K) = F_K(K, 1)$ and $w(K) = F_L(K, 1)$; and*
3. *$G^\star(K) = g^\star(K, K)$ for all $K$.*

The second condition does not state profit maximization of the firm explicitly but just uses the first-order conditions.[15] The third condition, labeled *consistency*, works as a market clearing condition: it requires that the evolution of the aggregate capital stock is consistent with the choices of individuals when they each hold the same level of capital.

Is the resource constraint met in our definition? It is. The representative consumer will, in a recursive competitive equilibrium, consume $K(1 - \delta + r(K)) + w(K) - G^\star(K)$, which, from the firm's problem and $F$ being CRS, equals $F(K, 1) - K' + K(1 - \delta)$, i.e., output minus investment.

[14] As an alternative to looking at time 0 to gain intuition, consider the planning problem of the economy under study: the state variable relevant to the planner will then be a state in the recursive equilibrium.

[15] Alternatively, state that $r(K)$ and $w(K)$ are such that $(K, 1)$ solves $\max_{k,\ell} F(k, \ell) - r(K)k - w(K)\ell$ for all $K$.

**Equilibrium characterization** Let us also derive the functional Euler equation of the agent. Taking first-order conditions and applying the envelope theorem (for a marginal change in $k$), we obtain (dropping $^\star$s for convenience) that,for all $(k, K)$,

$$u'\left((1-\delta+r(K))k+w(K)-g(k,K)\right)=$$

$$\beta u'\left(g(k,K)(1-\delta+r(G(K)))+w(G(K))-g(g(k,K),G(K))\right)\left[1-\delta+r(G(K))\right].$$

Given $r$, $w$, and $G$, this functional equation determines $g(k, K)$: the behavior of individual saving when the individual has $k$, possibly different from aggregate capital, $K$. Can $g$ be solved for explicitly? In general, no. However, when $u$ is a power function, again in line with our general balanced-growth requirements, it is possible to show that it takes the form $g(k, K) = \mu(K) + \lambda(K)k$.[16] Individual saving is *linear* in the own holdings of capital: the marginal propensity to save is $\lambda(K)$, i.e., independent of the level of $k$ (it only depends on the aggregate capital stock). This also means that if we were to consider several consumers and distribute capital among them, how we distribute it would not matter for aggregate saving. In other words, we have *aggregation* if $u$ is a power function. Thus, it is not restrictive to consider a representative agent. This result holds also if different agents have different endowments of labor: then $g_i(k, K) = \mu_i(K) + \lambda(K)k$, i.e., the intercept will vary across individuals but the slope will not. To conclude, if an economy with a power utility function has a nontrivial distribution of capital among agents, the aggregate law of motion will only depend on aggregate capital. That is, $K$ is still the aggregate state variable: the distribution of capital, though predetermined, is not relevant for understanding how prices are determined.

Turning to how one solves for aggregates, we can also evaluate the above Euler equation at $k = K$ and use the fact that the resource constraint will hold, along with the expression for the rental rate function, to obtain

$$u'\left(K(1-\delta)K+F(K,1)-G(K)\right)=$$

$$\beta u'\left(G(K)(1-\delta)+F(G(K),1)-G(G(K))\right)\left[1-\delta+F_K(G(K),1)\right].$$

This functional equation solves for the evolution of the capital stock. We note that it coincides with the functional equation of the planner's problem for the same economy: equation (4.25) of our optimization chapter. As in that case, this equation has to be solved numerically unless $u$ is logarithmic, $F$ is Cobb-Douglas, and $\delta = 1$.

Before concluding, let us consider the economy with valued leisure: agents have utility functions $u(c) - v(\ell)$. Now aggregate labor supply is endogenous. It is not predetermined in a given period, so it is not a state variable (neither for the individual nor for the aggregate). Thus, our state is still $(k, K)$. What is needed now, however, is a function $L = H(K)$ specifying how aggregate labor supply depends on the aggregate state. Similarly, on the individual level, we need $\ell = h(k, K)$ to denote the policy function for the choice of labor. The equilibrium, again for the representative-agent case, becomes

[16] See Appendix 5.A for a proof and a more extensive discussion of conditions under which the solution has this form.

**Definition 10** *A **recursive competitive equilibrium** for the economy with valued leisure consists of functions $r(K)$, $w(K)$, $G^\star(K)$, $H^\star(K)$, $V^\star(k,K)$, $g^\star(k,K)$, and $h^\star(k,K)$ such that*

1. *$V^\star(k,K)$ solves*

$$V(k,K) = \max_{k',\ell} u((1-\delta+r(K))k + w(K)\ell - k') - v(\ell) + \beta V(k', G^\star(K)) \quad \forall (k,K).$$

*and $k' = g^\star(k,K)$ and $\ell = h^\star(k,K)$ attain the maximum in this problem;*

2. *for all $K$, $r(K) = F_K(K, H^\star(K))$ and $w(K) = F_L(K, H^\star(K))$; and*

3. *$G^\star(K) = g^\star(K,K)$ and $H^\star(K) = h^\star(K,K)$ for all $K$.*

This is a straightforward extension of the definition without valued leisure. Notice that the agent does not use $H^\star$ in the maximization problem; $G^\star$ suffices, because $r$ and $w$ now capture how the labor input changes with $K$.

It is straightforward, but somewhat tedious, to derive the functional Euler equations; there will now be an intertemporal condition too determining labor supply. Again, evaluated at $k = K$, one finds that the conditions are identical to those of the corresponding planner's problem.

Will this economy deliver aggregation too, once $u$ is of the power form? As already alluded to above in Section 5.2.3, the answer is no, unless labor productivities differ across agents too and the ratio of the initial capital holding to labor productivity is the same across all agents.[17] If not, the aggregate state variable necessarily becomes the vector of capital holdings of all agents, not just the sum of these holdings.

## 5.5 Overlapping generations

We now consider overlapping generations models. The defining feature of these models is that agents live for a finite length of time but the economy continues after their death with new generations of agents. Historically, overlapping-generations (OG) models have played an important role in macroeconomics. They were first introduced by Allais (1947) and later used for a large variety of purposes: for understanding why fiat money has value and the potential need for a social security system (Samuelson, 1958) and for understanding government debt (Diamond, 1965). Interestingly, although most of his later work used dynastic settings, Lucas's path-breaking (1972) paper on the Phillips curve is using an OG model. We will revisit these applications later in the text. In this chapter, we will merely focus on some features that make OG market economies different than those studied above.

We will restrict attention to the simplest version of an OG setting: one where, each period, a cohort of people are born and then live for two periods only. That way, cohorts overlap, but

[17] In this case, this ratio will also stay the same over time.

only for one period. This is the so-called the two-period life OG setting. Clearly, the two-period life case is limited in its applicability: for example, it cannot be used for quantitative studies of the business cycle, as business cycles occur rather frequently (a time period in the OG model should perhaps be thought of as 25 or 30 years). However, it is of course possible to construct OG models where people live for an arbitrary (finite) number of periods and many of the special properties of such models are inherited from the two-period life case.

In our benchmark, we will assume—as is the case in the most commonly used OG model—that people, though they have children, do not give bequests or any other gifts to them. Relatedly, they express no altruism toward their offspring and thus maximize their own utility only, defined over consumption (and possibly leisure) in their two periods of life. The maximization problems are, then, conceptually simple; the finite-horizon settings studied in Section 4.2, in particular that in 4.2.1, can be immediately applied. The maximization problems, moreover, are often called *life-cycle models*, emphasizing that individuals go through different phases in life and, in particular, that young and old individuals have different time horizons. Here, we will limit heterogeneity to age and not consider further differences between people. Thus within each cohort, all agents are identical.

Defining equilibria for OG models is straightforward, too. As for the dynastic setting, there are three methods—sequence-based (AD and sequential-trade) and recursive definitions of equilibria—but here, for convenience only, we will focus exclusively on the sequential-trade setting. We will see that not only are the maximization problems simpler in OG models but computing equilibria is more straightforward too. However, what we will find is that equilibria for OG models can (but do not necessarily) have peculiar features. For example, equilibria may not be Pareto-optimal. In addition, there may be more than one equilibrium. In this chapter, we will mainly set things up; welfare properties will then be studied carefully in the next chapter and applications will be discussed in later chapters.

As before, we begin with the endowment case. We will then turn to the neoclassical growth model. Lastly, we will discuss introducing altruism and bequests. Throughout, we abstract from population growth.

### 5.5.1 The endowment economy

Let us assume that the cohort born at $t$ has utility given by

$$u_t(c_y,\, c_o) = u(c_y) + \beta u(c_o),$$

where we evaluate at two arbitrary consumption levels $(c_y, c_o)$ when young and old, respectively. The preferences of generation $t = -1$, who are old as time begins, are similarly represented by $u_{-1}(c) = u(c)$.

We consider endowment sequences given by $(\omega_{y,t}, \omega_{o,t+1})$ for cohort $t$: for all $t$, where $t$ is the time period, $\omega_{y,t}$ is the endowment of the young and $\omega_{o,t}$ the endowment of the old (who are of cohort $t-1$) in that time period.

A sequential equilibrium is defined in much the same way as in our endowment economy with dynastic agents.

**Definition 11** *A **sequential competitive equilibrium** is a set of sequences* $\left\{c^\star_{i,t}\right\}_{t=0}^\infty$*, for each* $i \in \{y, o\}$*,* $\left\{a^\star_{t+1}\right\}_{t=0}^\infty$ *and* $\{q_t\}_{t=0}^\infty$ *such that*

1. *for each* $t > 0$*,* $\left(c^\star_{y,t}, c^\star_{o,t+1}, a^\star_{t+1}\right)$ *solves*

$$\max_{c_y, c_o, a'} u(c_y) + \beta u(c_o) \quad s.t. \quad c_y + q_t a' = \omega_{y,t} \text{ and } c_o = \omega_{o,t+1} + a'$$

   *and* $c_{o,0} = \omega_{o,0}$*.*

2. *for all* $t \geq 0$*,* $c^\star_{y,t} + c^\star_{o,t} = \omega_{y,t} + \omega_{o,t}$*.*

The last requirement—goods market clearing—can equivalently, using consumers' budgets, be written $a^\star_{t+1} = 0$ for all $t$. To see this, begin in period 0 and roll forward: resource feasibility at time 0 means, when summing cohort 0's budget and cohort 1's first-period budget, that $a^\star_1$ must equal 0. Then the same procedure for next period delivers $a^\star_2 = 0$, and so on. Intuitively, this is obvious: any given cohort fundamentally only have endowment income, and saving (or borrowing) when young must mean that another cohort is on the other side of that transaction; but it cannot be the old who is now alive during their last period, and it cannot be the young of next cohort, since they have not been born yet.

**Equilibrium characterization** Having established $a^\star_{t+1} = 0$, a result that is true also in the dynastic model if all agents have the same endowments, we conclude that $(c^\star_{y,t}, c^\star_{o,t}) = (\omega_{y,t}, \omega_{o,t})$: autarky. These results are in line with what we found in the dynastic model. Also, as before, solving for the price only involves evaluating the consumer's first-order (Euler) condition, which—evaluated at autarky—reads

$$q_t u'(\omega_{y,t}) = \beta u'(\omega_{o,t+1})$$

at time $t$. If we impose $u(c) = \frac{c^{1-\sigma}-1}{1-\sigma}$, so that (in a slightly more general model) we obtain results consistent with balanced growth, we conclude that

$$q_t = \beta \left(\frac{\omega_{y,t}}{\omega_{o,t+1}}\right)^\sigma. \tag{5.24}$$

If the endowments are stationary, so that $\omega_{y,t}$ and $\omega_{o,t}$ do not depend on time, we obtain that $q_t$ is constant and equal to $\beta(\omega_y/\omega_o)^\sigma$. Thus, if the endowment when young does not equal the endowment when old, the interest rate will not just reflect the utility discount factor $\beta$ but also the shape of *life-cycle income*. In particular, if $\omega_y > \omega_o$, which makes sense if we identify $y$ with "working" and $o$ with "being retired," then we can even obtain $q > 1$, i.e., negative real interest rates, in this model. This is not possible to obtain in the dynastic endowment model (unless aggregate endowments fall over time at a constant, and high enough, rate).

The intuition for the possibility of negative interest rates in the OG model is that when life-cycle endowments decline over time (at a high enough rate), consumption is marginally

more valuable in the future than now in the absence of being able to smooth income over time. And in the two-period life OG model no such smoothing is feasible. In a 3-period OG model there could be borrowing/lending between the young and the middle-aged. It would be reasonable to think that the life-cycle endowment pattern then has $\omega_y < \omega_m > \omega_o$, where $\omega_m$ is the endowment of the middle-aged. The young then could borrow from the currently middle-aged, who want to lend. Thus, some smoothing would be obtained and the discount factor would play a more prominent role again; but full smoothing would not necessarily materialize and, thus, the OG model continues to give different predictions than does the dynastic model. Overall endowment growth would also affect the result in the direction of producing higher real interest rates.

Finally, note that it would be beneficial to transfer resources from the young to the old, in the case where $\omega_y > \omega_o$. If *all* young transfer to the current old then it is even possible for *all* generations to benefit. This kind of transfer can be thought of as a government-run pay-as-you-go pension scheme; in fact, this is a key early use of the OG model. Thus, we have an indication that the market is not efficient here. We will discuss this issue at much greater length in our welfare chapter: Chapter 6.

### 5.5.2 The neoclassical growth economy

In the two-period life OG model, the introduction of capital allows consumers to actually life-cycle save. They are endowed with labor income when young and labor income when old and, as before, have no initial assets. So if their income as young is higher than their income when old, they would buy capital when young and rent it out as, old agents, to firms. At time 0, the capital stock is thus owned by the old at that time.

For concreteness, we let the labor productivity (or alternatively time endowment) of the young and old at all times be $e_y$ and $e_o$, respectively, with $e_y + e_0 = 1$. We also restrict attention to a stationary production function. Growth in productivity is straightforward to introduce and it is not essential for the key points here. We have the following (where we use a representative agent within each cohort to economize on notation).

**Definition 12** *A **sequential competitive equilibrium** is a set of sequences $\left\{c^\star_{i,t}\right\}_{t=0}^{\infty}$, for each $i \in \{y, o\}$, $\left\{k^\star_{t+1}\right\}_{t=0}^{\infty}$, $\{r_t\}_{t=0}^{\infty}$, and $\{w_t\}_{t=0}^{\infty}$ such that*

1. *for each $t > 0$, $\left(c^\star_{y,t}, c^\star_{o,t+1}, k^\star_{t+1}\right)$ solves*

$$\max_{c_y, c_o, k'} u(c_y) + \beta u(c_o) \quad s.t. \quad c_y + k' = w_t e_y \text{ and } c_o = w_{t+1} e_o + (1 - \delta + r_{t+1})k'$$

   *and $c_{o,0} = w_0 e_o + (1 - \delta + r_0)k_0$;*

2. *for all $t$, $r_t = F_k(k^\star_t, 1)$ and $w_t = F_\ell(k^\star_t, 1)$, with $k^\star_0 \equiv k_0$; and*

3. *for all $t \geq 0$, $c^\star_{y,t} + c^\star_{o,t} + k^\star_{t+1} = F(k^\star_t, 1) + (1 - \delta)k^\star_t$.*

This definition is in line with the one for a dynastic economy with the one difference that there are two types of agents, only one of which saves in capital at any point in time.

**Equilibrium characterization** The focus is on the individual problem. The Euler equation is entirely standard-looking, so let us proceed immediately to evaluate it after expressing prices as a function of capital stocks:

$$u'\left(e_y F_\ell(k_t^\star, 1) - k_{t+1}^\star\right) =$$

$$\beta u'\left(e_o F_\ell(k_{t+1}^\star, 1) + (1 - \delta + F_k(k_{t+1}^\star, 1))k_{t+1}^\star\right)\left(1 - \delta + F_k(k_{t+1}^\star, 1)\right).$$

Conditional on a value for $k_t^\star$, this equation solves for $k_{t+1}^\star$. Therefore, we have a very different dynamic system than in the dynastic model, where we always obtained a second-order difference equation. That is, here, we can "solve forward": start with $k_0$, solve for $k_1^\star$, and so on.

Second, what does a steady state look like? Letting $\bar{k}$ denote a steady state, we obtain

$$u'\left(e_y F_\ell(\bar{k}, 1) - \bar{k}\right) = \beta u'\left(e_o F_\ell(\bar{k}, 1) + (1 - \delta + F_k(\bar{k}, 1))\bar{k}\right)\left(1 - \delta + F_k(\bar{k}, 1)\right).$$

That is, we do not, in general, obtain $\beta\left(1 - \delta + F_k(\bar{k}, 1)\right) = 1$, since consumption may not end up being fully smoothed. Again, thus, we see that OG models have qualitatively different implications for long-run interest rates: they can, depending on parameter values (intuitively, depending on the demand for saving given the life-cycle structure, and depending on firm's demand for capital) be either higher or lower than $1/\beta$ (in gross terms). We invite the reader to fully solve the model where $u$ is logarithmic, $e_y = 1$ and $e_o = 0$, $F$ is Cobb-Douglas, and $\delta = 1$ and verify that (i) capital's dynamics will be log-linear and converge monotonically to a steady state; and (ii) the gross steady-state interest rate will be $\alpha(1 + \beta)/(1 - \alpha)$, a number which is less than one if $\alpha$ is low enough.

Finally, we also note that although the solution for the model's dynamics involves only a first-order difference equation, it is not immediately obvious that, for each $k_t$, there is a unique value $k_{t+1}$ solving the Euler equation. In our dynastic model, we obtain a unique equilibrium so long as it can be obtained as a solution to a planner's problem, which we know to be unique under standard assumptions. Here, there is no immediate connection to a planner's problem.[18]

### 5.5.3 Some model comparisons

We have studied the OG model in its two-period life version. When consumers live for more than two periods, the model in some ways looks more like a dynastic model; as the time horizon gets longer, the life-cycle patterns can become less pronounced. Also, solving for equilibria in OG models with more than two-period lives also does involve difference equations that are of higher order than one. Still, however, they are conceptually different. The determination of the long-run real interest rate, for example, remains more complex than in the dynastic model, where very few parameters matter: without aggregate growth, the gross real interest rate is $1/\beta$, and with growth at rate $\gamma$, it is $(1 + \gamma)^\sigma/\beta$.

[18]Note also that a planner's objective in the OG setting would need to involve a social welfare function across cohorts.

**Random deaths** There are other variants of the OG model. One is the *perpetual-youth* (or, perhaps, *sudden-death*) model. There, all individuals face a constant probability of death between $t$ and $t+1$. Thus, a "lucky" individual can live very long, even forever. As a consequence, anybody alive at $t$ has the same expected remaining lifetime. An individual who dies will be replaced by a new-born individual, so as to avoid a shrinking population, but again not as part of a dynasty: this model shares with the basic OG setting that no individual cares about children. When an individual dies with unspent assets—which is typical—then these assets can be either seized by the government or given as "random bequests" to the surviving population. A third alternative—all have been used in the literature—is that agents can write a form of annuity contracts whereby they obtain a higher than the (safe) market interest rate if they survive, in return for losing all the money at death; a "bank" on the other side of this transaction would then on average, if the contract is written with many individuals, obtain a safe return. The perpetual-youth model inherits some of the characteristics of the OG model but, at the same time, is a move toward a dynasty model: as the probability of survival approaches one, the model's features approach those of the pure dynasty setting.

**Warm glow** One can also append an OG model with *bequest functions*. The idea here is that people do not exhibit altruism but, rather, they care about the act of giving money away (typically, to their children). This setting is referred to as one of *warm glow*: giving makes one happy (glow). To illustrate, consider the two-period life OG model. The preferences are now

$$u(c_y) + \beta u(c_o) + \varphi(b'),$$

where $\varphi$ (like $u$) is an increasing, strictly concave function and $b'$ is the amount of bequests given. Thus, the budget constraints for cohort $t$ (assuming a stationary endowment economy for simplicity) read

$$c_{y,t} + q_t a_{t+1} = \omega_y + b_t \text{ and } c_{o,t+1} + b_{t+1} = \omega_o + a_{t+1}.$$

Here, $b_t$ is the bequest inherited from cohort $t-1$ (the parents) and $b_{t+1}$ is the amount bequeathed to cohort $t+1$. Market clearing could again be expressed by requiring that total consumption equal total endowments period by period; alternatively, one could simply require $a_{t+1}$ to be zero at all times. Now notice that the old at time 0 have a non-trivial decision: maximize, by choice of $(c_{o,0}, b_0)$, $\beta u(c_{o,0}) + \varphi(b_0)$ subject to $c_{o,0} + b_0 = \omega_o$.

The warm glow OG model can be seen as a *behavioral* model: agents no longer have preferences over consumption goods (and leisure) but are endowed with utility functions defined by the act of giving. They are sometimes referred to as examples of impure altruism: it is not that the parent cares about the child—in fact, no features of the child's economic life appear in the bequest motive—but that the parent cares about giving per se. A model with pure altruism is instead one where the parent cares about the child *the way the child cares about themselves*, in which case one can no longer label the model behavioral since now preferences are defined over goods. If this is true for all cohorts, we can specify the utility

of cohort $t$ as follows:

$$u(c_{y,t}) + \beta u(c_{o,t+1}) + \tilde{\beta}\varphi(b_{t+1}),$$

where

$$\varphi(b_{t+1}) = u(c_{y,t+1}) + \beta u(c_{o,t+2}) + \tilde{\beta}\varphi(b_{t+2}),$$

and so on. Notice here that $\tilde{\beta}$, the weight on the child's indirect utility, does not have to equal $\beta$. This reveals a recursive structure, which is most easily described with dynamic programming notation. Assuming a constant interest rate $1/q$ for simplicity, we obtain

$$\varphi(b) = \max_{c_y, c_o, a', b'} u(c_y) + \beta u(c_o) + \tilde{\beta}\varphi(b'),$$

subject to $c_y + qa' = \omega_y + b$ and $c_o + b' = \omega_o + a'$.

In sum, we have seen the perpetual-youth model, which looks more and more like the dynastic model as the survival probability goes to one. The warm glow model, on the other hand, can be seen as a behavioral OG model, except in the very special case where $\varphi$ is (a constant times) the indirect utility function of the next cohort. Then, the model becomes a *dynastic life-cycle model*: one where there are life-cycle features but where parents care about children in a (purely) altruistic way. In this very special case, thus, the basic long-run features are exactly those of the simpler dynasty model we use in our benchmark. The long-run interest rate is now given by $1/\tilde{\beta}$, and so on.

# Chapter 6

# Welfare

The preceding chapter presented examples where the competitive equilibrium allocation is the same as the solution to a social planner's problem. These examples illustrate the First Welfare Theorem of economics, which gives conditions under which a competitive equilibrium is Pareto optimal. The First Welfare Theorem is a remarkable result—arguably the most impressive insight that economics has come up with so far. Market mechanisms can effectively coordinate the activities of large numbers of consumers and firms, who may be spread across the world and alive at different times. Amazingly, under the conditions of the First Welfare Theorem (FWT), the actions of these disparate and heterogeneous people are coordinated in a way that is socially optimal even though each individual is simply pursuing their private interests. This is Adam Smith's invisible hand.[1]

Under the FWT, a competitive equilibrium is Pareto optimal. We focus on the concept of Pareto optimality because it is widely used as a "minimal" welfare criterion. It is, in particular, silent on the distributional implications of an allocation. For example, an allocation with extreme inequality can nevertheless be Pareto optimal. At times, macroeconomists assume more specific social welfare functions that express a preference for more equal distributions of resources and we will comment upon this briefly.

While markets can work very well, they can also fail, which is to say that the competitive equilibrium may not be Pareto optimal. Such failures of the FWT can arise for a number of reasons, which we will discuss in detail. Our aim is to understand when market economies work well and when they do not. Indeed, much of modern macroeconomics is concerned with understanding the potential market failures that affect the economy and understanding the public policies that might lead to better outcomes. An understanding of the nature of market failure and how it can be corrected is an important part of evaluating the potential benefits of policy proposals.

After first reviewing the relationship between Pareto optimality and competitive equilibrium (the FWT) in an abstract, general setting, we map it into the more applied macroeconomic settings used in our text. We then go through the most common market failures that

[1] An inspiring description of the power of markets can be found in the short video available at `https://www.youtube.com/watch?v=67tHtpac5ws`

arise in macroeconomic analysis and indicate how each of these failures involve departures from the assumptions underlying the welfare theorem. We end the chapter with a short discussion of the role of government policy.

## 6.1 The First Welfare Theorem

The First Welfare Theorem says that a competitive equilibrium is Pareto optimal. To make this statement more precise and to give a sketch of the proof we will consider an abstract economy with many different goods. These goods could be different commodities or goods at different dates. Later we will describe how different specific economic environments relate to this abstract economy.

There is a set $\mathcal{I}$ of different consumers, each indexed by $i \in \mathcal{I}$, and a set $\mathcal{J}$ of firms, each indexed by $j \in \mathcal{J}$. We think of $\mathcal{I}$ and $\mathcal{J}$ as finite but we will consider them to be infinite in some extensions that we discuss below. Let $x$ be a vector of consumption levels of different goods that are traded. The length of the vector thus specifies how many markets there are in the economy. For now, we will think of the vector length as finite, but it will be relevant to consider the possibility of infinite vectors later. Moreover, the vector is, at least for now, allowed to have negative elements. Similarly, $y$ is a vector of production levels of the same goods and $\omega$ is a vector of exogenous endowment levels of the goods. Consumer $i$ is endowed with $\omega_i$ and consumes $x_i$ while firm $j$ produces $y_j$. Thus, our resource constraint reads

$$\sum_{i \in \mathcal{I}} x_i \leq \sum_{j \in \mathcal{J}} y_j + \sum_{i \in \mathcal{I}} \omega_i. \tag{6.1}$$

This inequality applies to each element of the vector so, for every good, the total amount consumed cannot exceed the endowment plus the amount produced by all the firms.

Let $X_i$ be a set of vectors that are feasible for agent $i$ to consume. For example, if there are two commodities, then $X_i = \mathbb{R}^2_+$ would rule out consuming a negative amount of either good. Firms have production possibility sets denoted by $Y_j$. We assume that consumers maximize utility taking their budget constraint, prices, and consumption possibility sets as given. Utility maximization means they choose a consumption bundle that is not preference-dominated (as defined by a preference ordering $\succeq_i$) by any other choice available to them. Similarly, firms maximize profits, taking prices as given, by choosing a production plan in their production possibility set $Y_j$. Firm profits are given to the consumers that own the firms with $\theta_{i,j}$ denoting consumer $i$'s share of firm $j$. Each firm is wholly owned by the consumers so $\sum_i \theta_{ij} = 1$. Finally, let $p$ be the vector of prices for each good.

**Key assumption** We will use the weakest assumption under which the First Welfare Theorem holds: local non-satiation (LNS). LNS expresses that each consumer can always be made better off by an infinitesimally higher consumption of some good.[2]

[2]This notion also presumes that the consumption possibility sets $X_i$ allow small movements in at least one desirable direction.

**A competitive equilibrium** A competitive equilibrium is a consumption allocation $\{x_i^\star\}_{\forall i}$, a production allocation $\{y_j^\star\}_{\forall j}$, and a price system $p^\star$ such that

1. for each $i \in \mathcal{I}$, the consumption choice $x_i^\star$ is in $X_i$ and there is no $x \in X_i$ such that $x \succ_i x_i^\star$ and $px \leq px_i^\star = p\omega_i + \sum_j \theta_{i,j} p y_j$;
2. for each $j \in \mathcal{J}$, $y_j^\star \in Y_j$ and there is no $y \in Y_j$ such that $py > py_j^\star$;
3. and the market for each good clears (equation (6.1) holds with equality).

**Theorem 6.1 (The First Welfare Theorem)** *An allocation that is part of a competitive equilibrium is Pareto optimal.*

**Proof.** The proof is by contradiction. So suppose there exists an allocation $\{\tilde{x}_i\}_{\forall i}$, $\{\tilde{y}_j\}_{\forall j}$ that is feasible (these values are all in their possibility sets and the resource constraint is satisfied) and that Pareto dominates the allocation of the given equilibrium. Then, by definition,

$$\forall i \in \mathcal{I} : \tilde{x}_i \succeq_i x_i^\star,$$
$$\exists \tilde{i} \in \mathcal{I} : \tilde{x}_{\tilde{i}} \succ_{\tilde{i}} x_{\tilde{i}}^\star.$$

Now the first property of a competitive equilibrium combined with LNS can be used to conclude that

$$p\tilde{x}_i \geq px_i^\star$$

holds for all $i$. If it were strictly cheaper to buy the new allocation, the consumer could have spent more to improve on the existing choice and hence that choice could not have been optimal for the consumer. Moreover, for $\tilde{i}$ it must be that

$$p\tilde{x}_{\tilde{i}} > px_{\tilde{i}}^\star,$$

again because otherwise the original allocation must not have involved consumer optimization.

Summing the budget constraints of all consumers we have

$$p\sum_{i\in\mathcal{I}} x_i^\star = p\omega + p\sum_{j\in\mathcal{J}} y_j^\star,$$

where $\omega \equiv \sum_i \omega_i$ and we have used the fact that $\sum_i \theta_{ij} = 1 \ \forall j$. By the arguments above, the alternative consumption allocation is more expensive

$$p\sum_{i\in\mathcal{I}} \tilde{x}_i > p\omega + p\sum_{j\in\mathcal{J}} y_j^\star.$$

Furthermore since $y_j^\star$ is profit maximizing for each $j$, it must be that $\sum_j p\tilde{y}_j \leq \sum_j py_j^\star$ so we have

$$p\sum_{i\in\mathcal{I}} \tilde{x}_i > p\omega + p\sum_{j\in\mathcal{J}} \tilde{y}_j.$$

In addition, the alternative allocation is resource feasible. Multiplying equation (6.1) by $p$ we obtain

$$p\sum_{i\in\mathcal{I}}\tilde{x}_i \leq p\omega + p\sum_{j\in\mathcal{J}}\tilde{y}_j.$$

These last two inequalities contradict one another. ■

**Mapping the setting into our macroeconomic models** The power of the abstract proof above is that it can be applied to a large number of contexts. Let us first consider a static macroeconomic model with one agent with endowments of capital and labor equal to $k$ and 1, and a neoclassical, constant-returns to scale production function that produces a consumption good. Preferences could be represented by a standard utility function, which satisfies LNS because it is strictly increasing. Then $x \in \mathbb{R}^3_+$ and a typical consumer choice would be $(c, 0, 0)$. Firms would have $(y, -k, -l) \in \mathbb{R}_+ \times \mathbb{R}^2_-$, the endowment vector would be $\omega = (0, k, 1)$, and the normalized price vector would be $(1, r, w)$.

The argument can easily be extended to include different types of goods. For example, if consumers value leisure we can simply treat them as "buying leisure" so that is just another good for them to consume. Other intermediate goods can be included too, as well as other resources, such as land. The theorem also applies to endowment economies by specifying the production possibility sets to only include the zero vector.

In a dynamic setting with $T < \infty$ time periods, goods and services at different dates are simply additional market goods and all the vectors become correspondingly longer. One can, in particular, define the vectors as simply containing $T$ times the number of elements in the static model. As we will see below, the argument also applies to infinite horizon models but with some additional caveats. In the next chapter we will discuss models with uncertainty where goods are indexed not just by time, but also by the state of the world. The First Welfare Theorem applies to those settings as well.

**Infinite horizon models** We start by considering a case with a finite number of infinitely-lived consumers. Later in the chapter we will discuss an overlapping-generations economy where time is infinite but each consumer has a finite life.

All the vectors in the proof above are infinite-dimensional but the competitive equilibrium is defined as before. A key point here is that for an equilibrium to exist—especially for it to satisfy its first property—it would have to be that the value of total expenditures of the consumer is finite, i.e., the dot product of the infinite price vector and the infinite quantity vector is finite. For an infinite sum, this would require, unless quantities go to zero, that prices fall sufficiently fast. Typically in our models, prices fall asymptotically at a geometric rate (equal to $\beta < 1$ if there is no growth), which for constant quantities imply a finite total value. As we shall see below, although this property holds in dynastic models, it does not hold in some overlapping-generations models. This means that overlapping-generation models can have very different welfare properties.

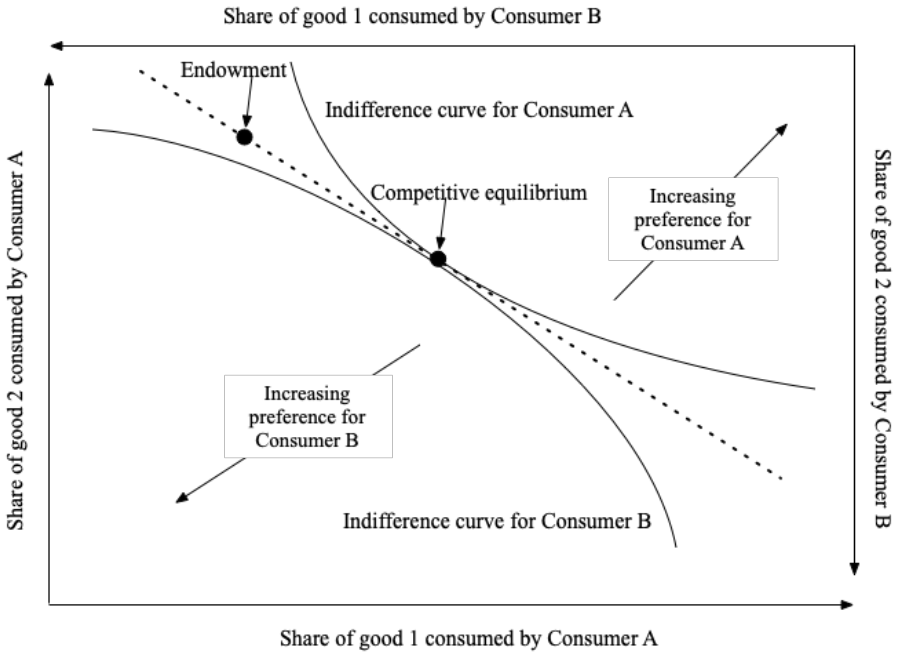
**Intuition** One issue underlying the First Welfare Theorem is the fact that all consumers and firms face the same prices. For all consumers the ratio of the marginal utilities of any two goods will equal the ratio of prices and, as the prices are the same, they all have the same ratio of marginal utilities. It is therefore impossible to make a marginal change in the consumption allocation that results in a Pareto improvement. This point can be illustrated with an Edgeworth box as shown in the top panel of Figure 6.1.

The figure depicts an endowment economy with two consumers and two goods. Their endowments are indicated by the dot. Given relative prices, we can draw a budget line as shown by the dashed line. Notice that each consumer will choose a consumption bundle that makes their indifference curve tangent to the budget line. When we impose market clearing, the points representing their consumption bundles in the Edgeworth box must coincide because what is not consumed by Consumer A must be consumed by Consumer B. Such a situation is marked as a competitive equilibrium in the figure. Notice that the indifference curves are tangent to the budget line and therefore also to each other. Because the indifference curves do not intersect, there is no way to move Consumer A to a higher indifference curve without harming Consumer B. Now consider the lower panel of the figure where we have imagined the two consumers face different prices. We will discuss several reasons this could occur, but a simple one is that Consumer A faces a tax on consuming one of the goods while Consumer B does not. If the consumers face different after-tax prices they will have different budget lines. As before each consumer chooses a consumption bundle where their indifference curve is tangent to their budget line. As the budget lines are different, the indifference curves now intersect and there are alternative allocations that yield higher welfare as shown by the shaded area in the figure.

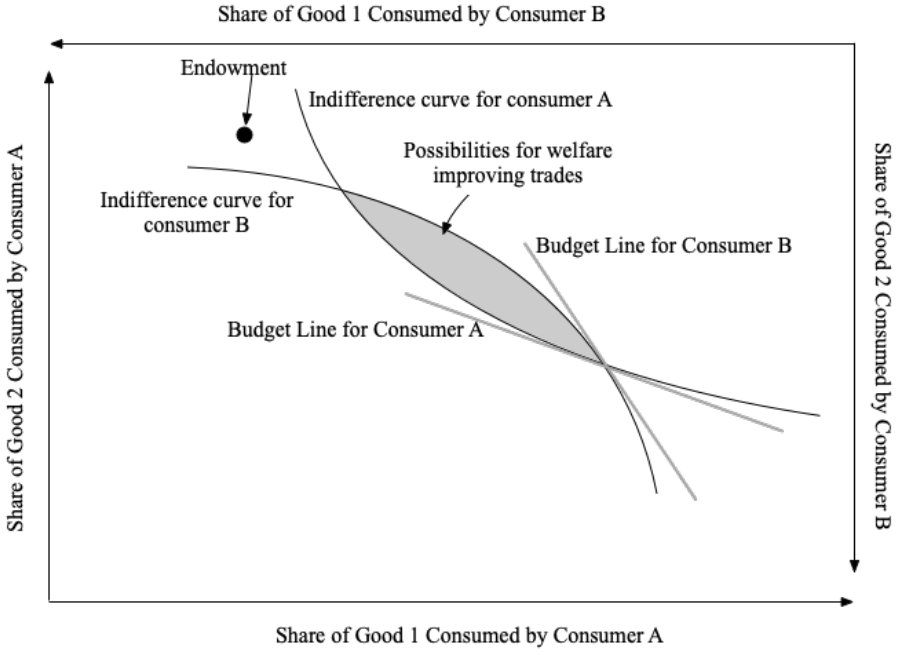
Similar logic applies to firm profit maximization. Profit maximization leads firms to set the marginal rate of transformation equal to the ratio of prices. As firms face the same prices as consumers, the marginal rate of transformation between any two goods will therefore be equal to the marginal rate of substitution between any two goods. A marginal change in the production allocation therefore cannot transform goods in a way that improves consumer welfare. This point can be visualized as point of tangency between the indifference curves of the consumers (remember they all have the same slope in equilibrium) and the production possibility frontier. A marginal change along the production possibility frontier does not increase consumer welfare.

These intuitive arguments are not as powerful as the more abstract proof above because they apply to marginal changes in allocations while the abstract proof is global. Nevertheless, they can be helpful in developing an intuitive understanding for when the First Welfare Theorem will hold.

How do these ideas apply to macroeconomic models? In the dynastic growth model with optimal saving, the Euler equation sets the marginal rate of substitution between goods at $t$ and $t+1$ equal to the relative price between them, i.e., the real interest rate. As all consumers face the same interest rate, they all have a common marginal rate of substitution between $t$ and $t+1$ goods. Moreover, as firms rent capital from households at a rental rate that is equal to the real interest rate, the marginal rate of transformation between goods at $t$ and




(a) Efficient equilibrium.




(b) Distorted market.

Figure 6.1: Edgeworth boxes depicting an endowment economy with two goods and two consumers. The top panel shows an efficient equilibrium in which both consumers face the same prices and have the same budget line. The lower panel shows a distorted market where the two consumers face different prices.

$t+1$ is equal to the households' marginal rates of substitution. Similarly, if we consider the model with elastic labor supply, the firm will produce at a point where the marginal product of labor is equal to the wage. The consumer will supply labor such that the marginal rate of substitution between consumption and leisure is equal to the wage. Eliminating the wage from these two optimality conditions gives us that the marginal rate of substitution between goods and leisure is equal to the marginal rate of transformation between goods and leisure.

## 6.2 Tracing out the Pareto frontier

One useful way to construct a Pareto-optimal allocation is to solve a social planner's problem in which the planner maximizes the weighted sum of the agents' utilities. A solution to such a problem must be Pareto optimal because if it were not, the Pareto dominating allocation would increase the planner's objective function. As we will explain, the solution to this planner's problem is closely related to the concept of competitive equilibrium.

For the sake of simplicity, assume a set of infinitely-lived consumers, indexed by $i \in \mathcal{I}$, live in an exchange economy. Each period, there is a single consumption good and each consumer receives an endowment $\omega_{i,t}$. Each consumer has preferences given by

$$U_i \equiv \sum_{t=0}^{\infty} \beta^t u(c_{i,t})$$

with $u'(c) > 0$ and $u''(c) < 0$. Let $p_t$ be the date-0 price of the date-$t$ good. A competitive equilibrium is a consumption allocation $\{c_{i,t}\}_{\forall i,t}$ and a price system $\{p_t\}_{\forall t}$ such that all consumers are maximizing utility taking prices as given and markets clear. Market clearing requires that $\sum_i c_{i,t} = \sum_i \omega_{i,t}$ for all $t$.

In the competitive equilibrium with date-0 trading, the consumers maximize their utility subject to the date-0 budget constraint. The Lagrangian of this problem is

$$\mathcal{L} = \sum_{t=0}^{\infty} \beta^t u(c_{i,t}) - \lambda_i \sum_{t=0}^{\infty} p_t (c_{i,t} - \omega_{i,t})$$

and the first-order condition for $c_{i,t}$ is

$$\beta^t u'(c_{i,t}) = \lambda_i p_t.$$

As the utility function is strictly concave, we can invert $u'(\cdot)$

$$c_{i,t} = (u')^{-1} \left(\beta^{-t} \lambda_i p_t\right), \tag{6.2}$$

which shows that knowledge of $\{\lambda_i\}_{\forall i}$ and $\{p_t\}_{\forall t}$ is enough to determine the entire consumption allocation. The Lagrange multiplier $\lambda_i$ captures the shadow value of date-0 wealth and will be decreasing in date-0 wealth.

Now consider a social planner that seeks to maximize a weighted sum of the utilities of the consumers. The planner's objective is

$$\sum_{i\in\mathcal{I}} \mu_i U_i,$$

where $\mu_i$ is the weight on consumer $i$'s utility. These weights are called Negishi weights or Pareto weights. The constraint on the planner is the resource constraint

$$\sum_{i\in\mathcal{I}} c_{i,t} \leq \sum_{i\in\mathcal{I}} \omega_{i,t},$$

which must hold at each date. The Lagrangian of this problem is

$$\mathcal{L} = \sum_{i\in\mathcal{I}} \mu_i \sum_{t=0}^{\infty} \beta^t u(c_{i,t}) - \sum_{t=0}^{\infty} \psi_t \sum_{i\in\mathcal{I}} (c_{i,t} - \omega_{i,t}),$$

where $\psi_t$ is the Lagrange multiplier on the resource constraint at date $t$. The first-order condition of this problem with respect to $c_{i,t}$ is

$$\mu_i \beta^t u'(c_{i,t}) = \psi_t.$$

Inverting $u'(\cdot)$ as above yields

$$c_{i,t} = (u')^{-1}\left(\beta^{-t}\frac{1}{\mu_i}\psi_t\right), \tag{6.3}$$

which shows that knowledge of $\{\mu_i\}_{\forall i}$ and $\{\psi_t\}_{\forall t}$ is enough to determine the entire consumption allocation.

There is a clear symmetry between equations (6.2) and (6.3). If $\mu_i = 1/\lambda_i$ and $\psi_t = p_t$, the two equations will give rise to the same consumption allocation. Let's assume that $\mu_i = 1/\lambda_i$. It must then be the case that $\psi_t = p_t$ because in the competitive equilibrium markets must clear and in the solution to the planner's problem the aggregate resource constraint must hold, which means in both cases the total consumption must equal the total endowment. By using equations (6.2) and (6.3) we can see that the only way we acheive the same total consumption at date $t$ is if $\psi_t = p_t$. It follows that if we have the right Negishi weights, the competitive equilibrium is optimal in the eyes of a planner that attaches more weight to those with higher date-0 wealth. This argument is closely related to the First Welfare Theorem: there are some Negishi weights that make the competitive equilibrium optimal for the planner, which implies the competitive equilibrium is Pareto optimal.

Now let's flip the argument in reverse. As we vary the Negishi weights, we will arrive at different consumption allocations as solutions to the planner's problem. Each one of these allocations is Pareto optimal because it solves a planner's problem—again, if it were not Pareto-optimal the planner would not choose it. By varying these weights, we can therefore map out many different Pareto-optimal allocations or in other words we can trace out the

Pareto frontier. Similarly, by varying date-0 wealth in the market economy we can generate different competitive equilibria with different sets of Lagrange multipliers $\lambda_i$. By choosing the appropriate distribution of date-0 wealth we can engineer a competitive equilibrium that mimics the solution to the planner's problem with particular Negishi weights and therefore gives rise to a particular Pareto-optimal consumption allocation. In this case, the consumers still trade with one another, but we redistribute resources between them to change the distribution of consumption.

This procedure of constructing a competitive equilibrium to deliver a particular Pareto-optimal consumption allocation is closely related to the Second Welfare Theorem. The Second Welfare Theorem begins with a Pareto-optimal allocation and then gives conditions under which there exists a competitive equilibrium delivering this allocation as an outcome. The conditions involve ensuring that consumers' and firms' maximization problems have well-defined solutions, which in turn in general necessitates assumptions of convexity (e.g., the consumer's utility function is globally concave). These additional assumptions are typically met in our macroeconomic applications so they are not a problem per se, but the statement of a general theorem is more cumbersome so we will not present it here.

Market equilibria, as observed in actual economies, have radically different consumption levels across agents, so viewed from the perspective of frictionless competitive equilibria and an additive social welfare function of the sort just described, they correspond to points on the Pareto frontier with high Negishi weights on the high-consumption agents. The social welfare function with Negishi weights is just an analytical tool that we can use to construct different Pareto-optimal allocations. To be clear, the argument we have made here is not saying that this distribution of consumption is desirable or just but simply that there are some Pareto weights that could make the planner choose that distribution of consumption.

## 6.3 Inefficient market outcomes

We will now very briefly, by means of simple examples, cover a number of commonly considered departures from the abstract frictionless economy considered above. For each case, we will comment on efficiency properties by describing how the proof of the First Welfare Theorem may or may not go through as well as explaining how the more intuitive marginal efficiency conditions may or may not hold. We begin with a case where the culprit is not the market but distortionary taxation and then look at externalities, missing markets, and monopoly power.

### 6.3.1 Taxes

First, let us consider lump-sum taxes (in a frictionless economy). Lump-sum taxes do not have to be equal across agents; the key is that the amount given to agent $i$ does not depend on the behavior of agent $i$. From the perspective of the marginal conditions characterizing optima, since the marginal conditions of competitive equilibria do not involve lump-sum taxes or transfers, equilibria remain optimal. From the perspective of our abstract proof,

all lump-sum taxes do is redistribute wealth (and thus "utils") across agents, i.e., move us along the Pareto frontier.[3]

Second, let us consider the two most commonly studied taxes in macroeconomic applications: taxes on labor earnings and taxes on capital income. Beginning with taxes on labor earnings, let us consider a tax that is proportional to earnings. Hence, instead of $w_t$ an agent receives $w_t(1-\tau_\ell)$ for each unit of hours worked. In an economy where consumers do not value leisure, this tax does not affect any first-order conditions; hence, it acts as a lump-sum tax and does not disturb the efficiency properties of equilibrium. However, if leisure is valued, the marginal rate of substitution between consumption and leisure will differ from the marginal rate of transformation between consumption and leisure by a factor $1-\tau_\ell$. Similarly, if capital income is taxed, the marginal rate of substitution between goods at $t$ and $t+1$ will differ from the marginal rate of transformation due to the tax.

**What goes wrong?** First, intuitively the firms and the households face different (after-tax) prices and therefore they may not exploit all of the possible trades that they should. Now looking at our abstract proof of equilibrium efficiency, what goes wrong in trying to use it? The key is that the tax appears in the equations. Suppose that the $p$ in the proof is the pre-tax prices. What would not hold is the point where we say the alternative consumption allocation must cost more: $p\sum_{i\in\mathcal{I}}\tilde{x}_i > p\omega + p\sum_{j\in\mathcal{J}} y_j^\star$. This inequality may not hold because the consumer is maximizing utility with respect to the after-tax prices.[4] Hence the proof does not in general go through. However, the proof is valid when leisure is not valued, because then the leisure chosen, which is an element in $x$, is zero, so the fact that $p$ is different for this good does not matter. This confirms the intuition that a proportional tax on an inelastic labor supply is non-distortionary.

Similarly, a proportional tax on capital income, e.g., with $r_t$ replaced by $r_t(1-\tau_k)$, will appear in the Euler equation and therefore make the marginal rate of substitution between goods at $t-1$ and $t$ differ from the corresponding marginal rate of transformation. In the proof of the First Welfare Theorem, the relative price of the time $t$ good is higher for consumers than for producers and hence the proof cannot be completed.

### 6.3.2 Externalities

An externality arises when one agent's activity has payoff relevance to other agents yet is not rewarded by the market. Let us use an example with a negative TFP externality: production by one firm damages the production carried out in other firms. Consider a static economy with a representative consumer who values leisure. In particular, the consumer maximizes

[3] To see this formally, one can consider the government to be one of the consumers in the economy. When we sum across the budget constraints of the consumers in the proof of the First Welfare Theorem, the lump-sum taxes will cancel out as the government's revenue equals the other consumer's tax payments.

[4] We cannot just interpret $p$ as the after-tax price because then when we say firms are profit maximizing, they are not maximizing profits with respect to $p$.

$u(c, \ell) = \log c - B\ell^{1+1/\theta}/(1 + 1/\theta)$ subject to

$$c = rk + w\ell$$

by choice of $(c, \ell)$. Output of a typical firm $j$ equals

$$A(\bar{y})k_j^{\alpha}\ell_j^{1-\alpha},$$

with $\bar{y} = (\sum_j y_j)/\mu$, where $\mu$ is the number of firms. Here, $A$ is a decreasing function, expressing a negative production externality: the higher is total production in the economy, the lower is the productivity of each firm. Let us also assume that $\mu$ is large enough that each firm $j$ ignores its own impact on $\bar{y}$; for simplicity, think of the set of firms as a continuum on $[0, 1]$, with $\mu = 1$, so that there is a notion of a representative firm. Hence, we can drop the subscript $j$ and the representative firm solves

$$\max_{k,\ell} \; A(\bar{y})k^{\alpha}\ell^{1-\alpha} - rk - w\ell,$$

thus taking $\bar{y}$ as given. In this static economy, output is determined by equilibrium in the labor market. From the consumer, we obtain

$$\frac{w}{rk + w\ell} = B\ell^{\frac{1}{\theta}}$$

and the firm's first-order conditions imply that

$$\frac{r}{w} = \frac{\alpha}{1-\alpha}\frac{\ell}{k}.$$

Combining the two equations, we obtain (with a small amount of algebra) that the equilibrium outcome for hours worked is given by the unique solution to

$$1 - \alpha = B\ell^{1+\frac{1}{\theta}}.$$

A striking feature of this solution is that it does not depend on the strength of the externality as captured by the function $A$: $A$ does not appear in the equation.[5]

The efficient allocation, on the other hand, is given by the solution to

$$\max_{\ell,y} \; \log y - B\frac{\ell^{1+\frac{1}{\theta}}}{1+\frac{1}{\theta}} \quad \text{s.t.} \quad y = A(y)k^{\alpha}\ell^{1-\alpha}.$$

Here, the first-order conditions (with $\lambda$ denoting the multiplier on the constraint) are

$$B\ell^{\frac{1}{\theta}} = \lambda(1-\alpha)A(y)k^{\alpha}\ell^{-\alpha} \quad \text{and} \quad \frac{1}{y} = \lambda\left(1 - A'(y)k^{\alpha}\ell^{1-\alpha}\right),$$

[5]This particular feature, which follows because income and substitution effects cancel in the labor-supply specification, makes the example stark but is not necessary for the arguments here.

which delivers

$$\frac{1-\alpha}{1-A'(y)k^{\alpha}\ell^{1-\alpha}} = B\ell^{1+\frac{1}{\theta}} \quad \text{and} \quad y = A(y)k^{\alpha}\ell^{1-\alpha}$$

as two equations in the two unknowns $\ell$ and $y$. The first of these equations can be compared to the equilibrium outcome: the equilibrium is not optimal, unless $A'(y) = 0$, i.e., unless there is no externality. Intuitively, in their input choices, firms do not take into account how their production hurts others, and they should.

**What goes wrong?** To see how our abstract proof of the First Welfare Theorem would go wrong in this case, first note that an equilibrium with externalities would be defined by letting the $Y_j$s—each firm's production possibility set—be endogenous and interdependent: $Y_j = Y_j((y_{j'})_{j'\neq j})$, where $(y_{j'})_{j'\neq j}$ is the vector of choices of other firms.[6] A key step in our abstract proof was that for each $j$, $p\tilde{y}_j \leq py_j^{\star}$, from profit maximization. This no longer follows, since the alternative allocation implies a different choice set for firm $j$, as given by $Y_j = Y_j((y_{j'})_{j'\neq j})$. In concrete terms, if other firms scale down their production relative to that in equilibrium, your choice set improves—your TFP increases—and your original choice was not optimal.[7]

The negative externality considered here is closely related to how the externalities due to climate change are usually modeled. There, the externality is usually assumed to affect TFP, as in our example here. Rather than occurring through overall production, however, the externality occurs through carbon emissions, which result from production of a specific good—energy derived from fossil fuels. Climate change is covered in Chapter 23.

### 6.3.3 Missing markets: an example with constraints on borrowing

A friction that is commonly studied in macroeconomic applications involves "missing markets": restrictions on trade in one way or another. For example, there is no market that allows workers to buy insurance against unfavorable changes in their salaries because of the moral hazard that workers will have incentives to collect the insurance payment rather than work hard. Another example is that households and firms may not be able to borrow much as they would like. Borrowing constraints are thought to be important constraints both for firms (for funding their day-to-day operations as well as long-term investments) and consumers (for purchasing homes and durable goods more generally). These borrowing constraints ultimately stem from features of the economic environment, such as the difficulty of enforcing repayment, but are sometimes modeled as simple constraints. We focus on the case of borrowing constraints here, but the logic applies more generally to settings where certain goods or agreements for borrowing or insurance are not traded.

Let's consider a two-agent dynastic endowment economy: a two-type special case of that presented in Section 5.3.1. The total endowment each period is constant at $\omega$, but let us

[6] For simplicity here we abstract from your own production lowering your own TFP.

[7] In the present example, equilibrium profits are zero, and if all other firms decrease their output below the equilibrium level, your profits can be made positive (and unboundedly large) by simply scaling up your input choices.

now assume that agent 1 is endowed with $2\omega/3$ in odd periods and $\omega/3$ in even periods; agent 2 thus has $\omega/3$ and $2\omega/3$ in odd and even periods, respectively (the odd-numbered agent is rich in odd periods). Let us also assume that both agents have logarithmic utility. The economy starts in period 0. In a competitive equilibrium with unrestricted markets we have (i) full consumption smoothing, so that $c_{1,t} = c_1$ for all $t$ and $c_{2,t} = c_2$ for all $t$; prices for goods at different dates satisfy $p_t = \beta^t$, i.e., the gross real interest rate is $1/\beta$; and (iii) $c_1 = \left(\frac{1}{3} + \frac{2\beta}{3}\right)\frac{\omega}{1+\beta}$ and $c_2 = \left(\frac{2}{3} + \frac{\beta}{3}\right)\frac{\omega}{1+\beta}$. Consumer 2 is richer, and hence consumes more, due to an endowment stream that is higher in present value because consumer 2 receives the larger endowment one period before consumer 1 does.

Note that the unrestricted competitive equilibrium has active borrowing and lending: agent 1 borrows in even periods and repays in odd periods. Suppose, then, that borrowing is simply not allowed. In terms of a sequence of budget constraints $c_t + q_t a_{t+1} = \omega_t + a_t$, where $a$ is asset holdings and $q$ is price of a one-period real bond delivering 1 unit of consumption next period, no borrowing means that the consumer is facing an additional constraint: $a_{t+1} \geq 0$ for all $t$. The consumer's maximization problem then reads

$$\max_{\{a_{t+1}\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t \log\left(\omega_t + a_t - q_t a_{t+1}\right) \quad \text{and} \quad a_{t+1} \geq 0 \quad \forall t.$$

Differentiation with respect to $a_{t+1}$ delivers

$$\frac{q_t}{c_t} = \beta \frac{1}{c_{t+1}} + \mu_t,$$

where $\mu_t$ is the Kuhn-Tucker multiplier on the borrowing constraint. This multiplier is non-negative: positive when the constraint binds and zero otherwise. Hence, we conclude that under a borrowing constraint, the Euler equation generally is an inequality constraint: $q_t u'(c_t) \geq \beta u'(c_{t+1})$. Intuitively, the marginal value of consumption today can be strictly higher than the consumer could obtain by consumption tomorrow, using market prices, but the consumer cannot increase consumption today further due to the borrowing constraint.

In our simple example, no borrowing means autarky: consumer 1 always consumes $2\omega/3$ in odd periods and $\omega/3$ in even periods, with the remainder of the total endowment consumed by agent 2.[8] The allocation is clearly not optimal, since consumption is not smoothed. What is the prediction for interest rates? In even periods, agent 2 is not constrained so we know that $\frac{q_t}{2\omega/3} = \beta\frac{1}{\omega/3}$ when $t$ is even. Hence, $q_t = 2\beta$: the real interest rate $1/q_t - 1$ is lower than under unrestricted borrowing, since the lender needs to be discouraged from lending.[9] The same logic applies in odd periods: $q_t = 2\beta$. We even see that the real interest rate will be negative provided that $\beta > 0.5$.

[8] As agent 1 is unable to borrow in period 0, agent 2 has nobody to lend to and cannot save. In period 1, the two agents face the same situation as period 0 with their roles reversed.

[9] Values of $q_t$ above $2\beta$ will also constitute equilibria, since all agents are then strictly constrained. However, we view the case with an interior solution as the interesting one, since it is the limiting equilibrium for economies with borrowing constraints $a_{t+1} \geq \underline{a}$ where $\underline{a} < 0$, with $\underline{a} \to 0$.

**What goes wrong?** Clearly, marginal rates of substitution are not equalized across agents here. From the perspective of our abstract proof, what goes wrong? As in the case with externalities, the sets from which agents are choosing will change: they will contain more restrictions and they will be endogenous. For example, without borrowing constraints, the main restriction on the consumption possibility sets is non-negativity. With borrowing constraints, it will include an additional constraint for each good: $c_0 \leq a_0 + \omega_0$, $q_0 c_1 \leq \omega_0 + a_0 - c_0 - q_0 \omega_1$, and so on. The equilibrium allocation can be dominated by an alternative allocation—say, one with full smoothing—simply because this alternative allocation is not within the agent's consumption possibility set.

Moreover, as we see here, the consumption possibility sets will now also depend on prices, thus departing conceptually from the abstract setup where the consumption possibility sets are exogenous. As in the externality example, there is an interdependence between the possibility sets the agents face. And just like the externality case, the agents will not take into account how their actions affect the possibility sets that other face. These are *pecuniary externalities*—one agent's choices affect the prices that appear in other agents' constraints. As a result of these pecuniary externalities, if a planner were to select the consumption allocation subject to the budget constraints and borrowing constraints, the planner may choose a different allocation than the competitive equilibrium because the planner would take account of the pecuniary externalities.

### 6.3.4 Lack of commitment

In the dynamic equilibrium models we looked at above, we always assumed that all agents engaged in intertemporal trade could commit to delivering on their promises. When such commitment is not available, the allocations are of course affected. One could, for example, imagine that it is possible to default on debt and that such defaults are not punished in any way. This would, given that lenders are rational, lead to de-facto borrowing constraints, as in Section 6.3.3 above. One could also imagine that default can occur but that it is punished, in which case intertemporal trade generally will occur. Note, however, that the punishment mechanism per se needs to be committed to and that it is not clear that punishment will be rational ex post.[10] When punishments are assumed in economic models, there is therefore a presumption that some agent (such as the government) has an ability to commit to it.

Macroeconomic models where consumers cannot fully commit, including where they do default in equilibrium, exist and are interesting cases of "endogenously incomplete markets," but they are not discussed in this text. A more common example involves government policy, such as the case where governments cannot fully commit to future tax policy; this case is discussed in Chapter 13 below. In the context of international economics, another important example is the case where default on sovereign debt (involving lending between countries) can occur; this is a central example in Chapter 22 on emerging markets.

[10] "Come home before midnight or I will kill you" may sound like a powerful and purposeful parental threat to a teenage child, but it is not credible.

### 6.3.5 Market power

A well-known case, and also one of great relevance for macroeconomics, is where some agents in the economy have market power. That is, they can, through their own behavior, affect the prices they are transacting at. This may occur on the individual level—you may be able to bargain for a higher salary—or at the firm level—the firm can set the price of its product, or try to bargain with their input suppliers over those prices. We will discuss examples of both of these occurrences in the book. When we study labor-market frictions, it will be natural to discuss bargaining. When we study growth and business cycles, we will study markups, i.e., how firms with market power set prices above marginal cost.

In macroeconomic models with market power, we will almost always assume that this power is quite limited: it is limited to an individual transaction and not affecting aggregates. Thus, for monopoly settings we most often study monopolistic competition, where each firm holds a monopoly over a specific good but faces competition with a continuum of imperfect substitutes, with each one playing a negligible role in the aggregate. Similarly, a worker will be assumed not to have any influence on aggregates in the wage negotiations. Clearly, there are examples where these assumptions are not appropriate (Apple can likely affect the entire market for smartphones, and perhaps even world GDP, in their pricing; particularly gifted vaccine researchers could be seen to have similar powers). But they are likely rare. So let us now briefly and compactly describe the basic, static model of monopolistic competition building on Dixit and Stiglitz (1977).

#### A model with monopolistic competition

Let us assume a representative consumer with preferences defined over a continuum of imperfectly substitutable goods and labor effort, $L$:

$$U\left(\left(c(i)\right)_{i=0}^{1}, L\right) = u\left(\left(\int_0^1 c\left(i\right)^{1-\frac{1}{\varepsilon}} di\right)^{\frac{\varepsilon}{\varepsilon-1}}\right) - v(L) \equiv u(C) - v(L).$$

This function, as we will see, implies a constant elasticity of substitution $\varepsilon \geq 0$ across different consumption goods. We will specialize $u$ to be logarithmic below in an example.

The consumer's budget is

$$\int_0^1 p\left(i\right) c\left(i\right) di = y,$$

where $y$ is income. We will now derive a price index by considering how the consumer should allocate goods in the cheapest way in order to reach a given level of $C$ . So consider

$$\min_{\left(c(i)\right)_{i=0}^{1}} \int_0^1 p\left(i\right) c\left(i\right) di \quad \text{s.t.} \quad \left(\int_0^1 c\left(i\right)^{1-\frac{1}{\varepsilon}} di\right)^{\frac{\varepsilon}{\varepsilon-1}} \geq C.$$

We obtain, with $\lambda$ denoting the multiplier for the constraint,

$$\begin{aligned} p(i) &= \lambda \frac{\varepsilon}{\varepsilon - 1} \left( \int_0^1 c(i)^{1-\frac{1}{\varepsilon}} di \right)^{\frac{\varepsilon}{\varepsilon-1}-1} \left(1 - \frac{1}{\varepsilon}\right) c(i)^{-\frac{1}{\varepsilon}} \\ &= \lambda \left( \int_0^1 c(i)^{1-\frac{1}{\varepsilon}} di \right)^{\frac{\varepsilon}{\varepsilon-1}-1} c(i)^{-\frac{1}{\varepsilon}} = \lambda \left( \frac{c(i)}{C} \right)^{-\frac{1}{\varepsilon}}. \end{aligned} \tag{6.4}$$

Multiply by $c(i)$ on both sides, sum across goods, and simplify to obtain

$$\int_0^1 p(i)\, c(i)\, di = \lambda C.$$

That is, $\lambda$ can be interpreted as a "unit price of $C$," the whole basket. What is $\lambda$ in terms of primitives? Often the multiplier is merely used to set up and solve a maximization problem but sometimes it carries important content, such as here, and then it is relevant to compute its value. To derive a formula for the unit price, use the expression for $p(i)$ again: raise it to $1 - \varepsilon$ and sum across $i$. This delivers

$$P \equiv \lambda = \left( \int_0^1 p(i)^{1-\varepsilon} di \right)^{\frac{1}{1-\varepsilon}}.$$

So $P$, the unit price, is itself a symmetric function that is increasing in all prices and homogeneous of degree one.

A key other implication of equation (6.4) is that it expresses an *inverse demand* function: it expresses a relation between the price of good $i$ and the demand for it, given an overall level of spending $PC$. The demand function itself becomes, after solving for $c(i)$ and using $\lambda = P$,

$$c(i) = C \left( \frac{p(i)}{P} \right)^{-\epsilon}.$$

Here we see that the price elasticity of demand is constant and equal to $\epsilon$. (One can replace $C$ by $y/P$, where $y$ is the consumer's income.)

The demand function is used as a central object in the definition of a monopolistically competitive equilibrium. We will now state it. We assume that one firm produces each kind of consumption good and that all production functions are identical and linear in labor: $c(i) = A\ell(i)$ for all $i$. We assume that labor supply is exogenous and equal to 1.

**A monopolistically competitive equilibrium** For the economy described above, a monopolistically competitive equilibrium is a consumption allocation $\{c(i)^\star\}_{\forall i}$, a labor allocation $\{\ell_i^\star\}_{\forall i}$, a price vector $\{p_i^\star\}_{\forall i}$, a profit vector $\{\pi_i^\star\}_{\forall i}$, and a wage $w^\star$ such that

1. $(\{c(i)^\star\}_{\forall i}, L^\star)$, where $L^\star \equiv \int_0^1 \ell^\star(i) di$, solves

$$\max_{(c(i))_{i=0}^1, L} u\left( \left( \int_0^1 c(i)^{1-\frac{1}{\varepsilon}} di \right)^{\frac{\varepsilon}{\varepsilon-1}} \right) - v(L)$$

subject to $\int_0^1 p^\star(i)c(i)di = w^\star L + \int_0^1 \pi^\star(i)di$.

2. for each $i$, $(p^\star(i), c(i)^\star, \ell(i)^\star)$ solves the maximization problem

$$\max_{p,c,\ell} \; pc - w^\star \ell \quad \text{subject to} \quad c = A\ell \quad \text{and} \quad p = P^\star \left(\frac{c}{C^\star}\right)^{-\frac{1}{\epsilon}},$$

where

$$C^\star = \left(\int_0^1 c^\star(i)^{1-\frac{1}{\varepsilon}} di\right)^{\frac{\varepsilon}{\varepsilon-1}} \text{ and } P^\star = \left(\int_0^1 p^\star(i)^{1-\varepsilon} di\right)^{\frac{1}{1-\varepsilon}}$$

and $\pi^\star(i)$ defines the maximum obtained.

Notice (i) that no market-clearing condition is needed as it is satisfied immediately as part of the firms' problems; (ii) that firms make profits, which accrue to the consumer; (iii) that firm $i$ solves a problem that does not depend on $i$, since all goods are symmetric here (still, we label the solutions with $i$); and (iv) that a key input into the firm's problem is the inverse demand function, $p^\star(c)$. We stated this last condition with knowledge of form for the inverse demand function for each good; this function, of course, has to be consistent with condition 1 of the definition and this was ensured in our derivations leading up to the definition. A monopolistically competitive equilibrium can thus alternatively be defined to include this demand function explicitly as an equilibrium object, with the added condition that it is consistent with consumer maximization.

To see what implications follow from this setup, let us solve the firm's problem. Substitute the constraints into the objective, so that it reads

$$\max_c \; PC^{\frac{1}{\varepsilon}} c^{1-\frac{1}{\varepsilon}} - \frac{w}{A} c,$$

where we have dropped the $i$ due to symmetry and $^\star$s for notational convenience. Clearly, this is a well-defined problem if $\varepsilon > 1$; if $\varepsilon \leq 1$, goods are not sufficiently substitutable, and the monopolist's problem does not have a solution.[11]

$$c = C \cdot \left(\frac{w}{A} \frac{1}{\left(1-\frac{1}{\varepsilon}\right) P}\right)^{-\varepsilon} \quad \text{and} \quad p = \frac{\varepsilon}{\varepsilon - 1} \frac{w}{A};$$

thus, $\mu \equiv \varepsilon/(\varepsilon - 1) > 1$ expresses that the firm charges a *markup* over marginal cost, $w/A$, that is constant in percent. Profits $\pi$ are thus given by $(\mu - 1)(w/A)c$.

The equilibrium is symmetric, so $c = C$. We can normalize one price, or a combination of prices, so we set $P = 1$. We then see that the equilibrium wage has to equal $w = A(1 - 1/\varepsilon)$:

[11] When $\varepsilon < 1$, infinite profits can be obtained by raising prices toward infinity; the reader is invited to show this by considering the implied maximization problem. The case $\varepsilon = 1$ is special: since revenues are then independent of the price, higher prices, and consequently lower quantities sold and therefore production costs, are always better and no profit maximum exists. Its supremum equals the revenue.

the wage is below marginal productivity (as $\varepsilon > 1$). Equilibrium work effort and consumption are then solved from $u'(C)w = v'(L)$. With a logarithmic $u$ we obtain

$$A(1 - 1/\varepsilon) = Cv'(C/A),$$

which has a unique solution for $C$ assuming $v(L)$ is convex and, hence, the right-hand side is strictly increasing in $C$.

It is easy to see that the outcome is inefficient. The planner will choose symmetry across goods and hence simply maximizes $u(C) - v(C/A)$. The outcome is the first-order condition

$$A = Cv'(C/A),$$

whose left-hand side is larger than in the monopolistic case and, hence, output and hours worked are too low in equilibrium.

What goes wrong with market power? Firms choose inputs taking into account how their demand is affected. As a result, firms tend to under-produce relative to the optimum, since that gives them a higher price. Thus, in equilibrium, firms transform labor into consumption goods at a marginal rate that is higher than the rate at which consumers value leisure relative to the consumption good. Our abstract proof of the First Welfare Theorem, moreover, cannot be used directly since an equilibrium with monopoly power is defined quite differently, as we have seen: some agents do not take $p$ as given.

### 6.3.6 Quantifying welfare losses

In applied macroeconomic analysis, the aim is most often quantitative. The researcher wants to go beyond qualitative statements like "the equilibrium is not optimal" to a quantitative one indicating *how much* worse, or better, one allocation is than another. There are numerous ways to do this and we will focus on the most common one here. We will use distortionary taxes in the context of a representative-agent economy as an example.

So suppose the government has an amount of expenditures—e.g., military purchases—that it needs to finance with taxes. One option, at least in theory, is to use lump-sum taxes. Another one is to tax labor earnings at a proportional rate. How much worse is it to use distortionary taxes? Let us focus on a static model because it is simple; the principles are the same in a dynamic context.

The procedure is simple. First compute an equilibrium for each of the two tax systems. Denote the resulting allocations of consumption and hours worked $(c_l, h_l)$ and $(c_d, h_d)$ ($l$ for lump-sum and $d$ for distortionary). Clearly, $u(c_l, h_l) > u(c_d, h_d)$. Now it is always possible, with standard utility functions, to find a $\Delta > 0$ such that $u(c_l(1-\Delta), h_l) = u(c_d, h_d)$. The $\Delta$ expresses how much, in percent, consumption needs to be decreased in the better allocation to generate the same utility as in the worse allocation, while maintaining the same hours choice.

The key in the example is that $\Delta$ has a real interpretation; while one could compute the difference in consumer "utils" between two allocations, such a measure would not have an interpretation as "utils" have no meaning per se. There are, of course, alternative ways

to define a $\Delta$. One could, for example, imagine both reducing consumption and raising hours worked at the same time. One could also define a $\Delta$ as the percentage increase in consumption in the worse allocation that would make it as good as the better allocation. This would deliver a different $\Delta$. One can, finally, define $\Delta$ as an amount of wealth that the consumer would need to be given as a lump-sum transfer in order to be indifferent between two allocations.

In dynamic models, the $\Delta$ is defined as a percentage change in consumption in all time periods. Consumption need not be constant over time in either of the allocations considered, but a unique $\Delta$ can still be defined: it is the percentage decrease in consumption applied in every period that would make the consumer indifferent with the worse allocation.

## 6.4 Overlapping generations

The efficiency properties in overlapping-generations (OG) models require a separate discussion; they have also been thoroughly studied in the literature. We will emphasize the key results and insights here only. A short summary of the key result is that, first, although there are no frictions—no markets are missing and distortionary taxes, monopoly power, and externalities are absent—equilibria can be, but are not necessarily, Pareto-inefficient. Second, there is a simple litmus test that will tell us whether an equilibrium is efficient or not; we will provide it (but not prove it).

Before proceeding, let us note that there are models that share properties with overlapping generations models, such as those where people die randomly and new people appear. Here, equilibria are not necessarily efficient either.

### 6.4.1 The endowment case

Let us consider a simple example for illustration: there is a representative consumer in each cohort who lives for two periods. Let us also use a concrete, very simple utility function: for every generation $t \geq 0$ preferences are represented by

$$u_t(c_y,\, c_o) = \log c_y + \log c_o$$

where we evaluate at two arbitrary consumption levels $(c_y, c_o)$. The preferences of generation $t = -1$, who are old as time begins, are similarly represented by $u_{-1}(c) = \log c$.

Let us consider stationary endowment sequences given by:

$$\begin{aligned} \omega_{y,t} &= \omega_y \\ \omega_{o,t} &= \omega_o. \end{aligned}$$

for all $t$, where $t$ represents the time period; thus, $\omega_{o,t}$ is the endowment of the old (who were born at $t-1$) at time $t$.

In the competitive equilibrium we consider, trading is sequential and there are no borrowing constraints. With $q_t$ being the price of a bond at $t$, the agent born at $t \geq 0$ solves

$$\max_{c_y,\, c_o} \log c_y + \log c_o$$

s.t.

$$c_y + q_t c_o = \omega_y + q_t \omega_o.$$

It is straightforward to solve this maximization problem. It delivers

$$c_{y,t} = \frac{1}{2}\left(\omega_y + q_t \omega_o\right) \tag{6.5}$$

$$c_{o,t+1} = \frac{1}{2}\left(\omega_y / q_t + \omega_o\right) \tag{6.6}$$

where we have now given the consumption choices sub-indexes for the time period in which they occur. Note that the consumer's saving when young is $\omega_y - c_{y,t} = \left(\omega_y - q_t \omega_o\right)/2$.

The old agent at time zero maximizes utility subject to the budget $c_{o,0} = \omega_o$ and hence the choice is trivially given by the budget.

Market clearing in this overlapping-generations economy for period $t = 0$ reads

$$c_{o,0} + c_{y,0} = \omega_y + \omega_o.$$

Since the old's choice is given, we conclude that $c_{y,0} = \omega_y$: the young also consumes the endowment. The bond price that clears the market is $q_0 = \omega_y / \omega_o$. It follows that $c_{o,1} = \omega_o$: the old at 1 will consume the endowment in the second period of life as well.

The argument is then repeated and we obtain, period by period, that

$$\begin{aligned} c_{y,t} &= \omega_y \\ c_{o,t} &= \omega_o \\ q_t &= \frac{\omega_y}{\omega_o}. \end{aligned}$$

This constant sequence supports the equilibrium where agents do not trade: the prices induce people to consume their initial endowments.

Let us now plug in specific numbers. Let $\omega_y = 3$ and $\omega_o = 1$. It follows that $q_t = 3$. Thus, the gross real interest rate is $1/3$; the net interest rate is thus - 67%.

Is this allocation Pareto-efficient? Consider the following alternative feasible allocation: for all $t$,

$$\begin{aligned} \tilde{c}_{y,t} &= 2, \\ \tilde{c}_{o,t} &= 2. \end{aligned}$$

That is, the alternative allocation $\tilde{c}$ is obtained from a chain of intergenerational goods transfers that consists of the young in every period giving a unit of their endowment to the old in that period. Notice that for all generations $t \geq 0$, this is just a modification of the timing in their consumption, since total goods consumed throughout their lifetime remain at 4. For the initial old, this is an increase from 1 to 2 units of consumption when old. It is clear, then, that the initial old strictly prefer the new allocation. We need to check what the remaining generations think about the change. It is clear that since utility is concave (the

log function is concave), this even split of the same total amount will yield a higher utility value: $\log 2 + \log 2 = 2 \cdot \log 2 = \log 4 > \log 3 + \log 1 = \log 3$.

We conclude that the competitive equilibrium, which we solved for uniquely, is not Pareto optimal: it is dominated by an allocation where each cohort gives a transfer when young and receives one when old. But why is the equilibrium not Pareto optimal? There is no friction: no agent is prevented from trading, there are no externalities or elements of market power. We will return to this question shortly, but first let us consider the reverse case: $\omega_y = 1$ and $\omega_o = 3$. Now, $q_t$ becomes $1/3$ each period; the net real interest rate is $+67\%$. Again, let us consider the alternative $(2, 2)$ allocation: is it a Pareto improvement? For all generations born at $t \geq 0$, the answer is yes, as in the previous example. However, the old at 0 will be made worse off. Hence, the proposed alternative is not a Pareto improvement. Is there some other, smarter alternative allocation that does the job? The answer is, in fact, no.

In the overlapping generations model, equilibria are sometimes Pareto-optimal and sometimes not. We will elaborate on the intuition later, but let us instead go back and revisit our abstract proof of the First Welfare Theorem in Section 6.1. The notation there can, as in our application, include infinite sequences, and the proof goes through in all parts, except possibly in one: when summing the budget over all agents, $\mathcal{I}$ is now infinite. Prices should now be viewed as given by the sequence $p = \{p_0, p_1, p_2, \dots\}$ where $p_t = \prod_{\tau=0}^{t-1} q_\tau$. In our given equilibrium where $q_t = 3$ for all $t$, then, the present value of the equilibrium allocation is $\omega_o + (\omega_y + \omega_o) + 3\,(\omega_y + \omega_o) + 3^2\,(\omega_y + \omega_o) + \dots$ and this sum is infinite. Hence, this proof strategy cannot be used. However, in the equilibrium where $q_t = 1/3$, the present value is finite, and the proof does go through. Thus, we in fact have a proof that there is no allocation that can Pareto-improve on the autarkic allocation $(c_{y,t}, c_{o,t}) = (1, 3)$ for all $t$!

When can equilibria where the present-value budget sum across all agents is infinite be Pareto-improved upon? We cannot rely on the standard proof of the First Welfare Theorem, but it turns out that there is a general theorem—one provided in Balasko and Shell (1980)—that gives us the answer. The theorem relies on some assumptions, but these assumptions are rather weak: aside from regularity conditions and a bounded sequence for total endowments, they mainly restrict the curvature of consumers' indifference curves away from the two extreme cases (linear and kinked). The Balasko-Shell result is: A competitive equilibrium in an endowment economy populated by overlapping generations of agents is Pareto optimal if and only if

$$\sum_{t=0}^{\infty} \frac{1}{p_t} = \infty.$$

The proof is quite involved and we refer the reader to the source. The two special cases we have looked at are consistent with the theorem: $q_t = 3$ for all $t$ implies that $\sum_{t=0}^{\infty} \frac{1}{p_t} = \sum_{t=0}^{\infty} 3^{-t} = 3/2$ is finite, and hence the equilibrium is not optimal; $q_t = 1/3$ for all $t$ implies that $\sum_{t=0}^{\infty} \frac{1}{p_t} = \sum_{t=0}^{\infty} 3^t$ is infinite, and hence the equilibrium is optimal. Now, however, we can also evaluate the middle case where $(\omega_y, \omega_o) = (2, 2)$. Here, since $q_t = 1$ implies an infinite sum,

our abstract proof cannot be used—as in the $(\omega_y, \omega_o) = (3, 1)$ case—but now the theorem tells us that equilibrium is optimal (a infinite sum of 1s is infinite).

The Balasko-Shell theorem can be applied also to non-constant sequences; indeed, it applies also for non-constant endowment sequences. An important observation, however, is that whether $\sum_{t=0}^{\infty} \frac{1}{p_t}$ is finite or not does not depend on anything but how $p_t$ behaves as $t$ literally goes to infinity. Thus, whenever the gross real interest rate $p_t/p_{t+1}$ converges, we know that the equilibrium is optimal if and only if the limit net real interest rate is equal to or above 0.[12] Relatedly, if the economy has a finite time horizon, no inefficiency can occur, no matter how the real interest rate evolves: the present value of the sum of all budgets is finite in this case, and the standard proof can be used. It is thus the combination of (i) infinite time and (ii) a corresponding infinite set of cohorts of consumers—each of which has a finite budget—that can make markets fail.

Third and finally, whenever the equilibrium is sub-optimal, a straightforward government policy of transferring resources from the young to the old will allow all generations to be made better off. This is an argument for the introduction of social security as a pure government-mediated transfer scheme: a "pay-as-you-go" system. The government works like a bank here: you give it money when young and get it back when old. What allows the government to achieve a better allocation than markets can deliver? On the one hand, the inefficiency of the overlapping generations model is a pure market failure, and in that sense a government can improve on it. But it does require an ability of the government to implement a sequence of transfers that stretches into eternity. If, in our simple $(\omega_y, \omega_o) = (3, 1)$ case transfers stopped at some point in time, the young in the very last period of transfers will be worse off by having given something away, with nothing in return.

Let us now provide some intuition for the Balasko-Shell result, and let us continue with our example and focus on the "toughest" case: $\omega_y = \omega_o = 2$, where a First Welfare Theorem cannot be proved the standard way and yet applies. First, we restrict attention to *stationary* allocations, i.e., allocations such that $c_{y,t} = c_y$ for all $t$ and $c_{o,t} = c_o$ for all $t$. So is there a stationary allocation that Pareto dominates (2, 2)? Figure 6.2 shows the resource constraint of the economy, plotted together with the utility level curve corresponding to the allocation (2, 2).

The shaded area is the feasible set; its frontier given by the line $c_y + c_o = 4$. It is clear from the picture with a tangency at (2, 2) (recall that the utility function is $\log c_y + \log c_o$) that it is not possible to find an alternative allocation that Pareto dominates this one. Now, let us however admit non-stationary allocations: could there be a non-stationary allocation that dominates (2, 2)? In order to implement such a non-stationary allocation, a chain of inter-generational transfers would require a transfer from young to old at some arbitrary point in time $t$. The agents giving away endowment units in their youth would have to be compensated when old. The question is how many units of goods would be required for this compensation.

[12] For an economy where endowments grow at some net rate $g$ in the limit, a similar theorem applies: the equilibrium is optimal if and only if the limit interest rate is equal to or above $g$.

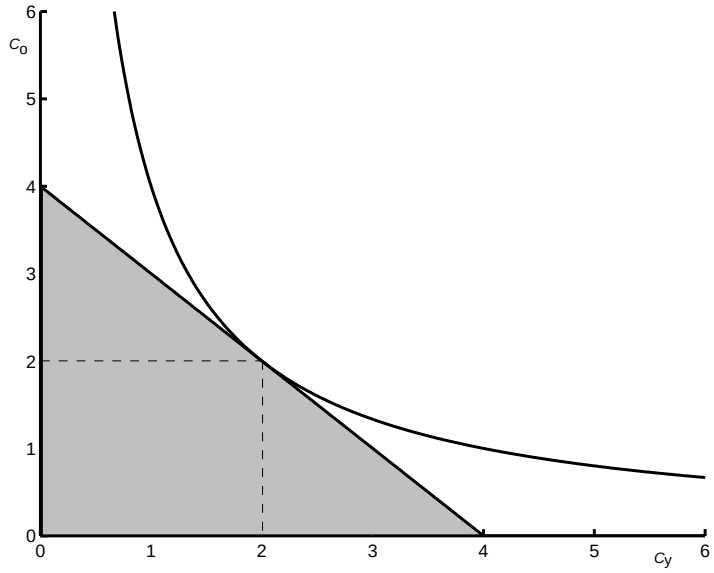



Figure 6.2: Pareto optimality of (2, 2) allocation

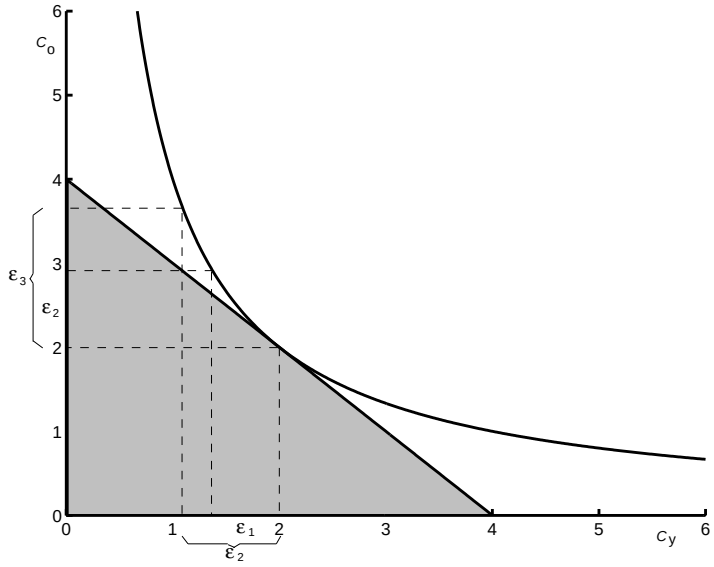



Figure 6.3: (2, 2) cannot be improved upon

Figure 6.3 illustrates that, given an initial transfer $\varepsilon_1$ from young to old at $t$, the transfer $\varepsilon_2$ required to compensate generation $t$ must be larger than $\varepsilon_1$, given the convexity of the indifference curves. This in turn will command a still larger $\varepsilon_3$, and so on. Is the sequence $\{\varepsilon_t\}_{t=0}^{\infty}$ thus formed feasible? No: eventually the transfer will exceed the young agent's endowment.

In the "simpler" case, where the equilibrium involves a gross real interest rate less than one, a constant transfer sequence is always possible: one can select another stationary allo-

cation and it is better for everybody. In the case where the gross real interest rate is above one, the argument is as illustrated in Figure 6.3, except even harder, because now the young needs to be compensated even more when old, given that their indifference curves have a higher slope (in absolute value). So no feasible better path exists here either.

### 6.4.2 Intertemporal production

Intertemporal production in an overlapping generations economy raises further issues. In some cases, the introduction of the possibility to save—in the form of "capital"—can help overcome an inefficiency; in others, it can lead to new market failures. We will keep the discussion brief and merely provide some illustrations.

Let us start with the endowment economy $(\omega_y, \omega_o) = (3, 1)$, where we know that the equilibrium is not optimal. Suppose that there is a simple storage technology allowing the consumers to save $k \geq 0$ units today and receive $\chi k$ units in return tomorrow. Clearly, if $\chi > 1/3$ they will: it gives them a higher return than in the original equilibrium. Is the resulting equilibrium allocation optimal? We can solve for individuals' optimal saving choices by simply setting $q_t = \chi > 1/3$ in equations (6.5)–(6.6). This is our equilibrium, which again is autarky, but with active individual storage.

If we regard the equilibrium allocation as a new endowment point and ask if Pareto improvements are possible, the answer is given by the Balasko-Shell theorem: improvements are possible if and only if $\chi < 1$. However, this is not the final answer: here, the answer is that the equilibrium is actually optimal when $\chi > 1$ but not when $\chi \leq 1$. In particular, $\chi = 1$ is not optimal. The reason is shockingly simple. In the equilibrium, all young agents store 1 unit and, hence, achieve the consumption allocation $(2, 2)$. So far so good; this is in fact the same allocation for people born at $t = 0$ and later as that with our social-security scheme. However, the old at 0 are still consuming only 1 unit, so a Pareto improvement can be obtained by instead carrying out the social-security scheme: the old at zero are strictly better off and no-one else worse off (everybody else is indifferent).

The market failure with storage here, and $\chi = 1$, is an example of *dynamic inefficiency*: it is possible, by means of different saving choices, to create more resources at at least one point in time without forsaking resources at any other point in time. The key insight here is that *oversaving* can occur in equilibrium. This is different than the market failure with fixed resources that we looked at before in our overlapping generations model: there, it was a matter of redistribution, and here, it is intertemporal production that is inefficient, again despite the absence of frictions.

As we shall see soon, the inefficiency in the special case with a one-for-one storage technology relies on the linearity of the production technology. If the intertemporal production technology is instead neoclassical, i.e., if it has decreasing returns to scale, the Balasko-Shell condition that the equilibrium is efficient if and only if, in the limit, the (gross) real interest rate is greater than or equal to 1, will be recovered. Let us therefore briefly revisit the neoclassical model here. Now let us interpret $(\omega_y, \omega_o)$ as the endowments of a cohort in labor

efficiency units. An agent's budget sets in period $t$ and $t+1$ are then

$$c_y + s = \omega_y w_t \quad \text{and} \quad c_o = s(1 - \delta + r_{t+1}) + \omega_o w_{t+1},$$

with $s$ denoting saving. If the utility function is strictly quasiconcave, the saving choice at $t$ is given uniquely by some function $h$:

$$s_t = h\left(w_t,\, r_{t+1},\, w_{t+1}\right)$$

Asset markets clear when $s_t = k_{t+1}$, where $k$ denotes the economy's total capital stock: the young buy the entire capital stock for the next period. Competitive pricing of inputs as usual implies $r_t = F_1\left(k_t,\, \omega_y + \omega_o\right)$ and $w_t = F_2\left(k_t,\, \omega_y + \omega_o\right)$. Thus, our equilibrium can be computed as the solution to the non-linear first-order difference equation

$$k_{t+1} = h\left(F_2(k_t, \omega_y + \omega_o), F_1(k_{t+1}, \omega_y + \omega_o), F_2(k_{t+1}, \omega_y + \omega_o)\right).$$

This equation implicitly determines $k_{t+1}$ as a function of $k_t$. We then have the following result that allows us to check whether the equilibrium savings choices are dynamically efficient.

**Theorem 6.2** *Define $R_t = 1 - \delta + F_1(k_t, \omega_t)$, where $\omega_t$ is the total labor endowment at $t$. Then $\{k_t\}_{t=0}^{\infty}$ is dynamically efficient if and only if*

$$\sum_{t=0}^{\infty} \left[ \prod_{s=1}^{t} R_s\left(k_s\right) \right] = \infty.$$

The theorem, whose assumptions are suppressed in the statement for brevity, relies in part on a proof of production efficiency that is similar in spirit to the proof of Pareto efficiency of an overlapping generations allocation with fixed resources.[13] The key point, however, is that it is the same condition as under fixed endowments: $\prod_{s=1}^{t} R_s\left(k_s\right) = 1/p_t$, where $p_t$ is our Arrow-Debreu price of consumption good at $t$ in terms of consumption good at 0.

To obtain intuition, it is instructive to restrict attention to steady states, i.e., solutions to

$$k_{ss} = h\left(F_2(k_{ss}, \omega_y + \omega_o), F_1(k_{ss}, \omega_y + \omega_o), F_2(k_{ss}, \omega_y + \omega_o)\right).$$

Clearly, from the theorem, a steady state is efficient if and only if $R = F'(k_{ss}, \omega_y + \omega_o) + 1 - \delta \geq 1$, that is, if the net interest rate is non-negative.[14] Let us start with the case of inefficiency. When $F'(k_{ss}, \omega_y + \omega_o) + 1 - \delta < 1$, an alternative saving plan is possible where savings at time $t$ are reduced by a small amount $\epsilon > 0$. This frees up resources for consumption at $t$. At $t+1$, there are now fewer resources available, but the reduction is less than $\epsilon$, given that

[13] The assumptions underlying it are that the sequence $\{R_t\}_{t=0}^{\infty}$ be uniformly bounded above and below away from zero and on the production-function curvature being bounded as follows: $0 < a \leq -f_t''\left(k_t\right) \leq M < \infty \;\; \forall t, \forall k_t$, where $f_t(k)$ is defined as $F(k/\omega_t, 1)$ and $F$ is assumed to have CRS.

[14] In a balanced-growth version of this economy, the efficiency requires the interest rate not to be below the growth rate.

the marginal return to saving was strictly less than one. However, suppose that at $t+1$, saving is again decreased by $\epsilon$ relative to the given steady state: then resources are freed up this period as well on net (there is less production, by less than $\epsilon$, but also less saving, by $\epsilon$). This procedure is repeated and, as a result, there is strictly more available to consume at all points in time beginning with period $t$. Now suppose there is a steady state with a gross return one or greater than one. If one were to try to create more resources at some point $t$ by reducing saving by $\epsilon$, the resources available at $t+1$ would be reduced by more than $\epsilon$.[15] The future reductions in saving required to keep resources at least as high as before will have to grow over time and will finally, become infeasible. The proofs of these statements, which are available Appendix 6.A, are the key behind understanding the logic of the theorem above.

In conclusion: in the overlapping generations model, equilibria—with or without production—can be efficient or inefficient. The key condition for efficiency, which holds rather generally, is that the asymptotic net real interest rate be non-negative or, in the case of growth, no less than the rate of growth.

## 6.5 Optimal government policy

We have now seen a number of examples of how markets may deliver inefficient outcomes. A natural suggestion in each of these cases is to propose a government policy to improve on the allocation. In the case of distortionary taxes, the origin of the inefficiency is in government policy itself, but in the other cases, what could we, as macroeconomic analysts, tell the government?

But let us start with the tax case, because it is still interesting. In particular, it is often argued, and for good reasons, that lump-sum taxes are hard to implement in practice. Thus, tax analysis could be carried out by comparing tax policies that are deemed feasible to implement. For example, in a dynamic model one can compare proportional taxes on labor earnings to proportional taxes on capital income: which is the better system from a welfare perspective, and by how much? Such an approach is referred to as Ramsey analysis, after Frank Ramsey's early work (in 1927). The approach is, however, somewhat problematic since it is often not clear which policies are feasible and which are not. Ramsey taxation will be studied in Chapter 13.

The case of externalities is well understood; here, Pigou (1920) suggested a tax, or transfer, that would counteract the distortion and cancel its effects exactly. In the example of pollution, the externality could be corrected by charging a per-unit output tax on firms equal to what the externality would be, evaluated at the optimal allocation, and it would deliver the optimal allocation as an equilibrium outcome. Similarly, positive externalities can be encouraged with per-unit subsidies where they occur. A different path here would be to follow Coase: assign property rights so that the side effects of agents' actions can be incorporated as market transactions. This path can, potentially, be easier to take, since

[15] If the net return is initially zero, it will raise above zero when we reduce saving.

there is no need for the government to compute what an appropriate tax rate would be: once property rights are assigned and enforced, the property value and the price its owner will charge for its use will be determined by markets and, in the absence of further frictions, lead to an optimal allocation. Sometimes, however, as in the case of damages due to climate change, ownership cannot be assigned: the earth's atmosphere is not possible to own.

Monopoly distortions that result in inefficiently low production are easy to handle in principle: one can, for example, subsidize production at a per-unit rate. Again, this requires a calculation of the appropriate tax rate, which requires much detailed knowledge. Therefore, whenever it is possible, anti-trust regulation can be used to minimize the presence of monopoly pricing. Monopoly power, finally, can play another role in the economy: it can provide incentives to invent new products if, namely, there is patent protection (or it is difficult for other reasons to imitate the product). Invention is typically costly so to incentive inventors one might wish to exclude others from using an invention. Thus, regulating against monopoly has drawbacks too; we will study this issue in Chapter 11.

### 6.5.1 Missing markets and the "chicken model"

What about the missing markets case? Recall the endowment economy where endowments alternated between agents and no borrowing was allowed. Here, a reasonably simple policy would seem to be available to the government: each period, tax the agent with a high endowment and transfer the proceeds to the other agent, so as to achieve full consumption smoothing. Both agents would then be better off, at least if the transfer is small enough. Similarly, in the context of missing insurance markets, the government could simply compensate people who received bad shocks and tax the luckier ones to finance the transfers. Is it reasonable to propose such a policy? This is not so clear. In reality, when a market is missing, it is usually missing for a reason. That reason could, for example, be problems of private information or moral hazard. If borrowing/lending and insurance are so beneficial, why do they not materialize, in the cases where they appear not to be present?[16] A consequence of this point is that it is entirely conceivable that there would be negative side effects of the governments intervention: those side effects that made markets missing in the first place.

The above considerations have given rise to the concept of *a "chicken model" of government*. The model here refers to an argument for government intervention in markets and goes as follows. Assume that (i) people like chicken; (ii) the market economy cannot produce chicken; and (iii) the government can produce chicken. The result then follows: the government should produce chicken. From our perspective, the lesson should be: in cases where government intervention is proposed, think about which friction is at work, and whether it is one that the government is likely to be able to deal with well, or better than the market. Sometimes the answer is likely yes, and other times no.

Clearly one kind of approach available here would be to try to model the causes of

[16] Think about whether you should write an insurance contract with your fellow graduate students, making sure that your post-graduation salaries are all the same, after taking transfers between you into account.

frictions, such as private information, explicitly. Thus, analysis following the work of Mirrlees (1971) has been used to study optimal taxes and transfers when markets appear to be functioning imperfectly. Another reason why some markets may not exist is a lack of commitment, as discussed in Section 6.3.4 above.

### 6.5.2 Redistribution policy

The discussion so far centered around minimizing frictions. In practice, a separate aim is often redistribution, i.e., the idea is not to Pareto-improve on the given allocation but rather to achieve a more equitable distribution of consumption even if some agents are made worse off. In macroeconomic models with heterogeneity, some of which will be studied later in this book, researchers often adopt a social welfare function to guide policy choices. In such cases, the welfare weights on different agents would represent the policymaker's preferences but, of course, not necessarily those of the researcher.

An often used social welfare function is an additive ("utilitarian") formulation. The most common assumption then is that the utilities of the agents are weighted equally. Clearly, equal weights would amount to equal consumption in an economy where direct, non-distortionary redistribution is available. Hence, equal weights embody a strong desire for equality. If taxes are distortionary, then equal weights still express the same desire but the optimal level of redistribution will typically not fully eliminate consumption inequality.

An argument for equally-weighted utilitarian social welfare functions that have been used is the "behind-the-veil-of-ignorance" notion. So imagine that a person, before they are born into a household somewhere in the world, possibly also without knowing what genetic skills they will have once born, is asked to consider potential distributional policies. Then redistribution could potentially be viewed as an optimal insurance scheme; in concrete terms, maximize $\pi u(c_A) + (1-\pi)u(c_B)$ subject to a resource constraint $\pi c_A + (1-\pi)c_B = \pi\omega_A + (1-\pi)\omega_B$, where $\omega_A > \omega_B$ would be the market incomes of the two types of agents; $\pi$ is the fraction of people who will be born as $A$ types. Clearly, this optimization problem embodies equal weights and delivers $c_A = c_B$. The insurance solution cannot be offered by markets, since the agents are not around to sign the contract before they are born, but governments can nevertheless carry out a policy which achieves the insurance outcome. Is this therefore a chicken model of government? Not so much; it is more seen a potential guiding philosophical principle with which you may agree or disagree.

# Chapter 7

# Uncertainty

Many aspects of economic life are not fully predictable. For example, at the aggregate level, technologies and policies can change in unpredictable ways. Life is even more uncertain at the individual level due to the unpredictability of income, health, and other events. We now introduce the main techniques that macroeconomists use to incorporate uncertainty into our analysis of the economy.

This chapter covers several issues. First, we introduce analytical tools that are helpful in analyzing stochastic economies. These tools include mathematical concepts related to stochastic processes as well as economic concepts related to decision making under uncertainty. We then use these tools to analyze the social planner's problem for a version of the neoclassical growth model with stochastic productivity. This important example is introduced in Section 7.3. We then discuss how agents trade with each other in an uncertain economic environment and present the competitive equilibrium of the stochastic neoclassical growth model. Finally, we briefly present an environment with incomplete insurance markets.

## 7.1 Stochastic processes

A **stochastic process** is a collection of random variables indexed by time. Suppose at each date $t$ there is a random outcome $X_t$. The collection of these random variables $\{X_t : t \in \mathcal{T}\}$ is a stochastic process, where the set $\mathcal{T}$ is the span of time we are interested in. When modeling an economy with uncertainty, we typically assume that some fundamental features of the economy follow an exogenous stochastic process. For example, we often assume that the level of productivity fluctuates over time and is modeled as an exogenous stochastic process. This randomness in fundamentals generates randomness in endogenous variables. Our economic model determines the stochastic process these endogenous variables follow. We will now introduce some general properties of stochastic processes before turning to a few of the main types of processes used by macroeconomists.

### 7.1.1 Properties of stochastic processes

When working with stochastic processes we often need to take expectations of them. One way to think about the expectation of a random variable $X_t$ that is part of a stochastic process is to imagine multiple realizations of the entire stochastic process $\{X_t^{(i)} : t \in \mathcal{T}, i \in 1, \cdots, I\}$, where $i$ indexes the different realizations. Taking the expectation across $i$ (i.e., add using probability weights) then gives the unconditional expectation. A related concept is a conditional expectation given information up to some point in time. For example we could suppose we have observed the realization of the stochastic process up to date $t$ and then imagine different possible realizations for $t+1$ and later dates. Taking the expectation of $X_{t+1}$ across these different realizations is then an expectation conditional on information through date $t$. We often use $\mathbb{E}_t[X_{t+1}]$ to denote such a conditional expectation. To be clear, $\mathbb{E}_t[X_{t+1}]$ means the expectation of $X_{t+1}$ conditional on all information available at date $t$, not just the realization of $X_t$ observed at date $t$. A very useful property of expectations is the **law of iterated expectations**, which for any dates $t < s < \tau$ says

$$\mathbb{E}_t\left[\mathbb{E}_s\left[X_\tau\right]\right] = \mathbb{E}_t\left[X_\tau\right].$$

A proof of the law of iterated expectations appears in the appendix.

In addition to expectations, we are often interested in the second moments of a stochastic process. These second moments are captured by the autocovariances. The $j$-th autocovariance is given by $\mathbb{E}\left[(X_t - \mu_t)(X_{t-j} - \mu_{t-j})\right]$ where $\mu_t$ is the unconditional expectation of $X_t$.

Many economic theories assume or imply stochastic processes that are stationary. A process is **covariance stationary** if neither the unconditional expectations nor the autocovariances depend on time $t$. In practice, stationarity means that the consequences of a shock to the process eventually fade. If this were not the case, as time goes by, the effects of all the shocks that have occurred would accumulate and the distribution of $X_t$ would change, e.g., become more and more dispersed, as $t$ increases.

In most cases we look at, a stationary process will be **ergodic**. For an ergodic process, observing a long time series allows us to understand the distribution of the stochastic process. For example, the unconditional expectation of $X_t$ is the expectation across different possible realizations of $X_t$. For an ergodic process, an average of a long time series $(1/T)\sum_{t=1}^{T} X_t$ will converge to $\mathbb{E}[X_t]$ as $T \to \infty$.[1] Ergodicity is useful because in practice we often have data on a single long time series.

A stochastic process is (first-order) **Markov** if its current value summarizes all the available information that is useful for predicting its future realizations. For example, the conditional distribution of $X_{t+1}$ conditional $X_t$ is the same as the conditional distribution of $X_{t+1}$ conditional on $\{X_\tau\}_{\tau \leq t}$, in which case knowledge of $X_\tau$ for $\tau < t$ does not add any relevant information beyond that contained in $X_t$. More formally, a Markov process satisfies $\Pr[X_{t+k} = x | X_\tau \forall \tau \leq t] = \Pr[X_{t+k} = x | X_t]$ $\forall t, k \geq 0$. In dynamic programming, it is convenient if a single state variable summarizes the process. For this reason, macroeconomists often work with Markov processes.

[1] See Hamilton (1994) for the conditions under which a stationary process is ergodic.

Over time, a deterministic sequence might converge to a particular value. A stochastic process, on the other hand, might continually be subject to random shocks and therefore not settle down to a particular value but nevertheless there is a sense in which it can converge. A stationary distribution (or invariant distribution) $\bar{\pi}(X)$ of a Markov process $X$ has the property that if $X_t \sim \bar{\pi}(X)$ then $X_{t+1} \sim \bar{\pi}(X)$.

### 7.1.2 Markov chains

Let $x_t$ be a random variable that takes on values in the discrete set $\mathcal{X} = \{\overline{x}_1, \overline{x}_2, \cdots, \overline{x}_N\}$. If $x_t$ follows a Markov process, the probability that a particular $x_{t+1}$ occurs depends only on $x_t$ and not on earlier values. A Markov process with a discrete state space is called a **Markov chain**. As $x_t$ takes on discrete values, we can summarize the process in terms of a **transition matrix**. Suppose the probability of moving from state $i$ to $j$ is given by $P_{ij}$. We can then collect these transition probabilities in a matrix $P$. Row $i$ represents the probabilities of states 1 through $N$ occurring next period conditional on the current state being $i$. As exactly one of these states will occur, these probabilities must sum to one. By this logic, each row of $P$ must sum to one.

In addition to the transition matrix, a full description of the stochastic process also requires knowing the initial distribution over the states. We will represent this distribution as a $1 \times N$ vector of probabilities $\pi_0$ such that the $i$th element of $\pi_0$ gives $\Pr[x_0 = \overline{x}_i]$.

Given the initial distribution $\pi_0$, the transition matrix determines the probability distributions for all $x_t$ for $t \geq 1$. We have

$$\Pr[x_1 = \overline{x}_j] = \sum_{i=1}^{N} \Pr[x_1 = \overline{x}_j | x_0 = \overline{x}_i] \times \Pr[x_0 = \overline{x}_i] = \sum_{i=1}^{N} P_{ij} \times [\pi_0]_i$$

where $[\pi_0]_i$ is the $i$th element of $\pi_0$. Notice that this sum is the product of $\pi_0$ against the $i$th column of $P$. Repeating this logic for each $j = 1, \cdots, N$ we have $\pi_1 = \pi_0 \times P$. This relationship generalizes to $\pi_{t+1} = \pi_t \times P$ and by repeated substitution we have

$$\pi_{t+k} = \pi_t \times P^k.$$

This is a very useful property of Markov chains: the conditional distributions $k$ periods ahead can be found by raising the transition matrix to the power $k$.

For a Markov chain, a stationary distribution $\bar{\pi}$ is one for which $\bar{\pi} = \bar{\pi} \times P$. If we transpose this definition we have $\bar{\pi}' = P' \times \bar{\pi}'$ and we can see that the stationary distribution is the eigenvector of $P'$ associated with a unit eigenvalue.[2]

To give an example with an economic interpretation, suppose workers can be employed or unemployed. Unemployed workers find jobs with probability $f \in (0, 1)$ and lose (or separate

[2] A transition matrix will always have a unit eigenvalue. As the rows of $P$ sum to one, we know $\mathbf{1} = P\mathbf{1}$, where $\mathbf{1}$ is a column vector of ones. This equation says $P$ has a unit eigenvalue and the eigenvalues of $P$ and $P'$ are the same.

from) jobs with probability $s \in (0,1)$. Then we can represent the transitions across the employed/unemployed states by the transition matrix

$$P = \begin{pmatrix} 1-s & s \\ f & 1-f \end{pmatrix},$$

where the first state represents being employed and the second represents being unemployed. As the stationary distribution has two probabilities that sum to one, there is only one unknown. Let $\bar{u}$ be the stationary (steady state) unemployment rate so that we have $\bar{\pi} = \begin{bmatrix} 1-\bar{u} & \bar{u} \end{bmatrix}$. The eigenvector of $P'$ associated with a unit eigenvalue solves

$$\begin{pmatrix} -s & f \\ s & -f \end{pmatrix} \begin{pmatrix} 1-\bar{u} \\ \bar{u} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Or, equivalently,

$$\bar{u} = s(1-\bar{u}) + (1-f)\bar{u}.$$

The steady state unemployment rate is equal to the mass of employed who separate plus the mass of unemployed who remain unemployed. Solving this equation yields

$$\bar{u} = \frac{s}{s+f}.$$

Notice that there is a unique stationary distribution of this economy. Suppose we start our economy with an unemployment rate $u_0$. As time goes by, it is also straightforward to check that the unemployment rate will converge to $\bar{u}$ regardless of what $u_0$ we start with.[3]

When will a Markov chain more generally converge to a unique stationary distribution? It is not always the case, as the following examples demonstrate. Consider a Markov chain with a transition matrix equal to the identity matrix. No matter what initial distribution we start with, the distribution will forever remain stationary so there is not a unique distribution although it converges immediately. As another example, consider the Markov chain with transition matrix given by

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

This Markov chain has a unique stationary distribution of [1/2, 1/2], but unless we start with that distribution, the Markov chain will never converge to it. To see this, suppose we start with probability mass $p \neq 1/2$ on the first state and $1-p$ on the second. The mass on the first state will oscillate between $p$ and $1-p$ forever.

A simple necessary condition for convergence and uniqueness is that from each state there is a positive probability of moving to any other state. This condition is actually substantially stronger than we need. Weaker conditions are as follows. State $j$ of a Markov chain is said to be reachable from state $i$ if there is some $n$ such that $\Pr(x_n = \bar{x}_j | x_0 = \bar{x}_i) > 0$. A

[3] We obtain $u_{t+1} = s(1-u_t) + (1-f)u_t = s + (1-f-s)u_t$. By repeated substitution we see that $u_t = s(1+\lambda+\lambda^2+\cdots+\lambda^t u_0)$, with $\lambda \equiv 1-f-s$, which will converge to $s/(s+f)$ since both $s$ and $f$ are strictly between 0 and 1.

Markov chain is said to be irreducible if all states are reachable from all other states. State $i$ of a Markov chain is said to be aperiodic if there is some $n$ such that for all $n' \geq n$ $\Pr(x_{n'} = \bar{x}_i | x_n = \bar{x}_i) > 0$. A Markov chain is aperiodic if all states are aperiodic. An irreducible Markov chain has a unique stationary distribution, $\bar{\pi}$, and if it is aperiodic then $\lim_{t\to\infty} \pi_0 P^t = \bar{\pi}$ for all $\pi_0$.

### 7.1.3 Autoregressive processes

We now turn our attention to stochastic processes with continuous distributions. A very common formulation is the autoregressive process of order one or **AR(1) process** for short. The stochastic process $x_t$ follows an AR(1) if it satisfies

$$x_t = \rho x_{t-1} + b\varepsilon_t + (1-\rho)\mu \tag{7.1}$$

where $x_t \in \mathbb{R}$, $\rho$, $b$, and $\mu$ are scalar coefficients, and $\varepsilon_t$ is a stochastic process that satisfies $\mathbb{E}_{t-1}[\varepsilon_t] = 0$, $\mathbb{E}_{t-1}[\varepsilon_t^2] = 1$, and $\mathbb{E}_{t-1}[\varepsilon_t \varepsilon_{t+s}] = 0$ for all $s > 0$.[4] We call the process (7.1) an AR(1) because $x_t$ depends on just one lagged value $x_{t-1}$.

We can calculate moments of this process by expressing $x_t$ as a moving average of past $\varepsilon$'s. Through repeated substitution we arrive at

$$\begin{aligned} x_t &= \mu + b\varepsilon_t + \rho b\varepsilon_{t-1} + \rho^2 b\varepsilon_{t-2} + \cdots \\ &= \mu + b\sum_{s=0}^{\infty} \rho^s \varepsilon_{t-s}. \end{aligned}$$

If $|\rho| < 1$, the effect on $x_t$ of shocks, $\varepsilon$, in the distant past vanishes and the process is stationary. Taking an unconditional expectation we see $\mathbb{E}[x_t] = \mu$ since $\mathbb{E}[\varepsilon_{t-s}] = \mathbb{E}[\mathbb{E}_{t-s-1}[\varepsilon_{t-s}]] = 0$ for all $s$. The unconditional variance of $x_t$ is given by

$$\operatorname{Var}[x_t] = \sum_{s=0}^{\infty} (b\rho^s)^2 \operatorname{Var}[\varepsilon_{t-s}] = \frac{b^2}{1-\rho^2},$$

where we have used the fact that the $\varepsilon$'s have unit standard deviation and are uncorrelated across time. Similarly, the covariance of $x_t$ and $x_{t+j}$ is

$$\begin{aligned} \operatorname{Cov}(x_t, x_{t+j}) &= \mathbb{E}[(x_t - \mu)(x_{t+j} - \mu)] \\ &= \mathbb{E}\left[\left(b\sum_{s=j}^{\infty} \rho^{s-j}\varepsilon_{t-s+j}\right)\left(b\sum_{s=0}^{\infty} \rho^s \varepsilon_{t+j-s}\right)\right] \\ &= b^2\left(\rho^j + \rho^{j+2} + \rho^{j+4} + \cdots\right) \\ &= \frac{b^2\rho^j}{1-\rho^2}. \end{aligned}$$

[4] The assumption that $\varepsilon_t$ has unit variance is a normalization as the parameter $b$ scales the effects of $\varepsilon_t$ on $x_t$.

The correlation between $x_t$ and $x_{t+j}$ is therefore $\rho^j$. In summary, the parameter $\mu$ determines the level of the process, the parameter $b$ determines the volatility of the process, and the parameter $\rho$ determines the persistence of the process.

### 7.1.4 Linear stochastic difference equations

We now will generalize our autoregressive specification to allow for vector-valued random variables. Let $x_t$ be a column vector in $\mathbb{R}^n$. Let $\varepsilon_t$ be a random variable in $\mathbb{R}^m$. The stochastic process $\varepsilon_t$ is assumed to satisfy

$$\mathbb{E}_t\left[\varepsilon_{t+1}\right] = 0 \tag{7.2}$$
$$\mathbb{E}_t\left[\varepsilon_{t+1}\varepsilon_{t+1}'\right] = I \tag{7.3}$$
$$\mathbb{E}_t\left[\varepsilon_{t+1}\varepsilon_{t+s}'\right] = 0 \quad \forall s > 1. \tag{7.4}$$

The second condition says the elements of $\varepsilon_t$ are uncorrelated with each other and have unit standard deviation. The third condition states that $\varepsilon_t$ is uncorrelated across time. We assume that $x_t$ follows a linear stochastic difference equation:

$$x_t = Ax_{t-1} + B\varepsilon_t + C. \tag{7.5}$$

The $n \times n$ matrix $A$ controls how $x_{t-1}$ affects $x_t$. If all the eigenvalues of $A$ are smaller than 1 in absolute value, then the effects of past shocks will eventually fade and $x_t$ will be a stationary process. The $n \times m$ matrix $B$ captures the effects of $\varepsilon_t$ on $x_t$.[5] Lastly, the $n \times 1$ vector $C$ affects the mean of $x_t$ as we describe next.

The unconditional expectation of $x_t$ is

$$\mu \equiv \mathbb{E}\left[x_t\right] = A\mathbb{E}\left[x_{t-1}\right] + C = A\mu + C.$$

Solving this equation yields $\mu = (I - A)^{-1}C$. Similarly, let $\Gamma(0)$ be the unconditional covariance matrix of $x_t$. The definition of a covariance matrix gives us

$$\begin{aligned}\Gamma(0) &= \mathbb{E}\left[(x_t - \mu)(x_t - \mu)'\right] \\ &= \mathbb{E}\left[A(x_{t-1} - \mu)(x_{t-1} - \mu)'A' + B\epsilon_t\epsilon_t'B'\right] \\ &= A\Gamma(0)A' + BB',\end{aligned} \tag{7.6}$$

where we have used the fact that $\varepsilon_t$ is independent of $x_{t-1}$.[6]

We are often interested in the behavior of a stochastic process following a particular event. For example, we might be interested in how the economy would behave following

[5] The assumption reflected in (7.3) that the $\varepsilon_t$ are uncorrelated with each other and have unit standard deviations is a normalization since the matrix $B$ can rescale their standard deviations and impart correlations across their effects.

[6] Equation (7.6) is a Lyapunov equation and can be used to solve for $\Gamma(0)$. Many software packages are available to solve such equations.

a TFP shock. In this thought experiment, we suppose no further shocks occur. Using $C = \mu - A\mu$, rewrite (7.5) as

$$x_t - \mu = A(x_{t-1} - \mu) + B\varepsilon_t.$$

Suppose there is a particular shock $\varepsilon_t$ at $t$ and then no future shocks. Repeated substitution yields

$$x_{t+h} - \mu = A^{h+1}(x_{t-1} - \mu) + A^h B\varepsilon_t.$$

The effect of $\varepsilon_t$ on $x_{t+h}$ is given by $A^h B\varepsilon_t$. This change in the future evolution of the process is called the impulse response of $x_t$ to the shock $\varepsilon_t$. The function $\mathcal{F}(h) = A^h B\varepsilon$ is the **impulse response function** of $x$ to the particular shock $\varepsilon$. The impulse response function tells us how $x$ responds to the shock as a function of the time since the shock has occurred. We saw some examples of impulse responses in the deterministic, non-linear Solow model in Section 3.5.2. If the Solow model is extended to include stochastic shocks, the results in this section would apply only to a linear approximation to the Solow model.

## 7.2 Choice under uncertainty

We will now begin our discussion of how agents make choices under uncertainty. In this section we start by introducing a framework for modeling uncertainty in a way that can introduce uncertainty without restricting ourselves to a specific stochastic process. We then discuss preferences over risky consumption outcomes. Risk aversion is an important aspect of preferences in an uncertain environment and we will demonstrate the implications of risk aversion through a portfolio choice problem.

### 7.2.1 Stochastic events

To incorporate uncertainty into economic theory, it is often convenient to define a stochastic event that determines all the risky outcomes. The idea here is that there are many different ways the world may take shape in the future and our uncertainty is that we do not know which of these "worlds" we live in. As our theories are typically dynamic, we need to allow for our uncertainty to resolve over time. Let $\omega_t \in \Omega_t$ be the stochastic event realized at date $t$ and let $\omega^t = \{\omega_0, \omega_1, \cdots, \omega_t\} \in \Omega^t$ be the history of events up to date $t$. To give an example, suppose your income each month can either be high or low. The event $\omega_t$ determines whether your income is high or low in month $t$ and $\omega^t$ gives a list of all the past events from which we can infer your past incomes. Figure 7.1 shows an example of how these stochastic events could unfold for $t = 0, 1, 2$ when there are two possible realizations of $\omega_t$ at each date: $\omega_t \in \{0, 1\}$.

The probability that history $\omega^t$ will be realized at date $t$ from the perspective of date 0 is given by $\pi_t(\omega^t)$. The conditional probability of $\omega^t$ given $\omega^\tau$ for $t > \tau$ is $\pi_t(\omega^t|\omega^\tau)$. An outcome at date $t$ is a function of the history up to date $t$. For example, the balance in your bank account reflects not just the randomness in your current income, but also the

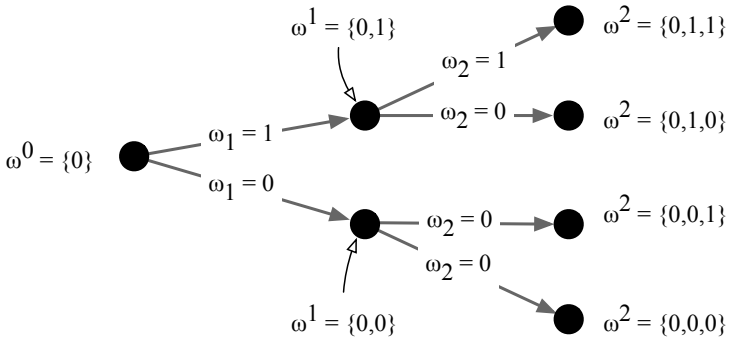



Figure 7.1: Example event tree.

fluctuations in your previous incomes (as well as the changes in spending they induce). We could write your assets as $a_t(\omega^t)$ to indicate that it depends on the whole history of events leading up to date $t$.

### 7.2.2 Expected utility and risk aversion

We now extend the preferences we introduced in Chapter 4 to incorporate uncertainty according to *expected utility theory*. Thus, utility over stochastic events is then a convex linear combination of a function $u(c)$, where $c$ is random, with the linear coefficients being the probabilities with which the different outcomes for $c$ are realized. Applied to the context of preferences over time, suppose $\{c_t(\omega^t) : \forall t, \omega^t\}$ and $\{\tilde{c}_t(\omega^t) : \forall t, \omega^t\}$ are two consumption processes. We will say the $c$ process is preferred to the $\tilde{c}$ process if and only if

$$\sum_{t=0}^{\infty} \sum_{\omega^t \in \Omega^t} \pi_t(\omega^t) \beta^t u(c_t(\omega^t)) > \sum_{t=0}^{\infty} \sum_{\omega^t \in \Omega^t} \pi_t(\omega^t) \beta^t u(\tilde{c}_t(\omega^t)).$$

In many situations, we will not write the stochastic events explicitly because the time subscript on the variables is sufficient to keep track of which histories they depend on. Using that notational convention, the above statement would be

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t u(c_t) > \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t u(\tilde{c}_t).$$

As we get started, however, it is useful to be explicit about the histories.

Risk aversion is a fundamental concept of choice under uncertainty that will allow us to explain, among other things, why consumers buy insurance and why savers hold diversified

portfolios. Risk aversion is the idea that a less risky consumption stream is preferred to a more risky stream with the same expected consumption. Mathematically, risk aversion is implied by Jensen's inequality when $u(\cdot)$ is concave. If $u(\cdot)$ is linear, then the consumer ranks consumption streams by the expected level of consumption and is said to be risk neutral.

The curvature of $u(\cdot)$ determines the level of risk aversion. The coefficient of absolute risk aversion is defined as $-u''(c)/u'(c)$. A more concave utility function leads to higher risk aversion. The second derivative is normalized by the first derivative to capture the change in curvature as utility changes.[7] The coefficient of absolute risk aversion refers to the attitude towards changes in consumption of a given (absolute) size. The constant absolute risk aversion (CARA) utility function is given by $u(c) = -\exp(-\alpha c)$. For this utility function, the coefficient of absolute risk aversion is $\alpha$ at all levels of consumption.

Alternatively, the coefficient of relative risk aversion measures a consumer's attitude to proportional changes in consumption. It is defined as $-cu''(c)/u'(c)$. Our power utility function

$$u(c) = \frac{c^{1-\sigma} - 1}{1 - \sigma},$$

motivated by its consistency with exact balanced growth, has a constant coefficient of relative risk aversion. Taking derivatives, we see the coefficient of relative risk aversion is simply $\sigma$. As the coefficient of relative risk aversion is the same at all levels of consumption, these preferences are also known as the constant relative risk aversion (CRRA) utility function.

In most macroeconomic applications, the CRRA function is used because, as we have discussed, it is the only one consistent with balanced, long-run growth. From the perspective here, using a CRRA function means that decisions about risks today and one hundred years ago—when the level of consumption was much lower—would have been made the same way if the risks were the same in percentage terms.

Recall from Section 4.2.4 that the elasticity of intertemporal substitution is the elasticity of consumption growth with respect to the real interest rate. More precisely, the consumption Euler equation for power utility implies $d\log[c_{t+1}/c_t]/dR_{t+1} = 1/\sigma$, thus implying that the elasticity of intertemporal substitution actually equals the inverse of the coefficient of relative risk aversion. The intuition here is that both of these features of the utility function are determined by the curvature of the utility function. If the utility function displays strong diminishing marginal utility, then a consumer will be unwilling to accept higher consumption in one state of the world in exchange for low consumption in another state of the world (i.e., risk aversion is high). Similarly, the consumer will be unwilling to accept low consumption in period $t$ in exchange for high consumption in period $t + 1$ (i.e., is unwilling to substitute intertemporally). Intuitively put, one single parameter guides the desire for smoothing consumption across time and states of nature.[8]

[7] In expected utility theory, a positive affine transformation of a utility function represents the same preferences. That is, the utility function $v(c)$ defined by $au(c) + b$ for constants $a > 0$ and $b$ is equivalent to the utility function $u$. Normalizing the coefficient of risk aversion by $u'(c)$ makes the coefficient of absolute risk aversion invariant to such an affine transformation of $u$.

[8] There is a generalization of the power-function preferences that allows one to separate risk aversion from intertemporal substitution using two separate parameters. This case will be discussed in Chapter 14 below.

### 7.2.3 Portfolio choice

A simple portfolio choice problem can help illustrate the differences between the two concepts of risk aversion introduced above. Suppose an individual has $W > 0$ units of wealth at date 0 to allocate between a risk-free asset and a risky asset. The assets will pay off at date 1 and the individual will consume the proceeds. The (gross) return on the risk-free asset is known to be $R^f$ while the return on the risky asset is unknown and denoted $Z$. Let $A$ be the assets invested in the risky asset while $W - A$ are invested in the risk-free asset. The investor's decision problem is

$$\max_A \mathbb{E}_0 \left[ u \left( R^f (W - A) + ZA \right) \right].$$

The first-order condition of this problem is

$$\mathbb{E}_0 \left[ u' \left( R^f (W - A) + ZA \right) \left( Z - R^f \right) \right] = 0.$$

Suppose the investor's utility function is the CRRA utility function $u(c) = c^{1-\sigma}/(1-\sigma)$. The first-order condition of the portfolio choice problem becomes

$$\mathbb{E}_0 \left[ \left( R^f (W - A) + ZA \right)^{-\sigma} \left( Z - R^f \right) \right] = 0.$$

Rearranging we arrive at

$$\mathbb{E}_0 \left[ \left( R^f \left( 1 - \frac{A}{W} \right) + Z \frac{A}{W} \right)^{-\sigma} \left( Z - R^f \right) \right] = 0,$$

where we have brought $W^{-\sigma}$ outside the expectation because it is known at date 0. This equation determines $A/W$ as a function of $\sigma$, $R^f$, and the distribution of the risky asset return. Notice that the solution for $A/W$ does not depend on $W$, which means that at any level of wealth, the investor will allocate the same fraction of savings to risky assets: the rich and the poor choose the same risk exposure.

Now suppose the investor has the CARA utility function $u(c) = -\exp(-\alpha c)$. The first-order condition then becomes

$$\mathbb{E}_0 \left[ \alpha \exp \left\{ -\alpha R^f W \right\} \exp \left\{ -\alpha \left( Z - R^f \right) A \right\} \left( Z - R^f \right) \right] = 0.$$

We can bring $\exp \left\{ -\alpha R^f W \right\}$ outside the expectation because it is known at date 0 to arrive at

$$\mathbb{E}_0 \left[ \exp \left\{ -\alpha \left( Z - R^f \right) A \right\} \left( Z - R^f \right) \right] = 0.$$

This equation gives a solution for $A$ that does not depend on $W$. In this case, the investor allocates a particular level of savings (an absolute number of goods or dollars) to risky assets regardless of their wealth. This means that the rich have a lower risky share than do the poor.

In the data, as we shall see in Chapter 14, the rich on average choose a higher risky share. Neither of these simple models can match this fact but the CRRA case is more in line with the data.

## 7.3 The stochastic growth model

We can now use the tools of choice under uncertainty and stochastic processes to analyze a stochastic version of the neoclassical growth model. As briefly discussed in Chapter 3 above, in this model, total factor productivity (TFP) is assumed to follow an exogenous stochastic process. The fluctuations in TFP then give rise to endogenous fluctuations in output, consumption, and investment; the model was first studied, as a planning problem, in Brock and Mirman (1972), and then became a workhorse framework macroeconomics, and we devote Chapter 12 below to it. Through this important example, we will discuss how our methods for dynamic optimization can be extended to allow for uncertainty.

### 7.3.1 A two-period economy

To begin, suppose the economy exists for two periods. In period 0, the level of TFP in period 1 is not known. Let $\omega_1$ be the stochastic event in period 1. Let $\pi_1(\omega_1)$ be the probability of $\omega_1$. TFP at date 1 is given by $A_1(\omega_1)$.

The economy is inhabited by a representative household with expected utility preferences that ranks consumption streams according to

$$U = u(C_0) + \beta \sum_{\omega_1 \in \Omega_1} \pi_1(\omega_1) u\left(C_1(\omega_1)\right), \tag{7.7}$$

where $C_1(\omega_1)$ is the level of consumption if $\omega_1$ occurs. In period 0, the economy is endowed with $K_0$ units of capital, which are used to produce $Y_0 = K_0^\alpha$ units of output. This output is then used for consumption and investment subject to the date-0 resource constraint

$$K_1 + C_0 = K_0^\alpha + (1-\delta)K_0. \tag{7.8}$$

In period 1, the value of $\omega_1$ becomes known and the economy produces $Y_1(\omega_1) = A_1(\omega_1)K_1^\alpha$. As there are no further periods, there is no reason to invest in capital so the resource constraint in period 1 is

$$C_1(\omega_1) = A_1(\omega_1)K_1^\alpha + (1-\delta)K_1. \tag{7.9}$$

The first important thing to note is that $A_1$, $Y_1$, and $C_1$ are all functions of the event $\omega_1$. From the perspective of date 0, agents do not know $A_1$ and therefore they cannot know how much will be produced or consumed. When we formulate a decision problem in date 0, the agents will not choose specific values for $Y_1$ and $C_1$, but rather they will choose a plan for how they will respond to each realization of $\omega_1$. This is an important feature of optimization under uncertainty: the choice variable is a contingent plan for actions following each history of stochastic events.

The planner's problem for this economy is to choose $C_0$, $K_1$, and $\{C_1(\omega_1) : \forall \omega_1\}$ to

maximize (7.7) subject to (7.8) and (7.9). We can form the Lagrangian as

$$\begin{aligned}\mathcal{L} = u(C_0) + \beta \sum_{\omega_1 \in \Omega_1} \pi_1(\omega_1) u\left(C_1(\omega_1)\right) - \lambda_0 \left[K_1 + C_0 - K_0^\alpha - (1-\delta)K_0\right] \\ - \sum_{\omega_1 \in \Omega_1} \lambda_1(\omega_1) \left[C_1(\omega_1) - A_1(\omega_1) K_1^\alpha - (1-\delta)K_1\right],\end{aligned}$$

where $\lambda_0$ and the $\lambda_1$'s are Lagrange multipliers on (7.8) and (7.9), respectively. As (7.9) must hold for each realization of $\omega_1$ we treat that as a separate constraint for each $\omega_1$. We therefore have separate Lagrange multipliers for each $\omega_1$ and we use the sum to include all of them in the Lagrangian.

Taking the first-order conditions for this problem, we have

$$\begin{aligned} u'(C_0) &= \lambda_0 \\ \lambda_0 &= \sum_{\omega_1 \in \Omega_1} \lambda_1(\omega_1) \left(\alpha A_1(\omega_1) K_1^{\alpha-1} + 1 - \delta\right) \\ \beta \pi(\omega_1) u'(C_1(\omega_1)) &= \lambda_1(\omega_1) \qquad \forall \omega_1. \end{aligned}$$

The second line is the first-order condition for $K_1$. On the right-hand side we have a sum over all possible realizations of $\omega_1$ because when the planner chooses $K_1$ they do not know which $\omega_1$ will occur so they need to take into account how $K_1$ affects output and consumption after each one. In contrast, the third line is the first-order condition with respect to $C_1(\omega_1)$ for a specific $\omega_1$ and there is one such equation for each $\omega_1$.

Combining the first-order conditions to eliminate the Lagrange multipliers we have

$$u'(C_0) = \beta \sum_{\omega_1 \in \Omega_1} \pi(\omega_1) u'\left(C_1(\omega_1)\right) \left(\alpha A_1(\omega_1) K_1^{\alpha-1} + 1 - \delta\right).$$

This is the stochastic consumption Euler equation for the planner. The left-hand side is the marginal utility loss in date 0 from saving one more unit. The right-hand side is the expected marginal utility gain in period 1 from saving one more unit. Notice that we sum over $\omega_1$ and weight the outcomes by $\pi_1(\omega_1)$ so we are taking an expectation. The term $\alpha A_1(\omega_1) K_1^{\alpha-1} + 1 - \delta$ is the marginal increase in resources from increasing $K_1$ and $u_C\left(C_1(\omega_1)\right)$ is the marginal utility of consuming more. The former is the return on saving, which is stochastic as it depends on TFP. The latter reflects the fact that the return on capital is valued differently after different realizations of $\omega_1$. This is due to diminishing marginal utility—when TFP is high, consumption will be high and the marginal value of consuming more is low. The uncertainty in TFP generates uncertainty in $C_1(\omega_1)$, which in turn generates uncertainty in marginal utility.

The Euler equation brings us to another important point about dynamic optimization under uncertainty that turns out to be general: the return to saving is evaluated differently in different states of the world. We do not just focus on the expected return, but instead a weighted average that accounts for the different value of resources in different situations.

### 7.3.2 An infinite-horizon economy

We now consider an infinite-horizon version of the model. At each date $t$ there is a realization $\omega_t$ and the date-0 probability of a history $\omega^t$ is given by $\pi_t(\omega^t)$. The representative household has preferences given by

$$U = \sum_{t=0}^{\infty} \sum_{\omega^t \in \Omega^t} \pi_t(\omega^t) \beta^t u\left(C_t(\omega^t)\right). \tag{7.10}$$

Unlike the two-period model, consumption at date $t$ now depends on the whole history of stochastic events. At each date, a stochastic TFP is realized and used to produce output

$$Y_t(\omega^t) = A_t(\omega^t) F(K_t(\omega^{t-1}), L_t(\omega^t)),$$

where $L_t(\omega^t)$ is the labor input and $A_t(\omega^t)$ is total factor productivity. The production function is twice continuously differentiable in $K$ and $L$, is strictly increasing and strictly concave in both arguments, and is constant returns to scale. Note that the capital that is used in production at date $t$, denoted $K_t$, is selected at date $t-1$ and therefore can only depend on the information that is available at the time it is selected. Therefore $K_t$ is a function of $\omega^{t-1}$, not $\omega^t$.

The economy is endowed with one unit of labor each period and we assume, for simplicity, that there is no preference for leisure so that labor supply is inelastic and equal to one. The aggregate resource constraint at date $t$ is

$$K_{t+1}(\omega^t) + C_t(\omega^t) = f\left(A_t(\omega^t), K_t(\omega^{t-1})\right), \tag{7.11}$$

where we have defined $f(A, K) \equiv AF(K, 1) + (1-\delta)K$. The economy begins with an initial endowment of capital, $K_0$. Negative capital holdings are not possible.

The planner's problem for this economy is to maximize (7.10) subject to (7.11) where the constraint applies to each $t$ and each $\omega^t$. The Lagrangian of this problem is

$$\mathcal{L} = \sum_{t=0}^{\infty} \sum_{\omega^t \in \Omega^t} \left\{ \pi_t(\omega^t) \beta^t u\left(C_t(\omega^t)\right) - \lambda_t(\omega^t) \left[ K_{t+1}(\omega^t) + C_t(\omega^t) - f\left(A_t(\omega^t), K_t(\omega^{t-1})\right) \right] \right\}.$$

When we take the first-order conditions of this problem a key point is that our choice of $K_{t+1}(\omega^t)$ will affect production in $t+1$ for all histories $\omega^{t+1}$ that are possible given that we have already reached $\omega^t$. For example, refer back to Figure 7.1 and suppose we have reached $\omega^1 = \{0, 0\}$ at date 1 and we are choosing $K_2(\omega^1)$. This choice of capital will affect production at date 2 for histories $\{0, 0, 0\}$ and $\{0, 0, 1\}$ because these are possible following $\omega^1$. This choice of capital will not affect production for histories $\{0, 1, 0\}$ or $\{0, 1, 1\}$ because these are not possible given $\omega^1$. The first-order condition for $K_{t+1}(\omega^t)$ is therefore

$$\lambda_t(\omega^t) = \sum_{\{\omega^{t+1} | \omega^t\}} \lambda_{t+1}(\omega^{t+1}) f_K\left(A_{t+1}(\omega^{t+1}), K_{t+1}(\omega^t)\right), \tag{7.12}$$

where the notation $\{\omega^{t+1}|\omega^t\}$ indicates that we sum over the histories $\omega^{t+1}$ that are possible given $\omega^t$. The first-order condition with respect to $C_t(\omega^t)$ is

$$\pi_t(\omega^t)\beta^t u'\left(C_t(\omega^t)\right) = \lambda_t(\omega^t).$$

Using this to eliminate the Lagrange multipliers in (7.12) we arrive at the consumption Euler equation for this problem

$$u'\left(C_t(\omega^t)\right) = \beta \sum_{\{\omega^{t+1}|\omega^t\}} \pi_{t+1}(\omega^{t+1}|\omega^t) u'\left(C_{t+1}(\omega^{t+1})\right) f_K\left(A_{t+1}(\omega^{t+1}), K_{t+1}(\omega^t)\right), \tag{7.13}$$

where we have defined the conditional probability $\pi_{t+1}(\omega^{t+1}|\omega^t) = \pi_{t+1}(\omega^{t+1})/\pi_t(\omega^t)$. The consumption Euler equation has the same interpretation as in the two-period economy. The right-hand side has a weighted average of the returns on savings with weights corresponding to the marginal utility of consumption in different states of the world.

In many applications, the time subscripts on variables are sufficient to indicate the history of events they depend on in which case we can rewrite (7.13) as

$$u'\left(C_t\right) = \beta \mathbb{E}_t\left[u'\left(C_{t+1}\right) f_K(A_{t+1}, K_{t+1})\right]. \tag{7.14}$$

The $\mathbb{E}_t$ indicates we are taking conditional expectation just as our sum over $\{\omega^{t+1}|\omega^t\}$ does.

We can now use equation (7.11), at $(t, \omega^t)$ and at all the nodes in the following period, to eliminate consumption from the Euler equation (7.13). In the deterministic model, we followed the corresponding method and arrived at a second-order difference equation in capital. Here, we obtain a second-order *stochastic* difference equation in capital, which holds at all nodes in the event tree. In the deterministic model, we also had a transversality condition—as a second-order difference equation and an initial value for capital leave one degree of freedom to choose capital—and we could then pin down a solution to the difference equation. With uncertainty, the situation has much of the same structure: a transversality condition must be added to determine a solution, and it has the same interpretation as before: it is a self-imposed constraint not to over-accumulate at infinity, expressed as an *expected* present value.[9]

Clearly, solving for a stochastic sequence of capital levels appears daunting. Closed-form solutions exist in special cases, but only under very special assumptions; one is when the utility function is logarithmic, the production function is Cobb-Douglas, and $\delta = 1$. For that case it is possible to verify that a constant saving rate is optimal. In other cases, one must apply numerical methods to solve the model. One such approach is to linearize the model around the steady state of the deterministic model; we show how to do this in Section 7.3.4. If non-linearities are believed to be important, a way forward is to solve the model numerically in a dynamic-programming version of the model. We now look at how it is formulated.

[9] In the case with stochastic shocks, there is still just one degree of freedom given the first-order conditions. To see this intuitively, the Euler equation at any node $(t, \omega^t)$ can be used to solve for $k_{t+1}(\omega^t)$ as a function of $k_t(\omega^{t-1})$ and an expression involving future values $k_{t+2}(\omega^t, \omega_{t+1})$ and substituted into the previous Euler equation. When repeated infinitely many times, we have a first equation involving $k_0$ and $k_1(\omega_0)$ only—the latter is the only remaining degree of freedom.

### 7.3.3 A recursive formulation

We will now analyze a recursive version of the same economy. To do so, we will assume that TFP follows a first-order Markov process so we only need to keep track of the most recent realization in order to know the distribution of its future realizations. As recursive modeling keeps track of the history of the economy explicitly through well chosen state variables, it is customary not to use histories of stochastic events ($\omega^t$) in this context. Following this convention, we let $\pi(A'|A)$ be the probability of $A'$ occurring next period given the current TFP $A$. The planner's problem can now be expressed as the following Bellman equation

$$V(A,K) = \max_{C,K' \geq 0} \left\{ u(C) + \beta \sum_{A'} \left[ \pi(A'|A) V(A',K') \right] \right\}$$

subject to

$$K' + C = f(A,K).$$

The difference compared to dynamic programming under certainty is that now the Bellman equation has an expectation over continuation values. In the next period, there will be a value of having states $(A',K')$ but $A'$ is not known yet. As in the case with certainty, the recursive formulation delivers the same solution as the sequential formulation of the problem. It may seem surprising that this works out since here we are only taking expectations one period into the future, while in the sequential formulation we take expectations of outcomes far in the future. However, the two sets of expectations are actually the same. In the recursive formulation $V(A',K')$ is a random variable that includes a term $\mathbb{E}[V(A,'' K'')|A']$. By the law of iterated expectations, the expectation of this term conditional on $A$ becomes $\mathbb{E}[V(A,'' K'')|A]$. The same logic applies for the value function at all future dates.

To analyze the recursive economy, we can proceed with the same steps as we would for the recursive economy without uncertainty. We can substitute the constraint in to the Bellman equation

$$V(A,K) = \max_{K'} \left\{ u\left(f(A,K) - K'\right) + \beta \sum_{A'} \left[ \pi(A'|A) V(A',K') \right] \right\}$$

and take the first-order condition with respect to $K'$ to obtain

$$u'(C) = \beta \sum_{A'} \pi(A'|A) V_K(A',K'). \tag{7.15}$$

The envelope condition gives us

$$V_K(A,K) = u'(C) f_K(A,K).$$

Using this to eliminate the derivative of the value function in (7.15) we obtain a similar consumption Euler equation as we had before

$$u'(C) = \beta \sum_{A'} \pi(A'|A) \left[ u'(C') f_K(A,K) \right].$$

We can then rewrite this Euler equation as a functional equation that determines the savings policy function. To do so, let $g(A,K)$ denote the choice of $K'$ as a function of states $(A,K)$. Then write the resource constraint as $C = f(A,K) - g(A,K)$. Substituting these definitions into the Euler equation we have

$$u'\left(f\left(A,K\right)-g\left(A,K\right)\right)=\beta\sum_{A'}\pi(A'|A)\left[u'\left(\underbrace{f(A',g(A,K))-g(A',g(A,K))}_{=C'}\right)f_K(A',g(A,K))\right]. \tag{7.16}$$

This equation must hold for all $(A,K)$ and it implicitly defines the function $g(A,K)$ that is the solution to the planner's problem. In the next section, we will use this functional Euler equation to derive a complete solution to the planner's problem.

### 7.3.4 Solving the model via linearization

Section 7.3.3 derived a functional Euler equation that the solution to the planner's problem must satisfy. In this model, productivity follows an exogenous stochastic process. We will now show how we can use a linear approximation to the functional Euler equation to derive a linear stochastic difference equation that the endogenous variables in the model must follow. We will then use the properties of linear stochastic difference equations from Section 7.1.4 to derive properties of the planner's solution.

We will now assume that TFP follows an AR(1) process given by $A' = \rho A + (1-\rho)\bar{A} + \varepsilon'$ with $\mathbb{E}_t[\varepsilon'] = 0$. We will approximate the behavior of the economy around the deterministic steady state, which is the same notion of a steady state we have studied before. Now that we have shocks in the model, the interpretation of the steady state changes. The economy will never converge to the steady state if it is constantly hit by shocks.[10] The deterministic steady state is the point the economy would converge to if all shocks take their unconditional expectation forever and the agents in the model expect this.[11]

We thus take a linear approximation of (7.16) around the steady state. For a variable $X$ we will use the notation $\hat{X}_t \equiv X_t - \bar{X}$, where $\bar{X}$ is the steady state value.[12] Our linear approximation (which is tedious but straightforward to derive) is

$$(f_K - g_K)\hat{K} + (f_A - g_A)\hat{A} = \mathbb{E}\left[\begin{array}{c} f_K\left((f_K-g_K)\left(g_K\hat{K}+g_A\hat{A}\right)+(f_A-g_A)\hat{A}'\right) \\ +\frac{u'}{u''}\left(f_{KK}g_K\hat{K}+f_{KK}g_A\hat{A}+f_{KA}\hat{A}'\right) \end{array}\middle| A\right],$$

[10] In addition, the average value of variables will not coincide with those at the deterministic steady state if the model is non-linear. As the shock variance becomes smaller and smaller, however, they will become increasingly similar and, in the limit where the shock variance is zero, coincide.

[11] That the agents perceive the environment to be deterministic is a subtle but important point. There is an alternative notion of a steady state in which the agents in the model perceive there to be risk but ex post all the shocks are realized at their unconditional means. Such a steady state is sometimes called a "stochastic steady state" or a "risky steady state."

[12] Previously in the text, we used this notation for deviations in logs; both linear and log-linear deviations are used in practice.

where all derivatives are evaluated at the steady state. Notice that $\hat{A}' = \rho\hat{A} + \varepsilon'$ so the distribution of $A'$ given $A$ is determined by the distribution of $\varepsilon$. We will write the expectation as summing over $\varepsilon'$ and substitute in for $A'$. Because this is a linear equation, we can pass the expectation operator inside the right-hand side to obtain

$$(f_K - g_K)\hat{K} + (f_A - g_A)\hat{A} = \beta \left[ \begin{array}{c} f_K \left( (f_K - g_K)\left( g_K\hat{K} + g_A\hat{A} \right) + (f_A - g_A)(\rho\hat{A} + \mathbb{E}[\varepsilon']) \right) \\ + \frac{u'}{u''} \left( f_{KK} g_K \hat{K} + f_{KK} g_A \hat{A} + f_{KA}(\rho\hat{A} + \mathbb{E}[\varepsilon']) \right) \end{array} \right].$$

As $\mathbb{E}[\varepsilon'] = 0$, it is as if there is no uncertainty and $\hat{A}'$ is treated as if it is known to be at its expected value $\rho\hat{A}$. This is a general feature of analyzing a stochastic economy through linearization known as **certainty equivalence**—once one linearizes the economy, only expected values matter. In particular, the variance of the exogenous shock does not influence outcomes.

Recall that equation (7.16) must hold for all values of $A$ and $K$. The equation above is a linear approximation to (7.16) and must hold for each $\hat{K}$ and each $\hat{A}$. The only way this can be true is if the coefficients on $\hat{K}$ on the left-hand side equal those on $\hat{K}$ on the right-hand side and, similarly, the coefficients on $\hat{A}$ match. Imposing that these coefficients match gives us two equations that allow us to solve for $g_K$ and $g_A$. Conveniently, one equation contains $g_K$ only. Therefore, starting with the coefficients on $\hat{K}$ we have

$$(f_K - g_K) = \beta \left[ f_K(f_K - g_K)g_K + \frac{u'}{u''} f_{KK} g_K \right].$$

Rearrange to obtain

$$g_K^2 - \left[ 1 + \beta^{-1} + \frac{u'}{u''}\frac{f_{KK}}{f_K} \right] g_K + \beta^{-1} = 0. \tag{7.17}$$

where we have used $\beta f_K = 1$, which follows from the steady state Euler equation. Equation (7.17) is a quadratic equation in $g_K$. The equation will have one root less than one and one root greater than one. To verify this, note that the coefficient on $g_K^2$ is positive so (7.17) is an upward facing parabola as shown in shown in Figure 7.2; the quadratic intersects the y-axis at $\beta^{-1}$, which is positive; and at $g_K = 1$ the quadratic takes the value $-\frac{u'}{u''}\frac{f_{KK}}{f_K}$, which is negative if the utility and production functions are both strictly increasing and strictly concave. The quadratic therefore takes the form shown in Figure 7.2 and has one root between 0 and 1 and one greater than 1. The relevant root is the smaller one, because it is the one that is consistent with the transversality condition. The root above one is explosive in that any initial condition or shock has an increasing effect on the economy.

Solving for $g_A$ is more straightforward because matching coefficients on $g_A$ leads to a linear relationship with the solution

$$g_A = \left( f_K - g_K + 1 - \rho + \frac{u'}{u''}\frac{f_{KK}}{f_K} \right)^{-1} \left[ (1-\rho)f_A - \rho\frac{u'}{u''}\frac{f_{KA}}{f_K} \right].$$

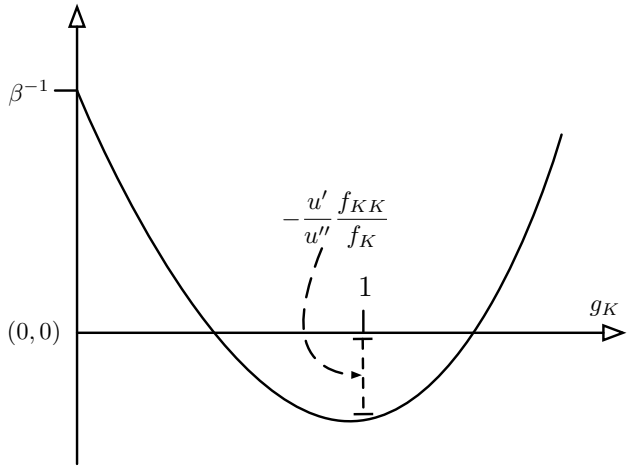



Figure 7.2: Quadratic equation to determine $g_K$ in the linearized stochastic growth model.

We have solved for the derivatives of the savings policy rule. The level of the policy rule is determined by the requirement that $(\bar{K}, \bar{A})$ is a steady state so we have

$$K' = \bar{K} + g_K(K - \bar{K}) + g_A(A - \bar{A}). \tag{7.18}$$

If we augment this equation with

$$A' = \bar{A} + \rho(A - \bar{A}) + \varepsilon', \tag{7.19}$$

we have a system of two linear stochastic difference equations and we can apply the techniques described in Section 7.1.4 to analyze the behavior of the economy.

For a numerical illustration of the properties of the stochastic growth model we will make some specific assumptions about the production function and the parameters of the model. We assume $f(A, K) = AK^\alpha + (1 - \delta)K$, with $\alpha = 0.3$, $\delta = 0.02$. The persistence of TFP is $\rho = 0.95$, the standard deviation of the innovations is 0.5%, and $\bar{A} = 1$. Finally, we assume $u(c) = \log(c)$ and $\beta = 0.99$.

With our parameterized solution to the model we can generate random draws of $\{\varepsilon_t\}_{t=0}^T$ and iterate equations (7.18) and (7.19) forward to simulate the behavior of the economy. Notice here that we are simulating the behavior of the state variables. Given the state variables it is straightforward to calculate the simulated path for output (using the production function) and the simulated path for consumption (using the aggregate resource constraint). Simulated paths for these variables are shown in Figure 7.3(a). Notice how TFP and output exhibit much more high-frequency variation than do capital and consumption. Consumption is smooth because of the diminishing marginal utility of consumption—when output is high, it is preferable to save some of the extra resources to consume them later rather than consume all of them when marginal utility is low. Capital is smooth because it reflects the accumulation of savings over many periods—one period of high investment will not make a big percentage difference to the capital stock. Panel (b) of Figure 7.3 shows the impulse response functions following a one standard deviation shock to TFP. $K_t$ is pre-determined

so it does not respond in the period the shock occurs. Therefore, on impact of the shock, output increases by the same percentage amount as TFP. Consumption increases, but not as much as output as some of the increase in output is directed to increased savings. Over time, capital increases and the increase in output exceeds that of productivity.

Figure 7.3(c) shows the results from simulating the economy for a long time and forming a histogram with the simulated data on the capital stock. The solid line in the figure shows the theoretical unconditional distribution of the capital stock as calculated from equation (7.6).[13] As the figure shows, the economy fluctuates in the vicinity of the steady state. Sometimes capital drifts higher, sometimes lower, but it tends to return towards the steady state level. Panel (d) of the figure shows why this is the case. The figure plots $g(A, K)$ as a function of $K$ for two levels of $A$—one high and one low. The dashed line is the 45-degree line. For low $K$ and high $A$, the savings policy rule is above the 45-degree line and the capital stock will increase. Similarly, for high $K$ and low $A$, the savings policy rule is below the 45-degree line and the capital stock will decrease. As $A$ fluctuates, the savings policy will shift up and down leading the capital stock to fluctuate. But note that for high $A$ and high $K$, or low $A$ and low $A$, the savings policy intersects the 45-degree line. These intersections imply that capital will not move out of this range (unless $A$ gets even higher or even lower).

## 7.4 Competitive market trade under uncertainty

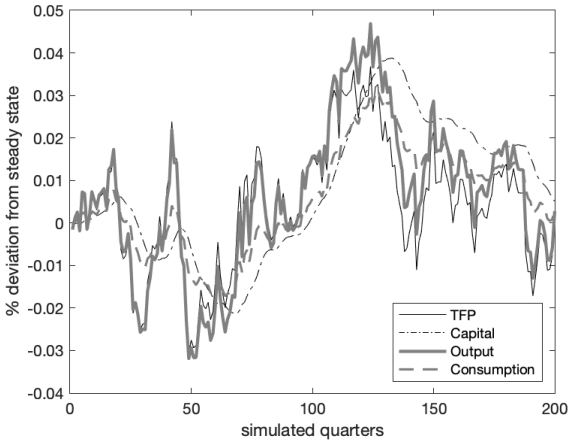
We will now begin to discuss market interactions in an uncertain economy and in the next section we will use these theoretical tools to analyze a decentralized equilibrium of the stochastic growth model. Before we get to that, we will first discuss how models of trade under uncertainty allow us to analyze the way agents insure themselves by sharing risks between them.

Broadly speaking, we can classify economic models of trade under uncertainty into two groups: complete markets models and incomplete markets models. In models with complete markets, agents can buy and sell goods with contracts tailored to every possible state of the world. They can write a contract for how they will behave after every possible history $\omega^t$. This means that any possible risk can be insured at some price. In incomplete markets models, some of these contracts are not available—some risks can not be insured against. In this chapter, we will mostly focus on complete markets models. We start here because it is simpler, not because it is more realistic, but we will introduce an incomplete markets environment in Section 7.6.

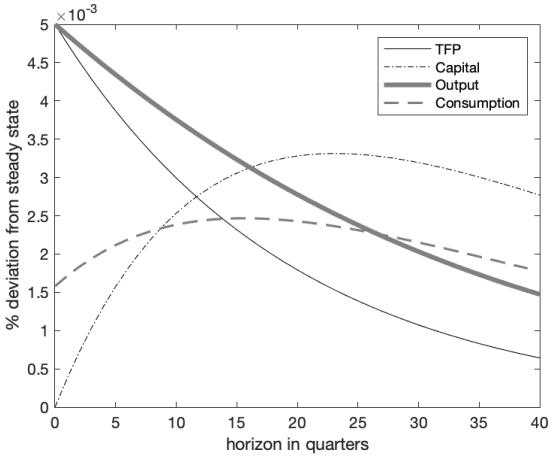
Suppose the economy is populated by a set $\mathcal{I}$ of infinitely-lived households; this set could be the continuum [0,1], as in most of Chapter 5, or it could be a different, perhaps smaller

[13] Specifically, we simulated the economy using Gaussian random variables for $\varepsilon$. As the dynamics of the economy are linear, the distribution of the state variables is also Gaussian. We use (7.6) to solve for the unconditional covariance matrix $A$ and $K$. We have plotted a Gaussian distribution with the unconditional variance of $K$ and a mean equal to the steady state capital stock.

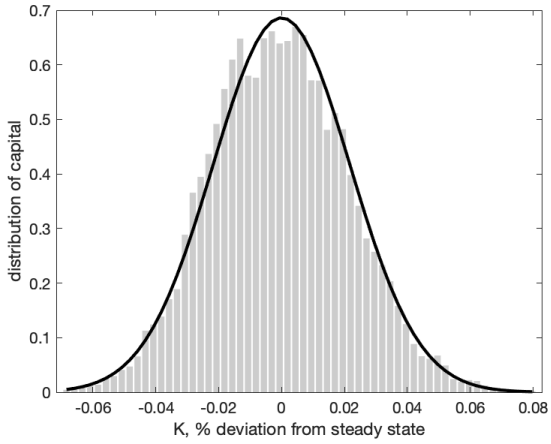
(a) Simulated sample paths




(b) Impulse response functions




(c) Distribution of $K$




(d) $g(K, A)$ for two levels of $A$

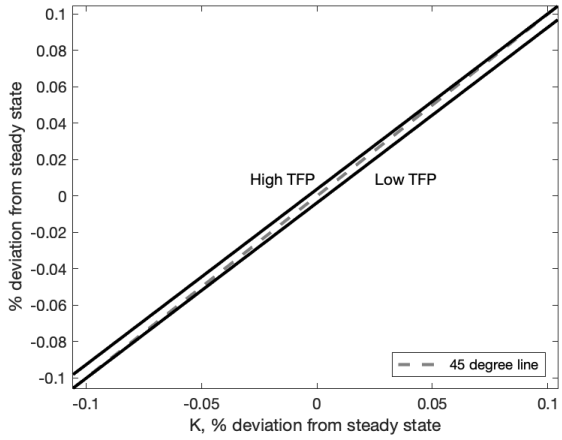



Figure 7.3: Numerical illustration of the stochastic growth model.

set. Each household has expected utility preferences given by

$$\sum_{t=0}^{\infty} \sum_{\omega^t \in \Omega^t} \pi_t(\omega^t) \beta^t u\left(c_{i,t}(\omega^t)\right),$$

where $c_{i,t}(\omega^t)$ is the consumption of household $i \in \mathcal{I}$ after history $\omega^t \in \Omega^t$. We do not spell out the nature of $\omega_t$ just yet, as it depends on the population details (the nature of the set $\mathcal{I}$). We assume $u$ is strictly increasing and strictly concave. Each household is endowed with a stochastic income stream that depends on the stochastic event. In particular let the income of household $i$ at date $t$ after history $\omega^t$ be $y_{i,t}(\omega^t)$. The good is perishable, meaning that it cannot be stored from one period to the next and must be consumed in the period it arrives in the economy. A consumption allocation is feasible if the total consumption of goods at $t$ and $\omega^t$ is equal to the total endowment of goods

$$\sum_{i \in \mathcal{I}} c_{i,t}(\omega^t) \leq \sum_{i \in \mathcal{I}} y_{i,t}(\omega^t).$$

We will begin by considering a market structure in which there is trade only at date 0. For each date $t$ and history $\omega^t$, there is a contract that says the seller will pay the buyer one unit of good at that date if that history has been realized, and nothing otherwise. These are called *Arrow securities*. We will denote the date-0 price of obtaining one unit of goods after history $\omega^t$ as $p_t(\omega^t)$. This price is denominated in terms of date-0 consumption.

The decision problem of household $i$ is to choose a contingent plan of $\{c_{i,t}(\omega^t) : \forall t, \omega^t\}$ to maximize the expected utility preferences subject to the budget constraint

$$\sum_{t=0}^{\infty} \sum_{\omega^t \in \Omega^t} p_t(\omega^t) c_{i,t}(\omega^t) = \sum_{t=0}^{\infty} \sum_{\omega^t \in \Omega^t} p_t(\omega^t) y_{i,t}(\omega^t).$$

This budget constraint says the date-0 cost of the consumption plan is less than the date-0 value of the income process. It is as if the agent, at date 0, sells claims to all their future income and then uses those funds to buy consumption goods to be delivered at future dates if particular histories occur.

Formally, we have the following.

**Definition 13** *An **Arrow-Debreu competitive equilibrium** is a set of stochastic sequences $\left\{c^\star_{i,t}(\omega^t) : \forall t, \omega^t\right\}$, for each $i \in \mathcal{I}$, and $\{p_t(\omega^t) : \forall t, \omega^t\}$ such that*

1. *for each $i$, $\left\{c^\star_{i,t}(\omega^t) : \forall t, \omega^t\right\}$ solves*

$$\max_{\{c_t(\omega^t):\forall t,\omega^t\}} \sum_{t=0}^{\infty} \sum_{\omega^t \in \Omega^t} \beta^t \pi(\omega^t) u(c_t(\omega^t)) \quad s.t. \quad \sum_{t=0}^{\infty} \sum_{\omega^t \in \Omega^t} p_t(\omega^t) c_{i,t}(\omega^t) = \sum_{t=0}^{\infty} \sum_{\omega^t \in \Omega^t} p_t(\omega^t) y_{i,t}(\omega^t)$$

2. *$\sum_{i \in \mathcal{I}} c^\star_{i,t}(\omega^t) di = \sum_{i \in \mathcal{I}} y_{i,t}(\omega^t) di$ for all $(t, \omega^t)$.*

To characterize the equilibrium, the Lagrangian of household $i$ is

$$\mathcal{L} = \sum_{t=0}^{\infty} \sum_{\omega^t \in \Omega^t} \beta^t \pi_t(\omega^t) u(c_{i,t}(\omega^t)) - \lambda_i \left[ \sum_{t=0}^{\infty} \sum_{\omega^t \in \Omega^t} p_t(\omega^t) \left( c_{i,t}(\omega^t) - y_{i,t}(\omega^t) \right) \right],$$

where $\lambda_i$ is the Lagrange multiplier on the date-0 budget constraint. The first-order condition of household $i$ with respect to $c_{i,t}(\omega^t)$ is

$$\beta^t \pi_t(\omega^t) u'(c_{i,t}(\omega^t)) = \lambda_i p_t(\omega^t) \tag{7.20}$$

We can understand several properties of the equilibrium consumption allocation from the first-order condition.

**Insurance** Consider equation (7.20) for two different values of $\omega^t$, call them $\omega^t$ and $(\omega^t)'$ and take the ratio of these two equations to arrive at

$$\frac{\beta^t \pi_t(\omega^t) u'(c_{i,t}(\omega^t))}{\beta^t \pi_t((\omega^t)') u'(c_{i,t}((\omega^t)'))} = \frac{\lambda_i p_t(\omega^t)}{\lambda_i p_t((\omega^t)')}.$$

If the prices satisfy $p_t(\omega^t) = \bar{p}_t \times \pi_t(\omega^t)$ with $\bar{p}_t > 0$, which we call **actuarially fair prices**, we have

$$\begin{aligned} u'(c_{i,t}(\omega^t)) =& u'(c_{i,t}((\omega^t)')) \\ c_{i,t}(\omega^t) =& c_{i,t}((\omega^t)'), \end{aligned}$$

where the second line follows from $u(\cdot)$ being strictly concave.[14] With actuarially fair prices, the households will buy full insurance and consumption does not depend on $\omega^t$. Full insurance is, of course, only feasible in equilibrium if total resources do not vary across the different values of $\omega^t$ so in general, prices have to adjust to reflect not just probabilities, but also relative scarcity, across states.

**Risk sharing** For some $\omega^t$, take the ratio of equation (7.20) for household $i$ and household $j$

$$\frac{u_c(c_{i,t}(\omega^t))}{u_c(c_{j,t}(\omega^t))} = \frac{\lambda_i}{\lambda^j}.$$

Now solve for the consumption of household $i$ in terms of that of household $j$

$$c_{i,t}(\omega^t) = u_c^{-1}\left( \frac{\lambda_i}{\lambda^j} u_c(c_{j,t}(\omega^t)) \right).$$

[14] In a static context, an actuarially fair gamble is one in which the expected payoff is equal to the cost of the gamble. To adapt this definition to a dynamic context, let's say prices are actuarially fair if the price of any payoff at date $t$ is equal to the price of any other payoff at date $t$ that has the same expected payoff. The condition $p_t(\omega^t) = \bar{p}_t \times \pi_t(\omega^t)$ imposes this definition. To see this, consider a set of payoffs at different histories that can arise at date $t$ given by $x_t(\omega^t)$. The date-0 value of such a portfolio of claims is $\sum_{\omega^t} x_t(\omega^t) p_t(\omega^t) = \bar{p}_t \mathbb{E}[x_t(\omega^t)]$, which only depends on the date and the expected payoff.

Goods market clearing requires $\sum_i c_{i,t}(\omega^t) = \sum_i y_{i,t}(\omega^t)$, so we obtain

$$\sum_i u_c^{-1}\left(\frac{\lambda_i}{\lambda^j} u_c(c_{j,t}(\omega^t))\right) = \sum_i y_{i,t}(\omega^t). \tag{7.21}$$

Equation (7.21) relates the consumption of household $j$ to the aggregate supply of goods and the Lagrange multipliers of all the households. Importantly, those Lagrange multipliers are constant across time so $c_{j,t}(\omega^t)$ varies over time as a function of aggregate income not as a function of $y_{j,t}$. This is a very important result for complete markets models: all idiosyncratic risk is insured away and consumption fluctuations only reflect aggregate risks.

**Aggregate and idiosyncratic risks** Aggregate risks lead to movements in aggregate variables such as aggregate income or aggregate consumption while idiosyncratic risks affect an individual's circumstances but do not affect the aggregate. One person becoming unemployed is an example of an idiosyncratic shock while an event that changes the unemployment rate is an example of an aggregate shock. If the set of consumers, $\mathcal{I}$, is finite then an individual shock, by definition, is an aggregate shock, albeit a small one if $\mathcal{I}$ has many elements. If $\mathcal{I} = [0, 1]$, then the set is (even uncountably) infinite and one individual's shock is truly idiosyncratic, unless it is synchronized/correlated with the shocks of others. To illustrate, an interesting case is precisely that where the shock is "employment" (say, with income $y_e$ for the individual) or "unemployment" (with income $y_u < y_e$). The date-$t$ shock $\omega_t$ would then specify a whole function taking, for each $i \in [0, 1]$, the value $e$ or $u$. The whole event tree would even be hard to imagine. A special case is where individuals' employment outcomes are independent draws with probabilities $\pi_e$ and $\pi_u = 1 - \pi_e$, respectively. Then if we can appeal to a law of large numbers there would be no aggregate uncertainty: aggregate resources would be a deterministic value $\pi_e y_e + \pi_u y_u$. However, each individual faces uncertainty, though in this case markets can allow full insurance. This kind of model, where the law of large numbers is assumed to hold, is often used in macroeconomics.[15] Now imagine that $\pi_e$ is random: an aggregate shock, which itself could, e.g., take on two values (say, high or low) as well vary over time. Then individuals' shocks are correlated, though if one conditions on the aggregate shock (high or low unemployment), their shocks can be thought of as purely idiosyncratic and uncorrelated. Our notation involving $\mathcal{I}$ and $\omega^t$ is abstract and meant to capture all these possibilities.

**Sequential trading** We can implement a complete set of markets with an alternative trading arrangement in which agents only trade securities that pay off in the next period and then trade again every period. This parallels our two ways to define equilibrium in deterministic contexts in Sections 5.2–5.3: we now merely have stochastic sequences.

For each event $\omega_{t+1}$ that can occur at $t + 1$, there is an asset traded at $t$ that pays one unit at $t + 1$ if that event occurs and zero otherwise. These are known as Arrow securities. Let $q_t(\omega_{t+1}|\omega^t)$ be the price at $t$ of a unit of consumption at $t + 1$ if event $\omega_{t+1}$ occurs. This

[15] Such an assumption involves mathematical subtleties; see, e.g., Uhlig (1996).

price can depend on the history leading up to date $t$, $\omega^t$, and is denominated in terms of consumption after history $\omega^t$. Let $a_{i,t+1}(\omega^{t+1})$ be the amount of this asset held by household $i$. The budget constraint of the household is then

$$c_{i,t}(\omega^t) + \sum_{\omega_{t+1}} q_t(\omega_{t+1}|\omega^t) a_{i,t+1}(\omega^{t+1}) \leq y_{i,t}(\omega^t) + a_{i,t}(\omega^t). \tag{7.22}$$

Financial wealth, $a_{i,t}(\omega^t)$, becomes a state variable for the household's problem. This wealth allows the household to consume more than its income stream in the current period and future periods. When financial wealth is negative, the household must consume less than its income either now or in the future.

While agents only trade assets that pay off one period in the future, they can use these asset prices to value payoffs further in the future. One unit of goods at $t+2$ after history $\omega^{t+2}$ has a value in date $t$ of $q_{t+1}(\omega_{t+2}|\omega^{t+1}) \times q_t(\omega_{t+1}|\omega^t)$. In this product, the first term discounts the unit of goods back to $t+1$ and the second term discounts it from $t+1$ to $t$. In general, we can define these discounts recursively as

$$\tilde{q}^t_{\tau+1}(\omega^{\tau+1}) = q_\tau(\omega_{\tau+1}|\omega^\tau)\tilde{q}^t_\tau(\omega^\tau)$$

with $\tilde{q}^t_t(\omega^t) = 1$. While the households do not trade assets for dates $\tau > t+1$ at date $t$, they do correctly anticipate the prices that will prevail in the future and the $\tilde{q}$ terms reflect these expectations.

At date-0, we assume that households have no financial wealth positive or negative because we assume that no trades have occurred prior to date-0 and so no household has a financial claim on any other household. Similar to models without uncertainty, the no Ponzi game constraint requires that the household could repay if it consumes nothing forever

$$a_{i,t}(\omega^t) \geq -\sum_{\tau=t}^{\infty} \sum_{\omega^\tau} \tilde{q}^t_\tau(\omega^\tau) y_{i,\tau}(\omega^\tau). \tag{7.23}$$

Notice that this constraint rules out Ponzi games: it is the "natural borrowing limit," discussed in Section 4.3.1, now applying state by state.

We can now write the household's problem as

$$\max_{\{c_{i,t}(\omega^t), a_{i,t}(\omega_t)\}_{\forall t,\omega^t}} \sum_t \sum_{\omega^t} \beta^t \pi_t(\omega^t) u(c_{i,t}(\omega^t))$$

such that (7.22) and (7.23) hold for all $t$ and $\omega^t$.

A competitive equilibrium is a consumption allocation $c_{i,t}(\omega^t)$ for all $i$, $t$, and $\omega^t$; asset positions $a_{i,t}(\omega^t)$ for all $i$, $t$, and $\omega^t$; a price system $q_t(\omega_{t+1}|\omega^t)$ for all $t$, $\omega_{t+1}$ and $\omega^t$ such that (i) for all $i$, the consumption-savings plan is optimal taking the prices, borrowing constraints, and $a_{i,0} = 0$ as given; (ii) for all $t$ and $\omega^t$, the goods markets clear: $\sum_i (c_{i,t}(\omega^t) - y_{i,t}(\omega^t)) = 0$; (iii) for all $t$ and $\omega^t$, the asset market clears $\sum_i a_{i,t}(\omega^t) = 0$.

**Definition 14** *A **sequential competitive equilibrium** is a set of stochastic sequences* $\left\{c^\star_{i,t}(\omega^t) : \forall t, \omega^t\right\}$ *and* $\left\{a^\star_{i,t+1}(\omega^t) : \forall t, \omega^t\right\}$ *for each* $i \in \mathcal{I}$*, and* $\{q_t(\omega^t) : \forall t, \omega^t\}$ *such that*

1. *for each* $i$, $\left(\left\{c^\star_{i,t}(\omega^t) : \forall t, \omega^t\right\}, \left\{a^\star_{i,t+1}(\omega^t) : \forall t, \omega^t\right\}\right)$ *solves*

$$\max_{\{c_t(\omega^t):\forall t,\omega^t\},\{a_{t+1}(\omega^t):\forall t,\omega^t\}} \sum_{t=0}^{\infty} \sum_{\omega^t \in \Omega^t} \beta^t \pi(\omega^t) u(c_t(\omega^t))$$

*subject to*

$$c_{i,t}(\omega^t) + \sum_{\omega_{t+1}} q_t(\omega_{t+1}|\omega^t) a_{i,t+1}(\omega^{t+1}) = y_{i,t}(\omega^t) + a_{i,t}(\omega^t) \tag{7.24}$$

*and (nPg)*

$$a_{i,t}(\omega^t) \geq -\sum_{\tau=t}^{\infty} \sum_{\omega^\tau} \tilde{q}^t_\tau(\omega^\tau) y_{i,\tau}(\omega^\tau) \tag{7.25}$$

2. $\sum_i \left(c^\star_{i,t}(\omega^t) - y_{i,t}(\omega^t)\right) = 0$ *and* $\sum_i a^\star_{i,t+1}(\omega^t) = 0$ *for all* $t$ *and* $\omega^t$.

Here, requirement *2* has two conditions (for each date and state); one of these implies the other.

As in the case of certainty, the equilibrium allocation under sequential trading is the same as the one that arises with date-0 trading. To prove this, the key step is to add the stochastic sequence of budget constraints, multiplied by the appropriate prices, to arrive at a time-zero consolidated constraint (after using the nPg constraint). Then, after seeing how the prices in the two settings map into each other, it becomes clear that the consumers solve the same problems in the two equilibrium definitions.

**Spanning and complete markets** Arrow securities are a convenient modeling device, but they are not recognizable as assets that we normally trade. Most real-world assets pay off in more than one state of nature. A system of markets would still be complete if we can construct portfolios of the available assets that have payoffs equivalent to a full set of Arrow securities.

Suppose there are $S$ states of the world that might be realized at date $t+1$ and there are $N$ assets traded at each date. We can construct the $S \times N$ payoff matrix $D$ that lists what each asset pays in each state. A portfolio is a vector $\theta \in \mathbb{R}^N$ that lists the weights on each asset. The $S \times 1$ vector listing the payoff of a portfolio $\theta$ in each state of the world is given by $D\theta$.

The range of the matrix $D$ is the space of all possible payoffs that can be constructed by making portfolios of the $N$ assets. Let

$$\mathcal{M} \equiv \left\{z \in \mathbb{R}^S : z = D\theta \text{ for some } \theta \in \mathbb{R}^N\right\}.$$

If $\mathcal{M} = \mathbb{R}^S$, the system of markets is complete. In this sense, a complete market means that one can construct a portfolio with any conceivable payoff vector. A system of markets will be complete if and only if $\text{rank}(D) = S$. If this rank condition is satisfied by the $N$ assets, we say the assets **span the payoff space**. If a system of markets is complete, there are portfolios $\{\theta^{\text{A}}_j\}_{j=1}^S$ such that $D\theta^{\text{A}}_j$ pays one unit if state $j$ occurs and zero units otherwise. The portfolios $\{\theta^{\text{A}}_j\}_{j=1}^S$ are the Arrow securities.

## 7.5 Competitive equilibrium in the growth model

We now use the framework of trade under uncertainty we just developed to define a competitive equilibrium for the stochastic growth model. We will state all of the assumptions here even though some of them were already introduced in Section 7.3.

We assume, for simplicity, that all individuals are identical. The representative household is endowed with $k_0$ units of capital and one unit of labor that is supplied inelastically. The intertemporal utility function is

$$\sum_t \sum_{\omega^t} \beta^t \pi_t(\omega^t) u(c_t(\omega^t)).$$

Output is produced according to

$$y_t(\omega^t) = A_t(\omega^t) F(k_t(\omega^{t-1}), 1),$$

where $F$ has the usual neoclassical properties. The aggregate resource constraint is

$$k_{t+1}(\omega^t) + c_t(\omega^t) = (1-\delta) k_t(\omega^{t-1}) + y_t(\omega^t).$$

Turning to markets, the representative household accumulates capital and rents it to a representative firm in a spot market at price $r_t(\omega^t)$. The total, gross, return is $r_t(\omega^t) + 1 - \delta$, thus also inclusive of the undepreciated capital. Similarly, the household rents its labor to the firm in a spot market at a price $w_t(\omega^t)$.

The firm's problem is exactly as we have discussed before and results in the first-order conditions

$$r_t(\omega^t) = A_t(\omega^t) F_k(k_t(\omega^{t-1}), 1) \tag{7.26}$$

and

$$w_t(\omega^t) = A_t(\omega^t) F_\ell(k_t(\omega^{t-1}), 1). \tag{7.27}$$

The household's budget constraint is

$$c_t(\omega^t) + k_{t+1}(\omega^t) = (r_t(\omega^t) + 1 - \delta) k_t(\omega^{t-1}) + w_t(\omega^t).$$

We could allow the household to trade Arrow securities contingent on $\omega^{t+1}$ but, without another party to trade with, the representative household must have a zero position in each security in equilibrium. Therefore, although the consumer faces incomplete markets here, a full set of state-contingent assets would not change the equilibrium allocation. The household cannot hold a negative capital position, $k \geq 0$, but we will assume this constraint does not bind.

The Lagrangian of the household's problem is

$$\mathcal{L} = \sum_{t=0}^{\infty} \sum_{\omega^t \in \Omega^t} \left\{ \beta^t \pi_t(\omega^t) u(c_t(\omega^t)) - \lambda_t(\omega^t) \left[ c_t(\omega^t) + k_{t+1}(\omega^t) - (r_t(\omega^t) + 1 - \delta) k_t(\omega^{t-1}) + w_t(\omega^t) \right] \right\}.$$

Taking the first-order conditions with respect to $c_t(\omega_t)$ and $k_{t+1}(\omega^t)$ we have

$$\begin{aligned}\beta^t \pi_t(\omega^t) u'(c_t(\omega^t)) &= \lambda_t(\omega^t) \\ \lambda_t(\omega^t) &= \sum_{\omega^{t+1}|\omega^t} (r_{t+1}(\omega^{t+1}) + 1 - \delta)\lambda_{t+1}(\omega^{t+1})\end{aligned}$$

and combining these we arrive at an Euler equation of

$$u'\left(c_t(\omega^t)\right) = \mathbb{E}_t\left[\beta u'\left(c_{t+1}(\omega^{t+1})\right)\left(r_{t+1}(\omega^{t+1}) + 1 - \delta\right)\right]. \tag{7.28}$$

A competitive equilibrium of this economy is a set of stochastic processes

$$\{r_t(\omega^t), w_t(\omega^t), c_t(\omega^t), k_{t+1}(\omega^t)\}_{\forall t, \omega^t}$$

such that $c_t(\omega^t)$ and $k_{t+1}(\omega^t)$ are optimal in the household's problem given the prices, the prices are set by competitive profit-maximizing firms in accordance with equations (7.26) and (7.27) and the resource constraint is satisfied.

**Definition 15** *A **sequential competitive equilibrium** is a set of stochastic sequences* $\{c_t^\star(\omega^t) : \forall t, \omega^t\}$ *and* $\left\{k_{t+1}^\star(\omega^t) : \forall t, \omega^t\right\}$ *and* $\{(r_t(\omega^t), w_t(\omega^t) : \forall t, \omega^t\}$ *such that*

1. $\left(\{c_t^\star(\omega^t) : \forall t, \omega^t\}, \left\{k_{t+1}^\star(\omega^t) : \forall t, \omega^t\right\}\right)$ *solves*

$$\max_{\{c_t(\omega^t):\forall t,\omega^t\},\{k_{t+1}(\omega^t):\forall t,\omega^t\}} \sum_{t=0}^{\infty} \sum_{\omega^t \in \Omega^t} \beta^t \pi(\omega^t) u(c_t(\omega^t))$$

*subject to the nPg constraint and*

$$c_{i,t}(\omega^t) + k_{i,t+1}(\omega^{t+1}) = (r_t(\omega^t) + 1 - \delta)k_t(\omega^{t-1}) + w_t(\omega^t) \tag{7.29}$$

2. *for all* $t$ *and* $\omega^t$,

$$r_t(\omega^t) = A_t(\omega^t) F_k(k_t^\star(\omega^{t-1}), 1) \quad \text{and} \quad w_t(\omega^t) = A_t(\omega^t) F_\ell(k_t^\star(\omega^{t-1}), 1)$$

3. *for all* $t$ *and* $\omega^t$,

$$k_{t+1}^\star(\omega^t) + c_t^\star(\omega^t) = (1 - \delta)k_t^\star(\omega^{t-1}) + A_t(\omega^t) F(k_t^\star(\omega^{t-1}), 1).$$

In this definition of equilibrium the last condition is superfluous. This is, as in the deterministic case, because a constant returns to scale production function is homogeneous of degree one and by Euler's theorem we therefore have

$$r_t(\omega^t)k_t(\omega^{t-1}) + w_t(\omega^t) = A_t(\omega^t)F(k_t(\omega^{t-1}), 1) = Y_t(\omega^t).$$

If we substitute this into the household budget constraint we arrive at the aggregate resource constraint, which is the goods market clearing condition.

The competitive equilibrium allocation is the same as we obtain from the planner's problem. If we substitute equation (7.26) into (7.28) and we arrive at the same Euler equation as we obtained in the planner's problem, equation (7.13). Furthermore, if we substitute into (7.26) and (7.27) into the household budget constraint, we obtain the same resource constraint as applies to the planner's problem.

The equivalence between the planner's solution and the competitive equilibrium allocation is not affected by adding uncertainty to the model. Indeed, there is nothing fundamentally different about an economy with uncertainty as compared to a deterministic one. Instead of indexing goods just by time, we now index goods by time and histories. As a result, the first and second welfare theorems continue to apply.

In Section 7.3 we used the functional Euler equation to solve the planner's problem for the stochastic growth model. Due to the equivalence between the competitive equilibrium and the planner's problem, the solution we found is also the solution to the competitive equilibrium. In fact, in Appendix 7.A we formulate a recursive competitive equilibrium and derive the exact same functional Euler equation as we found for the planner's problem.

## 7.6 An incomplete-market economy

In this chapter, we have mostly focused on competitive equilibria with complete markets. If a system of markets is incomplete, i.e., if some insurance contracts are not available, then the planner's problem and the competitive equilibrium will not coincide in general. In this case, the first welfare theorem fails to hold because some markets are missing (the markets for those insurance contracts). Here we will just give a brief introduction to incomplete-market models and return to this topic in later chapters.

Consider a consumer with preferences given by

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t u(c_t).$$

The consumer receives a stochastic income stream $y_t$ and can borrow or save in an asset that pays gross interest $1+r$, which is known and constant. Here we just describe the consumer's decision problem and we will take $r$ as given.

Letting $a_t$ be the agent's assets at the start of period $t$ and $q = 1/(1+r)$, the budget constraint is

$$qa_{t+1} + c_t = a_t + y_t.$$

We assume there is some lower limit to how much the consumer can borrow, $a_{t+1} \geq \underline{a}$, where $\underline{a}$ could be the natural borrowing limit (the amount the consumer could repay if they received the lowest possible income realization in all future periods) or some more restrictive borrowing constraint.

If we assume the endowment process is a first-order Markov process, the consumer's problem can be stated recursively as

$$V(a,y) = \max_{a' \geq \underline{a}} \{u(a+y-qa') + \beta \mathbb{E}\left[V(a',y')|y\right]\}.$$

The first-order condition of this problem is

$$u'(c)q = \beta \mathbb{E}\left[V_a(a', y')|y\right]$$

and the envelope condition is

$$V_a(a, y) = u'(c).$$

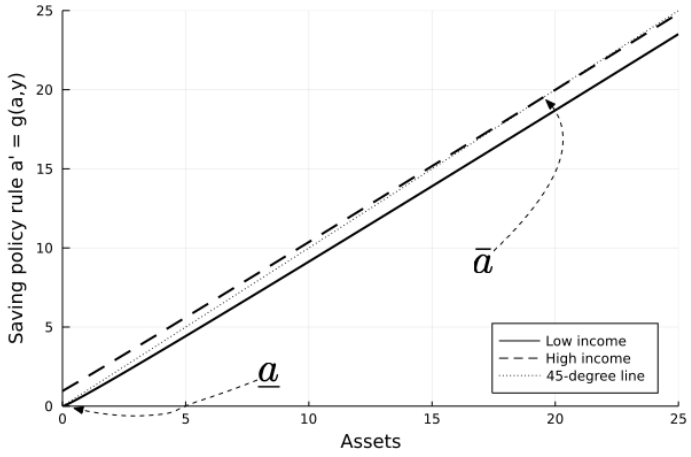
Combining these we have the Euler equation

$$u'(c) = (1 + r)\beta \mathbb{E}\left[u'(c')|y\right].$$

This is an example of an incomplete-market environment because the single asset cannot insure the consumer against the income risk. In other terms, using the single asset with a constant return, the consumer can borrow and save but the payoff of their portfolio is independent of their income in the next period. As a result, their consumption will fluctuate in response to the realizations in their individual endowment. However, the consumers can partially self insure. They can accumulate savings that they can spend down if they receive low endowments in the future. In this way the consumer can partially smooth their consumption. However, they will not in general choose to fully smooth their consumption. To see this, suppose a consumer wished to have a perfectly smooth consumption path. The only way they could do this is to set their consumption to a level that would be sustainable if they received the lowest possible income realizations at all future dates. If they choose a higher level, there would be a chance that they would be unlucky and have to reduce their consumption. But this extremely conservative plan will not be optimal even for a consumer with very high risk aversion. Instead, the consumer will choose to adjust their consumption level in response to the endowments they receive. In Chapter 9 we study problems of self insurance in much more detail.

The top panel of Figure 7.4 plots the decision rule $a' = g(a, y)$ for a version of this model in which income can take two values each period. When income is high, the consumer accumulates savings up to the point $\bar{a}$ where the upper line crosses the 45-degree line. When income is low, the consumer spends down their savings until it reaches the borrowing constraint. If the consumer begins with initial assets above $\bar{a}$, they will continuously spend down their assets regardless of their income until their assets reach $\bar{a}$. So in the long run the consumer will have asset holdings on the ergodic set $[\underline{a}, \bar{a}]$. Within this set, however, the consumer will sometimes be moving towards higher asset levels (when they have high income) and will sometimes be moving towards lower levels (when they have low income). Contrast this saving rule with what we found in our deterministic steady state endowment economy in Section 5.4.1 where the decision rules were $a' = g(a)$ so they are simply the 45-degree line. In Figure 7.4, the decision rule of high-income consumers has a slope less than 45 degrees but an intercept above the 45 degree line while the decision rule of low-income consumers is on the 45-degree line at $\underline{a}$ but has a slope less than 45 degrees. The fact that the decision rules are above and below the 45-degree line is the source of partial insurance—those with high income put some of their "extra" resources into savings while those with low income draw down their savings.

(a) Decision rule for net saving $g(a, y) - a$




(b) Simulated stationary distribution of assets

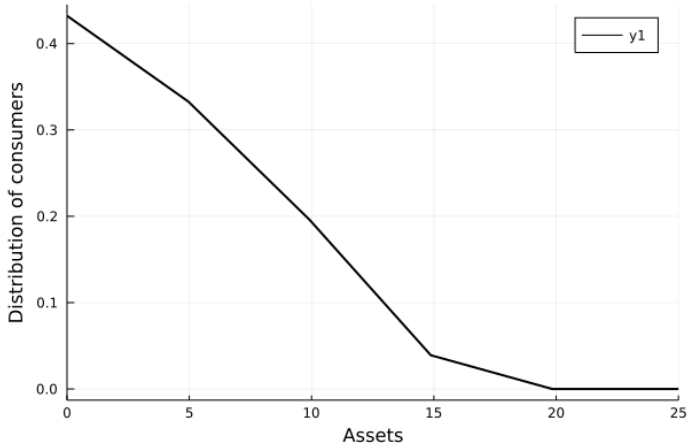



Figure 7.4: Decisions rules and stationary distribution for the incomplete-market consumption-savings problem.

Now suppose we simulate a large number of consumers each solving this decision problem and each receiving their own independent endowment process. The simulation produces a distribution of consumers over asset levels that reflects the fact that some consumers have been lucky and received many high endowments and some have been unlucky. Moreover, as consumption is a function of the consumer's assets and income, the simulation also produces a distribution of consumption levels. At a point in time, the distribution of asset holdings will depend in part on the distribution of asset holdings in the previous period. In the long run, the distribution will converge to a stationary distribution. The bottom panel of Figure 7.4 plots the stationary distribution of assets for our simulation.

In some contexts the complete market and representative agent assumptions are sensible simplifications of reality, but when they are not we often turn to incomplete-market models that build on the framework we have just presented. Such models can be used to study a wide range of issues and we will return to this topic in the chapters that follow.

# Chapter 8

# Empirical strategies and quantitative macroeconomics

# Part III

# Applications

# Chapter 9

# Consumption

*Gianluca Violante*

## 9.1 Introduction

Household consumption is a key determinant of welfare and, as a result, it plays a fundamental role in many areas of macroeconomics, such as growth, business cycles, inequality, taxation, and asset pricing. The growth of aggregate consumption over time is an unequivocal sign of augmented prosperity of a society. Because its fluctuations over the business cycle are costly to households, both fiscal and monetary policy go to great lengths in order to stabilize aggregate consumption expenditures. The distribution of consumption in the population is a credible measure of inequality in standard of living across households, more so than income. Crucial objectives of redistributive and social insurance policies are that of supporting a minimum level of consumption above poverty for all households, and that of limiting the pass through of income losses to household spending (and thus, to their well-being). Finally, consumption choices of investors over time and across states determine stochastic discount factors which price assets in financial markets.

In light of this centrality, it is not surprising that the theory and empirics of consumption choices has, historically, attracted so much attention from economists. As Deaton (1992) puts it, in the preface of his book, *attempts by economists to understand the saving and consumption patterns of households have generated some of the best science in economics.* The desire to microfound the empirical relation between consumption and income, which contradicted the simple Keynesian consumption function where expenditures are modelled as a linear function of current income, led to some of the first examples of forward-looking dynamic optimization (Modigliani and Brumberg, 1954; Friedman, 1957) and, since then, led to gradual enrichment of the optimization framework. At the same time, the ever wider availability of large microeconomic data sets on income and expenditures (first survey data, now administrative data) created fertile grounds for the application of state-of-the art econometric techniques to test models and quantify key magnitudes. In the last 30 years, this theoretical and empirical advances have been incorporated into general equilibrium models with heterogeneous households that constitute one of the main workhorses for the study of business cycle, inequality and government policy in macroeconomics.

This chapter offers an introduction to this vast literature. It emphasizes the theoretical advancements in this field, but it also makes an attempt to relate models to data. The chapter is organized as follows. Section 9.2 derives consumption allocations under two extreme financial market structures: autarky and complete markets. It then argues that the empirical evidence suggests a market structure in between these two which offers "partial insurance" against income shocks, and introduces an economy where only a non-state-contingent bond is traded. Section 9.3 studies in some depth the optimal intertemporal consumption/saving problem of a household who can save and borrow through this asset, the so called "income fluctuation problem". We start from the deterministic case, and then we analyze the stochastic case with random income fluctuations. The permanent income hypothesis, where certainty equivalence holds and risk plays no role for consumption allocation, is a special case of this environment. We then move beyond certainty equivalence and analyze environments where risk matters for consumption and saving. We analyze the two sources of precautionary saving, prudence and occasionally binding borrowing constraints, and explain

how they induce concavity in the consumption function. In Section 9.4, we combine a continuum of households facing income fluctuation problems, and study how they give rise to an endogenous joint distribution of income, consumption, and wealth. We then characterize the stationary equilibrium of these economies, first without and then with production. This last section extends the analysis of Section 7.6.

Because of space constraints, we omit developing a number of interesting and important topics related to consumption. For example, business cycles and asset pricing, life-cycle patterns, information frictions, consumer durables and housing, bequest, habits, and non-standard preferences. Some of these topics will be covered in later chapters, such as Chapters 12, 14 and 19. We also refer the reader to the surveys by Hall (1988), Muellbauer (1994), Browning and Lusardi (1996), Attanasio (1999), Browning and Crossley (2001), Campbell (2003), Attanasio and Weber (2010), Meghir and Pistaferri (2011), Piazzesi and Schneider (2016), Kaplan and Violante (2022), and the thematic books by Deaton (1992) and Jappelli and Pistaferri (2020) for additional material.

## 9.2 Consumption under autarky and full insurance

Consider an endowment economy with aggregate uncertainty, as the one outlined in Chapter 7 of the book. Let $\omega_t \in \Omega$ (a finite set) be the realization of a stochastic event (e.g., an aggregate shock) at date $t$. Let $\omega^t = \{\omega_0, \omega_1, ..., \omega_t\}$ be the history of events until time $t$, with $\omega^t \in \Omega^t \equiv \Omega \times \Omega \times ... \times \Omega$, the $t+1$ Cartesian product of $\Omega$. Each history $\omega^t$ has unconditional probability of occurring $\pi(\omega^t)$. We assume all households have rational expectations, i.e. they forecast by using the true probability distribution. The economy is populated by a continuum of measure one of infinitely-lived households indexed by $i$ who are endowed with stochastic income $y_{i,t}(\omega^t)$ such that

$$\int_0^1 y_{i,t}(\omega^t)\, di = Y_t(\omega^t),$$

where $Y_t(\omega^t)$ is the (random) aggregate endowment of the economy. Each realization $\omega_t \in \Omega$ corresponds to a particular value of the aggregate endowment and a particular distribution of it across households.[1]

Households are expected utility maximizers, with period utility $u(c_{i,t}(\omega^t))$, where $c_{i,t}(\omega^t)$ is consumption of individual $i$ upon realization of history $\omega^t$. We assume that $u$ satisfies standard properties, i.e. $u' > 0$ and $u'' < 0$. Let $C_t(\omega^t) = \int_0^1 c_{i,t}(\omega^t)\, di$ denote aggregate consumption. Aggregate feasibility implies that aggregate consumption equals the aggregate endowment along each history

$$C_t(\omega^t) = Y_t(\omega^t) \text{ for all } t, \omega^t \in \Omega^t.$$

[1] For example, with $I = 2$ and $\omega_t \in \{\omega^L, \omega^H\}$, we could have a configuration where $Y(\omega^L) = 2, Y(\omega^H) = 4$, and $y_1(\omega^L) = 2, y_2(\omega^L) = 0$, $y_1(\omega^H) = 1, y_2(\omega^H) = 3$. Thus, the aggregate endowment in state $H$ is larger than in state $L$, but type 1 is better off in state $L$ and type 2 in state $H$.

The individual consumption allocation that arises in the equilibrium of this economy depends on market arrangements. We consider two extreme benchmarks: autarky and full insurance.

Under autarky, there is no insurance market which allows individuals to trade across states, and no storage technology to transfer resources across time (e.g., the endowment is fully perishable). In this economy, an individual $i$ who receives a random stream of income shocks $\left\{\left\{y_{i,t}\left(\omega^t\right)\right\}_{\omega^t\in\Omega^t}\right\}_{t=0}^{\infty}$ has no other choice than consuming their income in every state:

$$c_{i,t}\left(\omega^t\right) = y_{i,t}\left(\omega^t\right), \text{ for all } t, \omega^t \in \Omega^t. \tag{9.1}$$

and equation (9.1) is also their budget constraint. Under autarky there is full pass-through of individual income shocks into consumption, or no consumption smoothing whatsoever.

Consider now the other end of the spectrum of market arrangements: complete markets (also called full insurance or full risk-sharing), introduced earlier in Section 7.4. Under this arrangement, households can trade a complete set of Arrow securities. This market structure allows every individual $i$ to achieve any transfer of income across states and across time, as long as these trades respect the time-zero Arrow-Debreu budget constraint

$$\sum_{t=0}^{\infty}\sum_{\omega^t\in\Omega^t} p_t\left(\omega^t\right)\left[c_{i,t}\left(\omega^t\right) - y_{i,t}\left(\omega^t\right)\right] = 0. \tag{9.2}$$

The Arrow-Debreu competitive equilibrium for this economy was defined in Section 7.4, where it is shown that for any pair of households $(i,j)$,

$$\frac{u'\left(c_{i,t}\left(\omega^t\right)\right)}{u'\left(c_{j,t}\left(\omega^t\right)\right)} = \frac{\lambda_i}{\lambda_j}, \text{ for all } t, \omega^t \in \Omega^t, \text{ and } (i,j), \tag{9.3}$$

with $\lambda_i$ and $\lambda_j$ representing Lagrange multipliers on their lifetime budget constraint.

Equation (9.3) illustrates the defining property of consumption allocations under full insurance: when households can trade a complete set of Arrow securities, *the ratio of marginal utility of consumption of any two households is constant across time and states.*

To make further analytical progress, note that in the special case of CRRA utility, where

$$u\left(c_{i,t}\left(\omega_t\right)\right) = \frac{c_{i,t}\left(\omega_t\right)^{1-\sigma}}{1-\sigma}, \text{ with } \sigma \in [0,\infty)$$

equation (9.3) becomes

$$\frac{c_{i,t}\left(\omega^t\right)}{c_{j,t}\left(\omega^t\right)} = \left(\frac{\lambda_i}{\lambda_j}\right)^{-\frac{1}{\sigma}},$$

which implies that the ratio of consumption allocations (not just marginal utility of consumption) between households is constant across states and over time. Summing over $i = 1, ..., I$ on both sides yields

$$c_{j,t}\left(\omega^t\right) = \left[\frac{\left(\lambda_j\right)^{-\frac{1}{\sigma}}}{\int_0^1 \left(\lambda_i\right)^{-\frac{1}{\sigma}} di}\right] C_t\left(\omega^t\right), \tag{9.4}$$

so, individual consumption is proportional to aggregate consumption (or aggregate endowment), with a coefficient of proportionality, in the square bracket, that depends on the relative values of $\lambda$.[2] Thus, individual consumption is not constant over time, but it is independent of the realization of the individual endowment. Unanticipated shocks to the aggregate endowment are the only source of individual variation in consumption. These fluctuations cannot be insured because, by definition, they are common across households.

### 9.2.1 Full insurance with preference heterogeneity

How does preference heterogeneity affect consumption allocations under complete markets? It is immediate to see from (9.3) that the hallmark of complete markets —constant ratio of marginal utility of consumption across households— is still valid even with heterogeneity in $u$. It is no longer necessarily true, however, that the ratio of consumption is also equalized. To see this, assume $u$ is CRRA and households (indexed by $i$) differ with respect to the curvature parameter $\sigma$

$$u^i\left(c_{i,t}\left(\omega_t\right)\right)=\frac{c_{i,t}\left(\omega_t\right)^{1-\sigma_i}}{1-\sigma_i}.$$

Consider an economy with two agents $i=1,2$. The individual FOCs conditions combined with the market clearing condition imply

$$\lambda_2 c_{1,t}\left(\omega^t\right)^{-\sigma_1}=\lambda_1\left[Y_t\left(\omega^t\right)-c_{1,t}\left(\omega^t\right)\right]^{-\sigma_2} \tag{9.5}$$

Figure 9.1 plots right-hand side and left-hand side as a function of $c_{1,t}$ for the case where type 1 is less risk averse than type 2 $(\sigma_1<\sigma_2)$. The two curves cross only once. Now consider a rise in the aggregate endowment. We know from (9.4) that when $\sigma_1=\sigma_2$ consumption would increase proportionately so to leave the ratio of consumption between the two households unchanged. With unequal risk aversion, instead, it is efficient for the planner to have the consumption of the least averse households (type 1 in our example) fluctuate more in response to aggregate shocks, as evident from Figure 9.1.

### 9.2.2 Empirical tests of the full insurance hypothesis

Abstracting from heterogeneity in preference for risk, one can combine the consumption allocation under autarky in (9.1) and complete markets in (9.4) to derive an encompassing empirical model for consumption growth that can be taken to the data:

$$\Delta\log c_{i,t}=\beta_1\Delta\log C_t+\beta_2\Delta\log y_{i,t}+\varepsilon_{i,t},$$

where $y_{i,t}$ is current individual income, $C_t$ is aggregate consumption and $\varepsilon_{i,t}$ is an error term independent of the two regressors.[3] The autarky hypothesis implies $\beta_1=0$ and $\beta_2=1$, i.e.

[2] We have derived this result in Section 5.2.1. There, our formulation did not include uncertainty and we have derived it directly from the equilibrium conditions.

[3] This equation can also be interpreted as the result of a log-linearization for more general preferences.

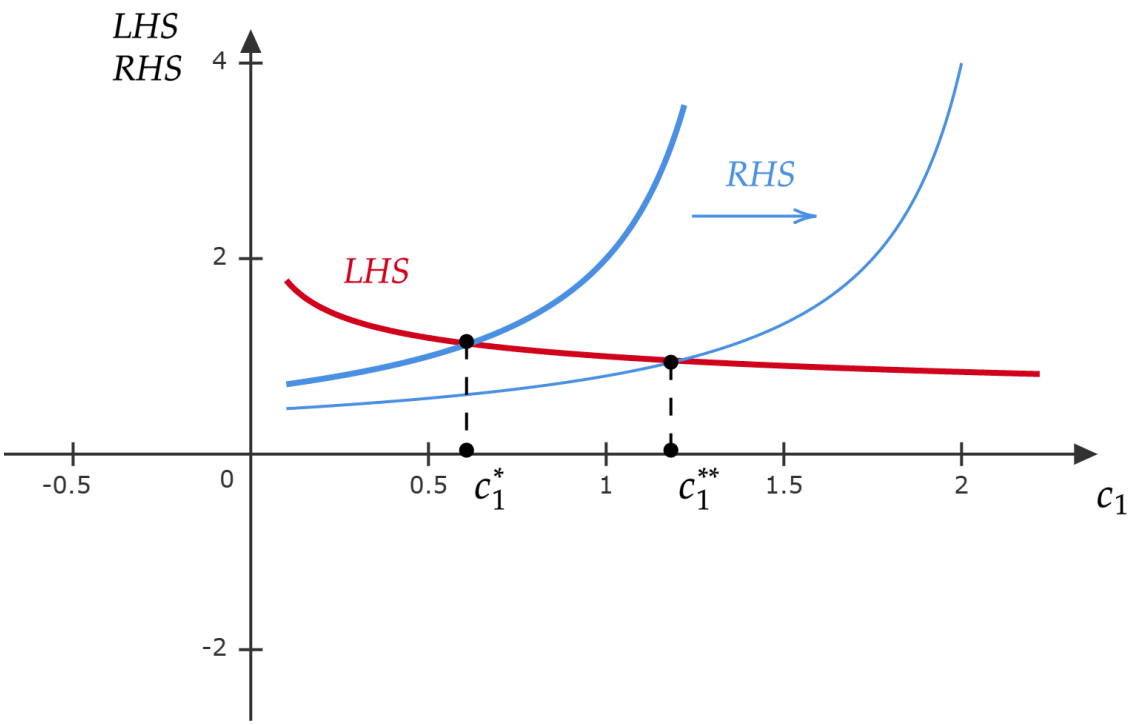



Figure 9.1: Plot of the left-hand-side (LHS) and right-hand-side (RHS) of equation (9.5).

**Notes**: Parameter values: $\sigma_1 = 0.25, \sigma_2 = 1, \lambda_1 = \lambda_2 = 1$. Aggregate endowment rises from $Y^* = 1.5$ to $Y^{**} = 2.25$. Agent 1, who is less risk-averse, absorbs most of the rise in the endowment.

individual consumption tracks individual income. The full risk-sharing hypothesis, instead, implies $\beta_1 = 1$ and $\beta_2 = 0$, i.e. individual consumption tracks aggregate endowment, but is independent of individual income.

In the early 1990s' a large empirical literature developed with the aim of testing these hypotheses using longitudinal micro data on consumption expenditures and income (Altug and Miller (1990), Mace (1991), Cochrane (1991), Nelson (1994), Townsend (1994), Attanasio and Davis (1996), Hayashi, Altonji, and Kotlikoff (1996), Jappelli and Pistaferri (2006)). Overall, the conclusion of this literature is that an empirically plausible model of consumption behavior lies somewhere in one in between full risk-sharing and autarky: individual income shocks are only *partially insured*.

**More details**

Mace (1991) used the short panel dimension of the Consumer Expenditure Survey (CES),

the main survey on consumption expenditures of US households, to test the relation between individual and aggregate consumption growth, and rejected the null hypothesis that $\beta_1 = 1$. Cochrane (1991), instead, tested the null hypothesis that individual consumption does not react to idiosyncratic income fluctuations ($\beta_2 = 0$). He used the Panel Study of Income Dynamics (PSID), the longest-running longitudinal dataset representative of the US population, to identify explicit events associated with income losses, such as days of work lost because of illness, involuntary job loss, weeks spent by jobless household heads looking for employment, days of work lost to strikes. He concluded that household food consumption expenditures (the only spending category well measured in this dataset at the time) are quite —although not fully— responsive to many of these indicators, a finding that contradicts the efficient risk sharing hypothesis. Attanasio and Davis (1996) analyzed the relation between changes in consumption and changes in relative earnings across demographic groups (defined by cohorts and education) in the U.S. in the 1980s. Their results indicate a "spectacular failure" of the full insurance hypothesis across groups.

According to the full risk-sharing hypothesis under CRRA, individual consumption moves in lockstep with aggregate consumption for all households. As a result, the household ranking (i.e., their relative position) in the consumption distribution remains constant over time. The hypothesis of lack of rank mobility in household consumption was tested, and amply rejected, by Fisher and Johnson (2006) and Jappelli and Pistaferri (2006).

Tests of perfect risk sharing were extended to the case of risk aversion heterogeneity by Mazzocco and Saini (2012) and Schulhofer-Wohl (2011). As discussed by these authors, preference heterogeneity can lead to a spurious rejection of this hypothesis. To see this, note that, as explained above, if households differ in their degree of curvature in utility, optimality conditions for the CRRA case imply that

$$\Delta \log c_{i,t} = \beta_{i,1} \Delta \log C_t + \beta_2 \Delta \log y_{i,t} + \varepsilon_{i,t} \tag{9.6}$$

where $\beta_i$ is larger the lower is risk aversion of household $i$ relative to its population average. Thus, if one estimates the misspecified equation (9.4) instead of (9.6), the error includes the term $(\beta_i - 1)\Delta \log C_t$. If earnings of low risk aversion (high $\beta_i$) individuals are more procyclical, then this omitted variable would induce a positive bias on the estimate of $\beta_2$. One reasons to expect that less risk-averse households will have more procyclical incomes is that more risk tolerant workers will choose occupations that carry more risk, both idiosyncratic and aggregate. Schulhofer-Wohl (2011) uncovers some evidence about this mechanism from respondents in the Health and Retirement Study.

### 9.2.3 Two approaches to partial risk sharing

How do we go about developing and quantifying economic theories of partial risk sharing? Economists have followed two methodologies. The first one —which we may call the *en-*

*dogenous incomplete markets* approach— is rooted in the tradition of microfoundation of macroeconomics. According to this approach, one should model explicitly the fundamental frictions that undermine the emergence of full insurance in the competitive equilibrium. Recall that a complete set of state-contingent securities will be traded if two assumptions hold: (i) perfect enforcement of contracts and (ii) perfect information. Deviations from these assumptions, such as limited commitment or private information (adverse selection or moral hazard), lead to partial risk sharing. There exist several examples of models that incorporate these frictions in the competitive equilibrium and analyze implications for the distribution of consumption allocations. For example, building on Kehoe and Levine (1993, 2001), Krueger and Perri (2006) relax the perfect contract enforcement assumption; Doepke and Townsend (2006) and Attanasio and Pavoni (2011) relax the perfect information assumption.

The principal strength of this methodology is that the market structure is not assumed exogenously, but it emerges in equilibrium and, as a result, is endogenous with respect to changes in primitives of the economy.[4] The main shortcomings of this strategy are two. First, the financial contracts that are traded in equilibrium are quite complex, e.g. they are history dependent and state-contingent, and thus very different from the simple ones we observe in reality. For example, in the limited enforcement economy default is an off-equilibrium threat, but it never actually happens, whereas households do default in the real world. Second, the empirical implications of these models for consumption allocations are often rejected by the data. Take, again, the limited commitment model as an example. Consumption always drifts down when the participation constraint of the household does not bind, and it jumps up when it does bind, which happens whenever income increases sufficiently (Kocherlakota, 1996). This pattern gives rise to an extreme and very counterfactual degree of left-skewness in the consumption distribution (Broer, 2013). This disconnect between theory and data represents a challenge for quantitative analysis of these models.

This criticism fueled the second methodology, which we can call the *exogenous incomplete markets* approach. This alternative view prescribes that we should only model the assets that we observe in the real world, e.g. non state-contingent bonds, risky publicly traded equity, housing, privately held firms, etc. The advantage of this perspective is that it leads to models that can be easily and naturally taken to the data. Its drawback is that the asset market structure is exogenously assumed and, as a result, it does not respond to changes in model parameters. To be precise, in an environment with exogenous incomplete markets, a shift in primitives (e.g., in the stochastic process of income shocks) does not lead to the addition or the disappearance of certain securities being traded, but it does impact the equilibrium prices at which existing assets are traded (and, possibly, also impacts the value of borrowing limits, depending on how they are specified). As a result, the pass-through of income shocks to consumption, an indicator of the degree of risk sharing, is also affected.

Before articulating the analysis of this approach, it is worth noting that there exist some results in the literature which show that the right combination of fundamental frictions

[4]For example, in the limited commitment economy of Krueger and Perri (2006), an increase in the size of idiosyncratic income risk reduces the value of autarky (and incentives to default), and increases equilibrium risk sharing, a force that contains the rise in consumption inequality caused by higher income uncertainty.

can give rise to a realistic market structure emerging endogenously in equilibrium. Notable examples are Allen (1985) and Cole and Kocherlakota (2001) who consider an environment with unobservable income shocks and hidden saving, and show that the constrained efficient allocations can be decentralized through a competitive asset market where households only trade a non-contingent bond.[5]

The canonical example of the exogenous incomplete market approach is the so-called *bond economy*, an environment where households are only allowed to trade a one period non state-contingent bond. This arrangement is reminiscent of a standard deposit/loan contract in the real world, where a saver receives an interest on their deposits every period, and a borrower repays interests on their loan every period without ever defaulting.[6] To fully understand the ad-hoc restrictions on market structure that we impose to obtain the bond economy, start from complete markets and consider the sequential formulation version of the Arrow-Debreu budget constraint (9.2) holding at every history $\omega^t \in \Omega^t$:

$$c_{i,t}\left(\omega^t\right) + \sum_{\omega_{t+1}\in\Omega} q_t\left(\omega_{t+1},\omega^t\right) a_{i,t+1}\left(\omega_{t+1},\omega^t\right) = y_{i,t}\left(\omega^t\right) + a_{i,t}\left(\omega^t\right) \tag{9.7}$$

where $a\left(\omega_{t+1},\omega^t\right)$ is an Arrow security purchased at date $t$ and state $\omega^t$ that pays one unit of consumption if state $\omega_{t+1}$ occurs next period, $q\left(\omega_{t+1},\omega^t\right)$ is the price of such Arrow security, and $a_{i,t}\left(\omega^t\right)$ are all the Arrow securities purchased at $t-1$ which pay in the current realized state $\omega^t$. In the bond economy, we force agents to trade only a non state-contingent asset. The budget constraint (9.7) is therefore replaced by the more restrictive

$$c_{i,t}\left(\omega^t\right) + q_t\left(\omega^t\right) a_{i,t+1}\left(\omega^t\right) = y_{i,t}\left(\omega^t\right) + a_{i,t}\left(\omega^{t-1}\right), \tag{9.8}$$

where $a_{t+1}\left(\omega^t\right)$ is a bond purchased at date $t$ and state $\omega^t$ that pays one unit of consumption next period, independently of the realization of the state $\omega_{t+1}$, and $q_t(\omega^t)$ is the price of such bond.

As the terminal condition, for now we only impose the no Ponzi game (nPg) condition and do not impose any further borrowing constraints. The nPg condition for this market structure is

$$\lim_{t\to\infty}\left(\prod_{s=0}^{t} q_s\left(\omega^s(\omega^t)\right)\right) a_{i,t+1}\left(\omega^t\right) \geq 0 \tag{9.9}$$

for all $\omega^t \in \Omega^t$, where $\omega^s(\omega^t)$ represents the sub-history of $\omega^t$ up to period $s$. A similar condition (for the deterministic case) has shown up in Section 5.3.1. Optimality implies the transversality condition (TVC) $\lim_{t\to\infty}\left(\prod_{s=0}^{t} q_s\left(\omega^s(\omega^t)\right)\right) a_{i,t+1}\left(\omega^t\right) \leq 0$ for all $\omega^t \in \Omega^t$, since a household with a positive limiting value of their wealth can improve their welfare

[5] Broer, Kapička, and Klein (2017) obtain similar implications from an economy which combines limited enforcement of contracts and private information about earnings.

[6] In reality such contracts are intermediated by banks. To the extent that the financial sector is competitive and banks solve a static maximization problem, financial intermediaries play no interesting role in the model and can be ignored. Chapter 17 discusses dynamic models of financial frictions where banks play a crucial role in determining equilibrium allocations.

by dissaving and consuming a bit extra.[7] Combining these two inequalities, we obtain the condition

$$\lim_{t\to\infty}\left(\prod_{s=0}^{t} q_s\left(\omega^s(\omega^t)\right)\right) a_{i,t+1}\left(\omega^t\right) = 0 \quad \text{for all} \quad \omega^t \in \Omega^t. \tag{9.10}$$

To simplify the analysis, in what follows we assume away fluctuations in the aggregate endowment $Y_t(\omega^t)$ which implies that the bond price is a constant, $p_t(\omega^t) = p$ and that we can omit the dependence on histories. Since the bond pays one unit of consumption in the next period, its rate of return is $r = 1/p - 1$. We can therefore reformulate the individual budget constraint of the bond economy (9.8) as

$$a_{i,t+1} = (1+r)\left(y_{i,t} + a_{i,t} - c_{i,t}\right). \tag{9.11}$$

This budget constraint (which reflects the market structure of the bond economy) is one of the cornerstones of the analysis of consumption and saving behavior in modern macroeconomics.[8] In this economy there are no explicit insurance markets, but, beyond borrowing when allowed, saving and dissaving act as a mechanism for *self-insurance*.[9] Throughout the rest of the chapter, we'll maintain this market structure.

## 9.3 Income fluctuation problems

We begin with a partial equilibrium analysis of the intertemporal decision problem of an infinitely-lived household who is subject to fluctuations in labor income, and every period must decide how much to consume and how much to save in a risk-free non-state contingent asset, possibly subject to a borrowing limit. The household takes the interest rate as given.

We analyze this *income fluctuation problem* (in the language of Schechtman and Escudero (1977)) first in the deterministic case and then in the stochastic case. A special case of this model, when the utility function is quadratic, is Hall (1978) formulation of Friedman (1957) permanent income hypothesis which displays certainty equivalence. We then generalize the model to settings where risk matters, either because of the presence of binding liquidity constraints or because the utility function displays prudence. For each of these models, we characterize the marginal propensity to consume.

[7] These concepts were introduced in Section 4.3.1.

[8] Note the timing convention implicit in the way we wrote this budget constraint: income is paid and consumption is chosen at the beginning of the period, and thus interests accrue on savings defined as $y_t + a_t - c_t$. The alternative timing convention, which we'll sometimes use in this chapter, is that income is paid and consumption decisions are made at the end of the period, which leads to the formulation of the budget constraint: $a_{t+1} = y_t + (1+r)a_t - c_t$.

[9] The expression "self-insurance" refers to the fact that individuals are insuring against future shocks by dissaving and saving, i.e. by trading intertemporally with themselves, and not by trading state-contingent insurance contracts with others, as such contracts are not available in this market structure.

### 9.3.1 Deterministic case

We begin by abstracting from income uncertainty. Consider the problem of an individual who faces deterministic (i.e., perfectly known ex-ante) income fluctuations $\{y_t\}_{t=0}^{\infty}$, has to choose optimally how to allocate consumption $c_t$ over time, and can only save through a risk free bond $a_t$. This individual solves

$$\begin{aligned} & \max_{\{c_t\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t u(c_t) \\ & s.t. \\ a_{t+1} &= (1+r)(y_t + a_t - c_t) \end{aligned} \tag{9.12}$$

This problem is analogous to the consumption-saving model from Chapter 4, but extended to incorporate time-varying deterministic income levels $y_t$.

By solving the Lagrangean associated to this problem, it is easy to see that the household first order condition yields

$$\frac{u'(c_t)}{\beta u'(c_{t+1})} = 1 + r. \tag{9.13}$$

This *consumption Euler equation* has the standard interpretation: the marginal rate of substitution between consumption today and consumption next period (intended as two different goods) equals the relative price of consumption today relative to consumption next period, or the interest rate.[10] It also has a variational interpretation: the value of a unit of consumption today is its marginal utility. Shifting such unit to next period yields $(1+r)$ units valued at the discounted marginal utility tomorrow. Optimality requires the household to be indifferent, hence the equal sign.

Because of the strict concavity of $u$, the Euler equation implies that the slope of the optimal individual consumption profile between $t$ and $t+1$ is increasing in $\beta$ and in $r$. Both higher patience and higher return on saving induce the household to save more today and postpone consumption to the future, which tilts upward the consumption profile.[11] Turning to the special case of CRRA utility with curvature parameter $\sigma$, (9.13) becomes

$$\frac{c_{t+1}}{c_t} = [\beta(1+r)]^{\frac{1}{\sigma}}. \tag{9.14}$$

This formulation clarifies that the extent to which variation in $\beta$ and $r$ translates into steeper or flatter consumption paths depends on the elasticity of intertemporal substitution, $EIS = 1/\sigma$. See Section 4.2.4 for details on the derivation of the $EIS$.

[10] To see why the interest rate is the intertemporal price of consumption, rewrite the budget constraint (9.8) in nominal terms with the price of the final good multiplying quantities at $t$ and $t+1$

$$p_{t+1}a_{t+1} = p_t a_t + p_t y_t - p_t c_t,$$

then divide through by $p_{t+1}$ and compare with (9.8).

[11] These statements are about the relative consumption across the two periods. Whether the *level* of consumption at $t$ increases or decreases depends on the relative strength of income and substitution effect due to the increase in $r$ (and thus on the level of wealth $a_t$), as for any other change in relative prices.

To sum up, in this environment, households want to smooth consumption with respect to deterministic income fluctuations, and thus they save when income is high relative to its mean and dissave when it is low. The extent of consumption smoothing (i.e., how close $c_t$ and $c_{t+1}$ are to each other) depends on the product $\beta(1+r)$ and on the willingness of households to substitute intertemporally, determined by $1/\sigma$. The closer is $\beta(1+r)$ to 1 and the lower is the elasticity of substitution, the more consumption is smoothed across periods.

**Marginal propensities to consume**

Let $R \equiv 1+r$ and iterate forward the budget constraint to obtain, after imposing condition (9.10)

$$c_t + \frac{1}{R}c_{t+1} + \frac{1}{R^2}c_{t+2} + ... = a_t + \sum_{j=0}^{\infty}\left(\frac{1}{R}\right)^j y_{t+j}.$$

Using the Euler equation (9.14) to substitute $c_{t+j}$ for all $j > 0$ on the left hand side as a function of $c_t$, and collecting terms, we arrive at:

$$c_t = \left(1 - R^{-1}(\beta R)^{\frac{1}{\sigma}}\right)\left[a_t + \sum_{j=0}^{\infty}\left(\frac{1}{R}\right)^j y_{t+j}\right]. \tag{9.15}$$

This expression is useful to introduce the concept of *marginal propensity to consume* (MPC). The MPC out of wealth $a_t$ (or, equivalently, out of a transitory change in income $y_t$) is defined as $\partial c_t/\partial a_t$.[12] Differentiating (9.15), we obtain that

$$MPC = 1 - R^{-1}(\beta R)^{\frac{1}{\sigma}}. \tag{9.16}$$

Two special cases are of interest. First, recall that in the equilibrium of a representative agent model without growth, $\beta R = 1$ and thus $MPC = 1 - \beta$. Let $\beta \equiv 1/1+\rho$, where $\rho$ is the discount rate. Since $\rho$ is small relative to 1, $MPC \simeq \rho$.[13] Thus, in the representative agent model, the marginal propensity to consume is approximately equal to the discount rate. Second, assuming log-utility ($\sigma = 1$) in the individual problem, without imposing $\beta R = 1$, by following similar steps one obtains $MPC \simeq r$.

Expression (9.15) shows that in this simple model optimal consumption is linear in wealth. In addition, equation (9.15) is an incarnation of Friedman's permanent income hypothesis: the term in the square bracket is the sum of financial wealth $a_t$ plus human wealth (the present value of future income), i.e. total wealth. Optimal consumption equals a constant fraction of total wealth. This equation also illustrates one of the key concept in Friedman's theory of consumption: the marginal propensity to consume out of transitory and permanent changes in income are different. Consider the case $\beta R = 1$ where the MPC out of transitory income is $1 - R^{-1}$. From (9.15) , it is easy to see that a permanent change in income of one unit, i.e. an increase of one unit in $y_{t+j}$ for all $j \geq 0$, leads to a change in human wealth

[12] By transitory we mean a change in $y_t$ which leaves unchanged income $y_{t+j}$ at any $j > 0$.

[13] Another way of deriving this result is to define $\beta = \exp(-\rho)$ and use the approximation $\exp(x) \simeq 1+x$.

equal to $1/\left(1-R^{-1}\right)$ and, thus a change in consumption equal to exactly 1 unit. Thus, the MPC out of permanent income is 1, and much larger than the MPC out of transitory income. Next, we explore the permanent income hypothesis in more detail.

### 9.3.2 Permanent income hypothesis

We now reintroduce income uncertainty and rewrite the household problem with the conditional expectation operator in the objective function as

$$\begin{aligned} & \max_{\{c_t\}_{t=0}^{\infty}} \mathbb{E}_t \sum_{t=0}^{\infty} \beta^t u\left(c_t\right) \\ & s.t. \\ a_{t+1} \quad = \quad & (1+r)\left(y_t+a_t-c_t\right) \end{aligned} \tag{9.17}$$

We also make two additional assumptions: quadratic utility

$$u\left(c_t\right)=b_1 c_t-\frac{1}{2} b_2 {c_t}^2, \qquad b_2>0, \quad c_t<b_1/b_2,$$

and $\beta R=1$. From the consumption Euler equation implied by (9.17), jointly with these two assumptions, we obtain

$$c_t=\mathbb{E}_t c_{t+1}. \tag{9.18}$$

This is the well known result of Hall (1978) that consumption is a martingale (or a random walk).[14] From the law of iterated expectations and the martingale property of the optimal consumption allocation:

$$\mathbb{E}_t c_{t+2}=\mathbb{E}_t\left[\mathbb{E}_{t+1} c_{t+2}\right]=\mathbb{E}_t c_{t+1}=c_t$$

and, more in general:

$$\mathbb{E}_t c_{t+j}=c_t, \text{ for any } j \geq 0. \tag{9.19}$$

If we iterate forward $J$ times on budget constraint in (9.17), and apply the conditional expectations to deal with uncertain future realizations of income and consumption, we obtain

$$\sum_{j=0}^{J}\left(\frac{1}{R}\right)^j \mathbb{E}_t c_{t+j}=a_t+\sum_{j=0}^{J}\left(\frac{1}{R}\right)^j \mathbb{E}_t y_{t+j}+\left(\frac{1}{R}\right)^{J+1} \mathbb{E}_t a_{t+J+1}$$

Taking the limit as $J \to \infty$ and using condition (9.10), the last term goes to zero. Using the martingale property (9.19) into the left hand side, we obtain

$$c_t=\frac{r}{1+r}\left[a_t+\sum_{j=0}^{\infty}\left(\frac{1}{1+r}\right)^j \mathbb{E}_t y_{t+j}\right]. \tag{9.20}$$

[14] A stochastic process $\{x_t\}$ is a random walk when it satisfies, at every $t$, $\mathbb{E}_t x_{t+j}=x_t$ for any $j>0$.

This expression illustrates the stochastic version of the permanent income hypothesis: optimal consumption is, again, linear and equals the annuity value of total wealth, i.e. financial wealth plus human wealth.[15]

If one assumes that income is deterministic and solves (9.17) , one obtains $c_{t+1} = c_t$ from the Euler equation and, by iterating forward on the budget constraint,

$$c_t = \frac{r}{1+r}\left[a_t + \sum_{j=0}^{\infty}\left(\frac{1}{1+r}\right)^j y_{t+j}\right] \tag{9.21}$$

Comparing (9.21) to (9.20) demonstrates that consumption satisfies *certainty equivalence*: in order to obtain the solution of the stochastic problem (9.17), it suffices solving the deterministic problem and applying conditional expectations to the exogenous variables $\{y_{t+j}\}_{j=0}^{\infty}$ in place of the variables themselves. Put differently, no higher moment of the income process, beyond the mean, matters for the dynamics of consumption. This property descends directly from the linear-quadratic structure of the problem. Any deviation from quadratic objective and linear constraints breaks certainty equivalence.

From (9.20), the change in consumption at time $t$ equals

$$\Delta c_t = c_t - c_{t-1} = c_t - \mathbb{E}_{t-1}c_t = \frac{r}{1+r}\left[\varpi_t - \mathbb{E}_{t-1}\varpi_t\right], \tag{9.22}$$

where $\varpi_t$ is total wealth defined as

$$\varpi_t \equiv a_t + \sum_{j=0}^{\infty}\left(\frac{1}{1+r}\right)^j \mathbb{E}_t y_{t+j},$$

and the last term on the right hand side is the innovation, or unexpected change, in permanent income at time $t$, i.e. the difference between realization and conditional expectation

$$\begin{aligned}\frac{r}{1+r}\left[\varpi_t - \mathbb{E}_{t-1}\varpi_t\right] &= a_t - \mathbb{E}_{t-1}a_t + \sum_{j=0}^{\infty}\left(\frac{1}{1+r}\right)^j \left[\mathbb{E}_t y_{t+j} - \mathbb{E}_{t-1}\left(\mathbb{E}_t y_{t+j}\right)\right] \\ &= \sum_{j=0}^{\infty}\left(\frac{1}{1+r}\right)^j \left(\mathbb{E}_t - \mathbb{E}_{t-1}\right) y_{t+j}, \end{aligned} \tag{9.23}$$

where we have used the law of iterated expectations $\mathbb{E}_{t-1}\left(\mathbb{E}_t y_{t+j}\right) = \mathbb{E}_{t-1}y_{t+j}$, and the fact that $a_t = \mathbb{E}_{t-1}a_t$, since there is no uncertainty at time $t$ (after $y_t$ is realized) about the

[15] The annuity value is exactly $r/(1+r)$, i.e. that portion of wealth that, when consumed every period, keeps asset holdings constant. To see this, abstract from income $y_t$, and note that

$$a_{t+1} = (1+r)(a_t - c_t) = (1+r)\left(a_t - \frac{r}{1+r}a_t\right) = a_t.$$

evolution of wealth into $t+1$. Combining (9.22) and (9.23), we arrive at

$$\Delta c_t = \frac{r}{1+r} \sum_{j=0}^{\infty} \left( \frac{1}{1+r} \right)^j (\mathbb{E}_t - \mathbb{E}_{t-1}) \, y_{t+j}. \tag{9.24}$$

This equation contains another useful result: under the permanent income hypothesis, the change in consumption between $t-1$ and $t$ is proportional to the revision in expected future income due to the new information (the "news") accruing in that same time interval.

### Permanent and transitory income shocks

To make further progress, we need to make some assumptions on the income process. We choose a specification that is very common in labor economics, at least since Abowd and Card (1989). We model labor income as the sum of two orthogonal components, $y_t^p$ which follows a martingale with i.i.d. innovation (or shocks) $v_t$, and $u_t$ which is also an i.i.d. shock

$$\begin{aligned} y_t &= y_t^p + u_t, \\ y_t^p &= y_{t-1}^p + v_t. \end{aligned} \tag{9.25}$$

In addition, we assume that $\mathbb{E}(v_t) = \mathbb{E}(u_t) = 0$ and that the two shocks are orthogonal, $u_t \perp v_\tau$ for all pairs $(t, \tau)$. If we let $x_t$ denote either shock, our assumptions imply that $\mathbb{E}_t(x_{t-j}) = x_{t-j}$ for $j \geq 0$, and $\mathbb{E}_t(x_{t+j}) = 0$ for $j > 0$.

Combining these two equations in (9.25), we obtain the representation

$$y_t = y_{t-1} + u_t - u_{t-1} + v_t. \tag{9.26}$$

Using this income process into (9.24), after some algebra, we obtain

$$\Delta c_t = \frac{r}{1+r} u_t + v_t, \tag{9.27}$$

which establishes that households adjust their consumption responding to the annuity value of transitory shocks and to the full value of permanent shocks. This finding is analogous to the one for the deterministic economy: the pass-through of income shocks to consumption depends on their expected duration. Remarkably, this version of the bond economy is quite close to full insurance with respect to transitory shocks. In conclusion, borrowing and saving through a non-state contingent asset offers ample opportunity for consumption smoothing as long as shocks are not too persistent.

**Using the PIH to learn about the nature of the rise in inequality**

Income inequality increased substantially in the 1980s and the 1990s in many developed countries. Much of the public, including many policymakers and economists, interpreted this trend as indicating widening differentials in standard of living across households. This interpretation, however, is open to the criticism that current income may not reflect the long-run level of resources available to a household, and hence their welfare.

Blundell and Preston (1998) showed how one can use the permanent income hypothesis, together with data on the joint cross-sectional distribution of income and consumption, to learn whether the rise in cross-sectional income inequality is of a transitory nature and hence not too worrisome, or of a permanent nature and thus detrimental for inequality in household welfare.

Consider the model of Section 9.3.2. From equations (9.26) and (9.27), we have that the evolution of income and optimal consumption for an individual $i$ at time $t$ are given by

$$\begin{aligned} y_{i,t} &= y_{i,t-1} + u_{i,t} - u_{i,t-1} + v_{i,t} \\ c_{i,t} &= c_{i,t-1} + \frac{r}{1+r} u_{i,t} + v_{i,t}. \end{aligned}$$

Now, compute the cross-sectional variance of consumption and the cross-sectional covariance between consumption and income for all individuals belonging to a cohort $k$, assuming that $r \simeq 0$. Then, one obtains:

$$\Delta var_{k,t}(c) = \Delta covar_{k,t}(c, y) \simeq var_t(v). \tag{9.28}$$

In other words, by tracing the change over time of the within-cohort variance of consumption, or covariance between consumption and income, one can estimate the change over time in the variance of the permanent component of income. Blundell and Preston concluded that the bulk of the rise in UK income inequality was driven by the permanent component, a result that is consistent with the idea that skill-biased technical change (and rising college wage premium), is a key driving force of the recent changes in income distribution. Although based on a different methodology, these empirical findings are reminiscent of those in Attanasio and Davis (1996) discussed in Section 9.2.2. In particular, they also represent a rejection of full insurance since, from the perspective of the efficient risk sharing hypothesis, consumption should not react to idiosyncratic income shocks, no matter their persistence.

### Saving for the rainy days

Define household savings $s_t$ as capital income plus labor income net of consumption expenditures

$$s_t = \frac{r}{1+r} a_t + y_t - c_t. \tag{9.29}$$

Combining (9.29) with (9.20) we obtain an expression for saving only as a function of current and future expected income

$$s_t = y_t - \frac{r}{1+r} \sum_{j=0}^{\infty} \left( \frac{1}{1+r} \right)^j \mathbb{E}_t y_{t+j}.$$

Unfolding this summation on the right hand side, we obtain

$$\begin{aligned} s_t &= y_t - \frac{r}{1+r}y_t - \frac{r}{1+r}\left[\left(\frac{1}{1+r}\right)\mathbb{E}_t y_{t+1} + \left(\frac{1}{1+r}\right)^2 \mathbb{E}_t y_{t+2} + ...\right] \\ &= \frac{1}{1+r}y_t - \frac{r}{1+r}\left(\frac{1}{1+r}\right)\mathbb{E}_t y_{t+1} - \frac{r}{1+r}\left[\left(\frac{1}{1+r}\right)^2 \mathbb{E}_t y_{t+2} + ...\right] \\ &= \frac{1}{1+r}\mathbb{E}_t \Delta y_{t+1} + \left(\frac{1}{1+r}\right)^2 \mathbb{E}_t y_{t+1} - \frac{r}{1+r}\left[\left(\frac{1}{1+r}\right)\mathbb{E}_t y_{t+2} + ...\right] \end{aligned}$$

and, using the same approach on terms $t+j$ with $j > 1$ we obtain

$$s_t = -\sum_{j=1}^{\infty}\left(\frac{1}{1+r}\right)^j \mathbb{E}_t \Delta y_{t+j}.$$

This expression shows that savings are equal to the discounted sum of expected declines in income. If a household expects income to fall in the future, in anticipation they will save and these savings will allow them to smooth consumption once income actually declines. For example, under a mean reverting income process, households who experience a sequence of positive (above mean) income shocks will save because they understand that this abundance is only transient. Similarly, those who experience a sequence of negative (below mean) income shocks will dissave. Campbell (1987) calls this behavior *saving for the rainy day*. Or, dissaving in the expectation of a sunny day.

### 9.3.3 Borrowing constraints

To derive the results in the previous section, we have abstracted from borrowing constraints. But, to what extent can we safely ignore borrowing limits, as if they were never binding? The answer depends on the income process.

Suppose we impose the ad-hoc no-borrowing constraint $a_{t+1} \geq 0$ on the household problem. When income $y_t$ is a random walk, from (9.20) we obtain

$$c_t = \frac{r}{1+r}a_t + y_t$$

since $\mathbb{E}_t y_{t+j} = y_t$ for all $j \geq 0$. Substituting this expression for consumption into the budget constraint yields $a_{t+1} = a_t$. In this case, wealth is constant, so if a households starts with a positive wealth level, the zero debt limit will never bind.

Suppose now that $y_t$ follows an i.i.d. shock with mean $\bar{y}$. From (9.20) we have

$$c_t = \bar{y} + \frac{r}{1+r}(a_t + y_t) \tag{9.30}$$

since $\mathbb{E}_t y_{t+j} = \bar{y}$. Substituting into the budget constraints yields $\Delta a_{t+1} = y_t$. In this case, wealth follows a random walk, which means that starting from any initial level of wealth there

is always a positive probability that at some point in the future the household will hit the borrowing limit in finite time. For example, consider an individual who receives an income realization equal to $y_t < \bar{y}$ and whose initial wealth is $a_t = ry_t$. It is easy to see that if the individual wanted to consume its optimal unconstrained level in equation (9.30), they would enter the next period with negative wealth, which would violate the no-borrowing limit. As a result, the individual would have to consume below its unconstrained level. More broadly, the borrowing constraint is likely to bind whenever $a_t$ is small, $y_t$ follows a mean-reverting process, and the individual is hit by a low enough shock (relative to its mean). Because income is expected to revert back to its higher mean, consumption smoothing dictates that the household should dissave and keep their consumption high. In some states, however, dissaving is not enough, borrowing would be necessary and the constraint binds. In this case, the optimal consumption choice is constrained.

These examples highlight the fact that, in general, borrowing limits cannot be ignored and their presence affects optimal consumption and saving choices, as we explain next. We now examine the more general consumption-saving problem with stochastic income $y_t$ and an ad-hoc zero-borrowing constraint.

Consider the problem of a household facing a no-borrowing constraint. Stated in a sequential formulation, we have

$$\begin{aligned} &\max_{\{c_t, a_{t+1}\}_{t=0}^{\infty}} \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t u(c_t) \\ &\text{s.t.} \\ &a_{t+1} = R(a_t + y_t - c_t) \\ &a_{t+1} \geq 0 \end{aligned} \tag{9.31}$$

where $u' > 0$ and $u'' < 0$. The first order condition of this problem yields the modified Euler equation:

$$u'(c_t) = \beta R \mathbb{E}_t [u'(c_{t+1})] + \lambda_t,$$

where $\lambda_t$ is the multiplier on the borrowing constraint at date $t$. Because $\lambda_t \geq 0$, the Euler equation can be rewritten as

$$u'(c_t) \geq \beta R \mathbb{E}_t [u'(c_{t+1})], \tag{9.32}$$

where the strict inequality holds when the constraint is binding. Assume $\beta R = 1$. If the constraint is not binding, the household's choice is interior, dictated by the Euler equation which states that marginal utility today is equated to expected marginal utility next period. If, instead, the constraint is binding, current marginal utility is higher and consumption is lower: the household problem has a corner solution and is not determined by the Euler equation, but by the borrowing limit and the budget constraint. Since $a_{t+1} = 0$, from the budget constraint, we obtain $c_t = a_t + y_t$, i.e. the household consumes all their cash in hand. The multiplier $\lambda_t$ plays a role akin to the interest rate, in that it increases the cost of consuming today relative to postponing consumption to the future (Hall, 1988).

There are two key differences between the model without and with borrowing constraints. The first one is that, absent borrowing constraints, what matters for current consumption is

only the expected discounted value of labor income, not its timing (see equation (9.20)). In the presence of borrowing constraints, instead, the timing of income matters for consumption. Consider two random income streams with the same discounted expected value, one that is decreasing and another one that is increasing: a household facing the latter is more likely to be constrained and their consumption path may differ from the consumption path of a household facing the decreasing stream. The second difference is that, for a constrained household, the marginal propensity to consume is 1 for both transitory and permanent shocks: no matter what the duration of the shock is, a small change in current income increases current consumption one for one. Thus constrained (or hand-to-mouth) households display higher MPC out of transitory income relative to unconstrained ones.

### Natural borrowing limit

So far, when modelling borrowing constraints we have specified ad-hoc limits of the type $a_{t+1} \geq -\bar{a}$. Another type of constraint is the so called "natural" borrowing limit. In Section 4.3.1, we discussed the notion of the natural borrowing limit in the context of a deterministic model. In the Online Appendix to Chapter 4, we have shown that, in the deterministic model, the natural borrowing limit is equivalent to the no-Ponzi-game (nPg) condition. In what follows, we extend this notion to an environment with uncertainty. Here, the natural borrowing limit is the lowest asset position consistent with non-negative consumption in every state and every period.

Start from the case where $\{y_t\}_{t=0}^{\infty}$ is deterministic. Iterating forward on the budget constraint and imposing $c_t \geq 0$ for all $t$, we obtain

$$a_{t+1} \geq -\sum_{j=0}^{\infty} \left(\frac{1}{1+r}\right)^j y_{t+1+j}.$$

Assume now that income is stochastic and the lowest possible realization is $y_{\min}$, always occurring with positive probability. Then, to guarantee positive consumption at every $t$, debt cannot exceed the present value of $y_{\min}$, and thus the natural borrowing limit becomes

$$a_{t+1} \geq -\left(\frac{1+r}{r}\right) y_{\min}.$$

Borrowing up to this value, or more, would mean that there is a positive probability that a state occurs where consumption must be zero. This statement is easy to prove. Suppose the household has borrowed exactly up to this limit, i.e. $a_t = -\left(\frac{1+r}{r}\right) y_{\min}$ and they are hit by an income realization $y_t = y_{\min}$. Then from the budget constraint:

$$c_t = \underbrace{-\left(\frac{1+r}{r}\right) y_{\min}}_{a_t} + y_{\min} - \frac{a_{t+1}}{1+r} \leq -\frac{1}{r} y_{\min} - \frac{1}{1+r} \underbrace{\left[-\left(\frac{1+r}{r}\right) y_{\min}\right]}_{\text{max that can be borrowed}} = 0$$

It follows that, if the utility function satisfies the Inada condition $u(0) = -\infty$, an optimizing consumer will never borrow up to the natural borrowing limit because doing that

is inconsistent with expected utility maximization. Thus, one can always safely assume an interior solution for the Euler equation. This is, instead, not true for ad-hoc debt limits which always affect the optimal solution in some region of the state space, as explained above. Finally, note that if $y_{\min} = 0$, then the no-borrowing constraint is also the natural borrowing limit.[16]

### 9.3.4 Precautionary saving

The term *precautionary saving* refers the additional amount of saving a household chooses in response to a rise in the uncertainty about future income. The certainty equivalence property of consumption allocation under quadratic utility implies that the consumption function is linear in wealth and in the expected present value of future income. Thus, a mean preserving spread of the income shock distribution (i.e., an increase in dispersion around the same mean) does not impact saving: no precautionary motive for saving is present in this environment.

In what follows we establish two forces that lead to precautionary saving behavior. The first one is the presence of occasionally binding borrowing constraints, and the second is prudence. Prudence is a property of the utility function that mathematically corresponds to a positive third derivative, i.e. $u''' > 0$. Under both forces a rise in income uncertainty leads to a rise in the level of savings. As first uncovered by Zeldes (1989) through numerical simulations, in these settings the consumption policy function is concave in wealth, a property that has important implications for the marginal propensity to consume.

#### Precautionary motive with borrowing constraints

To isolate the role of borrowing constraints for precautionary saving, consider again the quadratic specification for period utility, which does not display prudence because $u''' = 0$. Continue assuming that and $\beta R = 1$. Suppose households face a borrowing limit $a_{t+1} \geq -\underline{a}$. Then,

$$c_t = \begin{cases} \mathbb{E}_t c_{t+1} & \text{if } a_{t+1} > -\underline{a} \\ y_t + a_t + \frac{\underline{a}}{R} & \text{if } a_{t+1} = -\underline{a} \end{cases}$$

The first line describes the optimal unconstrained intertemporal consumption allocation when the constraint is not binding, and the second line the consumption allocation if the constraint is binding, in which case the individual consumes all their resources. The above pair of conditions can be written in compound form as

$$c_t = \min \left\{ y_t + a_t + \frac{\underline{a}}{R}, \mathbb{E}_t c_{t+1} \right\}. \tag{9.33}$$

[16] Throughout this discussion, we have assumed that $r > 0$. As we will see in Section 9.4, $r < 0$ can be an equilibrium outcome of economies populated by a continuum of households facing an income fluctuation problem. In such a case, it is convenient (and realistic) to assume the existence of a wedge $\chi > 0$ between the lending rate $r$ and the borrowing rate $r^b$ which can be interpreted as a linear intermediation cost of the financial sector. The interest rate entering the natural borrowing limit is $r^b = r + \chi$, and if $\chi$ is large enough, $r^b > 0$ and the natural borrowing limit is well defined again.

Assume the constraint is not already binding at date $t$. Then

$$c_t = \mathbb{E}_t c_{t+1} = \mathbb{E}_t \left[ \min \left\{ y_{t+1} + a_{t+1} + \frac{a}{R}, \mathbb{E}_{t+1} c_{t+2} \right\} \right]. \tag{9.34}$$

Suppose that the uncertainty about $y_{t+1}$ increases but its mean does not change. Very low realizations of income $y_{t+1}$ become more likely, and thus the borrowing constraint is also more likely to bind next period. This reduces the value of $\mathbb{E}_t \left[ \min \left\{ y_{t+1} + a_{t+1} + \frac{a}{R}, \mathbb{E}_{t+1} c_{t+2} \right\} \right]$ and of current consumption $c_t$. As a result, savings increase. Intuitively, when households face borrowing limits which can potentially bind in the future, they understand that if they were to receive bad income realizations, they would be pushed towards the constraint and be forced to consume their income without the ability of smoothing consumption, which reduces their welfare. To prevent this scenario, they save more. This is the first source of precautionary saving motive.

An implication of the presence of borrowing constraints is that the consumption function is concave in wealth, as illustrated in Figure 9.2.

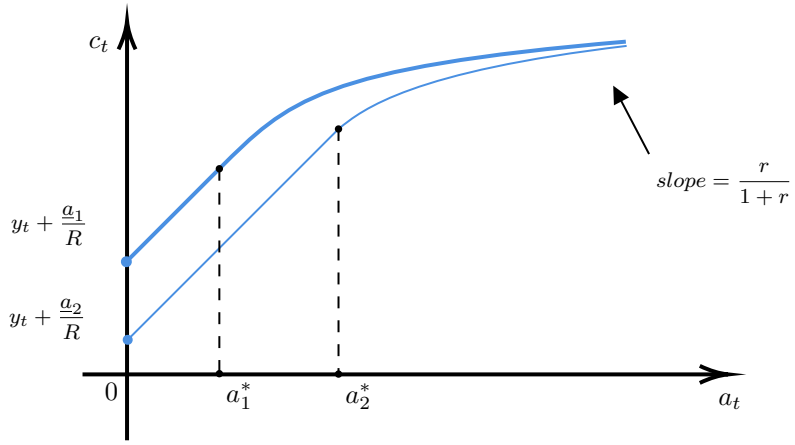



Figure 9.2: Decision rule for consumption in the presence of a borrowing constraint $\underline{a}$ as a function of wealth $a_t$, for a given realization of income $y_t$.

**Notes**: The function is linear, with slope equal to 1, until $a^*$, after which it becomes concave. For large enough wealth, its slope converges to $\frac{r}{1+r}$. The two lines correspond to different values of the borrowing limit.

If we fix labor income $y_t$ to a sufficiently low value, it is clear from (9.33) that there exists a level of wealth $a_t = a^*$ below which the constraint binds and $c_t = a_t + y_t + \frac{a}{R}$. In this region, the slope of the consumption function with respect to wealth is 1. For sufficiently high levels of $a_t$, it becomes extremely unlikely (or impossible, depending on the process for $y_t$) that the constraint will bind in the future and consumption asymptotes to the linear unconstrained solution (9.20) with slope $r/(1+r) < 1$. For intermediate levels of wealth, the borrowing limit could be binding again, and consumption is depressed by the precautionary motive relative to the unconstrained optimum. In this range, the consumption function is

strictly concave and the marginal propensity to consume is bracketed between $r/(1+r)$ and 1.

The figure also illustrates what happens when the credit limit tightens from $\underline{a}_1$ to $\underline{a}_2 < \underline{a}_1$. Obviously, households are constrained for a wider range of $a_t$. For unconstrained households, consumption falls across the board because of the stronger precautionary saving motive, and their marginal propensity to consume (the slope of the consumption function) increases. In the limit as wealth keeps growing, the two consumption functions converge because the borrowing constraint becomes irrelevant.

An important implication of this discussion is that there is a range of wealth levels for which the individual MPC is larger than in the PIH, either because the constraint binds (and the consumption function has slope equal to 1) or because it might bind with some probability in the future (and the consumption function is strictly concave). We return on this observation in Section 9.4.2.

### Precautionary motive with prudence

To illustrate the second source of precautionary saving, consider a simple two-period consumption-saving problem (Leland, 1968; Sandmo, 1970)

$$\begin{aligned} \max_{\{c_0,c_1,a_1\}} & \; u(c_0) + \beta \mathbb{E}_0 [u(c_1)] \\ \text{s.t.} & \; c_0 + a_1 = y_0 \\ & \; c_1 = Ra_1 + y_1 \end{aligned}$$

where $y_0$ is given, and $y_1$ is stochastic. Again, to isolate this second force from the first one we just described, we assume away any borrowing constraint at $t = 0$. The Euler equation for this problem is

$$u'(y_0 - a_1) = \beta R \mathbb{E}_0 [u'(Ra_1 + y_1)] \tag{9.35}$$

an equation in one unknown, $a_1$. The left-hand-side is increasing in $a_1$ since $u'' < 0$, and the right-hand-side is decreasing for the same reason, hence the solution for $a_1$ is uniquely determined. What happens to optimal consumption at $t = 0$ as future income $y_1$ becomes more risky? Consider a mean-preserving spread of $y_1$. If $u'$ is convex then, by Jensen's inequality, the value of the right-hand-side will increase. Graphically (see Figure 9.3), the left-hand-side is unchanged and the right-hand-side shifts upward, inducing a rise in optimal savings $a_1$.[17] One way to understand this result is that the hike in future consumption uncertainty reduces welfare. Increasing savings today raises the expected value of future consumption which compensates for higher variance. Sibley (1975) and Miller (1974) extend this proof to a multi-period model with a finite $T$, and i.i.d. income shocks.

The convexity of the marginal utility corresponds precisely to the condition $u''' > 0$, or prudence. Prudence is a property of preferences, like risk aversion: risk-aversion refers to the curvature of the utility function, prudence refers to the curvature of the marginal utility function. Kimball (1990) defines the index of absolute prudence as $-u'''(c)/u''(c)$

[17] This result is a simple application of Stiglitz and Rothschild (1970)

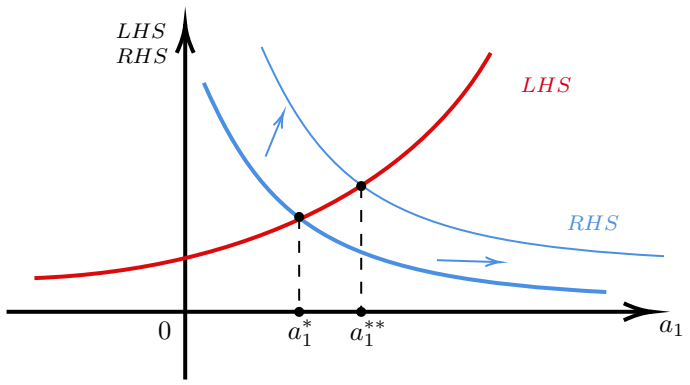



Figure 9.3: Left hand side (LHS) and right hand side (RHS) of the Euler equation (9.35).

**Notes**: It shows that when a mean-preserving spread in future income occurs, the RHS shifts outward and the optimal amount of saving in period zero $a_1$ increases.

and relative prudence as $-u'''(c)\,c/u''(c)$ in a conceptually similar way to the Arrow-Pratt index of absolute and relative risk-aversion. There exists also an interesting relationship between the two: any utility function in the decreasing absolute risk aversion (DARA) class, which includes CRRA, displays prudence. To see this, let $\alpha(c)$ be the coefficient of absolute risk aversion, then

$$\alpha(c) = \frac{-u''(c)}{u'(c)} \Rightarrow \alpha'(c) = \frac{-u'''(c)\,u'(c) + [u''(c)]^2}{[u'(c)]^2}.$$

Since with DARA $\alpha'(c) < 0$, then we have that

$$-u'''(c)\,u'(c) + [u''(c)]^2 < 0 \Rightarrow u'''(c) > \frac{[u''(c)]^2}{u'(c)} > 0.$$

Intuitively, a rise in uncertainty reduces the certainty-equivalent income next period and, with DARA utility, it effectively increases the degree of risk-aversion of the agent, inducing them to save more.

**Role of prudence: an analytical expression**

Blanchard and Mankiw (1988) derive an analytical expression that illustrates how risk affects optimal consumption and saving. Take a second-order approximation of $u'(c_{t+1})$ around the value $c_{t+1} = c_t$:

$$u'(c_{t+1}) \simeq u'(c_t) + u''(c_t)(c_{t+1} - c_t) + \frac{1}{2}u'''(c_t)(c_{t+1} - c_t)^2.$$

Substitute this approximation into the right-hand side of the Euler equation $u'(c_t) = \beta R \mathbb{E}_t u'(c_{t+1})$, divide both sides by $c_t^2$, and use the approximation

$$(\beta R)^{-1} = \frac{1+\rho}{1+r} = \exp\left(\log \frac{1+\rho}{1+r}\right) \simeq 1 + \rho - r.$$

Rearranging terms, we obtain

$$\mathbb{E}_t\left(\frac{c_{t+1}-c_t}{c_t}\right) \simeq EIS(c_t) \cdot (r-\rho) + \frac{1}{2} P(c_t) \cdot \mathbb{E}_t\left[\left(\frac{c_{t+1}-c_t}{c_t}\right)^2\right] \tag{9.36}$$

where

$$EIS(c_t) \equiv -\frac{u'(c_t)}{c_t u''(c_t)}, \text{ and } P(c_t) = -\frac{u'''(c_t)\, c_t}{u''(c_t)}$$

are, respectively, the elasticity of intertemporal substitution and the coefficient of relative prudence. Equation (9.36) features the two determinants of expected consumption growth. The first term captures the standard intertemporal consumption smoothing motive, active as long as $u$ is strictly concave. The second term contains the role of risk, captured by the expected variability of consumption growth (measured by the second uncentered moment): the stronger is relative prudence, the more consumption growth will respond to changes in risk. Higher uncertainty in future consumption growth, for given expected income, implies higher saving through the precautionary motive.

Similarly to the case where debt limits are binding, one can prove that even in absence of constraints, the consumption function is concave in wealth if the utility function belongs to the hyperbolic absolute risk aversion (HARA) class and displays positive third derivative.[18] First, recall that the envelope condition of the household problem yields

$$u'(c(a)) = V'(a) \tag{9.37}$$

where $V$ denotes the value function and where, to ease notation, we have omitted the dependence on labor income. It can be proved that, under general conditions, the value functions inherits the assumptions on $u$, i.e. it is increasing, concave and differentiable—see the discussions in Section 4.4. Differentiating both sides with respect to $a$ gives

$$c'(a) = \frac{V''(a)}{u''(c(a))} > 0, \tag{9.38}$$

which implies that optimal consumption is strictly increasing in wealth. Differentiating one more time, we obtain

$$c''(a) = \frac{V'''(a)\, u''(c(a)) - V''(a)\, u'''(a)\, c'(a)}{[u''(a)]^2}. \tag{9.39}$$

[18] A utility function belongs to the HARA class if its coefficient of absolute risk aversion is an hyperbola (or, equivalently, absolute risk tolerance is an affine function of wealth). Many common utility functions, such as quadratic, CRRA and CARA are special cases of HARA.

Using (9.37) and (9.38) into (9.39), we conclude that the consumption function is concave ($c'' \leq 0$) whenever

$$\frac{V'''(a)\,V'(a)}{V''(a)^2} \geq \frac{u'''(a)\,u'(a)}{u''(a)^2}. \tag{9.40}$$

This is the content of Lemma 2 in Carroll and Kimball (1996). If $u$ belongs to the HARA class, then one can show that the right hand side of (9.40) is equal to a constant $\kappa \geq 0$, i.e. prudence is $\kappa$ times higher than risk aversion.[19] Assuming HARA, Carroll and Kimball (1996) and Jensen (2018) also prove that, under a finite optimization horizon, the inequality in (9.40) is true and therefore the consumption function is concave. Strict concavity arises for $\kappa > 0$, which holds for the CRRA class, with the exception of $\kappa = 1$ corresponding to CARA utility.[20]

Once again, the intuition is that the precautionary saving motive is declining in wealth. Consider the deviation in optimal consumption in the presence of uninsurable risk relative to the no-uncertainty case in which the consumption function is linear. For large enough wealth, precautionary saving approaches zero and the consumption function asymptotes the no uncertainty case. As wealth falls, precautionary saving keeps rising and consumption keeps falling, hence the concavity.

**Capital income uncertainty**

What if the income uncertainty refers to capital income, i.e. to the rate of return on saving $r$, instead of labor income? Does an increase in uncertainty still induce more saving? Consider a problem analogous to the one analyzed earlier in this section. The household at date $t = 0$ receives an income $y_0$ which they can either consume or save, i.e $y_0 = c_0 + a_1$. In the second period, consumption is simply $c_1 = (1 + r)\,a_1$, but $r < -1$ is now stochastic. The Euler equation corresponding to this problem is

$$u'(y_0 - a_1) = \beta \mathbb{E}_0 \left[(1 + r)\,u'((1 + r)\,a_1)\right].$$

Note that the random interest rate is now inside the expectation. As before the left-hand side is increasing in $a_1$ and the right-hand side decreasing in $a_1$, hence the solution is unique. What happens to optimal saving $a_1$ after a mean-preserving spread in $r$? The answer is not obvious ex-ante. On the one hand, the precautionary saving force would suggest that savings optimally expand. On the other hand, with higher savings, the household becomes even more exposed to risk, and thus reducing saving curtails risk. This latter force was not present when we analyzed labor income risk.

[19] Specifically, quadratic utility is the knife-edge case, corresponding to $\kappa = 0$, constant absolute risk aversion (CARA) corresponds to $\kappa = 1$, and constant relative risk aversion (CRRA) utility functions satisfy $\kappa > 1$.

[20] In fact, Cantor (1985) proved that the optimal solution to the consumption/saving problem of a CARA consumer facing uninsurable labor income risk (and no borrowing constraints) displays consumption which is linear in wealth. Compared to the no-uncertainty case, uninsurable risk only affects the intercept of the consumption function by lowering it. The slope (i.e., the constant MPC) is the same.

Under prudence, however, the right-hand side is still a convex function of $r$ (it is the product of a linear and a convex function), and thus with this additional assumption we can conclude that precautionary saving behavior emerges even in the face of higher volatility of returns.

### Buffer-stock saving

Consider a version of the intertemporal consumption problem where households have a finite horizon, utility is CRRA with curvature parameter $\sigma > 0$, log income follows a stochastic process which is the sum of a permanent shock $y_t^P$ and a transitory shock $u_t$ (as in Section 9.3.2, but in logs), both Normally distributed. In addition, there is positive probability that income will be zero at any $t$ (capturing, e.g., an unemployment shock) and, as a result, the natural borrowing limit is zero. Finally, the discount rate $\rho$ exceeds the interest rate $r$.[21]

Optimal consumption in this setting, called "buffer-stock model", is fully characterized in a series of papers by Carroll, Hall, and Zeldes (1992); Carroll (2001). What makes this model particularly simple to analyze is that the individual state space can be reduced to one state variable only, $x_t$, the ratio of cash in hand $a_t + y_t$ to permanent income $y_t^P$. Similarly to our derivation of equation (9.36), one can show that

$$\mathbb{E}_t \Delta \log c_{t+1} \simeq \frac{1}{\sigma}(r - \rho) + \frac{\sigma}{2}\mathbb{E}_t var\left(\Delta \log c_{t+1}\right),$$

where $var$ denotes the variance.[22] As $x_t \to 0$, expected consumption growth approaches infinity because $c_t$ approaches zero: when $x_t = 0$ from the budget constraint $-c_t = a_{t+1}/R \geq 0$, and thus $c_t = 0$. As $x_t \to \infty$, uncertainty about future labor income becomes essentially irrelevant and expected consumption growth approaches $\sigma^{-1}(r - \rho) < 0$.

Figure 9.4 illustrates the relation between consumption growth and normalized wealth $x$. The figure highlights that there exists a level of wealth $x^*$ that is dynamically stable. If $x_t > x^*$ the negative intertemporal saving motive outweighs the precautionary saving motive. Consumption growth is negative, and wealth converges back toward $x^*$. If $x_t < x^*$, the precautionary motive dominates, consumption growth is positive, and wealth rises toward $x^*$. In sum, households have a *target level of assets* $x^*$ which is above the level of the no uncertainty case, i.e., they hold additional wealth as a buffer stock. Labor income shocks take households periodically above or below this target level, and households adjust their optimal consumption in order to quickly move back to $x^*$.

[21] As we prove in Section 9.4, $r < \rho$ is an equilibrium outcome in an economy populated by a continuum of this type of consumers.

[22] For a full derivation, see for example Jappelli and Pistaferri (2017), chapter 6.

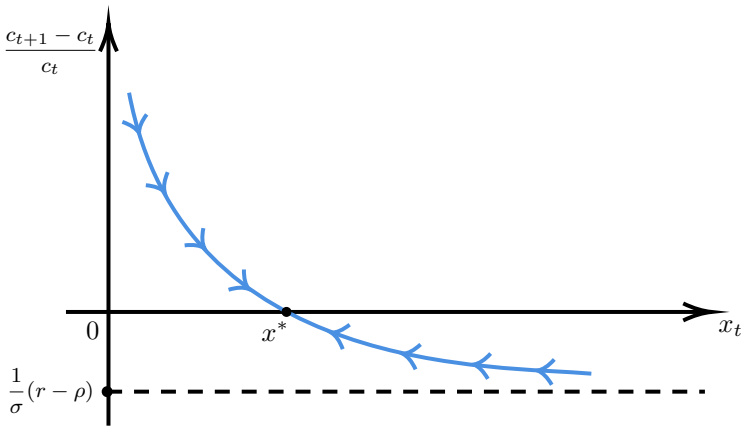



Figure 9.4: Consumption dynamics in the buffer stock model as a function of cash-in-hand $x$.

**Notes**: There exists an optimal target level of cash in hand $x$ toward which the household wants to move.

**Excess sensitivity and liquidity constraints**

According to the permanent income hypothesis, consumption growth between $t-1$ and $t$ should only depend on the innovation in income accruing over the same period. Any past income change should not affect current consumption growth. Empirical tests of this prediction routinely lead to a rejection (e.g., Flavin (1981)) of this hypothesis and to evidence of so-called "excess sensitivity." Deaton (1992) and Jappelli and Pistaferri (2017)) discuss various econometric difficulties with this type of tests. Campbell and Mankiw (1989) also found evidence of excess sensitivity on aggregate time series data, and interpreted the size of the coefficient on past income as the share of "hand-to-mouth" consumers, i.e. households that are either myopic or subject to liquidity constraints.

Zeldes (1989) links explicitly excess sensitivity tests to the presence of credit constraints. If we extend the derivation in (9.36) to the case where household face a borrowing limit, and focus on isoelastic utility with curvature parameter $\sigma$, we obtain the following Euler equation for consumption:

$$\Delta \log c_{t+1} = \frac{1}{\sigma}(r-\delta) + \frac{\sigma}{2}\text{var}_t(\Delta \log c_{t+1}) + \frac{1}{\sigma}\log(1+\lambda_t) + \varepsilon_t$$

where $\lambda_t$ is the multiplier on the borrowing limit, and $\varepsilon_t$ is a forecasting error orthogonal to consumption growth. Note that the multiplier acts as an omitted variable, if not always equal to zero.

Zeldes splits his empirical sample of PSID households into low-wealth (likely constrained) and high-wealth (likely unconstrained) households, and adds lagged income as a regressor to that equation. He argues that, if liquidity constraints bind, one would

estimate a significant negative coefficient on lagged income for the low-wealth sample. The reason is that, upon receiving a high income realization, those households would want to save some of it which relaxes the constraint and lowers the value of the multiplier $\lambda_t$. Zeldes' empirical estimates are consistent with the view that liquidity constraints bind for a subgroup of the population, but not for all.

**Comparative statics on the consumption function**

Figure 9.5 plots the optimal consumption function obtained by solving numerically the income fluctuation problem (9.31). In particular, we have assumed that $u$ is CRRA with risk aversion equal to 2, $\beta = 0.95$, $R = 1.02$, the borrowing limit is zero, and income follows a two-state Markov chain with values $\{0.5, 1.5\}$, and probability that each income state persists into next period equal to 0.8.

The six panels illustrate how the solution to the household problem changes as we change one parameter at the time. Panel (A) shows that households with a low income realization consume less on average for two reasons. First, mechanically, they have less resources. Second, when they are not constrained they are more concerned about hitting the constraint in the future and thus have a stronger precautionary motive. Note that, as wealth grows the households becomes a permanent-income consumer, and the consumption gap converges to the expected difference in permanent labor income. In particular, for wealth large enough the slope of the consumption function is the same. Turning to panel (B), more impatient households (low discounting) consume more and, as long as they are unconstrained, have a higher MPC, even for high values of wealth. Recall the expression for the MPC under full insurance in equation (9.15). Panel (C) shows that, since $\beta R < 1$ households who are more willing to substitute intertemporally consume more. In addition, as seen from (9.15), their MPC is larger even for high levels of wealth. Panel (D) illustrates that households who face a higher rate of return on saving optimally accumulate more assets and consume less for a given level of wealth (but they will be richer on average). Panel (E) shows that a mean-preserving spread in income risk lowers consumption because it expands precautionary saving. For the same reason, the credit constraint binds for a smaller wealth range. Finally, panel (F) depicts optimal consumption under two values for the credit limit. Households who face a loose constraint save less for precautionary reasons. This result indicates the existence of substitutability between precautionary saving and ability to borrow to smooth consumption when needed.

### 9.3.5 Bounds on wealth accumulation

As discussed, there are two forces that determine saving rates in the stochastic income fluctuation problem: intertemporal and precautionary motive. The latter is always positive. Whereas the first is positive or negative depends on whether $\beta R$ is above or below one. It is therefore intuitive that unless $\beta R$ is small enough, households will tend to keep accumulating savings, and an upper bound on wealth might not exist. This result represents a threat to existence of equilibrium once we combine a continuum of households and let them trade in

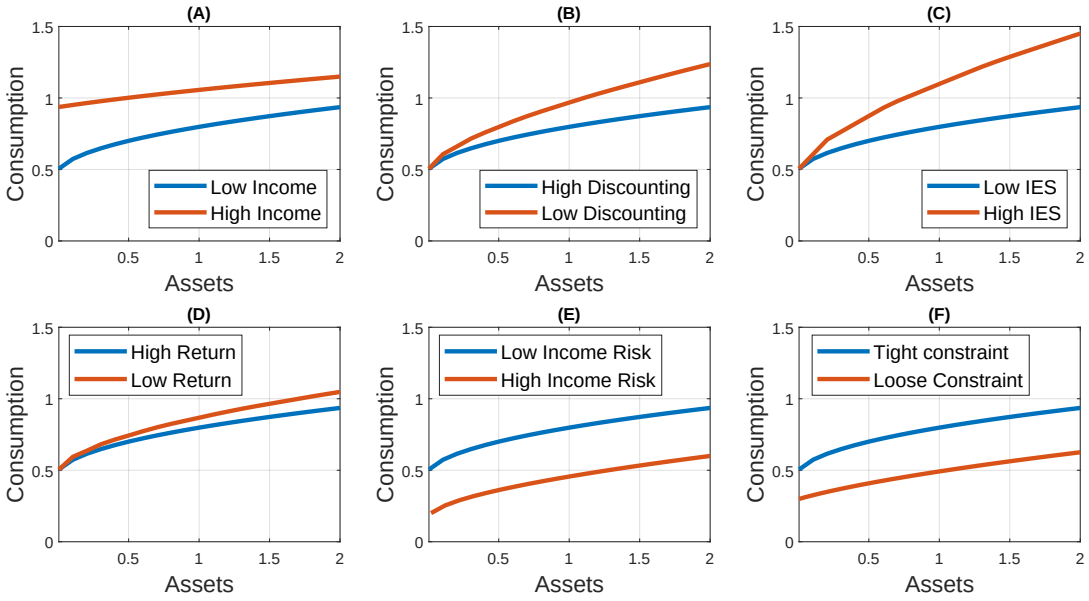



Figure 9.5: Comparative statics of the optimal consumption function with respect to various parameters in the income fluctuations problem.

financial markets: if assets can grow without limit, the state space is not compact and an equilibrium might not exist.[23]

It turns out that a sufficient condition for wealth accumulation to be bounded in the stochastic version of the income fluctuation problem is $\beta R < 1$ (Schechtman and Escudero, 1977; Clarida, 1987; Huggett, 1997).[24] We now provide an informal version of the proof of this result based on Schechtman and Escudero (1977): besides $\beta R < 1$, the proof assumes i.i.d. income shocks. We use a recursive formulation where 'prime' denotes next-period variables and $u_c$ the marginal utility of consumption. Let $x = Ra + y$ be cash in hand, and $c(x)$ the consumption decision rule –a function of the individual state variable $x$. From the Euler Equation:

$$u_c\left(c\left(x\right)\right) = \beta R \mathbb{E}\left[u_c\left(c\left(x'\right)\right)\right], \tag{9.41}$$

where $x' = Ra'\left(x\right) + y'$ is cash in hand next period, given that today's cash in hand is $x$ and that next period income realization is $y'$. Let $\bar{x}' = Ra'\left(x\right) + \bar{y}$ be the cash in hand associated to the maximum realization of income $\bar{y}$ next period (hence maximum cash in hand next period), given today's cash in hand $x$. We can write:

$$u_c\left(c\left(x\right)\right) = \beta R \mathbb{E}\left[u_c\left(c\left(x'\right)\right)\right] = \beta R \frac{\mathbb{E}\left[u_c\left(c\left(x'\right)\right)\right]}{u_c\left(c\left(\bar{x}'\right)\right)} u_c\left(c\left(\bar{x}'\right)\right). \tag{9.42}$$

[23] A lower bound for the asset space is always guaranteed by the ad-hoc or natural borrowing limit.

[24] The counterpart sufficient condition in the deterministic version of the income fluctuation problem is $\beta R \leq 1$. It is a weaker condition because the precautionary saving motive is absent. See Ljungqvist and Sargent (2018) for a thorough discussion of both the deterministic and the stochastic case.

Suppose we can show that

$$\lim_{x\to\infty}\frac{\mathbb{E}\left[u_c\left(c\left(x'\right)\right)\right]}{u_c\left(c\left(\bar{x}'\right)\right)}=1. \tag{9.43}$$

Then, since $\beta R<1$ by assumption, for $x$ large enough the Euler equation (9.42) yields

$$u_c\left(c\left(x\right)\right)=\beta R u_c\left(c\left(\bar{x}'\right)\right)<u_c\left(c\left(\bar{x}'\right)\right).$$

Concavity of $u$ and monotonicity of $c$ with respect to $x$, proved in (9.38) imply that $\bar{x}'\left(x\right)<x$. This would conclude the proof, because we would have demonstrated that cash in hand does not increase forever: when $x$ is large enough, $x'<x$ for sure.

We need to establish conditions under which the limit in (9.43) holds. Consider the marginal utility function $u_c\left(c\left(x'\right)\right)$ and compute a first-order Taylor approximation around $x'=\bar{x}'$ :

$$u_c\left(c\left(x'\right)\right)\simeq u_c\left(c\left(\bar{x}'\right)\right)+u_{cc}\left(c\left(\bar{x}'\right)\right)c_x\left(\bar{x}'\right)\left(x'-\bar{x}'\right).$$

Taking expectations of both sides

$$\begin{aligned}\mathbb{E}\left[u_c\left(c\left(x'\right)\right)\right] &\simeq u_c\left(c\left(\bar{x}'\right)\right)-u_{cc}\left(c\left(\bar{x}'\right)\right)\mathbb{E}\left[\bar{x}'-x'\right]c_x\left(\bar{x}'\right)\\ &= u_c\left(c\left(\bar{x}'\right)\right)-u_{cc}\left(c\left(\bar{x}'\right)\right)\mathbb{E}\left[\bar{y}-y'\right]c_x\left(\bar{x}'\right).\end{aligned} \tag{9.44}$$

In the first line we used that $\bar{x}'$ is deterministic since it is implied by the specific income realization $\bar{y}$. In the second line we used the fact that $x'\equiv Ra'+y'$ which implies $\bar{x}'-x'\equiv\bar{y}-y'$. Dividing equation (9.44) by $u_c\left(c\left(\bar{x}'\right)\right)$ we obtain:

$$\frac{\mathbb{E}\left[u_c\left(c\left(x'\right)\right)\right]}{u_c\left(c\left(\bar{x}'\right)\right)}\simeq 1+\alpha\left(c\left(\bar{x}'\right)\right)\left[\bar{y}-\mathbb{E}\left(y'\right)\right]c_x\left(\bar{x}'\right),$$

where $\alpha\left(c\left(\bar{x}'\right)\right)$ is the coefficient of absolute risk aversion evaluated at consumption level $c\left(\bar{x}'\right)$. Since both $\bar{y}-\mathbb{E}\left(y'\right)$ and $c_x\left(\bar{x}'\right)$ are positive and finite (recall that the $c$ function is concave), a sufficient condition for the limit in (9.43) to hold is that

$$\lim_{x\to\infty}\alpha\left(c\left(x\right)\right)=0. \tag{9.45}$$

Condition (9.45) requires that absolute risk aversion decreases with wealth, i.e. that $u$ belongs to the DARA class. DARA means that the household is less and less concerned about income uncertainty as they get rich because they are less risk averse, so they will consume more and accumulate less precautionary wealth. As wealth increases, the intertemporal dissaving motive (present because of $\beta R<1$) eventually overcomes the precautionary saving motive, and wealth will decrease.[25]

To conclude, the condition $\beta R<1$ is sufficient to guarantee a bounded asset space in the household problem. In the next section, we'll prove that this inequality is also a feature of the stationary competitive equilibrium in an economy with a continuum of households.

[25] Huggett (1993) generalizes this proof to a 2-state Markov chain for the income process, under CRRA utility.

## 9.4 Heterogeneous-agent incomplete-market models

Now that we have analyzed in some depth the income fluctuation problem, we want to characterize the equilibrium of economies populated by a continuum of households subject to uninsurable income shocks who can only trade a risk-free bond. This class of heterogeneous-agent incomplete-market models has become a workhorse of modern macroeconomics. These models feature an endogenous distribution of consumption and wealth across households. In sharp contrast with its complete-market counterpart, this framework features partial consumption insurance, and mobility within the consumption distribution. In addition, its equilibrium allocations are not socially efficient, which makes policy analysis interesting.

A variety of important questions have been addressed in this framework, for example: what is the fraction of wealth accumulated because of the precautionary motive (Aiyagari, 1994)? How much of the observed wealth inequality can one explain through income inequality (Castaneda, Diaz-Gimenez, and Rios-Rull, 2003)? What are the optimal levels of taxation and public debt (Aiyagari and McGrattan, 1998; Domeij and Floden, 2006; Heathcote, 2005a)? How does heterogeneity change the transmission of aggregate shocks compared to the representative agent model (Krusell and Smith, 1998; Kaplan and Violante, 2018)? Can market incompleteness help in generating a large equity premium (Storesletten, Telmer, and Yaron, 2001) (Krueger and Lustig, 2010)? Are welfare costs of business cycles higher with incomplete markets (Krusell, Mukoyama, Şahin, and Smith Jr, 2009)?[26]

In what follows, we describe two versions of the model, an endowment economy and a production economy, and discuss existence and uniqueness of the stationary equilibrium.

### 9.4.1 An endowment economy

Consider the endowment economy introduced in Section 7.6. Time is discrete and indexed by $t$. There is no aggregate uncertainty. The economy is populated with a continuum of measure one of infinitely lived, ex-ante identical households. Households have time-separable preferences over streams of consumption (the final numeraire good)

$$U = \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t u(c_t),$$

where the period utility function $u$ satisfies $u' > 0, u'' < 0$, and the discount factor $\beta \in (0, 1)$. The expectation is over future sequences of idiosyncratic shocks, conditional on the information set at time $t = 0$. Each household faces stochastic fluctuations in their endowment of the final good $y_t \in Y$, a finite set. This stochastic process follows a first-order ergodic N-state Markov chain with transition probabilities $\pi(y', y) = \Pr(y_{t+1} = y' \mid y_t = y)$. Since shocks are i.i.d across consumers, a law of large numbers holds: $\pi(y', y)$ is also the fraction of agents in the population subject to this particular transition from $y$ to $y'$ between time $t$

[26] For surveys of applications of this framework, see Heathcote, Storesletten, and Violante (2009); Quadrini and Ríos-Rull (2015); De Nardi and Fella (2017); Benhabib, Bisin, and Luo (2019); Kaplan and Violante (2022).

and $t+1$ (Uhlig, 1996). Let $\Pi^*(y)$ be the a unique invariant (i.e. limiting) distribution of endowments. The household budget constraint at time $t$ is

$$c_t + a_{t+1} = R_t a_t + y_t.$$

Wealth $a_t$ takes the form of a one-period non-state-contingent asset, with return $R_t \equiv 1 + r_t$ independently of the individual realization $y_t$. At every $t$, agents face the ad-hoc borrowing limit

$$a_{t+1} \geq -\underline{a}$$

where $\underline{a}$ is a parameter which we assume is more stringent than the natural debt limit. We begin by positing that this asset is in zero aggregate net supply, as in Huggett (1993), and traded among households in a competitive financial market.

**Recursive formulation**

We now restate the economy in recursive form. The dynamic programming version of the household problem described above is

$$\begin{aligned} V(a,y) &= \max_{c,a'} \left\{ u(c) + \beta \sum_{y' \in Y} \pi(y',y) V(a',y') \right\} \\ &\quad s.t. \\ c + a' &= Ra + y \\ a' &\geq -\underline{a} \end{aligned} \tag{9.46}$$

where $v$ denotes the value function, and the pair $(a,y)$ is the individual state vector.

Let $\lambda$ be the probability distribution of agents over states. Let $\bar{a}$ be the maximum asset holding in the economy, and for now assume that such upper bound exists. Let $A \equiv [-\underline{a}, \bar{a}]$ be the asset space. Let the state space $S$ be the Cartesian product $A \times Y$. Let the $\sigma$-algebra $\Sigma_s$ be defined as $B_A \otimes P(Y)$ where $B_A$ is the Borel sigma-algebra on $A$ and and $P(Y)$ is the power set of $Y$. The space $(S, \Sigma_s)$ is a measurable space. Let $\mathcal{S} = (\mathcal{A} \times \mathcal{Y})$ be the typical set of $\Sigma_s$. For any element of the sigma algebra $\mathcal{S} \in \Sigma_s$, $\lambda(\mathcal{S})$ is the measure of agents in the set $\mathcal{S}$.

How can we characterize the way individuals transit across states over time? We need a transition function. Define $Q((a,y), \mathcal{A} \times \mathcal{Y})$ as the (conditional) probability that an individual with current state $(a,y)$ transits to the set $\mathcal{A} \times \mathcal{Y}$ next period. Formally, $Q : S \times \Sigma_s \to [0,1]$, and

$$Q((a,y), \mathcal{A} \times \mathcal{Y}) = \mathbb{I}_{\{a'(a,y) \in \mathcal{A}\}} \sum_{y' \in \mathcal{Y}} \pi(y',y) \tag{9.47}$$

where $\mathbb{I}_{\{.\}}$ is the indicator function, and $a'(a,y)$ is the saving policy. This transition function can be used to construct a sequence of distributions as

$$\lambda_{n+1}(\mathcal{A} \times \mathcal{Y}) = \int_{A \times Y} Q((a,y), \mathcal{A} \times \mathcal{Y}) \, d\lambda_n \tag{9.48}$$

where $d\lambda_n$ is short for $\lambda_n(da, d\varepsilon)$. We are now ready to proceed to the definition of a stationary, or steady-state, equilibrium.

### Stationary equilibrium

A recursive stationary competitive equilibrium is: (a) value function $V : S \to \mathbb{R}$, (b) policy functions $a' : S \to \mathbb{R}$, and $c : S \to \mathbb{R}_+$, (c) an interest rate $r^*$, and (d) a stationary measure $\lambda^*$ such that:

- Given $r$, the policy functions $a'$ and $c$ solve the household's problem (9.46) and $V$ is the associated value function
- The asset market clears: $\int_{A\times Y} a'(a, y)d\lambda^* = 0$ at the interest rate $r^*$
- The goods market clears: $\int_{A\times Y} c(a, y)d\lambda^* = \sum_{i=1}^{N} y_i \Pi^*(y_i)$
- For all $(\mathcal{A} \times \mathcal{Y}) \in \Sigma_s$, the invariant probability measure $\lambda^*$ is the fixed point of (9.48) and satisfies
$$\lambda^*(\mathcal{A} \times \mathcal{Y}) = \int_{A\times Y} Q((a, y), \mathcal{A} \times \mathcal{Y}) d\lambda^*,$$
where $Q$ is the transition function defined in (9.47).

In the stationary equilibrium, households optimize, markets clear, and the distribution of agents across states is invariant, i.e. this probability measure will reproduce itself permanently. Households, however, will exchange places and move upward and downward within this joint distribution of income and wealth, since they face different sequences of endowment shocks.

What guarantees that a steady-state exists and is unique? We start from existence.

**Existence.** Let us express the excess aggregate demand function for assets in financial markets as
$$A(r) = \int_{A\times Y} a'(a, y; r)d\lambda_r^*,$$
where we made explicit the dependence of the policy function and the stationary distribution on the interest rate $r$ –think of $r$ as a parameter for now. One of the conditions that yield existence is that $A(r)$ is continuous in $r$. $A(r)$ is continuous if both $a'$ and $\lambda_r^*$ are, in turn, continuous in $r$.

The theorem of the maximum implies that if $u$ is continuous, $u' > 0$ and $u'' < 0$, the solution to the household problem is unique and the policy function $a'(a, \varepsilon; r)$ is continuous in $r$.

Stokey, Lucas and Prescott (Theorems 12.12 and 12.13) lay out conditions under which $\lambda_r^*$ is continuous in $r$. Essentially, the stationary distribution must exist and be unique. We refer the interested reader to Stokey, Lucas and Prescott for the exact conditions needed and to Huggett (1993) and Aiyagari (1994) for proofs that these assumptions are met in this

environment. Most of these proofs are straightforward. We highlight two conditions that deserve more discussion, one for existence and one for uniqueness.

Existence of an invariant distribution requires compactness of the state space, i.e. a finite upper bound for wealth $a$. As explained, $\beta R < 1$ is a sufficient to obtain a finite upper bound for $a$, but $R$ is endogenous so this restrictions cannot be assumed. It must hold in equilibrium, and later in this section we show that it is true.

Uniqueness of the invariant distribution is guaranteed, for example, if the transition function $Q\left((a,y),\mathcal{A}\times\mathcal{Y}\right)$ satisfies, for any given $r$, the "monotone mixing condition" (and a few regularity conditions). This condition states that there is a positive probability that a household moves from the bottom to the top of the asset space (and viceversa) in finite time. Namely, the economy must have enough upward and downward mobility within the income and wealth distribution. To see why this condition is satisfied in our model, suppose the household starts from $(\bar{a}, y_{\max})$ and receives a long stream of the worst realization of the shock $y_{\min}$. If $y$ follows a stationary (i.e., mean reverting) process, the household will keep decumulating wealth to smooth consumption until reaching some neighborhood of the lower bound. Symmetrically, suppose the household starts with $(-\underline{a}, y_{\min})$ and receives a long stream of the best shock $y_{\max}$. Knowing that sooner or later the shock will revert toward its mean, they will keep accumulating wealth until they reach some neighborhood of the upper bound.

Having proved that the $A(r)$ function is continuous, we need to argue that it crosses zero at least once at some finite value for $r$. For $\beta(1+r)=1$, we know from Section 9.3.5 that households will keep accumulating wealth without bound, so $A\left(\frac{1}{\beta}-1\right)=+\infty$. For low enough (possibly negative) values of $r$, every household would want to borrow. For example, it is clear that if $r=-1$, it is optimal to borrow up to the limit since one would never have to repay, and thus $A(-1)=-\underline{a}$. Since $A(r)$ is continuous, it will cross zero at least once and an equilibrium exists. This logic is represented graphically in Figure 9.6. Note that the equilibrium interest rate $r^*$ will always lie below its complete market counterpart $1/\beta-1$. As a result, the condition $\beta R^* < 1$ holds in equilibrium, which puts a limit to wealth accumulation and confirms that the asset space is bounded above.

**A solution to the risk-free rate puzzle**

The observation that the equilibrium interest rate is lower than under full insurance is important for asset pricing, as emphasized by Huggett (1993). The representative agent model can only generate a high equity premium for very large levels of risk aversion (Mehra and Prescott, 1985) which, under CRRA utility, imply very low values for the intertemporal elasticity of substitution. In this situation, households have a very sharp desire to smooth consumption. When income has positive growth on average, as in the data, consumption smoothing dictates a strong desire to borrow against future income which pushes up the real rate well above observed values.

The cost of solving the equity premium puzzle with complete markets is, therefore, the "risk-free rate puzzle": the return on a risk-free bond is too high compared to the

historical data. Huggett (1993) showed that the presence of uninsurable idiosyncratic income risk can help solving the puzzle. As soon as one deviates from complete markets, the precautionary saving motive sets in (and a strong one if risk aversion is high) and, with it, a desire to save that pulls down the equilibrium interest rate on safe assets and better aligns it to the data.

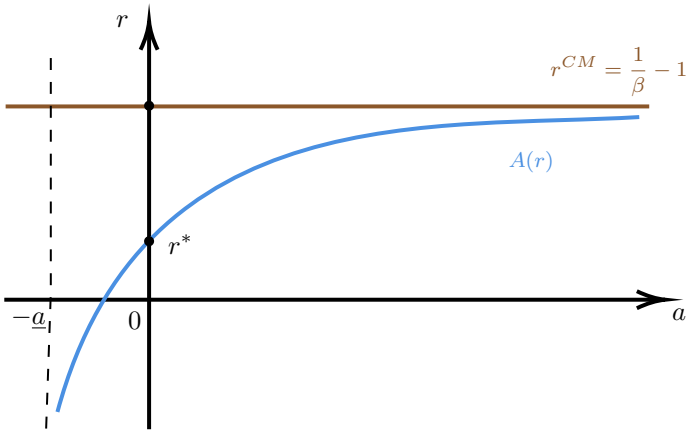



Figure 9.6: Equilibrium of the endowment economy where assets are in zero net supply.

**Notes**: $A(r)$ denotes aggregate asset holdings of the household sector as a function of the interest rate in the incomplete-market model. $r^{CM}$ denotes the complete markets equilibrium real rate, and also the infinitely elastic demand for assets in the complete market model. $r^*$ is the equilibrium real rate in the model with incomplete markets.

**Uniqueness.** Having proved that the equilibrium exists, we ask: is the equilibrium unique? Uniqueness is guaranteed if the function $A(r)$ is monotonically increasing in $r$. This property is hard to prove in general because changes in $r$ have both income and substitution effects on savings: the relative dominance between the two could switch at a certain level of assets, so $a'(a, \varepsilon; r)$ may not be monotone in r. Achdou, Han, Lasry, Lions, and Moll (2022) state a sufficient condition for CRRA utility in a continuous time version of this model. Let $\sigma$ be the coefficient of risk aversion. Recall that the smaller is $\sigma$ the larger is the income effect and a higher $r$ increases savings. Uniqueness is obtained if $\sigma < 1$.

### Assets in positive supply

We conclude this section by extending this logic to an endowment economy where assets are not in zero net supply, but are one-period real bonds (or real balances of fiat money) issued by the government. This is the economy studied by Bewley (1980, 1983). Let $B$ be this

amount, fixed in steady-state. The government budget constraint is

$$rB = T \tag{9.49}$$

where $T$ is the lump-sum tax (if $r > 0$) levied on households to finance interest payments. This is a new equilibrium condition determining $T$. The household budget constraint in (9.46) is modified as

$$c + a' = Ra + y - T,$$

and the equilibrium condition in the asset market, represented graphically in Figure 9.7, becomes

$$\int_{A\times Y} a'(a, y; r) d\lambda_r^* = B.$$

The analysis of equilibrium is virtually the same as in the previous section, with the only modification that we have an additional endogenous variable, $T$, and an additional equation, (9.49).

We conclude by noting that, in this economy, the government would have an incentive to provide more liquidity to households by increasing the level of debt $B$. Returning to Figure, as $B$ increases, the economy moves closer to the full insurance outcome. In the economy with production that we analyze next, however, expanding government debt also entails a cost, because it crowds out capital and production. Aiyagari and McGrattan (1998) analyze this trade-off in the determination of the optimal quantity of debt.

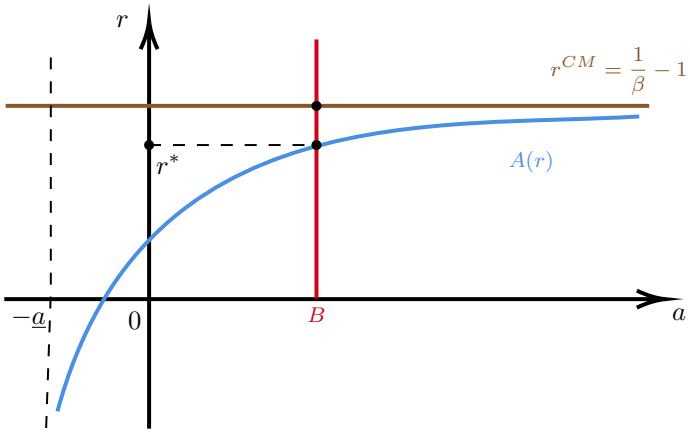



Figure 9.7: Equilibrium of the endowment economy where assets are in fixed positive supply.

**Notes**: $A(r)$ denotes aggregate asset holdings of the household sector as a function of the interest rate in the incomplete-market model. $r^{CM}$ denotes the complete markets equilibrium real rate, and $r^*$ the equilibrium real rate in the model with incomplete markets.

### 9.4.2 A production economy

We now follow Aiyagari (1994) and introduce production in this economy. We re-interpret $y$ as efficiency units of labor (i.e., individual labor productivity) and assume that households supply labor inelastically, with each individual time endowment normalized to 1. The asset $a$ represents financial claims to physical capital. Households supply labor and capital to a representative firm. The firm produces the final good with CRS production function $F(K, L)$, which uses capital and efficiency units of labor as inputs and satisfies $F_K > 0, F_L > 0, F_{KK} < 0, F_{LL} < 0$ as well as standard Inada conditions, $\lim_{K\to\infty} F_K = 0$ and $\lim_{K\to 0} F_K = +\infty$. Physical capital depreciates geometrically at rate $\delta \in (0, 1)$. Firms act competitively by maximizing profits taking prices as given. The final numeraire good can be used for consumption and investment, and is traded in a competitive good market. The labor market and capital market are also competitive and clear, respectively, at the wage rate $w$ (per efficiency unit) and interest rate $r$.

The household budget constraint in problem (9.46) becomes

$$c + a' = Ra + wy$$

and everything else in the household problem is unchanged. We now state the new definition of equilibrium.

A recursive stationary competitive equilibrium is: (a) value function $V : S \to \mathbb{R}$, (b) policy functions $a' : S \to \mathbb{R}$, and $c : S \to \mathbb{R}_+$, (c) firm choices $L$ and $K$, (d) prices $r^*$ and $w^*$, and (e) a stationary measure $\lambda^*$such that:

- Given $(r, w)$, the policy functions $a'$ and $c$ solve the household's problem (9.46) and $V$ is the associated value function
- Given $(r, w)$, the firm chooses optimally its capital stock $K$ and its labor input $L$, i.e., $r + \delta = F_K(K, L)$ and $w = F_L(K, L)$
- The labor market clears: $L = \sum_{i=1}^{N} y_i \Pi^* (y_i)$ at the wage $w^*$
- The asset market clears: $\int_{A\times Y} a'(a, y) d\lambda^* = K$ at the interest rate $r^*$
- The goods market clears: $\int_{A\times Y} c(a, y) d\lambda^* + \delta K = F(K, L)$
- For all $(\mathcal{A} \times \mathcal{Y}) \in \Sigma_s$, the invariant probability measure $\lambda^*$ satisfies

$$\lambda^* (\mathcal{A} \times \mathcal{Y}) = \int_{A\times Y} Q ((a, y), \mathcal{A} \times \mathcal{Y}) d\lambda^*,$$

Production changes the shape of the aggregate demand for capital from firms. To characterize it, we need to consider the firm's problem. From the optimal choice of the firm, we obtain $K(r)$ implicitly from $F_K(K, L) = r + \delta$. It is immediate to see that for $r = -\delta$, then $K \to +\infty$, while for $r \to +\infty$, $K \to 0$, given our assumptions on $F$, in particular the Inada conditions. Thus, firms' demand for capital is a continuous, strictly decreasing function of

the interest rate $r$.[27] Figure 9.8 illustrates the equilibrium in this model. Existence follows from the same conditions discussed above.

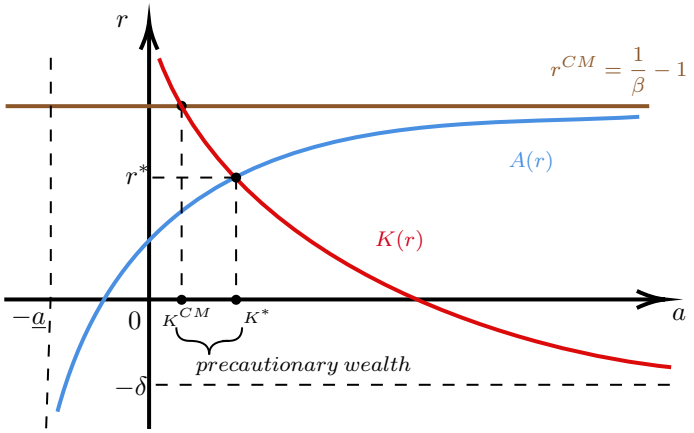



Figure 9.8: Equilibrium of the production economy.

**Notes**: $A(r)$ denotes aggregate asset holdings of the household sector as a function of the interest rate in the incomplete-market model. $r^{CM}$ and $K^{CM}$ denote, respectively, the complete markets equilibrium real rate and capital stock. $r^*$ and $K^*$ are their incomplete markets counterparts. The difference between $K^*$ and $K^{CM}$ is the equilibrium amount of precautionary wealth.

**Quantitative analysis of the model.**

Aiyagari (1994) calibrated the model to the the US economy. One of the key ingredients of the model, the stochastic process of earnings shocks (in particular, their persistence and volatility), is estimated from longitudinal micro data on individual labor income. Aiyagari reached three important conclusions from his quantitative analysis, all results that have shaped the literature for decades.

First, he argued that this incomplete market structure is quite effective at insuring income risk. By saving, dissaving, and borrowing (i.e., through self-insurance), an individual can cut consumption variability by about half compared with autarky. Second, he pointed out that one can use the model to quantify the amount of aggregate wealth held by households because of precautionary reasons. The capital stock in complete markets is the value $K^{CM}$ where the aggregate demand for capital by firms crosses the infinitely elastic supply of capital of the representative consumer at $r^{CM} = 1/\beta - 1$. The difference between $K^*$ and $K^{CM}$ is therefore the additional precautionary stock of capital. He concluded that, in his baseline calibration, the precautionary motive augments the aggregate saving rate by 3 percentage points, but calibrations featuring higher risk aversion or more income volatility can raise the aggregate saving rate by 10 percentage points or more. Third, he observed that the model can

[27] For example, if $F(K, L) = K^{\alpha} L^{1-\alpha}$, then $K(r) = \left(\frac{\alpha L}{\delta + r}\right)^{\frac{1}{1-\alpha}}$.

generate an equilibrium wealth distribution that is more positively skewed (mean > median) and more dispersed than the income distribution, as in the data. Compared to its empirical counterpart, though, the model's wealth distribution features too much wealth accumulation at the bottom and too little wealth concentration at the top. Chapter 19 describes the progress made in the literature on wealth inequality since Aiyagari's observation.

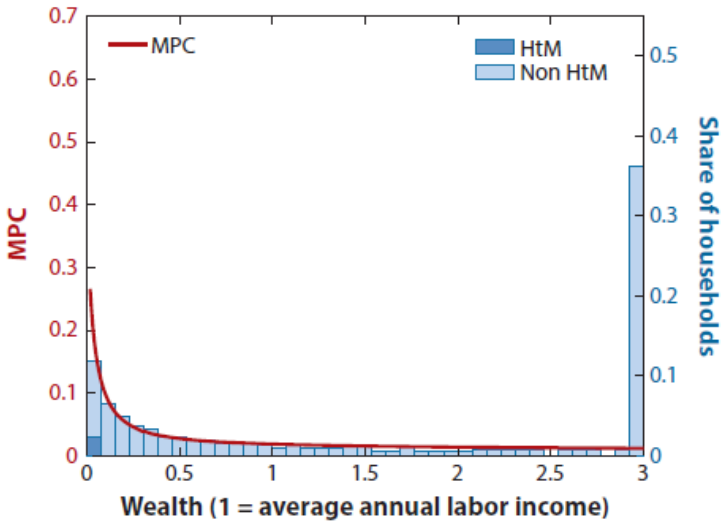



Figure 9.9: MPC, i.e. the slope of the consumption decision rule (red curve), as a function of wealth jointly with the distribution of wealth (blue bars).

**Notes**: The dark blue bar near zero represents the share of constrained (or hand-to-mouth) households in the calibrated model. This figure is reproduced from Kaplan and Violante (2022)

**The marginal propensity to consume: models vs data**

In Section 9.3.1 we showed that, under the permanent income hypothesis (PIH), the marginal propensity to consume (MPC) out of wealth and unanticipated transitory changes in income equals $1 - R^{-1}(\beta R)^{\frac{1}{\sigma}}$. In the equilibrium of a representative agent economy, where $\beta R^* = 1$, the MPC therefore roughly equals the real interest rate, or around 1% on a quarterly basis. In addition, according to the PIH, the MPC out of anticipated income changes is zero.

There is a vast body of work that estimates the MPC out of (more or less anticipated) transitory changes in income using various approaches. Some of the most convincing evidence comes from the observation that when households receives fiscal stimulus pay-

ments from the government (like in the last three recessions, for example), they spend on average around 15-30% of this transfer on nondurable goods in the first quarter after receipt (e.g., Parker, Souleles, Johnson, and McClelland (2013)). These estimates indicate that, empirically, the MPC might be 20 times as large in the data than what implied by the naive PIH.

The heterogeneous-agent incomplete-market model we analyzed has the potential to match the data better. As discussed in Section 9.3.4, for low enough wealth levels the slope of the consumption function can be much steeper than $r$. The key question is: in the equilibrium of a plausibly calibrated model, what is the share of agents with wealth holdings in that range? The answer is not many for two reasons. First, in order for the model to match the aggregate wealth level observed in the U.S., the model discount factor has to be high. Second, precisely because being on a constraint induces excessive consumption fluctuations that are costly in terms of welfare, optimization leads households to save and keep away from their credit limit. Overall, the quarterly MPC can be boosted up to 5% or so, but no more. Figure 9.9, reproduced from Kaplan and Violante (2022) illustrates this result.

The literature has studied how to modify the baseline model to generate levels of average MPC more in line with the data. We refer to Kaplan and Violante (2022) for a comprehensive survey, and here only discuss two of them. First, one can introduce heterogeneity in discount factors across consumers. Households with low discount factors display steeper consumption functions (recall our comparative statics of Section 9.3.4). In addition, because of impatience, they tend to dissave and thus hold small amounts of wealth. With enough households of this type, the average MPC of the economy can be large. Second, one can take the view that the relevant notion of wealth for short-term consumption smoothing is *liquid wealth*, e.g., cash and bank deposits but not housing or retirement accounts. Empirically, liquid wealth is only a small fraction of total wealth. A model with two types of assets, a liquid one and an illiquid one which can only be accessed by paying a transaction cost, is able to yield averages MPC in line with the data. Interestingly, this type of models features *wealthy hand-to-mouth* consumers, i.e. consumers who do have substantial wealth locked in high-return illiquid assets, but hold small amounts of liquid wealth and thus are highly responsive to transitory income changes (Kaplan and Violante, 2014; Kaplan, Violante, and Weidner, 2014).

# Chapter 10

# Labor supply

*Richard Rogerson and Gianluca Violante*

# Chapter 11

# Growth

*Timo Boppart and Pete Klenow*

# Chapter 12

# Real business cycles

*Kurt Mitman*

# Chapter 13

# Government and Public Policies

*Marina Azzimonti, Jonathan Heathcote, and Kjetil Storesletten*

## 13.1 Introduction

The government has a large impact on economic outcomes through fiscal policy, monetary policy, and regulation policy. In this chapter, we focus on fiscal policy; in particular, on taxes, government spending, and debt. After presenting a summary of how governments tax, spend and borrow in practice, we turn to theory and discuss how fiscal policy choices impact the competitive equilibrium allocation, and how to frame the problem of optimizing over those policy choices. Monetary policy is discussed in Chapter 16. Our discussion centers on developed economies, with a particular focus on the United States. Chapter 22 introduces fiscal policy in emerging markets.

In Chapter 6, we discussed conditions under which the First Welfare Theorem holds. When markets are complete and competitive and there are no public goods or externalities –i.e., there are no "market failures"– competitive equilibrium allocations are Pareto optimal. Why then, do governments intervene in the economy? There are three main rationales.

The first is that there are public goods, such as national defense, that the market cannot provide because there is no way to restrict the enjoyment of public goods to those households who choose to pay for them. There are other goods and services, like education and healthcare, that are not pure public goods, but whose consumption confers large positive externalities. For example, if my neighbors are vaccinated, they are less likely to make me sick. Thus, absent government involvement, education and healthcare might be under-consumed.

A second reason governments intervene is that markets are not complete and competitive, in part because of private information frictions. For example, it might be difficult to buy private unemployment insurance, or annuities that insure against longevity risk. Thus, there may be a role for the government to provide public unemployment insurance, or to fund a public pension system. It is also possible that absent government intervention, the economy might occasionally get stuck in an inefficiently depressed equilibrium because of frictions in private markets. Thus, the government intervened during the Global Financial Crisis in 2008, bailing out a range of financial institutions to avoid a cascade of bankruptcies, and cutting taxes to try to boost consumer confidence.

The third rationale for government intervention is redistribution. Market economies tend to generate substantial income inequality, as discussed in Chapter 9 (and later in Chapter 19). This inequality may be Pareto efficient, but taxing the rich in order to fund transfers to the poor will generate a more equal allocation of resources, and one that a majority of households might prefer. In many economies, transfers account for most of total government spending. We discuss redistribution in Section 13.7.1.

The impact of the government on equilibrium allocations depends not just on how much the government wants to spend, but also on how the government pays for that spending. In practice, the taxes that households and firms are required to pay depend on the choices they make about how much to work and earn, how much to consume versus save, and how much to invest. Thus the tax system distorts all those choices, ultimately reducing output. We explore the effects of distortionary taxation by adding proportional capital and labor income taxes to a standard neoclassical growth model. Note that higher public consumption or higher transfers necessitates higher tax rates, implying larger distortions to

private sector choices and lower efficiency. There is significant cross-country variation in total government spending and in the extent of redistribution through the tax and transfer system. That suggests societies differ in how they view the trade-off between the benefits of a more equitable distribution of resources or higher public good provision, versus the efficiency costs of higher and more distortionary taxes (see Figure 13.1).

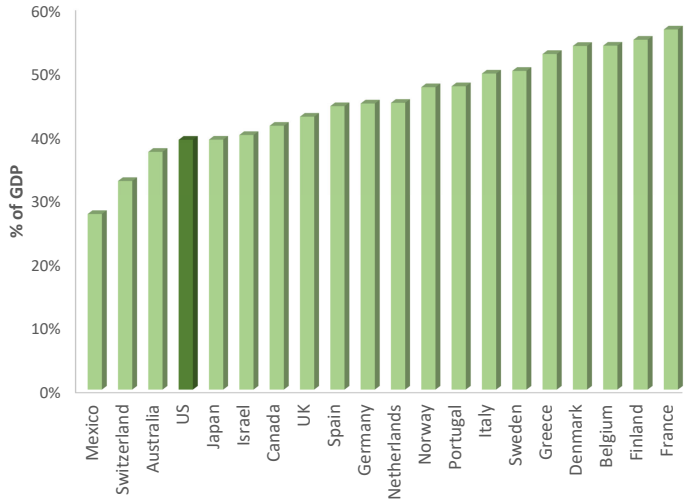



Figure 13.1: Government Spending across Countries (avg. 2010-2019).

## 13.2 Public Finance: An Overview of the Data

The government spends money on publicly-provided goods and services $G_t$, such as education and defense, and makes transfers $T_t$ to individuals and corporations, such as food stamps and agricultural subsidies. These expenditures are financed out of tax revenues $Rev_t$, collected through taxes on goods and services (sales and excise taxes), taxes on income (income and payroll taxes), property taxes, and taxes on corporate profits. When revenues are insufficient to cover expenditures, the government borrows from domestic households and firms or from international lenders. Denoting the stock of debt at the start of period $t$ by $B_{t-1}$, net borrowing is equal to $B_t - B_{t-1}$. We denote the nominal interest rate on public debt by $i_t$, so $i_t B_{t-1}$ is interest payments. The government budget constraint can be written as

$$\underbrace{G_t + T_t + i_t B_{t-1}}_{\text{Expenditures}} = Rev_t + \underbrace{B_t - B_{t-1}}_{\text{Borrowing}}. \tag{13.1}$$

When expenditures – including interest payments – exceed revenues, we say that the government runs a *deficit*. In that case, $B_t > B_{t-1}$ and public debt rises. When expenditures are lower than revenues, the government runs a *surplus* and debt decreases. The stock of

debt at any point in time, then, is the cumulative sum of net deficits run by a government through history. We now review some facts about the evolution of the main components of the government budget constraint to frame the topics discussed in this chapter.

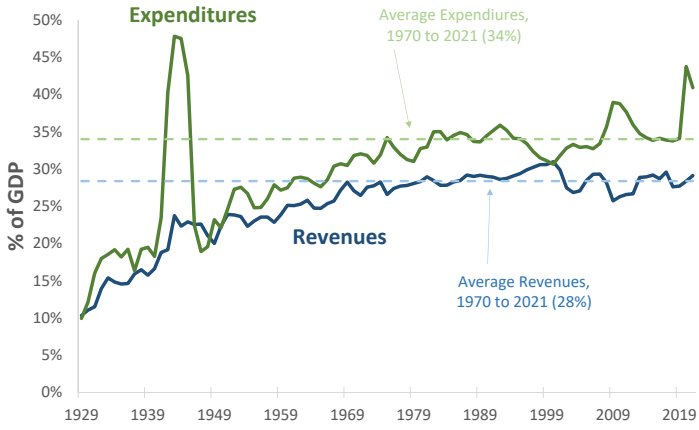



Figure 13.2: Revenues and Outlays, as percentages of GDP.

Figure 13.2 shows revenues and expenditures as percentages of GDP for the U.S. between 1929 and 2021. The series, which incorporate all levels of government (federal, state, and local), were obtained from the NIPA tables constructed by the Bureau of Economic Analysis (see Appendix 13.A.1 for details). Three key points can be drawn from this figure. First, both revenue and spending exhibit upward trends between 1929 and 1970, and then stabilize. Second, expenditures tend to exceed revenues, implying that the U.S. government typically runs deficits. Between 1970 and 2021 expenditures and revenues averaged 34 percent and 28 percent of GDP respectively. Third, expenditures jump during periods of war or recession, while revenues typically fall. Large increases in expenditure are evident during World War II, the Great Recession of 2007-2009, and the COVID-19 recession in 2020.

Figure 13.3 describes how the sources of tax revenue in the United States have changed over time. Over the post-war period, income and social insurance tax revenues increased significantly, and now account for around 11 and 7 percent of GDP, respectively. Revenue from sales and import taxes have been relatively constant at about 8 percent of GDP, while revenue from corporate taxes has declined and is now less than 2 percent of GDP.

The theoretical literature on the impact of taxes differentiates between taxes on labor income, on capital income, and on consumption.

In Section 13.3 we discuss the impact of these taxes on labor supply, investment, and savings choices. Note, however, that this simple theoretical categorization does not map cleanly into the empirical partition of taxes: in particular, while labor earnings are part of the base for U.S. personal income taxes, income taxes taxes also apply to income accruing to capital, including unincorporated business income, dividend, interest, and rental income, and capital gains.

The government can postpone taxation by using public debt to finance expenditure. Does

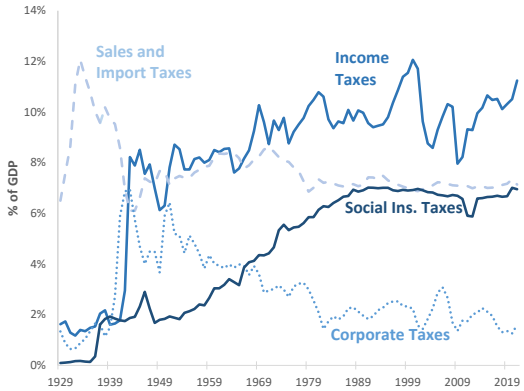



Figure 13.3: Taxes by Category, as percentages of GDP.

it matter whether the government finances spending out of current taxation versus whether it issues debt which must be repaid out of future tax revenue? In Section 13.4 we show that when taxes are lump-sum, the timing of taxes is irrelevant. This famous result is known as "Ricardian Equivalence." However, in the more realistic case in which taxes are distortionary, the timing of taxes does matter. What is then the optimal timing of taxes? In Section 13.5, we formalize the problem of a benevolent government that chooses a sequence for taxes and for debt to maximize social welfare, following a formulation known as the "Ramsey problem." Solving this problem we demonstrate an important "tax smoothing" result: it is optimal to finance temporary shocks such as wars, recessions or pandemics mostly by issuing debt. Evidence that governments do in fact smooth taxes over time is presented in Figure 13.4 (left panel), showing the total deficit (expenditures minus revenues, solid line), the primary deficit (which is the deficit excluding interest payments, dark bars), and interest payments (light bars). The U.S. government borrowed heavily during wars and recessions, particularly during WWII and the COVID-19 pandemic. The right panel shows the stock of debt. In the U.S., most public debt is issued by the Federal government, the result of balanced budget rules written into State constitutions. In other countries, a significant part of borrowing is done by sub-national units.

While borrowing is largest during recessions and wars, we also see that the U.S. has run persistent deficits. This raises the question of how much debt is sustainable. One way to approach this question is to compare projections of future government deficits to the deficit levels that are consistent with a stable debt to GDP ratio.

Let $D_t$ denote the primary deficit at date $t$ (government spending excluding interest payments minus revenue). The government budget constraint (eq. 13.1) can then be written, in nominal terms, as

$$B_t = B_{t-1} \cdot (1 + i_t) + D_t.$$

Dividing through by nominal GDP at $t$ gives

$$b_t = b_{t-1} \cdot \frac{1 + i_t}{(1 + \gamma_t)(1 + \pi_t)} + d_t,$$

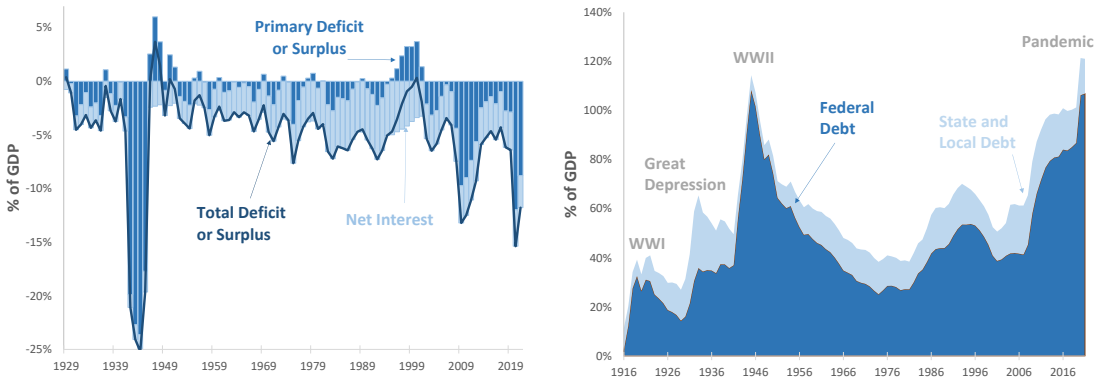



Figure 13.4: Left: Total Deficits, Primary Deficits, and Net Interest Outlays. Right: Total Debt. All as percentages of GDP.

where lower case letters denote values relative to nominal GDP, and where $\gamma_t$ and $\pi_t$ denote the growth rates of real GDP and the price level between $t-1$ and $t$. Let $1+r_t = (1+i_t)/(1+\pi_t)$ denote the *ex post* gross real interest rate between $t-1$ and $t$. Thus, the debt to GDP ratio evolves according to

$$b_t = b_{t-1} \cdot \frac{1+r_t}{1+\gamma_t} + d_t.$$

It is clear from this equation that the value of the real interest rate relative to the real growth rate is critical for the dynamics of public finances. When the primary deficit $d_t$ is zero, the debt to GDP ratio will rise when $r_t > \gamma_t$, and will fall when $r_t < \gamma_t$.

One can ask what size primary deficit is consistent with a constant debt to GDP ratio. Debt will rise over time ($b_t > b_{t-1}$) if and only if

$$d_t > \frac{\gamma_t - r_t}{1+\gamma_t} \cdot b_{t-1}. \tag{13.2}$$

At the time of writing (September, 2023) U.S. government debt held by the public is around 100 percent of annual U.S. GDP – i.e., $b_{t-1} = 1.0$. The growth rate of real GDP in the United States varies over time, but has averaged around 3 percent per year in the post-War period, suggesting $\gamma_t = 0.03$. The interest rate on 10 year inflation-protected government bonds is currently a little over 2 percent, suggesting $r_t = 0.02$.

Plugging these numbers into our debt substainability equation suggests that the largest primary deficit consistent with debt not rising is approximately 1 percent of GDP. How does this compare to the actual primary deficit? The primary federal deficit in fiscal year 2022 was 3.6 percent of GDP, and the Congressional Budget Office (CBO) is forecasting primary deficits over the next 10 years of around 3.0 percent of GDP.[1] Thus, the U.S. debt

[1]See Table 1.1 here: https://www.cbo.gov/publication/58946.

to GDP ratio is likely to continue to grow. But will debt explode? Perhaps surprisingly, the arithmetic suggests not. In particular, given constant values for $r$ and $\gamma > r$, any size primary deficit is consistent with a stable debt to GDP ratio, as long as that ratio is large enough. For example, suppose $r_t$ and $\gamma_t$ are expected to remain constant at values of 2 and 3 percent respectively. A 3 percent of GDP primary deficit is then consistent with a stable debt to GDP ratio of 309 percent of GDP (in terms of equation (13.2), $0.03 = \frac{0.01}{1+0.03} \times 3.09$). However, we should be very cautious about this calculation. As debt rises, the equilibrium real interest rate is likely to rise – investors will demand higher returns to buy all that debt. And once the differential between $\gamma$ and $r$ changes sign, stabilizing the debt to GDP ratio will require primary surpluses rather than deficits.[2]

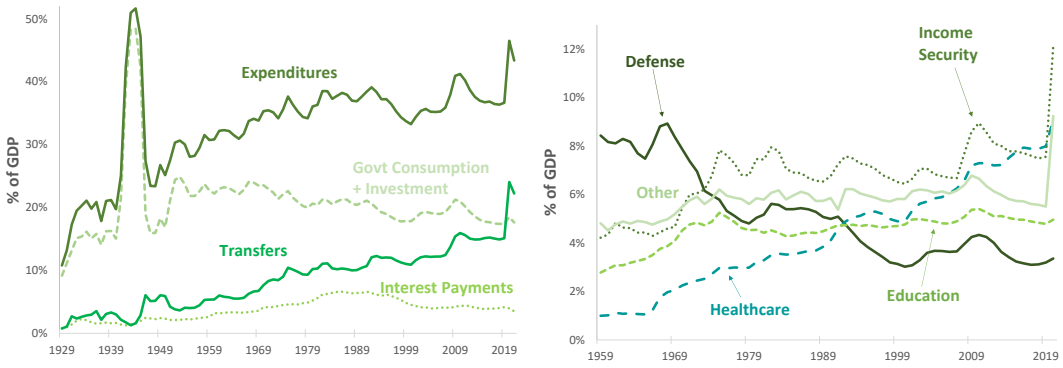



Figure 13.5: Expenditures by Category and Function, as percentages of GDP.

The growth in the size of the U.S. government coincides with the creation (and expansion) of the Social Security system and the unemployment insurance program following the Great Depression, as well as the increase in public investment after WWII. This is illustrated by the left panel of Figure 13.5, which decomposes expenditures into the three sub-components shown on the left-hand side of the government budget constraint in equation (13.1). Early on in the sample, government consumption and investment constituted the largest portion of expenditures and drove most of the trend. Over time, transfers expand significantly, overtaking $G_t$ during the COVID-19 pandemic. In the data, transfers include both redistributive and insurance programs. Redistributive policies will be studied in Section 13.7.

The right panel of the figure shows the evolution of expenditures by function for selected items from 1959 and onward. Public education spending is relatively constant, accounting for 5 percent of GDP. Defense peaks in 1959 at 8 percent, but decreases significantly thereafter, to around 3 percent today. Health-care expenditure (including Medicare and Medicaid programs), on the other hand, show a steady rise, and now exceed 8 percent of GDP. Finally, "income security" spending, which includes unemployment insurance, retirement programs, disability and welfare, fluctuates around 7 percent of GDP. Spending on these items rises in

[2]See Hall and Sargent (2020) and Blanchard (2023) for more on debt sustainability. When debt burdens become large, countries sometimes choose to default (as discussed in Chapter 22).

recessions and decreases in booms, and as such these programs are typically referred to as "automatic stabilizers."

## 13.3 The effects of distortionary taxes

The objective of this section is to show how proportional taxes affect equilibrium allocations and prices. We do this in the context of the neoclassical growth model. Capital depreciates at rate $\delta$. Households are infinitely-lived, discount at rate $\beta$, and enjoy utility each period from consumption and hours worked given by $u(c_t, \ell_t)$. They save in the form of capital, and rent capital and labor services to competitive firms at rates $w_t$ and $r_t$. Firms produce according to a constant returns to scale production function $y_t = f(k_t, \ell_t)$. The government finances government consumption $G_t$ and transfers $T_t$ (which may be positive or negative) using proportional taxes on consumption, on labor income, and on rental income net of depreciation, $\tau_{c,t}$, $\tau_{\ell,t}$, and $\tau_{k,t}$. For now we assume no government debt (we introduce it in Section 13.4). The resource constraint is

$$C_t + G_t + K_{t+1} = f(K_t, L_t) + (1 - \delta)K_t. \tag{13.3}$$

The government budget constraint is

$$G_t + T_t = \tau_{c,t}C_t + \tau_{\ell,t}w_tL_t + \tau_{k,t}(r_t - \delta)K_t. \tag{13.4}$$

The budget constraint for a representative household is

$$(1 + \tau_{c,t})c_t + k_{t+1} = (1 - \tau_{\ell,t})w_t\ell_t + k_t + (1 - \tau_{k,t})(r_t - \delta)k_t + T_t. \tag{13.5}$$

A government policy is a sequence $\{\tau_{c,t}, \tau_{k,t}, \tau_{\ell,t}, G_t, T_t\}_{t=0}^{\infty}$.

**Definition 13.1** *: A competitive equilibrium given a policy $\{\tau_{c,t}, \tau_{k,t}, \tau_{\ell,t}, G_t, T_t\}_{t=0}^{\infty}$ is a sequence of allocations $\{C_t, L_t, K_{t+1}\}_{t=0}^{\infty}$ and prices $\{w_t, r_t\}_{t=0}^{\infty}$ such that*

*i. Given policy and prices, the sequence $\{C_t, L_t, K_{t+1}\}_{t=0}^{\infty}$ maximizes household lifetime utility $\sum_{t=0}^{\infty}\beta^t u(c_t, \ell_t)$ subject to budget constraints of the form (13.5) for all $t$, initial capital $k_0 = K_0$, and a borrowing constraint $k_{t+1} \geq 0$ $\forall t$.*

*ii. The allocation $\{L_t, K_t\}_{t=0}^{\infty}$ is a solution to the firm profit maximization problem at each date $t$, with $\ell_t = L_t$ and $k_t = K_t$ in equilibrium*

$$\max_{k_t, \ell_t} \{f(k_t, \ell_t) - w_tk_t - r_tk_t\}.$$

*iii. The government budget constraint eq. (13.4) is satisfied at each date $t$.*

At each date $t$, the first order conditions that define optimal saving and labor supply decisions for a household are

$$\begin{aligned} \frac{1+\tau_{c,t+1}}{1+\tau_{c,t}} \cdot \frac{u_c\left(c_t, \ell_t\right)}{u_c\left(c_{t+1}, \ell_{t+1}\right)} &= \beta\left[1+(1-\tau_{k,t+1})(r_{t+1}-\delta)\right], \qquad (13.6) \\ -u_\ell\left(c_t, \ell_t\right) &= \frac{1-\tau_{\ell,t}}{1+\tau_{c,t}} \cdot w_t \cdot u_c\left(c_t, \ell_t\right). \end{aligned}$$

The conditions for profit maximization are

$$\begin{aligned} w_t &= f_\ell(k_t, \ell_t), \\ r_t &= f_k(k_t, \ell_t). \end{aligned}$$

In order to discuss the distortionary effects of taxation, it is useful to compute the Pareto optimal allocation. Because this is a representative agent economy, the efficient allocation can be found by solving the problem of a benevolent planner that maximizes lifetime utility subject to the resource constraint

$$C_t + G_t + K_{t+1} = f(K_t, L_t) + (1-\delta)K_t.$$

The first-order conditions to this problem are

$$\begin{aligned} \frac{u_c\left(C_t, L_t\right)}{u_c\left(C_{t+1}, L_{t+1}\right)} &= \beta\left[1+f_k(K_{t+1}, L_{t+1})-\delta\right] \\ -u_\ell\left(C_t, L_t\right) &= f_\ell(K_t, L_t) \cdot u_c\left(C_t, L_t\right). \qquad (13.7) \end{aligned}$$

Comparing across these two sets of conditions, we can see how taxes change households incentives to save and to work. Absent taxes, households equate the inter-temporal marginal rate of substitution between consumption at $t$ and at $t+1$ to one plus the marginal product of capital, net of depreciation. With taxes, households care instead about the gross after-tax return to saving, which is given by

$$\frac{1+\tau_{c,t}}{1+\tau_{c,t+1}}\left[1+\left(1-\tau_{k,t+1}\right)\left(r_{t+1}-\delta\right)\right].$$

Holding $r_{t+1}$ constant for a moment (in equilibrium $r_{t+1}$ will depend on taxes) it is clear that taxes on rental income depress the after-tax return to saving, and that rising consumption taxes ($\tau_{c,t+1} > \tau_{c,t}$) work in the same direction. Similarly, taxes on labor income depress the return to working, as do taxes on consumption. Because taxes change workers' incentives to save and to work, they distort equilibrium allocations, and will typically reduce capital, labor supply, and output relative to the solution to the planner's problem.

### 13.3.1 Long-run distortions

Suppose we consider a steady state of the economy in which tax rates and allocations are constant. In such a steady state the household first-order conditions simplify to

$$\begin{aligned} 1 &= \beta\left[1+(1-\tau_k)(r-\delta)\right], \qquad (13.8) \\ -u_\ell\left(c, \ell\right) &= \frac{1-\tau_\ell}{1+\tau_c} \cdot w \cdot u_c\left(c, \ell\right), \qquad (13.9) \end{aligned}$$

where

$$\begin{aligned} w &= f_\ell(k,\ell), \\ r &= f_k(k,\ell). \end{aligned} \tag{13.10}$$

From the first of these it is immediate that a higher capital income tax $\tau_k$ must increase the steady state equilibrium rental rate for capital $r$. If the production function $f$ has a Cobb-Douglas form, $f(k,\ell) = k^\alpha \ell^{1-\alpha}$, then $r = f_k(k,\ell) = \left(\frac{k}{\ell}\right)^{\alpha-1}$, and thus a higher rental rate corresponds to a lower capital-labor ratio. In particular,

$$\frac{k}{\ell} = \left(\frac{\alpha\left(1-\tau_k\right)}{\rho+\delta\left(1-\tau_k\right)}\right)^{\frac{1}{1-\alpha}} \tag{13.11}$$

where $\rho = \frac{1-\beta}{\beta}$ denotes the household's rate of time preference. Note that because a higher $\tau_k$ depresses the steady state capital-labor ratio, it will depress the steady state wage $w$, in addition to raising $r$. In contrast, labor and consumption taxes have no impact on the pre-tax prices $r$ and $w$.

The effect of taxes on labor supply will depend on the preference specification, via the marginal utility terms in eq. (13.9). An increase in $\tau_\ell$ affects labor supply in the current period directly by reducing the after-tax wage, inducing $\ell$ to fall via a substitution effect. At the same time, because the individual becomes poorer when labor income declines, consumption shrinks, which results in a higher marginal utility of consumption, incentivizing the agent to work more via an income effect. The total effect of an increase of labor income taxes on $\ell$ is therefore ambiguous, depending on the relative strength of substitution and income effects.

For an illustrative example we consider a particular utility function, made famous by Greenwood, Hercowitz, and Huffman (henceforth, GHH):

$$u\left(c,\ell\right) = \ln\left(c - \frac{\ell^{1+\frac{1}{\phi}}}{1+\frac{1}{\phi}}\right).$$

The GHH functional form is particularly tractable because consumption drops out of the the first-order condition for hours worked; thus, this utility function can be described as one in which there are no income effects.[3] Steady state labor supply is given by

$$\ell = \left(\frac{1-\tau_\ell}{1+\tau_c}\right)^\phi w^\phi \tag{13.12}$$

Note that hours worked depend only on the return to working, where $\phi$ defines the elasticity of hours to after-tax wages. Higher labor income or consumption taxes depress hours worked. Higher capital income taxes also depress hours, via their negative impact on $w$.

[3]See Appendix 13.A.2 for an alternative utility function with income effects.

### 13.3.2 Tax incidence

In addition to simplifying the algebra, another feature of the GHH specification is that the steady state of the representative agent model specification is identical, at the aggregate level, to an alternative decentralization in which there are two household types: (i) workers, who rent labor services but own no capital, and (ii) capitalists, who own and rent out capital but who do not work. In what follows we focus on this worker-capitalist specification, because it allows for a discussion of tax incidence, namely the issue of who pays different sorts of taxes (in a representative agent setting, the representative household pays all taxes).[4]

From eqs. (13.10) and (13.12), we can solve for hours worked as a function of the capital to labor ratio

$$\ell = \left[\frac{1-\tau_\ell}{1+\tau_c}(1-\alpha)\left(\frac{k}{\ell}\right)^{\alpha}\right]^{\phi}$$

which, combined with eq. (13.11), gives an expression for steady state output

$$y = \left(\frac{k}{\ell}\right)^{\alpha}\ell = \left[\frac{1-\tau_\ell}{1+\tau_c}(1-\alpha)\right]^{\phi}\left[\frac{\alpha\left(1-\tau_k\right)}{\rho+\delta\left(1-\tau_k\right)}\right]^{\frac{\alpha(1+\phi)}{1-\alpha}} \tag{13.13}$$

Note that all three tax rates affect the level of steady state output, and that the level of output is decreasing in each tax rate. Capital income taxes depress the capital labor ratio, but their impact on output is amplified by the fact that a lower capital-labor ratio means lower wages, which in turn depress labor supply.

Consider the case with no transfers ($T = 0$). In steady state, workers consume

$$c_w = \frac{1-\tau_\ell}{1+\tau_c}w\ell = \frac{1-\tau_\ell}{1+\tau_c}(1-\alpha)y,$$

while capitalists consume

$$c_k = \frac{1-\tau_k}{1+\tau_c}(r-\delta)k = \frac{\rho}{1+\tau_c}\left[\frac{\alpha(1-\tau_k)}{\rho+(1-\tau_k)\delta}\right]y.$$

From these expressions, it is clear that labor taxes directly depress the consumption of workers, while capital income taxes directly depress the consumption of capitalists. Consumption taxes depress the consumption of both types. Note, however, that all three types of taxes indirectly depress the consumption of both types via their impact on equilibrium output.

If we are designing a tax system, we would like to know more about how effective different sorts of taxes are in terms of raising revenue, relative to how distortionary they are in terms of depressing output. To make further progress on this question in a tractable way, we

[4] Why is the steady state of the worker-capitalist model identical, given the GHH utility specification, to the steady state of the representative agent specification? The logic is that the level of consumption appears in neither the steady state first-order condition for saving, nor in the first order condition for hours worked. Thus the distribution of aggregate consumption between workers and capitalists has no impact on either steady state capital or steady state hours worked.

now make two additional assumptions. First, we temporarily rule out consumption taxes by setting $\tau_c = 0$. Second, we assume that the government has to devote a fraction $g$ of aggregate output to government purchases: $G = gY$. Thus, the steady state government budget constraint is

$$\begin{aligned} gy &= \tau_\ell wl + \tau_k(r-\delta)k \\ &= \tau_\ell(1-\alpha)y + \tau_k\rho\frac{\alpha}{\rho+(1-\tau_k)\delta}y \end{aligned}$$

From this budget constraint we can immediately solve for the locus of budget-balancing pairs $(\tau_\ell, \tau_k)$ :

$$\tau_\ell = \frac{1}{1-\alpha}\left[g - \tau_k\rho\frac{\alpha}{\rho+(1-\tau_k)\delta}\right] \tag{13.14}$$

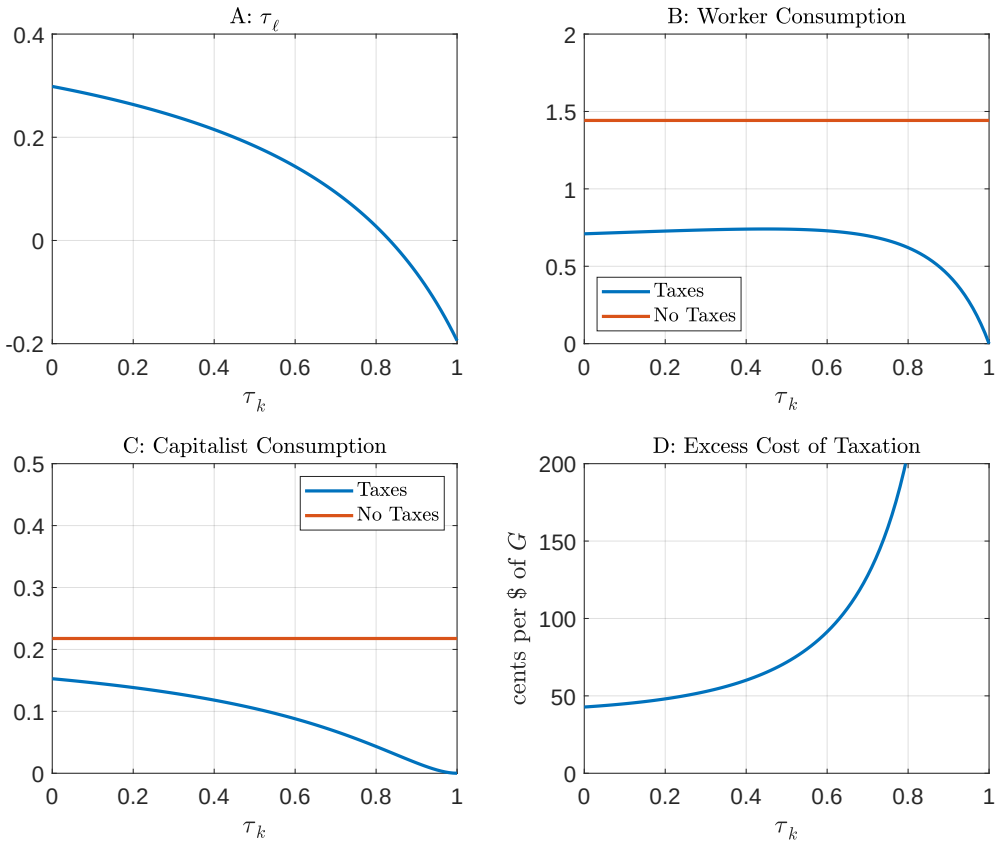



Figure 13.6: How Allocations Vary with $\tau_k$.

Panel A of Figure 13.6 plots how the labor tax rate $\tau_\ell$ varies with capital income tax rate $\tau_k$ according to eq. (13.14). The parameters used to construct the plot are $\alpha = 0.33$, $\delta = 0.07$, $\beta = 0.97$, $\phi = 1$ and $g = 0.2$. For each $\tau_k$ and the corresponding value for $\tau_\ell$ given in Panel A, Panels B and C plot consumption of workers and capitalists, respectively. For comparison we also plot the levels of the two types' consumption for an economy without taxes. Note that as $\tau_k \to 1$, consumption of both types converges to zero. When $\tau_k$ is restricted to be non-negative, steady state consumption of capitalists is maximized at $\tau_k = 0$, while consumption of workers is a hump-shaped function of $\tau_k$. Workers' utility depends on hours worked in addition to consumption, but given this utility function the consumption-equivalent argument of flow utility in equilibrium is proportional to consumption:

$$c_w - \frac{\ell^{1+\frac{1}{\phi}}}{1+\frac{1}{\phi}} = \frac{1}{1+\phi}(1-\alpha)(1-\tau_\ell)y = \frac{1}{1+\phi}c_w.$$

One way to define the deadweight cost of taxation is to ask by how much is total consumption reduced by taxes per unit of government consumption that the taxes finance. Let $c_{w,\tau=0}$ and $c_{k,\tau=0}$ denote the steady state consumption levels of workers and capitalists when $\tau_k = \tau_\ell = 0$, and define the excess cost of taxation as private consumption lost net of public consumption financed, measured per dollar of such spending:

$$\text{Excess Cost} = \frac{\frac{1}{1+\phi}\left(c_{w,\tau=0} - c_w\right) + c_{k,\tau=0} - c_k - gy}{gy}$$

If this ratio is equal to zero, then steady state utility in consumption units is reduced by one for each unit of government purchases. Panel D of Figure 13.6 plots the excess cost as $\tau_k$ varies (in the background $\tau_\ell$ is adjusted with $\tau_k$ to balance the government budget constraint). When $\tau_k = 0.2$, each dollar of government consumption comes with an excess cost of 50 cents, meaning that total private consumption is reduced by \$1.50 for each dollar of public goods provided. Clearly, the excess cost of taxation is always positive. Furthermore, the excess cost of taxation is increasing and convex in the capital tax rate $\tau_k$.

### 13.3.3 Tax reform

Panel D of Figure 13.6 suggests that the excess cost of taxation is minimized when $\tau_k = 0$. The tax rate on capital income in the U.S., however, ranges from 10% to 37% (the ordinary income tax brackets in 2022, applied to capital held less than a year). From Panels B and C of the same figure, we see that if capital taxes were reduced to zero and labor taxes were raised to support the same $g = G/GDP$ ratio, then capitalists' consumption in steady state would be much higher without much reduction in workers' consumption. These findings might suggest that eliminating capital income taxes would be a good idea. However, the steady state associated with a lower $\tau_k$ has a larger capital stock, and if capital taxes are reduced it will take time for the economy to accumulate this extra capital. Additional capital accumulation will come at the cost of reduced current consumption. It is therefore important to analyze *transitional dynamics* in addition to steady states when considering tax reforms.

We illustrate this with a simple example, using the worker-capitalist framework from the previous section and again assuming no transfers, $T_t = 0$.[5]

We start from a situation in which $\tau_k = 0.25$, a midpoint of the current tax brackets. Labor taxes are set to $\tau_\ell = 0.2529$, obtained from eq. (13.14) to sustain $g = 0.2$ given the parameters used in Section 13.3.2. We assume that the economy is in steady state until period 10, and that a switch to $\tau_k = 0$ is implemented, unexpectedly and permanently, in period 11. At that date, we increase the labor tax to $\tau_\ell = 0.2985$ so that eq. (13.14) still holds at $g = 0.2$ in the new steady state. We allow $G_t$ to vary during transition to balance the government budget date by date given constant tax rates. While the initial and final steady states can be characterized analytically, the evolution of $k_t$ during transition requires the use of computational methods. We assume that the economy has reached the final steady state by date $T$. Knowing the initial and final conditions, $k_0$ and $k_T$, respectively, all we need is a sequence $\{k_t\}_{t=1}^{T-1}$ consistent with the first-order conditions of workers and capitalists at dates $\{0, ..., T-2\}$. This is a system of $T-1$ inter-temporal first-order conditions and $T-1$ unknowns, which can be solved using a standard non-linear system of equations root-finding routine.[6]

The left panel of Figure 13.7 displays the evolution of capital and labor income taxes (exogenous parameters to the model), as well as the endogenous evolution of $G_t$. Public spending changes in response to the reform because prices and allocations change in equilibrium (as seen in the right panel), in turn affecting government revenues (recall that $G/Y$ is identical in the initial and final steady states). The elimination of capital income taxes encourages capital accumulation, whereas the increase in labor income taxes discourages labor supply upon impact.[7] Over time, capital grows, and because this positively affects wages, labor supply gradually recovers, ending up only slightly below the initial steady state level. GDP tracks labor supply in the short run, and capital in the long run, declining right after the reform and recovering slowly over time. Aggregate consumption decreases initially, both because output is low, and because lower capital taxes are stimulating saving and investment. Subsequently, income rises and investment slows, pushing consumption back to a value close to the initial steady state.

While agents are eventually better off as the economy becomes more efficient (they work less than in the initial steady state but enjoy similar consumption), the exercise highlights that transitions can be painful. Whether the reform is beneficial for society overall depends on how much weight is assigned to capitalists versus workers. Capitalists are definitely better off, as their income always grows, whereas workers may be significantly worse off, as their tax

[5] While the steady state of the representative agent model (RA) is identical to the worker-capitalist environment (WK), transitional dynamics are not the same. In the RA model, the first order condition with respect to capital includes the disutility of labor (due to non-separability between $c$ and $\ell$), whereas this is absent in the WK environment, since capitalists set $\ell = 0$. In the example computed above, the difference is numerically insignificant. However, it could be sizable with other preference specifications.

[6] The values for consumption in the inter-temporal first-order condition can be substituted out using the budget constraint of capitalists, $(1 + \tau_{c,t})c_k + k_{t+1} = k_t + (1 - \tau_{k,t})(r_t - \delta)k_t$. It is important to verify that $T$ is large enough that the economy has indeed converged to the new steady state.

[7] The latter is an artifact of the specific utility function used, since it exhibits no income effects. In Appendix 13.A.2, we re-compute this experiment using an alternative specification.

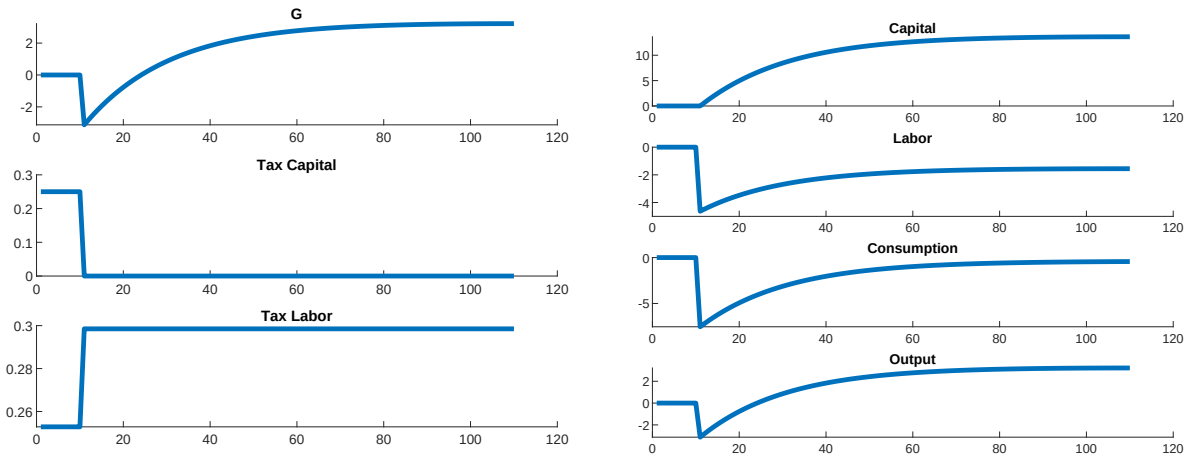



Figure 13.7: Eliminating capital income taxes.

**Notes**: Except for the tax rates, variables are plotted as percentage deviations from their values in the pre-reform steady state.

burden increases (see, e.g., Domeij and Heathcote 2004.) Even when considering a representative household, whether public spending is valued or not is important in this calculation. The reform is more desirable if agents derive utility from government consumption – which rises over time – than if public spending is entirely wasteful.

### 13.3.4 The Laffer Curve

Another question of interest is: what combination of tax rates maximizes steady state tax revenue? A government that is fighting a war and wants to purchase the largest possible number of tanks might be especially interested in answering this question. When the planner only has access to taxes on labor earnings and rental income, and when tax rates must be positive, the revenue maximizing pair of tax rates is given by

$$\begin{aligned} \tau_\ell &= \frac{1}{1+\phi} \\ \tau_k &= 0. \end{aligned}$$

Note that the higher is the Frisch elasticity of labor supply, $\phi$, the lower is the revenue-maximizing labor tax rate.[8]

[8]These rates are the solution to the problem:

$$\max_{\tau_\ell \geq 0, \tau_k \geq 0} \left\{ \tau_\ell (1-\alpha) y + \tau_k \rho \frac{\alpha}{\rho + (1-\tau_k)\delta} y \right\},$$

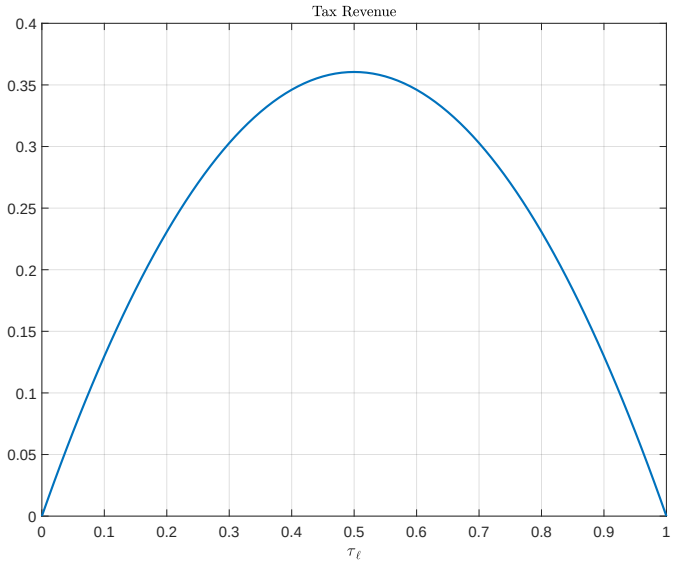



Figure 13.8: The Laffer curve

Figure 13.8 plots tax revenue as a function of $\tau_\ell$, holding fixed $\tau_k$ at $\tau_k = 0$. This type of plot was popularized by Arthur Laffer in the 1970s, and is thus called a Laffer curve. Revenue is hump-shaped in the tax rate, and for tax rates above the revenue-maximizing rate, raising rates reduces revenue, because the tax base shrinks faster than the rate increases. Such a situation is known as "being on the wrong side of the Laffer curve." To see why the Laffer curve must be hump-shaped, it is enough to observe that at $\tau_\ell = 0$, no revenue is raised because no taxes are levied, while at $\tau_\ell = 1$ no revenue is raised because hours worked and output are equal to zero, and thus there is nothing to tax.

### 13.3.5 Theories of $G$

So far, we have assumed that public spending is exogenously given and generates no benefits to society; revenues are "thrown into the ocean." However, as shown in Figure 13.5, public expenditures are composed of government consumption, public investment, transfers, and interest payments. Here, we briefly describe theories of government consumption and public

where output $y$ is given by eq. (13.13). The first-order condition with respect to $\tau_\ell$ gives the solution

$$\tau_\ell = \frac{1 - \tau_k \frac{\rho\phi}{1-\alpha} \cdot \frac{\alpha}{\rho+(1-\tau_k)\delta}}{1+\phi}$$

Given this value for $\tau_\ell$, tax revenue is declining in $\tau_k$ at $\tau_k = 0$, indicating that $\tau_k = 0$ is the revenue-maximizing rate when tax rates must be positive.

investment. Interest payments will be described in the next section, after we introduce debt into the model, and transfers and welfare programs will be described at the end of the chapter.

First, let us focus on how to model government consumption. We typically assume that the government has the technology to provide public goods that are valued by society. These include defense, law and order, taking care of parks and common areas, sanitation, etc. A key assumption is that the government uses resources to produce these goods, which then provide utility to agents. We typically assume that agents derive utility from public and private goods, $u(c, g)$, with $u_g > 0$ and $u_{gg} \leq 0$. The first best solution (e.g., when the government has access to lump-sum taxation) prescribes equating the marginal utility of private consumption to the marginal utility of public consumption: $u_c(c, g) = u_g(c, g)$. When taxes are distortionary, the government must take into account, in addition, the deadweight losses associated with taxation.

A second strand of theories of government spending focuses on the role of the government to provide key infrastructure such as roads, bridges, and schools. Letting $k_g$ denote 'public capital,' the neoclassical growth model augmented to include public investment involves a production function,

$$f(k, k_g, \ell) = Ak^{\alpha}\ell^{1-\alpha}k_g^{\theta},$$

and a law of motion for $k_g$,

$$k_{g,t+1} = i_{g,t} + (1 - \delta_g)k_{g,t},$$

with $i_{g,t}$ denoting public investment in period $t$, and $\delta_g$ the rate of depreciation of public capital. The parameter $\theta$ controls the elasticity of output with respect to public capital. If $\theta > 0$, there are increasing returns to scale. Estimates of $\theta$ vary, from as low as 0.05 (see Leeper, Walker, and Yang (2010)) to as high as 0.39 (Aschauer (1989)). If revenue can be raised in a non-distortionary way, then it is optimal to equate the marginal products of private and public capital, net of depreciation, $f_{k,t+1} - \delta = f_{g,t+1} - \delta_g$.

## 13.4 Government debt and Ricardian Equivalence

We now add government debt to the analysis. To simplify the presentation, we abstract from capital, as this reduces the number of state variables. Using a two-period example, we present conditions under which debt is irrelevant. This result is known as "Ricardian Equivalence." We then introduce distortionary taxation and explain how the government can use debt to smooth out tax distortions over time. The optimal sequence of taxes solves what is known as the "Ramsey Problem."

Consider a two-period representative household model. The government can issue debt in period $t = 0$ and can levy lump-sum taxes or hand out lump-sum transfers in periods $t = 0$ and $t = 1$. Suppose that the government contemplates a lump-sum transfer $T_0$ in the first period, financed by issuing government debt $B_0$, which will be paid back by levying a lump-sum tax $\tau_1$ in the second period.

The representative household has utility defined over consumption and hours worked in periods 0 and 1 given by

$$u(c_0, \ell_0) + \beta u(c_1, \ell_1).$$

Households choose labor supply, consumption and saving in each period, taking as given exogenous wages, $w_0$ and $w_1$, and an exogenous return to saving, $r_0$. Abstracting from other taxes, the budget constraints in the two periods are

$$\begin{aligned} c_0 + b_0 &= w_0 \ell_0 + T_0 \\ c_1 &= w_1 \ell_1 + (1 + r_1)\, b_0 - \tau_1, \end{aligned}$$

where $b_0$ is the amount of debt the household buys in the first period, and $r_1$ is the interest paid on that debt. Dividing the second equation through by $1 + r_1$ and adding it to the first equation expresses the budget constraint in present value form:

$$c_0 + \frac{c_1}{1 + r_1} = w_0 \ell_0 + \frac{w_1 \ell_1}{1 + r_1} + T_0 - \frac{\tau_1}{1 + r_1}.$$

Note that the value for debt purchases, $b_0$, drops out of the present-value version of the budget constraint. In addition, note that given a promised interest rate of $r_1$, the second period tax $\tau_1$ will have to satisfy $\tau_1 = (1 + r_1)\, B_0 = (1 + r_1)\, T_0$. Thus, the tax and transfer terms in the present value budget constraint must sum to zero, regardless of the size of the initial transfer $T_0$. Since the tax and transfer scheme does not affect the lifetime budget constraint, the household's optimal allocation of consumption and hours must be identical to that in the case of zero taxes and transfers ($T_0 = \tau_1 = B_1 = 0$). The household must therefore respond to the initial transfer $T_0$ by increasing savings by exactly $T_0$. This extra savings will (i) exactly match the additional supply of government bonds issued, and (ii) provide exactly enough second period income to pay the expected lump-sum tax $\tau_1$. This neutrality result is an example of *Ricardian Equivalence*. Note that this result hinges on the assumptions that taxes are lump-sum, that households face no credit constraints, and that the households who get the transfers are the same ones that must repay the debt. The result does not hinge on there being no capital. Using the same logic, it is possible to show that the result extends to an infinite-horizon economy (see also Barro 1974 and Heathcote 2005b).

## 13.5 Ramsey Taxation

It is traditional in public finance to assume that the government cannot impose lump-sum taxes. Why? In a representative agent economy, there is no reason not to impose lump-sum taxes: such taxes are a distortion-free way to raise revenue. But in practice, actual households differ widely in terms of their income. Some households are so poor that they could not afford to pay a moderate lump-sum tax. Equally importantly, many people would find it unfair if the poor were expected to pay as much tax as the rich. Thus, the literature has focused on taxes that are *proportional* to income (see Chamley (1986) and Judd (1985)

for early examples, and more recently Straub and Werning (2020)). Of course, one could consider making taxes a more complicated function of income – and we shall do so shortly – but proportional is simple, and simplicity can be viewed as a virtue.

But even if taxes are proportional to income, there is no need to tax different types of income at the same rate. In particular, earned income (income from labor) can be taxed at a different rate to unearned income (income generated by wealth). And if the government can save or borrow by issuing government debt, then it can also choose how tax rates should vary over time. We now consider a simple two period model and ask how a government that seeks to maximize the welfare of a representative agent should optimally set proportional taxes.

The timing assumptions are as follows. The government moves first, and announces tax rates for both periods, which we label periods 0 and period 1. To start, we assume that the government commits to these tax rates, and does not have the ability to deviate at $t = 1$ from the policies announced at $t = 0$. Later we will discuss how the analysis might change if the government does not have this commitment power.

A tax plan is *feasible* if there is a competitive equilibrium characterized by allocations and prices such that (i) those allocations are optimal choices for households and firms given prices and the tax rates described in the plan, (ii) the government budget constraint is satisfied, and (iii) markets clear.

A tax plan is *optimal* if it is feasible and the associated competitive equilibrium maximizes the welfare of the representative household. The optimal tax plan is called the *Ramsey plan*, and the associated equilibrium the *Ramsey equilibrium*.[9] In general, lots of different fiscal plans will be feasible, but only one will be optimal. There are two different approaches in the literature to solving for the Ramsey plan.

The first and more intuitive approach is to work with the full set of equilibrium equations and variables, and to conceptualize the planner choosing tax rates to maximize household welfare, internalizing how changes in tax rates will affect all equilibrium variables. This is called the dual approach.

An alternative approach, called the primal approach, is sometimes easier to implement (see Atkeson, Chari, and Kehoe, 1999). Under the primal approach, we think of the planner as choosing equilibrium allocations for consumption and hours directly, subject to two sets of constraints. The first set of constraints ensure that allocations are technologically feasible. The second set of constraints ensure that there exists a set of tax rates such that the allocation is the competitive equilibrium given those taxes. These second constraints are called the *implementability constraints*. No equilibrium prices or tax rates appear in the primal problem. Once one has solved the primal problem, one can back out the tax rates that decentralize the solution in a final step.

[9] After Frank Ramsey, who wrote a handful of important papers in economics and more in the fields of mathematics and philosophy, before his death in 1930 at the age of 26.

### 13.5.1 The primal approach to optimal taxation: A simple example

The best way to understand how the primal approach works is to consider a simple example economy. Consider, in particular, the following two period model. There is a representative household with utility defined over consumption and hours worked in periods 0 and 1 given by

$$u(c_0, \ell_0) + \beta u(c_1, \ell_1).$$

The representative household supplies labor to a representative firm that produces output according to

$$\begin{aligned} y_0 &= A_0 \ell_0, \\ y_1 &= A_1 \ell_1, \end{aligned}$$

where $A_t$ denotes potentially time-varying labor productivity. Because labor markets are assumed to be competitive, equilibrium wages equal productivities:

$$\begin{aligned} w_0 &= A_0, \\ w_1 &= A_1. \end{aligned} \tag{13.15}$$

Households (and the government) can save or borrow using a storage technology that converts one unit of output at date 0 into $1 + r$ units of output at date 1, where $r$ is an exogenous constant.[10] Let $b_{t-1}$ denote household wealth at the start of period $t$. Assume that $b_{-1} = 0$.

The government must finance exogenous expenditures $g_0$ and $g_1$ in periods 0 and 1. It can raise tax revenue via proportional taxes on labor income at rates $\tau_0$ and $\tau_1$, and by taxing income from wealth in period 1 at rate $\tau_{b,1}$.[11] A *tax plan* is a vector $\{\tau_0, \tau_1, \tau_{b,1}\}$.

Let $b_{g,0}$ denote government savings in the storage technology at date 0. The government budget constraints for periods 0 and 1 are

$$\begin{aligned} g_0 + b_{g,0} &= \tau_0 w_0 \ell_0, \\ g_1 &= \tau_1 w_1 \ell_1 + \tau_{b,1} r b_0 + (1 + r) b_{g,0}. \end{aligned}$$

Note that government savings at $t = 0$ delivers income at $t = 1$. If $b_{g,0} < 0$, the government is borrowing. These two constraints can be combined to give

$$g_0 + \frac{g_1}{1 + r} = \tau_0 w_0 \ell_0 + \frac{\tau_1 w_1 \ell_1}{1 + r} + \frac{\tau_{b,1} r b_0}{1 + r}.$$

[10] One interpretation might be that the economy is small and open, and $r$ is the world interest rate.

[11] The household has no wealth at date 0, so a tax on wealth or income from wealth at date 0 would not raise any revenue. If the household did have wealth at date 0, the government would like to tax that wealth, since that would effectively amount to a non-distortionary lump-sum tax. In the spirit of not allowing for lump-sum taxation, it is typically assumed that the Ramsey planner cannot tax initial wealth, or that there is an upper bound on the feasible initial tax rate.

Given taxes and wages, the representative household solves

$$\begin{aligned} & \max_{\{c_0,\ell_0,c_1,\ell_1,b_0\}} u(c_0,\ell_0)+\beta u(c_1,\ell_1) \\ & s.t. \\ c_0 &= (1-\tau_0)w_0\ell_0 - b_0, \\ c_1 &= (1-\tau_1)w_1\ell_1 + (1+r(1-\tau_{b,1}))b_0, \end{aligned}$$

where again the two budget constraints can be collapsed to give

$$c_0 + \frac{c_1}{1+r(1-\tau_{b,1})} = (1-\tau_0)w_0\ell_0 + \frac{(1-\tau_1)w_1\ell_1}{1+r(1-\tau_{b,1})}.$$

The first-order conditions that characterize the solution to the household's problem are

$$\begin{aligned} u_{c,0}w_0(1-\tau_0) &= -u_{\ell,0}, \\ u_{c,1}w_1(1-\tau_1) &= -u_{\ell,1}, \\ u_{c,0} &= \beta(1+r(1-\tau_{b,1}))u_{c,1}, \end{aligned} \tag{13.16}$$

where $u_{c,t}$ denotes the marginal utility of consumption in period $t$.

Resource feasibility in this economy can be summarized by a single equation, which states that the present value of private plus public consumption is equal to the present value of output

$$c_0 + g_0 + \frac{c_1+g_1}{1+r} = A_0\ell_0 + \frac{A_1\ell_1}{1+r}. \tag{13.17}$$

What about the implementability constraints? An allocation is a competitive equilibrium if it satisfies the three first-order conditions from the household problem, the two equilibrium expressions for wages, and the household and government budget constraints. We now show that these six equations can be collapsed into a single implementability condition. The idea is to take the household lifetime budget constraint, and to use the household first-order conditions to substitute out for $(1-\tau_0)w_0$, $(1-\tau_1)w_1$, and $1+r(1-\tau_{b,1})$. In particular, the first-order condition for saving implies that, in any competitive equilibrium, it must be the case that

$$1+r(1-\tau_{b,1}) = \frac{u_{c,0}}{\beta u_{c,1}},$$

while those for labor supply imply

$$(1-\tau_t)w_t = -\frac{u_{\ell,t}}{u_{c,t}}.$$

After these substitutions, the household lifetime budget constraint can be written as

$$c_0 + \frac{c_1}{\frac{u_{c,0}}{\beta u_{c,1}}} = -\frac{u_{\ell,0}}{u_{c,0}}\ell_0 - \frac{u_{\ell,1}}{u_{c,1}}\ell_1\frac{1}{\frac{u_{c,0}}{\beta u_{c,1}}}, \tag{13.18}$$

or, after multiplying through by $u_{c,0}$, as

$$u_{c,0}c_0 + \beta u_{c,1}c_1 = -u_{\ell,0}\ell_0 - \beta u_{\ell,1}\ell_1.$$

This is the implementability condition. It can be shown that any allocation that satisfies the resource constraint and the implementability condition can be implemented by some feasible tax plan (see, e.g., Atkeson et al., 1999).

It should be clear that the implementability condition embeds the household optimality conditions for labor supply and savings in addition to the household budget constraint. But what about the government budget constraint? We do not have to worry separately about that, because if the resource constraint and the household budget constraint are satisfied, the government budget constraint must also be satisfied, by Walras Law.

The *Ramsey problem* is to maximize lifetime utility for the representative agent, subject to the resource and implementability constraints. Writing this problem as a Lagrangian, with multipliers $\lambda$ and $\mu$ on the resource and implementability constraints, we have

$$\begin{aligned}
&\max_{c_0,c_1,\ell_0,\ell_1} \{u(c_0,\ell_0) + \beta u(c_1,\ell_1) \\
&+\lambda\left(A_0\ell_0 + \frac{A_1\ell_1}{1+r} - c_0 - g_0 - \frac{c_1+g_1}{1+r}\right) \\
&+\mu\left(u_{c,0}c_0 + \beta u_{c,1}c_1 + u_{\ell,0}\ell_0 + \beta u_{\ell,1}\ell_1\right)\}
\end{aligned}$$

The first-order conditions are

$$\begin{aligned}
u_{c,0} - \lambda + \mu u_{c,0} + \mu u_{cc,0}c_0 + \mu u_{\ell c,0}\ell_0 &= 0, \\
u_{\ell,0} + \lambda A_0 + \mu u_{\ell,0} + \mu u_{c\ell,0}c_0 + \mu u_{\ell\ell,0}\ell_0 &= 0, \\
\beta u_{c,1} - \frac{\lambda}{1+r} + \beta\mu u_{c,1} + \beta\mu u_{cc,1}c_1 + \beta\mu u_{\ell c,1}\ell_1 &= 0, \\
\beta u_{\ell,1} + \frac{1}{1+r}\lambda A_1 + \beta\mu u_{\ell,1} + \beta\mu u_{c\ell,1}c_1 + \beta\mu u_{\ell\ell,1}\ell_1 &= 0.
\end{aligned}$$

where, for example, $u_{cc,0}$ denotes $\partial u_{c,0}/\partial c_0$.

These four first-order conditions alongside eqs. (13.17) and (13.18) constitute six equations that can be used to solve for the six unknowns $(c_0, c_1, \ell_0, \ell_1, \lambda, \mu)$. Thus, one can solve for the Ramsey allocation. Consider, in particular, a separable utility function of the form

$$u(c,\ell) = \frac{c^{1-\sigma}}{1-\sigma} - \frac{\ell^{1+\frac{1}{\phi}}}{1+\frac{1}{\phi}}.$$

In this case, the cross derivative terms drop out, and the second derivatives simplify to

$$u_{cc,t}c_t = -\sigma u_{c,t}, \;\; u_{\ell\ell,t}\ell_t = \frac{1}{\phi}u_{\ell,t}.$$

Thus, the first-order conditions can be written as

$$\begin{aligned} u_{c,0}(1+\mu-\mu\sigma) &= \lambda \\ u_{\ell,0}\left(1+\mu+\frac{\mu}{\phi}\right) &= -\lambda A_0 \\ \beta u_{c,1}\left(1+\mu-\mu\sigma\right) &= \frac{1}{1+r}\lambda \\ \beta u_{\ell,1}\left(1+\mu+\frac{\mu}{\phi}\right) &= -\frac{1}{1+r}\lambda A_1 \end{aligned}$$

Comparing the first and the third, it is immediate that, at an optimum

$$\beta(1+r)u_{c,1} = u_{c,0}.$$

Comparing the first and the second, we see that

$$u_{c,0}A_0 \cdot \frac{1+\mu-\mu\sigma}{1+\mu+\frac{\mu}{\phi}} = -u_{\ell,0}$$

Similarly, the third and the fourth give

$$u_{c,1}A_1 \cdot \frac{1+\mu-\mu\sigma}{1+\mu+\frac{\mu}{\phi}} = -u_{\ell,1}$$

Comparing these expressions to the first order conditions for saving and for working in the original economy (eqs. 13.16), and noting that, in equilibrium, $w_0 = A_0$ and $w_1 = A_1$, it is clear that the only way both sets of first order conditions can be satisfied at the same allocation is if

$$\begin{aligned} \tau_{b,1} &= 0, \\ \tau_0 &= \tau_1 = 1 - \frac{1+\mu-\mu\sigma}{1+\mu+\frac{\mu}{\phi}} = \frac{\mu\left(\frac{1}{\phi}+\sigma\right)}{1+\mu+\frac{\mu}{\phi}}. \end{aligned}$$

Thus, this simple example illustrates two classic results in the Ramsey taxation literature. First, the government should commit to neither tax nor subsidize income from savings (see Chamley 1986 and Judd 1985). Second, the labor tax rate should be constant over time, and will be positive as long as either $g_0$ or $g_1$ is strictly positive (so that revenue must be raised). This result is described as *tax smoothing*, and the idea is that because distortions from taxes increase with the tax rate in a convex fashion, constant tax rates are preferable to time-varying tax rates (see Barro (1979) and Lucas and Stokey (1983)).

### 13.5.2 Time consistency

Let us assume that parameters are such that the household saves in period 0 under the Ramsey plan. One parameter configuration that would deliver this is $\beta(1+r) = 1$ – so that

the Ramsey allocation features $c_1 = c_1$ – and $A_0 > A_1$, so the Ramsey allocation features $A_0\ell_0 > A_1\ell_1$.

Recall that our analysis above presumed that the government announced a tax plan at date 0 and stuck to the plan at date 1. Suppose now that we give the planner the ability to redesign taxes in period 1. At $t = 1$ the planner can raise revenue either by taxing labor earnings (as promised in the original plan) or by taxing household income from savings. What combination of taxes would a benevolent planner choose? The answer is that such a planner would set $\tau_1 = 0$ and $\tau_{b,1}$ as high as necessary to fund required government purchases. The reason is that once period 1 rolls around, household wealth $b_1$ is already determined, and taxes on income from wealth are effectively a lump-sum tax. In contrast, taxes on labor earnings are distortionary. Because the planner would like to deviate from the Ramsey plan at date 1, given the chance to do so, the plan is said to be *time inconsistent* (see Kydland and Prescott (1977)).

There are many policy questions where time consistency arises as a central issue. For example, Chapter 22 focuses on debt policy. If governments can commit to repaying debt, they will be able to borrow cheaply. But those promises to repay may not be time consistent, in the sense that once debt has been accrued the government might be better off defaulting.

Does the fact that the Ramsey plan described above is time inconsistent mean that we should not take it too seriously as a practical policy prescription? It is certainly useful to know what policy would be optimal given commitment and to understand how the planner might be tempted to deviate from the Ramsey plan. It might be possible to design institutions in such a way that the planner has more commitment power – for example, by writing a constitution that precludes frequent tax changes. At the same time, there is a large literature that attempts to characterize the best time consistent policies (see Klein, Krusell, and Ríos-Rull (2008)).

## 13.6 Debt and pensions with overlapping generations

We close this chapter by studying fiscal policy in a non-dynastic economy and return to the two-period overlapping-generations endowment economy discussed in Section 5.5. We extend this model to incorporate government debt, a pay-as-you-go (PAYG) pension system, and taxes. To simplify the exposition, we assume a small open economy with access to a global bond market.

The population grows at a constant rate $n$. Let $N_t = (1+n)^t$ denote the size of the newborn young population at date $t$. The share of young people is $N_t/(N_t + N_{t-1}) = (1+n)/(2+n)$. Only the young work and their endowment of efficiency units grows at rate $\gamma$. Thus, the labor income of an individual born in period $t$ is $y_t = (1+\gamma)^t \omega$ and aggregate labor income is $Y_t = y_t N_t = (1+n)^t (1+\gamma)^t \omega$.

The government operates a PAYG pension system and provides a public good $G_t$. The pension system pays $p_t$ to every old individual, financed by taxing labor income at rate $\tau_p$. The pension per retiree is therefore

$$p_t = (1+n)\,\tau_p y_t.$$

Spending on the public good is assumed to be a fixed fraction of GDP: $G_t = gY_t$. Government spending is financed by taxing labor income by a flat tax $\tau_t$ and by issuing debt. We abstract from taxes on capital income. The government budget constraint is given by

$$G_t + p_t N_{t-1} + (1+r) B_{t-1} = (\tau_p + \tau_t) Y_t + B_t,$$

where $B_t$ is the issuance of new debt that matures next period and the interest rate $r$ is exogenous and fixed. Because the pension system is self-financing, the budget can be expressed as follows,

$$g + \frac{1+r}{(1+n)(1+\gamma)} b_{t-1} = \tau_t + b_t,$$

where $b_t$ is the debt to GDP ratio at the end of period $t$.

Individuals maximize discounted utility, $u(c_{y,t}) + \beta u(c_{o,t+1})$, subject to budget constraints when young and old,

$$\begin{aligned} c_{y,t} + a_t &= (1 - \tau_t - \tau_p) y_t, && (13.19) \\ c_{o,t+1} &= (1+r) a_t + p_{t+1}, && (13.20) \end{aligned}$$

where $a_t$ denotes saving at $t$, and where, in equilibrium, $p_{t+1} = (1+n)(1+\gamma)\tau_p y_t$. The sequence for optimal consumption can then be computed from two equations: a lifetime budget constraint and an Euler equation:

$$\begin{aligned} c_{y,t} + \frac{c_{o,t+1}}{1+r} &= \left[1 - \tau_t - \tau_p + \frac{(1+n)(1+\gamma)}{1+r}\tau_p\right] y_t && (13.21) \\ 1 &= (1+r)\beta \frac{u'(c_{o,t+1})}{u'(c_{y,t})}. && (13.22) \end{aligned}$$

Note, first, that a pension system is equivalent to issuing a particular form of government debt. To see this, consider an individual who has no pension tax or transfer ($\tau_p = p_{t+1} = 0$) but who is forced to purchase $b_{p,t}$ government bonds with a promised return $r_p$. The budget constraints for this individual would be

$$\begin{aligned} c_{y,t} + a_t + b_{p,t} &= (1 - \tau_t) y_t, && (13.23) \\ c_{o,t+1} &= (1+r) a_t + (1+r_p) b_{p,t}. && (13.24) \end{aligned}$$

If we set $b_{p,t} = \tau_p y_t$ and $1 + r_p = (1+n)(1+\gamma) \approx (1+n+\gamma)$, then the budget constraints with "pension debt" (13.23-13.24) are equivalent to those with a pension system, (13.19-13.20). Note that the return on forced saving in the PAYG pension system is equal to the growth rate of output. Therefore, the pension system increases the present value of household income for the young if and only if wage growth exceeds the interest rate, i.e., iff $\gamma + n > r$ (this condition determines whether the right-hand side of eq. (13.21) is increasing in $\tau_p$). This insight has an important implication: if the interest rate is larger than the growth rate of output $n + \gamma$ (a "normal" scenario) then the pension system is effectively a tax on the young generation. The generation who are old when the system is first introduced

gain because they receive benefits without having paid taxes when they were young. But the current young generation and all future generations lose because they get a higher return on private savings than on pension contributions. However, if the interest rate is *lower* than the wage growth rate $n + \gamma$, then all generations gain from introducing a pension – both the initial old generation and all generations of young. In this case, introducing a PAYG pension system is Pareto improving. This corresponds to an equilibrium featuring dynamic inefficiency, as discussed in Section 6.4.1.

A major change in this overlapping-generations model relative to the standard infinite-horizon model is that Ricardian equivalence no longer holds. In particular, the timing of taxes and transfers now matters for the distribution of consumption across cohorts, and for the trajectory of aggregate consumption. To see this, consider a one-time transitory tax holiday in period $t$ in an economy with zero initial debt. Thus, $\tau_t = 0$ and $B_t = gY_t$. Moreover, assume that the government finances the repayment of this debt by increasing taxes in period $t + 1$ and does not issue debt thereafter. Note, first, that this "tax holiday" increases the present value of consumption for generation $t$ and lowers the present value of consumption for generation $t + 1$; see equation (13.21). Thus, the debt-financed tax cut shifts the tax burden from generation $t$ to generation $t + 1$. The result is that aggregate consumption will increase in period $t$ and fall in period $t + 2$.[12] Thus, Ricardian equivalence breaks down. This is different from the case we studied in Section 13.4. There, Ricardian equivalence held because the government's debt policy did not affect the present value of taxes for the representative household. Here, in contrast, debt policy reshuffles the tax burden across generations.

This analysis has a number of implications for actual policies. Real-world pension systems are not purely PAYG and many countries have accumulated pension funds. However, public pension savings are relatively small: the U.S. Social Security Administration is scheduled to deplete its trust fund by 2033.

Two factors have strained public pension systems in OECD countries. First, the number of retirees relative to the number of workers has increased and will continue to increase in coming decades. Population aging is driven by both lower mortality (retirees living longer) and by lower fertility. This can be interpreted as a lower $n$ in the model above. Second, the productivity growth rate has fallen in recent decades (secular stagnation). For example, the U.S. growth rate of GDP per capita – $\gamma$ in our model – fell from 2.3% between 1950 and 2000 to 1.2% between 2000 and 2020.

The analysis above suggests that pension promises are a form of debt. What is the total level of effective government debt for the U.S. federal government? A narrow definition of debt corresponds to the value of government bonds outstanding. For the U.S., federal debt held by the public was 94% of GDP in the second quarter of 2023.[13] A more comprehensive definition includes the implicit debt in the pension system, i.e., the present value of future

[12] The effect on aggregate consumption in $t + 1$, $C_{t+1} = c_{0,t+1}N_t + c_{y,t+1}N_{t+1}$, is ambiguous because the tax cut will increase $c_{0,t+1}$ and lower $c_{y,t+1}$.

[13] Debt held by the public excludes the holdings of government debt by Federal government entities such as the Social Security Trust Fund, but includes debt held by the Federal Reserve.

federal pension promises. For the U.S. federal government this measure of debt is $65.9 trillion, or almost 2.5 times annual GDP.[14] This massive figure excludes the future costs of Medicare (the federal health care program for retirees). These two measures of government debt—94% versus 94+245=339%—are strikingly different. The *de facto* debt burden therefore depends on how seriously one should take promises about financial debt versus promises of future pension benefits. An outright default on nominal debt is ruled out by the U.S. Constitution. However, the government could increase surprise inflation and thereby inflate away some of the debt. Pension promises, in contrast, do not enjoy any constitutional protection and the government is always free to reduce social security benefits or raise the age at which people are eligible to collect them.

## 13.7 Taxes and transfers as instruments for redistribution

An important function of government is to redistribute and provide social insurance. To this end, the tax and transfer system includes a wide array of taxes, social insurance programs and means-tested benefits at different levels of government (federal, state, and local). To illustrate the extent of redistribution embedded in the U.S. system, Figure 13.9 plots pre- versus post-government income for each percentile of the pre-government income distribution. Pre-government income is income before taxes and transfers. Post-government income is disposable income, defined as pre-government income plus transfers minus taxes. Each dot in the plot shows average pre- and post-government income for one percentile of the pre-government household income distribution. The relationship is approximately linear, except at the lowest income percentiles. This suggests that the U.S. tax- and transfer system can be well approximated by a log-linear function:

$$y - T(y) \quad = \quad \lambda y^{1-\tau}, \tag{13.25}$$

where $y$ is pre-government household income, $T(y)$ is taxes minus transfers, and $y - T(y)$ is disposable income. The parameter $\lambda$ controls the level of taxation, while the parameter $\tau$ can be interpreted as a measure of tax progressivity. To see this, note that when $0 < \tau < 1$, the tax system is *progressive* in the sense that the marginal tax rate $T'(y)$ is larger than the average tax rate $T(y)/y$ for any positive income level. Conversely, when $\tau < 0$, the marginal tax rate is lower than the average tax rate, $T'(y) < T(y)/y$, implying that taxes are *regressive*. When $\tau = 0$, the tax system is flat, with a constant marginal tax rate $T'(y) = T(y)/y = 1 - \lambda$.

The right panel of Figure 13.9 plots average net tax rates, defined as taxes minus transfers divided by pre-government income. The picture illustrates that the average net tax rate is increasing with income. The U.S. tax and transfer system can therefore be said to be *progressive*.

[14] Source: 2023 OASDI Trustees Report, Table VI.F2.

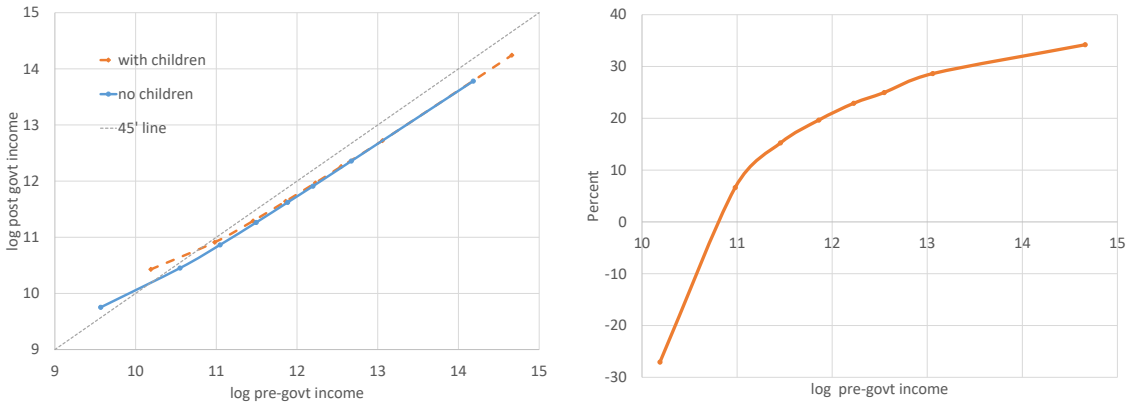



Figure 13.9: Left plot: Scatter plot of pre-government income against and post-government income for percentiles of U.S. households. Right plot: Average net tax rates by household income, defined as taxes minus transfers as a share of income, for households with children.

**Source**: Congressional Budget Office (CBO), 2016.

### 13.7.1 A macro model of progressivity

We now illustrate the effects of tax progressivity on inequality and the macro economy using a simple static model of redistribution. The economy is populated by a unit continuum of individuals indexed by $i$. The utility function $u$ is

$$u\left(c_i, \ell_i, G\right) = \log c_i - \frac{\ell_i^{1+\frac{1}{\phi}}}{1+\frac{1}{\phi}},$$

where $c_i$ and $\ell_i$ are consumption and labor supply of individual $i$.

The tax and transfer system is assumed to take the log-linear form described in equation (13.25). The government's budget must be balanced, which imposes a constraint on the set of feasible fiscal policy choices $(\tau, \lambda, G)$.

The aggregate resource constraint dictates that output is spent on either private consumption or on public goods:

$$Y = \int_0^1 c_i \, di + G.$$

Individuals differ with respect to labor productivity. Their labor income is $w_i \ell_i$, where $w_i$ is individual $i$'s productivity. Individuals have no wealth, so consumption must equal disposable income:

$$c_i = \lambda \left(w_i \ell_i\right)^{1-\tau}.$$

Taking a first-order condition with respect to hours worked, one can solve in closed form for the equilibrium allocation. Hours worked and consumption are given by

$$\log \ell_i = \frac{\log(1-\tau)}{1+\frac{1}{\phi}}, \tag{13.26}$$

$$\log c_i = \log \lambda + (1-\tau)\frac{\log(1-\tau)}{1+\frac{1}{\phi}} + (1-\tau)\log w_i. \tag{13.27}$$

Hours worked are falling in $\tau$ but are independent of the individual wage, $w_i$. Progressivity ($\tau > 0$) reduces hours because workers internalize that if they increase hours they will face a higher marginal tax rate, depressing the after-tax return. In the limit as $\tau \to 1$, workers anticipate that disposable income will equal $\lambda$, irrespective of hours worked, and thus hours will shrink to zero. Hours are independent of the wage because the utility function is in the balanced growth class.

Consumption is increasing in individual productivity, $w_i$. Tax progressivity dampens the pass-through from wages to consumption: a one percent increase in wages translates to a $1-\tau$ percent increase in consumption. Thus, tax progressivity reduces consumption inequality: the variance of log pre-government earnings is $var(\log w)$, while the variance of log consumption is $(1-\tau)^2 var(\log w)$. In conclusion, this simple economy illustrates the fundamental trade-off between efficiency and redistribution in a setting with an empirically plausible tax- and transfer system: higher progressivity reduces hours worked (lower efficiency) but also reduces consumption inequality (more redistribution). For more discussion of this trade-off, see Heathcote, Storesletten, and Violante (2017).

# Chapter 14

# Asset prices

*Monika Piazzesi and Martin Schneider*

# Chapter 15

# Money

*Andreas Hornstein and Per Krusell*

## 15.1 Introduction

So far, this textbook has exclusively discussed *real* variables. In the models covered, money simply is not present, and "dollars" have no special meaning. In microeconomics, on which the macroeconomic models here are based, the label money is sometimes used but merely to denote a numéraire. That is, the only prices that matter are relative prices between goods and services traded. In macroeconomics more broadly, however, money is very frequently in focus. First, people are concerned with *inflation* and the notion that it may erode the purchasing power of people's income. This mechanism, moreover, maybe a more serious problem for some people than for others, so inflation affects the degree of inequality in society. Second, one of the main macroeconomic policy tools available to governments is *monetary policy*, conducted by *central banks* by controlling the stock (supply) of money or the nominal interest rate, i.e., the rate of exchange between dollars at different points in time (overnight, or from year to year). Historically, monetary policy was perhaps primarily used as a source of revenue (*seigniorage*): the printing of new money (notes and coins) is cheap and allows the government to help finance its operations. Today, the monetary policy conducted by central banks is more about maintaining price stability and stabilizing business cycles: to limit fluctuations in macroeconomic activity.

In most developed countries up until 2021, the need to control inflation from becoming too large had almost disappeared from public debate, since inflation rates had been maintained at well under 5 percent for three decades and the (explicit or implicit) target for inflation of 2 percent that most economies maintain had even been difficult to reach, with episodes nearing deflation during and around the Great Recession period. Then quite suddenly, at least at first glance as a result of world events—e.g., supply-chain challenges in the aftermath of the 2020 coronavirus pandemic and Russia's 2022 invasion of Ukraine—many prices, such as those for goods and later energy, rose sharply, and significantly and inflation rates were again well above 5 percent. Today, inflation control is again an important topic in developed economies.

Meanwhile, many less developed and emerging-market economies have struggled with inflation rates remaining at high levels, disrupting daily life in major ways. When inflation runs out of control completely, we speak of *hyperinflation*, which may seem merely like an intellectual curiosity to those who were fortunate enough not to have experienced one, but which quite clearly is disastrous for the economy. Thus, first-order questions for the present chapter include: what determines the price level and causes inflation, how is human welfare affected by inflation, and how can stable prices be maintained? These questions fundamentally hover around the "dollar bill": what is it really for and how does the market and government policy jointly determine its value?[1] We will therefore introduce theories *of* money, in an effort to explain the role money has in the economy. We will also discuss less ambitious theories —they do not attempt to explain money's role—but are at least theories *with* money that are frequently used. With these frameworks, we will discuss the determination of inflation and discuss a variety of options for monetary policy.

[1] Like Barbie, the dollar bill asks itself "What was I made for?".

A particularly salient phenomenon in monetary economics is *indeterminacy*: the idea that there are many (competitive) equilibria in a given economic environment and that these equilibria may be associated with different real allocations. In a basic sense, this should perhaps not be surprising: money's value today ought to depend on its value in the future, which in turn depends on its value after that, and so on. Can hyperinflations, for example, simply be unfortunate equilibria in economies where there is also an equilibrium with a stationary price level? We will see that in many economic environments, the answer is yes. Thus, it seems that at least in theory, the view held by many, including Milton Friedman, that (hyper)inflation can only result as a consequence of the central bank allowing the money supply to increase (a lot), is not correct.[2] Whether hyperinflations are associated only with excessive money growth empirically is, however, not so much in focus in the chapter: the main purpose here is merely to cover basic monetary theory.

The need for theory with (or of) money in monetary economics is not a foregone conclusion, however: in the framework that is most frequently employed in policy-making settings today, the New Keynesian model developed in Chapter 16, money is not even present. That setting hence abstracts from monetary aggregates and instead focuses entirely on the role of nominal stickiness (of prices and wages) and how it can make monetary policy have significant real effects. In contrast, in the present chapter, prices will always be assumed to be flexible, so that the discussion can be focused sharply on more basic issues. However, we will demonstrate how the New Keynesian model without money can be motivated, namely, as the "cashless limit" of a model economy with money.

The chapter begins in Section 15.2 with the first fundamental model *of* money: the overlapping-generations model, with which Samuelson (1958) made a case that intrinsically useless paper money—*fiat* money—could have value under some circumstances. The idea here is that in the overlapping-generations setting when the young's endowments and preferences are such that they want to save, in the absence of capital or other assets, there is no one they can lend to. However, the presence of fiat money can allow this saving. When fiat money has value, it is therefore an example of an asset bubble: money is an asset, without a dividend, and when people then accept money in exchange for goods and services it must be that people believe in its value only because others do. In this model, money's function is that of a *store of value*, and it has value so long as there is no other asset that dominates money in return, e.g., bears interest. Moreover, the section demonstrates that there is equilibrium indeterminacy: there are also hyperinflationary equilibria as well as an equilibrium where money never has value.

Next, Section 15.3 looks at models with infinitely lived agents. There, we first show that money cannot have value, even when other assets are missing. The intuition behind this result is that in a finite-horizon economy, money quite trivially cannot have value in the very last period, since it is fiat: at that time, no seller of goods or services will accept money as payment. Hence, it is never valued earlier either. Subsection 15.3.1 shows that

[2]Milton Friedman's famous 1963 speech in India includes the assertion "inflation is always and everywhere a monetary phenomenon." Friedman's message, which was based on empirical observation, was that money printing, i.e., central bank policy, is a necessary and sufficient condition for inflation to occur.

this intuition survives also with an infinite horizon but, most importantly, the section sets up a general consumer budget constraint with money as well as government bonds. Next, we move to models *with* money, where a reduced-form liquidity value of money is assumed directly. Section 15.3.2 thus begins with the cash-in-advance model where money is assumed to be needed to purchase goods. The motivation is money's second role: the medium-of-exchange role. Similarly, in money-in-transactions-costs and money-in-the-utility-function models, real money balances are assumed to save on transactions costs and to give direct utility, respectively. These models feature rate-of-return dominance: other assets, that bear interest, do not yield these kinds of liquidity services (by assumption). The simple cash-in-advance model we present can also be seen as a motivation for the *quantity theory*: $M/P$ is a constant times real output. Not only undergraduate textbooks but also some advanced research papers use a demand for money formulation that is simply an assumed quantity equation.[3] Using the reduced-form models, in Section 15.3.3 we then briefly discuss optimal monetary policy (absent any stabilization concerns), including the well-known *Friedman rule*. We also show that in the models with reduced-form liquidity as well, there can be equilibrium indeterminacy.

We then move to several other important conceptual topics. One is the cashless limit mentioned above: the idea that it is possible to use nominal quantities without even having a money stock present in the model. In fact, one of the main purposes of this chapter is to prepare the ground for Chapter 16 on New Keynesian models, where we then simply begin by assuming that money is absent but that prices are sticky in nominal units. The cashless limit is non-trivial and involves the notion of *monetary policy rules*, such as interest-rate rules. In particular, when the interest rate is set as an increasing function of the price level, a range of equilibria are eliminated. This insight underlies the use of *Taylor rules* in New Keynesian models and in practical policy-making, where interest rates respond to the price level or the inflation rate. Finally, yet another policy rule that has received significant attention involves the interaction between fiscal and monetary policy: the *fiscal theory of the price level*. We explain its origin and the intuition behind why it can be seen as a mechanism for eliminating equilibria and, yet, is controversial.

Before briefly discussing another way to motivate money as a store of value—namely based on there being a limited set of other assets available in the economy, the topic of Section 15.4—we then look at multiple monies, i.e., exchange rates, in Section 15.5. The idea is not to venture into international economics; rather, we discuss the implications of our basic theories of money for the relative values of different fiat currencies that might even circulate within a given economy. This section therefore also involves a short discussion of crypto-currency.

In Section 15.6, we finally look at theories *of* money as a medium of exchange. Here, the idea is that search frictions among traders of goods/services with an *absence of double coincidence of wants* can be seen as a deep friction motivating a value for fiat money.

[3] See, e.g., Mankiw and Reis (2002).

## 15.2 Money in overlapping-generation models

Historically, various kinds of monies have circulated, often in the form of real objects of intrinsic value (such as precious metals), with a variety of price stability and longevity outcomes. However, today money—as defined by notes and coins— is fiat, i.e., it has no intrinsic value, and it is not "backed" (say, by gold). Rather, people voluntarily choose to accept money as a means of payment for goods and services because they expect money to have real value later when they want to use it. Money is thus an asset, but not one that promises anything real of direct value. Asset pricing, as discussed in Chapter 13, should thus be useful, but we need to specify more clearly what the potential future benefits of money might be for someone considering accepting it today.

Undergraduate textbooks mention three roles of money: it is a store of value, a medium of exchange, and a unit of account. In this first part of the chapter, we look at money as a store of value. For that, we begin with the overlapping-generations model, which already Samuelson (1958) realized could be used to explain why money—under some conditions—will have equilibrium value, despite being fiat. That model satisfies Neil Wallace's *dictum* (see Wallace (1998)): the use of money should not be an assumption, but an outcome, and in fact one among many, since another plausible outcome is that people do not believe it will ever have value. Thus, let us now revisit the overlapping-generations model considered in Chapter 5. We proceed using simple examples, which can be easily elaborated on and extended.

### 15.2.1 An endowment economy

So first consider an economy without production. People live for two periods and there is a representative agent per generation. The endowment vector of any agent is $(\omega_y, \omega_o)$, i.e. the economy is stationary, and preferences are assumed to be logarithmic: $\log c_y + \log c_o$ for all cohorts, except for the old at the beginning of time (time 0) whose utility is simply strictly increasing in $c_o$.

There is fiat money in the environment: an amount $M$ of perfectly divisible and intrinsically useless objects. We assume that the initial old ones own these. The core question now is whether fiat money can have market value. Clearly, the idea here is that money can be used as a *store of value*. As such, it may be valuable since the endowment economy does not allow saving (and the young have no one to lend to that will be able to pay back).

Let $p_{mt}$ denote the value of a unit of money at time $t$ in terms of consumption goods at time $t$. Thus, $p_{mt} = 0$ for all $t$ would be a *non-monetary* equilibrium, where no one values money. Also, let $P_t \equiv 1/p_{mt}$ be the "price level" at time $t$, that is, the price of a unit of consumption goods at time $t$ in terms of money.

Assume for the moment that $p_{mt}$ is positive and finite, that is, $0 < P_t < \infty$. Then, the maximization problem of the generation $t$ agent is

$$\max_{c_y, c_o, M'} \log c_y + \log c_o \tag{15.1}$$

$$\text{s.t.} \quad c_y + \frac{M'}{P_t} = \omega_y, \quad c_o = \omega_o + \frac{M'}{P_{t+1}}, \quad \text{and} \quad M' \geq 0.$$

The last inequality is natural: money can only be held in positive amounts. This is unlike other assets we discussed so far; for example, we thought of it as being possible to borrow by issuing bonds, that is, holding negative amounts of bonds.

The agent of generation $-1$ has a trivial decision and simply sets $c_{o,0} = \omega_o + M'/P_0$.

We will continue to assume, for now, that $P_t$ is finite and positive. We can then combine the constraints from (15.1) to describe the available budget set for the consumer, without explicitly involving money. This delivers

$$c_y + \frac{c_o}{P_t/P_{t+1}} = \omega_y + \frac{\omega_o}{P_t/P_{t+1}} \quad \text{and} \quad \omega_y - c_y \geq 0. \tag{15.2}$$

The implied budget set is presented in Figure 15.1; the inequality constraint, which reflects monetary holdings being non-negative, implies that consumption when young cannot exceed endowments when young. As can be seen, the gross real return on money is $p_{m,t+1}/p_{m,t} = P_t/P_{t+1} \equiv 1/(1+\pi_{t+1})$. Here $\pi_{t+1}$ denotes the inflation rate between $t$ and $t+1$.[4]

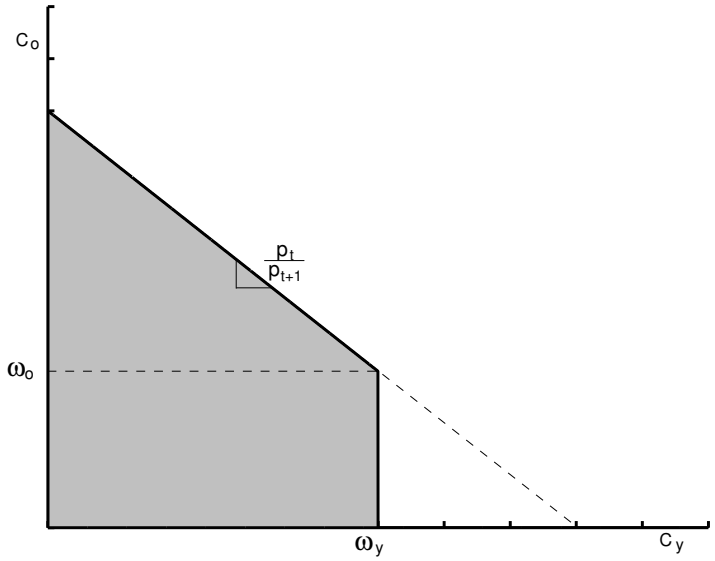



Figure 15.1: Budget set in the economy with fiat money

Solving the maximization problem is straightforward. Since the consumer's indifference curves are strictly convex and strictly decreasing we realize from the figure that either the solution is a point on the downward-sloping budget line that is tangent to the indifference curve—and, hence, the inequality constraint is slack—or it is right at the kink of the budget set, with $c_y = \omega_y$.[5] To find out which case applies, first ignore $\omega_y - c_y \geq 0$ and solve the

[4] From first-order Taylor approximation it follows that net real return on one dollar invested in money is $\simeq -\pi_{t+1}$ (for small values of $\pi_{t+1}$).

[5] Alternatively, use Kuhn-Tucker optimization; here, the Kuhn-Tucker multiplier will be zero in the former case and positive in the latter.

first-order conditions combined with the budget. The solution becomes

$$\begin{aligned} c_y &= \frac{1}{2}\left(\omega_y + \omega_o \frac{P_{t+1}}{P_t}\right), \\ c_o &= \frac{1}{2}\left(\omega_y + \omega_o \frac{P_{t+1}}{P_t}\right)\frac{P_t}{P_{t+1}}. \end{aligned}$$

Now check to make sure that $c_y \leq \omega_y$. This amounts to

$$\omega_y \geq \omega_o \frac{P_{t+1}}{P_t} \quad \Longleftrightarrow \quad \frac{P_t}{P_{t+1}} = \frac{p_{m,t+1}}{p_{mt}} \geq \frac{\omega_o}{\omega_y}.$$

That is, if the gross real return on money is at $\omega_o/\omega_y$ or above, the solution is valid. The smaller is the ratio $\omega_o/\omega_y$, the larger is the consumer's desire to save to smooth consumption over time, but of course the consumer is also dissuaded from saving if the return on money is low. If $p_{m,t+1}/p_{mt} < \omega_o/\omega_y$, the solution is $c_y = \omega_y$ and $c_o = \omega_o$ (and $M_{t+1} = 0$): the consumer would want to buy a negative amount of money (borrow), but of course cannot. In sum, individual money demand by the young at $t$ as a function of the price levels at $t$ and $t+1$ is

$$\frac{M_{t+1}}{P_t} = \max\left\{\frac{1}{2}\omega_y - \frac{1}{2}\omega_o \frac{P_{t+1}}{P_t}, 0\right\}. \tag{15.3}$$

Given the demand function, we can solve for equilibrium, which amounts to the young buying the entire money stock from the old:

$$M_{t+1} = M \ \forall \ t.$$

There are two cases to consider. In one, money would have no value at any point in time. This would amount to $p_{mt} = 0$ (or an infinite value of $P_t$) for all $t$.[6] Recall that, in this case, (15.2) is not valid. Instead, the consumer is simply unable to save because no matter how much money they acquire, this money will be worthless when they are old. Hence, there is an equilibrium where money does not have value.

If $P_t$ is instead finite (and positive), so that money has real value, then we obtain

$$P_{t+1} = P_t \frac{\omega_y}{\omega_o} - \frac{2M}{\omega_o}. \tag{15.4}$$

This is a first-order difference equation where the initial value $P_0$ is not given: it is endogenous (this is the whole point!). Thus, if we can find a solution to this difference equation where $P_t$ is positive at all points in time, we have a monetary equilibrium. For this, consider the following three cases: (i) $\omega_y > \omega_o$; (ii) $\omega_y < \omega_o$; and (iii) $\omega_y = \omega_o$. First, we can see that there is a constant solution to this difference equation: $P_t = \bar{P} = 2M/(\omega_y - \omega_o) > 0$. Whether

[6] In this case, $M_{t+1} = M$ can still be assumed to hold: if money is free (and of no use), we can assume it is passed on in full from generation to generation. Expressed in terms of the price level, this looks like an equilibrium that does not satisfy standard conditions ($P_t$ is infinite), but the appropriate price is the price of money in terms of goods, which is simply zero.

this value is positive or negative clearly depends on the case we are in: in case (i), it is and in case (ii) it is not; case (iii) is a knife edge, where the value is zero.

Is the constant solution of case (i) the only possible solution or could there be non-stationary equilibria, where the price level changes over time? Cases (i) and (ii) are depicted graphically in Figure 15.2, where dynamics of equation (15.4) can be analyzed.

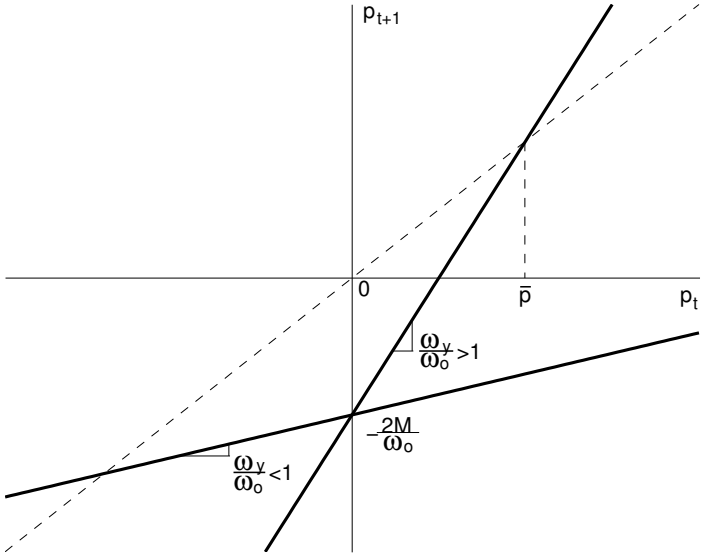



Figure 15.2: Price dynamics

Case (i) is the one where the solid line has a slope greater than one. Consider a $P_0 < \bar{P}$. We see, following the kind of Solow-picture dynamics seen in Chapter 3, that no such value can be part of an equilibrium: sooner or later, $P_{t+1}$ will be below zero, which cannot be an equilibrium. On the other hand, if $P_0 > \bar{P}$, we see that the price dynamics will involve monotonically higher and higher price levels, ad infinitum. If we translate this into the sequence for the real value of money, the $p_m$'s, we see that money will gradually lose value and, in the limit, have zero value. This is an endogenous hyperinflation; we see that $P_{t+1}/P_t = \omega_y/\omega_o - 2M/(P_t\omega_o)$. I.e., the gross rate of inflation will rise over time and, in the limit, become $\omega_y/\omega_o$.

Case (ii), on the other hand, is one where no monetary equilibrium exists: from the figure, for any positive $P_0$'s, $P_t$ will eventually be negative, though how long this would take depends on how high a $P_0$ is selected. Case (iii), not included in the figure, is a knife-edge case where no monetary equilibrium exists either. This case would be represented by a solid line of slope 1, with a negative intercept with the y-axis. Again, like in case (ii) any positive $P_0$ would eventually lead to a negative price.

Note that the amount of nominal units in the economy have no real import: it is possible to rewrite the equilibrium difference equation as $\frac{1}{m_{t+1}} = \frac{1}{m_t}\frac{\omega_y}{\omega_o} - \frac{2}{\omega_o}$, where $m_t \equiv p_{mt}M$. I.e., the equilibrium determines the real value of the total money stock, independently of how

many fiat money units are available.

To summarize: if $\omega_y > \omega_o$, there is a continuum of equilibria indexed by an initial price level in the set $[2M/(\omega_y - \omega_o), \infty)$. Of these different initial values of money, only one—the stationary one—gives money a positive value in the limit: all the others involve hyperinflation where money's value goes to zero.

### 15.2.2 Welfare comparisons across equilibria

This model satisfies Wallace's dictum: money is not a "primitive" in that it may or may not have value. Moreover, there are important observations to make regarding welfare. In short, money can become a store of value when a store of value is missing, as we shall now argue.

Recall the welfare characterization of overlapping-generations equilibria from Chapter 5: the Pareto efficiency of equilibria is all about the asymptotic marginal rate of substitution between consumption when young and when old or, equivalently, the gross marginal return on saving between consecutive periods. Namely, if this return is above or equal to one, the equilibrium is efficient, and if it is below one, it is inefficient.[7] Hence, in the $\omega_y \leq \omega_o$ case, where money cannot have value, there is Pareto efficiency. When money can have value ($\omega_y > \omega_o$), all equilibria (non-stationary monetary as well as non-monetary) are Pareto inefficient, except for the stationary monetary equilibrium: the long-run return on saving is $\omega_o/\omega_y < 1$ in the former cases and 1 in the stationary monetary equilibrium.

Is it the case that monetary equilibria, when they exist, are helpful from a welfare perspective? In particular, do they provide a Pareto improvement on the autarky allocation? Clearly, the stationary monetary equilibrium does: the initial old are happier in it, since they can sell their money for a positive amount, and all the other cohorts obtain utility equal to $2\log((\omega_y + \omega_o)/2)$, which exceeds $\log \omega_y + \log \omega_o$.[8] Notice that the fully smoothed allocation is better than autarky for all cohorts also in the case where $\omega_y < \omega_o$ but money cannot help attain this allocation. Moreover, if it could, it would make the initial old worse off.

In fact, all the non-stationary monetary equilibria do improve on autarky as well. Moreover, the monetary equilibria are ranked in the following sense: the smaller is $P_0 \geq \bar{P}$, the higher is utility for all generations. We will leave these results as an exercise for the reader; the key insight is that a higher return on saving is better for the consumer in this case.

### 15.2.3 Extensions: a neoclassical growth economy and policy

The basic results of the previous section extend in various ways. The essential ingredient in obtaining a value of money is the lack of a (good) store of value. Put differently, for money not to have value here, either people don't need a store of value—such as when $\omega_y < \omega_o$—or a "good enough" alternative store of value is available. So, consider the overlapping-generations version of the neoclassical growth economy of Chapter 5, where people can save

[7] The non-monetary equilibria described here do not have a return on saving since there is no active savings vehicle. However, we can introduce borrowing and lending, which have to equal zero in equilibrium. The gross interest rate therefore needs to adjust to $\omega_o/\omega_y$ to make the young choose exactly zero borrowing.

[8] Make sure that you can show this!

using physical capital. Clearly, capital would then compete with money and potentially rule out monetary equilibria.

Now the characterization of equilibria is somewhat more cumbersome than in the endowment case. However, consider any such equilibrium, and for the preferences assume $\log c_y + \log c_o$, just as above. Recall that an equilibrium is still Pareto inefficient if, as time goes to infinity, the limit for the gross interest rate is below one. This, moreover, is an outcome that is possible in the neoclassical model. It's another good exercise to try to derive a condition under which you can see this to be true.[9] In such a case, fiat money can have value and real effects on economic activity. Capital will not be competed out entirely if production is Cobb-Douglas, so money and capital will both be used and both give a gross return of 1 asymptotically.

It is straightforward to introduce money printing into the model, e.g., as lump-sum transfers. If the government increases the money supply at a gross rate of $1+\mu$ every period, then the same type of equilibrium characterization as above obtains, with the difference that the stationary monetary equilibrium is now replaced by one where the real return on money is below one—it is $1/(1+\mu)$—and where the total value of the money stock is lower. Money printing can also be used as a source of income for the government (seigniorage) to pay for its expenditures or debts. One can thus also introduce government bonds easily. If they are in positive supply, bonds would compete with money as a store of value. Open market operations can be studied, whereby the government would change its outstanding stock of bonds over time, which in general will also affect allocations and the value of money. However, the effects on allocations only materialize if the present-value budget of a cohort is changed; otherwise, a Ricardian-equivalence theorem applies, as studied in Chapter 13. We will return to some of these issues in the next section.

Thus we can construct versions of the overlapping-generations model for which valued money co-exists with alternative means of storage if it pays the same rate of return as these alternative assets. However, the overlapping generations model still has trouble explaining why people in actual economies hold money despite there being other assets that dominate it in return.

## 15.3 Money in dynastic models

Dynastic models are different from overlapping-generations models in that their competitive equilibria, absent other frictions, are Pareto optimal. Thus, at least in this sense, money is not needed (as a store of value or otherwise). We will thus begin to demonstrate that indeed, money cannot have value in the basic dynastic settings. Then we introduce, one by one, additional assumptions that will give money value.

[9] To find at least one example, look at a case where there is a closed-form solution, such as when $\omega_o = 0$ and $\delta$, the rate of depreciation of physical capital, is 1.

### 15.3.1 Fiat money has no value

We now show that in a standard infinite-horizon environment with dynastic households fiat money has no value if it only only serves as a store of value. For this and the following sections, we will consider variations of a representative agent production economy with labor or one with endowments only (both covered in Chapter 5). Unlike in the section on overlapping generations, we will now explicitly consider a more general structure with government bonds as well as changes in the stock of money.

The household's preferences over consumption, $c$, and leisure, $\ell$, and the production of consumption through labor are described by

$$\sum_{t=0}^{\infty} \beta^t u\left(c_t, \ell_t\right) \tag{15.5}$$

$$c_t = z\left(1 - \ell_t\right). \tag{15.6}$$

The planning solution would hence maximize the stated utility subject to this constraint; the solution is time-independent and characterized by $zu_c(c, \ell) = u_\ell(c, \ell)$.

The household's budget constraint, in real terms, is

$$c_t + q_t a_{t+1} + \frac{M_{t+1}}{P_t} + \frac{B_{t+1}}{P_t} = w_t\left(1 - \ell_t\right) + a_t + \frac{M_t}{P_t} + \left(1 + i_{t-1}\right)\frac{B_t}{P_t} - \tau_t.$$

Here, $a$ is a real asset; its discount rate is $q_t = 1/(1 + r_t)$, where $r_t$ is the net real return between $t$ and $t+1$. $M$ and $B$ are nominal holdings of money and nominal bonds, respectively; $P$ is the price level. The nominal net interest rate on bonds held between time $t$ and $t + 1$ is $i_t$. The government imposes a lump sum tax, $\tau_t$.

As before, it proves more convenient to use the notation $p_m \equiv 1/P$ than $P$: $p_m$ is the real value of a unit of money. Hence an equilibrium where money has no value would have $p_{m,t} = 0$ for all $t$. Using this notation, the consumer's constraints are given by

$$c_t + q_t a_{t+1} + p_{m,t} M_{t+1} + p_{m,t} B_{t+1} = w_t\left(1 - \ell_t\right) + a_t + p_{m,t} M_t + p_{m,t}\left(1 + i_{t-1}\right) B_t - \tau_t \tag{15.7}$$

$$M_t \geq 0. \tag{15.8}$$

The household is allowed to borrow or lend in the real asset and nominal bonds, but cannot borrow by issuing money. The government's budget constraint (where we abstract away from real expenditures) is given by

$$p_{m,t}(M_{t+1} - M_t) + p_{m,t} B_{t+1} = p_{m,t}\left(1 + i_{t-1}\right) B_t - \tau_t. \tag{15.9}$$

In equilibrium, the representative household holds outstanding fiat money and nominal government debt and real assets are zero, $a_t = 0$. Thus, the resource constraint is satisfied: $c_t = z(1 - \ell_t)$.

Let $m_t \equiv p_{m,t}M_t$ and $b_t \equiv p_{m,t}B_t$ denote the real value of money and bonds and define $\hat{a}_t = a_t + m_t + b_t(1 + i_{t-1})$ to be total real wealth at the beginning of the period. By using this definition, along with some algebra, we can rewrite the budget as

$$c_t + q_t\hat{a}_{t+1} + \left[1 - q_t\frac{p_{m,t+1}}{p_{m,t}}\right]\frac{p_{m,t}}{p_{m,t+1}}m_{t+1} + \left[\frac{p_{m,t}}{p_{m,t+1}} - q_t\left(1 + i_t\right)\right]b_{t+1} = w_t\left(1 - \ell_t\right) + -\tau_t + \hat{a}_t. \tag{15.10}$$

In this budget constraint, $\hat{a}$ (and no other variable) allows saving from $t$ to $t+1$. The terms involving $\frac{p_{m,t}}{p_{m,t+1}}m_{t+1} = M_{t+1}/P_t$ and $b_{t+1}$ are "static": they are losses/gains at $t$ to the extent the expressions in square brackets are positive/negative. Thus, first, the household would not want to hold money if its real return, $p_{m,t+1}/p_{m,t}$, is below the real interest rate $1/q_t$. Second, the household would be able to attain unbounded consumption for given wealth and prices if the real return on bonds, $(1 + i_t)p_{m,t+1}/p_{m,t}$ were not equal to the real interest rate: if the return is higher (lower), the consumer could obtain unbounded resources by raising (lowering) $b_{t+1}$ without bound. Thus, for an equilibrium with positive money holdings to exist, the following needs to hold:

$$q_t = \frac{p_{m,t}}{p_{m,t+1}} = \frac{p_{m,t}}{p_{m,t+1}}\frac{1}{1 + i_t}. \tag{15.11}$$

The no-arbitrage condition means that we can write the budget constraint in more compact form as

$$c_t + q_t\hat{a}_{t+1} + \frac{i_t}{1 + i_t} \cdot \frac{p_{m,t}}{p_{m,t+1}}m_{t+1} = w_t\left(1 - \ell_t\right) - \tau_t + \hat{a}_t. \tag{15.12}$$

We see immediately that money cannot be held in positive amounts—it cannot have real value, i.e., $p_m$ will have to equal 0—if the nominal interest rate on bonds is positive. If $i_t > 0$, money would then be dominated in return by bonds and be a pure loss too hold in positive amounts. Second, if $i_t$ is zero at all times (or nominal bonds are not available in the economy), then money (or bonds, which are now equivalent to money) cannot have value either.[10] To see this, note that the consumer's optimization problem collapses to a problem with one asset with real return $1/q_t$. We know that for this problem the transversality condition is a necessary condition for optimality. Applying the condition to the real value of money holdings we obtain

$$0 \geq \lim_{T\to\infty} q_0q_1 \ldots q_Tm_T = \lim_{T\to\infty} \frac{p_{m,0}}{p_{m,1}}\frac{p_{m,1}}{p_{m,2}} \ldots \frac{p_{m,T-1}}{p_{m,T}}p_{m,T}M_T = p_{m,0}\lim_{T\to\infty} M_T. \tag{15.13}$$

Thus, if fiat money is not vanishing (i.e., being withdrawn) in this economy, money cannot have value earlier on. In other words, for money to have value in this economy it must disappear in the limit. We will return to this point.

To conclude, note that with bonds, we can rewrite the expression for their return as the *Fisher equation*:

$$1 + i_t = \left(1 + r_t\right)\frac{P_{t+1}}{P_t}, \tag{15.14}$$

[10] Recall that nominal interest rates at zero were observed for an extended period in the aftermath of the Great Recession, so this case is not just a theoretical one.

which states that the gross nominal interest rate equals the gross real interest rate, $1 + r_t = 1/q_t$, times the gross inflation rate. Here it can be seen as an arbitrage condition involving real and nominal bonds.

### 15.3.2 Fiat money with reduced-form liquidity value has value

There are different ways to give value to money by assuming, in a reduced-form way, that it matters to consumers. We now briefly look at them, one by one. Because government bonds, from the perspective of the consumer, are identical to borrowing and lending—both $b$ and $a$ can be held in positive as well as negative amounts and hence will yield the same return—we only keep $a$ in the frameworks. We will sometimes refer to the nominal interest $i_t$, which as before means the money return at $t+1$ on a nominal bond bought at $t$.

#### The cash-in-advance model

We first impose the constraint that money has to be used in transactions; in particular, we assume that goods can only be purchased using money. This constraint, commonly known as a cash-in-advance (CIA) constraint, from Clower (1967), gives rise to an equilibrium where money has value and is dominated in return by other assets.

Let us first incorporate the CIA into the household's budget constraint

$$c_t \leq p_{m,t} M_t \tag{15.15}$$

$$q_t a_{t+1} + p_{m,t} M_{t+1} = a_t + p_{m,t} M_t - c_t + w_t (1 - \ell_t) - \tau_t. \tag{15.16}$$

The first expression is the CIA constraint and states that money is needed to buy goods. The second expression states that money that is not spent today can be saved holding money or the asset $a$. The two expressions combined also state that current earnings from the sale of goods are not available contemporaneously for consumption.

Consider now the modified consumption-savings problem of the representative agent

$$\max_{\{c_t, \ell_t, m_{t+1}\}} \sum_{t=0}^{\infty} \beta^t u(c_t, \ell_t) \tag{15.17}$$

$$\text{s.t.} \quad c_t \leq m_t, \tag{15.18}$$

$$c_t + q_t a_{t+1} + \frac{p_{m,t}}{p_{m,t+1}} m_{t+1} = a_t + m_t + w_t (1 - \ell_t) - \tau_t, \tag{15.19}$$

where we have again replaced nominal money balances with real balances. The first-order conditions for the household's problem are

$$u_c(c_t, \ell_t) = \lambda_t + \mu_t \tag{15.20}$$

$$u_\ell(c_t, \ell_t) = \lambda_t w_t \tag{15.21}$$

$$\lambda_t q_t = \beta \lambda_{t+1} \tag{15.22}$$

$$\lambda_t \frac{p_{m,t}}{p_{m,t+1}} = \beta (\lambda_{t+1} + \mu_{t+1}), \tag{15.23}$$

where $\beta^t\mu_t$ and $\beta^t\lambda_t$ are the Lagrange multipliers on the CIA constraint (15.18) and on the budget constraint (15.19), respectively.

Now consider a stationary equilibrium for a constant money stock $M$ and labor productivity, $z$, and let us consider a general equilibrium: then, the household holds all of the money, net assets are zero, $a = 0$, and consumption and leisure are $c = z\ell$. From the FOCs for consumption and leisure it follows that the Lagrange multipliers are constant, and from the FOC for assets it follows that the discount rate on assets is equal to the discount factor, $q = \beta$.

If the CIA constraint is binding then the price of money is constant, $c = p_m M$, which as before delivers a nominal interest rate such that $q(1 + i) = 1$, as there is no inflation. From the FOC for real balances it then follows that the Lagrange multiplier on the CIA is positive, $\mu > 0$. Combining the FOC for consumption and leisure we obtain

$$\frac{u_\ell(c, c/z)}{u_c(c, c/z)} = z\frac{\lambda}{\lambda + \mu} = \beta z, \tag{15.24}$$

where we have also used the Euler equation for real balances, (15.23). Equation (15.24) pins down the steady-state consumption level, from which all the remaining equilibrium variables can be obtained.

Notice that the CIA must be binding in the stationary equilibrium. To see why, suppose it is not binding, that is, $c < p_{m,t}M$ and $\mu = 0$. From the first-order condition for real balances it then follows that the return on money is equal to the interest rate. But this implies that the price of money is increasing over time. So the CIA constraint remains satisfied, but the TVC is now violated as shown in the previous section.

Notice, now, that compared to the same economy without the CIA constraint, the marginal rate of substitution between leisure and consumption is less than the marginal rate of transformation: $\beta z < z$. But this means that in the CIA economy leisure is higher and consumption is lower. Effectively, the real wage is lower in the CIA economy because today's labor income can only be used tomorrow and the return on saving the labor income is less than the real interest rate. Is there a way to fix the problem?

Suppose the government imposes a nominal lump sum tax on the representative household, $T_t$, to be paid using money. Also, assume that money is withdrawn at a constant rate, $M_{t+1}/M_t = \gamma < 1$. Then the implied real lump sum tax is $\tau_t = p_{m,t}(M_t - M_{t+1}) = (1 - \gamma)p_{m,t}M_t$. With a binding CIA constraint the price of money increases as the money stock shrinks, $M_t p_{m,t} = c_t$, and the discount rate for real balances is $\gamma$. We can still find a stationary equilibrium where the value of the total money stock is constant: the nominal stock changes at a gross rate $\mu$, but the value of each unit of money changes at $1/\mu$; for $\mu > 1$, this means a constant rate of inflation and for $\mu < 1$ a constant rate of deflation. We can, in particular, choose $\gamma = \beta$ such that the return on assets and money is equalized. For this policy the TVC is satisfied since the nominal money stock is vanishing in the limit. This result—that withdrawing money from the economy at the rate of discount makes the equilibrium optimal, and constitutes optimal monetary policy—is known as the *Friedman rule*. Milton Friedman cast this policy in terms of "paying interest on money", which is

something the central bank in principle could engineer. With the money stock shrinking at rate $\beta$, people are not constrained in the use of money, and similarly if money paid interest and were therefore identical to bonds, they would not be constrained either.

The CIA model of money does not satisfy Wallace's dictum: though we have not demonstrated it, money will always be valued here because it has been hard-wired to be equal to consumption (in real terms, and hence its value could not be zero, or else consumption would be zero). Money is a store of value in the CIA model but one that is worse than bonds and other assets: its value derives from it being required in order to buy goods. Money is thus used as a medium of exchange, though the exchange is not explicitly modeled. One way to describe the CIA constraint is that it imposes a *quantity equation* of sorts, with a velocity of 1: the equation $VM = Py = Pc$ is met by assumption with a $V = 1$. The CIA model is an often-used framework and arguably it can be used to address some questions in monetary economics that do not involve endogenous changes in the way money is used in exchange. We very briefly look at the quantity equation from an empirical perspective in Section 15.3.2.

The CIA constraint is sometimes also used to motivate the notion of "aggregate demand" used in many undergraduate textbooks: a downward-sloping relationship with output on the x-axis and the price level, $P$, on the y-axis. On the surface and due to the language, this demand function looks like it is just taken from microeconomics, but recall that such relations in microeconomics involve relative prices between goods and $P$ is not such a price: $P$ is the average dollar price of all goods and services or, put in terms of the one-good economy we use as a benchmark macroeconomic model, it is the dollar price of a unit of the good. Thus, to the extent $P$ is a relative price, it is not comparing goods but it is comparing goods and dollar bills, the latter of which is not valued in utility or production. However, to the extent it reflects money's value in exchange as described by a CIA constraint, this relation can be seen as an argument behind why a relation like the aggregate demand curve may exist.

Besides a CIA constraint there are other ways to introduce money into a standard representative agent economy without frictions. Next, we consider two prominent examples. The first example posits that transactions are costly, but less so if one holds money. The second example simply represents any inherent value to money by incorporating it into the utility function.

### Digression: the quantity theory in the data

Defining characteristics of the Quantity Theory of Money (QTM) are the presence of a stable demand for money and that money growth and inflation move one for one. If there is a stable nominal demand for money that is proportional to the nominal transaction volume, that is, the product of real activity and prices, then price inflation will move one-for-one with money growth over the long run if money is neutral over the long run, that is, real activity is independent of money growth over the long run.

For an empirical evaluation of QTM one thus needs measures of the money stock, the transaction volume, quantities and prices, and the opportunity cost of holding money. From the point of view of money as a means to execute transactions, one can think of various

measures. Standard measures of money published by central banks include M1, consisting of currency and checkable demand deposits, and M2, which adds savings deposits to M1. Regulatory changes and technical advances may affect what should be included in a measure of money, see Lucas and Nicolini (2015). For example, before the 1980s regulation Q in the U.S. imposed limits on the ability of banks to pay interest on accounts. Once Regulation Q was abolished banks started to offer new interest-bearing liquid accounts, e.g., money market deposit accounts with limited transaction features. In the mid-1990s IT improvements made moving funds between demand deposits and money market (SWEEP) accounts easier, making the latter close payment substitutes. Payments from bank accounts have also been made easier using electronic transfers that can be initiated using mobile phones. Bitcoin and other digital currencies may represent additional future means of payment. Thus what should be included in a measure of money changes over time, which affects the stability of money demand as defined by a fixed measure of money. Furthermore, the M1 and M2 measures listed are simple sums of their various components, but these components may well differ in their ability to perform transactions. Similar to the aggregation of various final goods into an aggregate output measure like GDP, Divisia indices have been proposed for constructing aggregate money stocks; see Barnett (1980). The standard measure of transactions in the literature is GDP, but obviously, many transactions precede the purchases of final goods in the economy: there usually are multiple stages of production. Thus using GDP as a sufficient statistic for the transaction volume implicitly assumes that the production structure is not changing much over time. Finally, various short-term interest rates have been used for the opportunity cost of holding money. For the U.S. that includes the Federal Funds rate, the rates on short-term commercial paper or short-term U.S. Treasuries. In the following, we study how well the QTM holds up for the U.S. for the period 1901 to 2023 using M2 as a measure of the money stock, GDP as a measure of transactions, and the 6-month commercial paper rate as a measure of the opportunity cost of holding money.[11]

We first consider the long-run relation between M2 growth and GDP inflation.[12] For this purpose we calculate long-run movements of M2 growth and inflation as 15-year symmetric moving averages. In Figure 15.3 we plot the filtered M2 growth rates and GDP inflation for annual data from 1902–2023 for five sub-samples: 1902–1928, 1929–1954, 1955–1983, 1984–2005, 2005–2023. The first period precedes the Great Depression, the second period covers the Great Depression and World War II, the third period covers the post-WW-II period including the inflationary 1970s, the fourth period covers what has been called the Great Moderation and the adoption of inflation targeting among advanced economies, and the fifth period covers the Great Recession and the policy of Quantitative Easing (QE). We see that inflation moves roughly one-for-one with money growth but that there is notable variation in that relation across subsamples. On the one hand, during the period covering high inflation in the late 1960s and 1970s, inflation appears to respond strongly to changes in money growth. On the other hand, during the Great Moderation and inflation targeting, the response of inflation to changes in money growth is much weaker, and during the QE

[11] The use of other measures of money and interest rates produces very similar results.

[12] This discussion follows Sargent and Surico (2011).

period, inflation appears to decline as money growth increases.

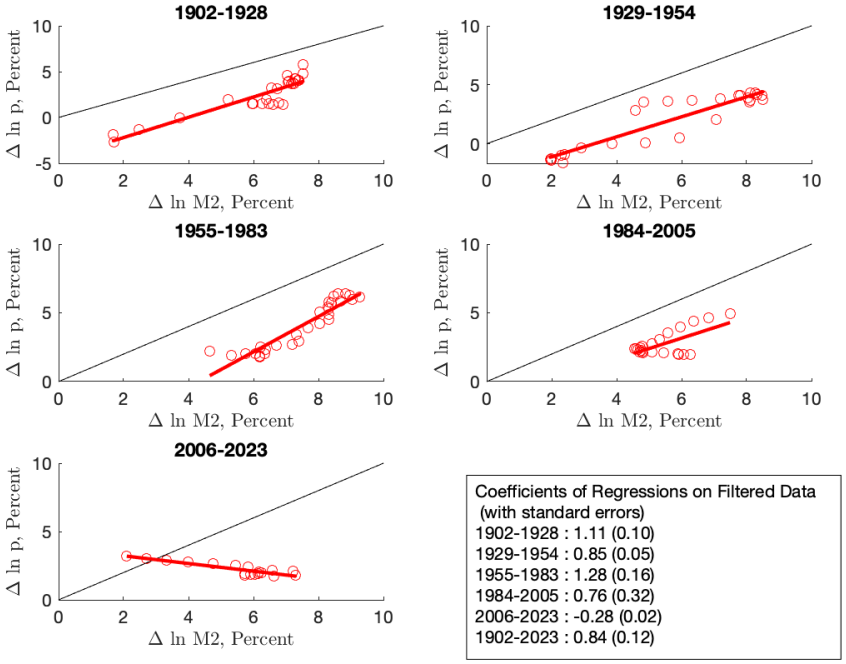



Figure 15.3: Inflation and money growth

**Notes**: Each panel plots the filtered GDP inflation rates against the filtered M2 growth rates (red circles), and contains the 45-degree line (thin black line) and the fitted values from an OLS regression of inflation on money growth (thick red line). The estimated OLS coefficients for the sub-samples are in the lower right-hand corner with heterogeneity and auto-correlation corrected standard errors in parentheses.

We now consider the long-run stability of money demand. In Figure 15.4 we plot the M2-GDP ratio and the 6-month commercial paper rate for both the actual data and their 15-year symmetric moving averages. At first inspection, Figure 15.4 seems to provide evidence for the long-run stability of money demand despite the large and persistent deviations of the variables from their long-run trends. Whenever the filtered short rate increases, the filtered M2-GDP ratio declines. This stable relationship for the filtered data disappears, however, when we look at the sub-samples as in Figure 15.3. Now, the OLS regression coefficients of the M2-GDP ratio on the short rate can be either positive or negative, and they are usually not significant.

A cautious summary is that the evidence on QTM is mixed.

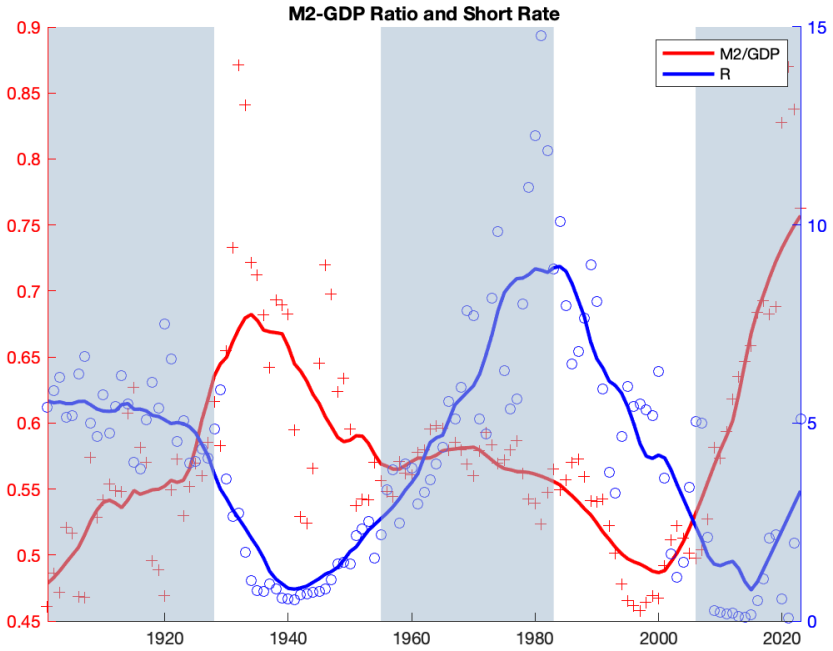



Figure 15.4: Inflation and money growth

**Notes**: We plot annual data for the ratio of M2 to nominal GDP (red crosses) and the 6-month Commercial Paper (CP) rate (blue circles). The solid red and blue lines are the 15-year symmetric moving averages of respectively the M2-GDP ratio and the CP rate. Shading and its absence sets apart the five subsamples from Figure 15.3.

### Money reduces transactions costs

Consider the budget constraint (15.19) but assume that the purchase of consumption goods involves a real cost in terms of the consumption good

$$c_t + q_t a_{t+1} + \frac{p_{m,t}}{p_{m,t+1}} m_{t+1} = a_t + m_t + w_t (1 - \ell_t) - \tau_t - \psi \left( \frac{c_t}{m_{t+1} p_{m,t} / p_{m,t+1}} \right) c_t, \quad (15.25)$$

where $\psi(x) = \kappa x^\eta$ is an increasing function, $\kappa, \eta > 0$. Thus transaction costs are increasing in consumption, decreasing in real balances, which we recall are $M_{t+1}/P_t = m_{t+1} p_{m,t}/p_{m,t+1}$, and homogeneous of degree one in consumption and real balances jointly. Here current real balances facilitate transactions, unlike in the CIA model which requires real balances carried over from the previous period to make transactions.

Using a subscript $t$ on $\psi$ and its derivative to indicate evaluation at $\frac{c_t}{m_{t+1} p_{m,t}/p_{m,t+1}}$, we

obtain first-order conditions for the household's problem as follows.

$$u_c(c_t, \ell_t) = \lambda_t \left[1 + \psi_t' \cdot \frac{c_t}{m_{t+1} p_{m,t}/p_{m,t+1}} + \psi_t\right] \tag{15.26}$$

$$u_\ell(c_t, \ell_t) = \lambda_t w_t \tag{15.27}$$

$$\lambda_t q_t = \beta \lambda_{t+1} \tag{15.28}$$

$$\lambda_t \frac{p_{m,t}}{p_{m,t+1}} \left[1 - \psi_t' \cdot \left(\frac{c_t}{m_{t+1} p_{m,t}/p_{m,t+1}}\right)^2\right] = \beta \lambda_{t+1}, \tag{15.29}$$

where $\beta^t \lambda_t$ is the Lagrange multipliers on the budget constraint (15.25). Combining the FOCs for the real asset and real balances, (15.28) and (15.29), we obtain

$$1 - \frac{q_t}{p_{m,t}/p_{m,t+1}} = \psi_t' \cdot \left(\frac{c_t}{m_{t+1} p_{m,t}/p_{m,t+1}}\right)^2. \tag{15.30}$$

Recalling the Fisher equation (15.14) for the nominal interest rate and using the functional form for $\psi$ we arrive at an expression,

$$\frac{M_{t+1}}{P_t} = c_t \left[\frac{1}{\kappa\eta} \frac{i_t}{1 + i_t}\right]^{-1/(1+\eta)}, \tag{15.31}$$

which relates the demand for money to the value of transactions and the opportunity cost of holding money, the nominal interest rate. This expression is in line with the usual expressions for money demand, unlike in the simple CIA model above, for which money demand is interest-inelastic.[13] That is, real money demand is equal to consumption (or output, if output is defined net of transactions costs) times the inverse of velocity, an expression that is decreasing in the nominal interest rate.

We have so far described this model only from a single consumer's perspective. To solve for a general equilibrium, we would need to specify production; we could assume linear production as in the previous model with $w_t = z$. As before, we would then assume that the consumer holds the entire money stock and that $a_t$ is zero at all times. The budget constraint would then reduce to the economy's resource constraint, which says that consumption equals production minus transactions costs. Finally, the Fisher equation is used, relating inflation and the nominal interest rate to the real interest rate; in the absence of inflation, in a stationary equilibrium, this would boil down to $q(1 + i) = 1$.

### Money in the utility function

We have just assumed that real balances are for some reason helpful in executing consumption purchases. We could take an even more reduced form approach and assume that real balances are valuable, period. For this purpose, we simply include real balances in the utility function

[13] More elaborate versions of the CIA model also feature interest-elastic money demand.

of the representative agent, rather than modifying the budget constraint. So, assume that preferences are

$$\sum_{t=0}^{\infty} \beta^t \left[ u\left( c_t, m_{t+1} \frac{p_{m,t}}{p_{m,t+1}} \right) + v\left( \ell_t \right) \right]; \tag{15.32}$$

here, the second argument of $u$ is simply the real amount of money purchased at $t$, $M_{t+1}/P_t$, but expressed in terms of $m$ and $p_m$. The specific functional form used is convenient but of course not the only possibility. The FOCs for the household's consumption-savings problem with preferences (15.32) and budget constraint (15.19) are

$$u_{c,t} = \lambda_t \tag{15.33}$$

$$v'_t = \lambda_t w_t \tag{15.34}$$

$$\lambda_t q_t = \beta \lambda_{t+1} \tag{15.35}$$

$$\lambda_t \frac{p_{m,t}}{p_{m,t+1}} = u_{m,t} \frac{p_{m,t}}{p_{m,t+1}} + \beta \lambda_{t+1}, \tag{15.36}$$

where we have suppressed the arguments of the derivatives of $u$ and $v$ for compactness. Another equation is also relevant: our budget constraint (15.19) does not explicitly feature nominal government bonds paying interest $i_t$ between $t$ and $t+1$, but that is because of the no-arbitrage condition between real and nominal bonds that yield the Fisher equation, $q_t(1+i_t) = p_{m,t}/p_{m,t+1}$. Using this equation together with equations (15.35) and (15.36), we obtain $1 = u_{m,t}/u_{c,t} + 1/(1+i_t)$. This delivers an expression that is very similar to equation (15.30):

$$\frac{u_{m,t}}{u_{c,t}} = \frac{i_t}{1+i_t}. \tag{15.37}$$

If we assume that $u(c,m)$ is a homothetic function, that is, it is a monotone transformation of a homogeneous of degree one function, we again obtain a money demand equation that relates the ratio of real balances to consumption, $m/c$, to the opportunity cost of holding real balances, $i/(1+i)$, similar to (15.31). We can then again treat this equation as a (static) money demand relationship much like that in the transactions-costs case. And as in that case, the stationary general equilibrium allocation is obtained by adding assumptions on production and market clearing.

### 15.3.3 Policy and the value of money in the reduced-form models

We will now illustrate some key features of the reduced-form models of money. For most of the discussion, we will limit attention to the case with money in the utility function.

#### Stationary equilibria: quantity theory and optimality

All of our reduced-form models of fiat money give rise to a well-defined demand for real balances. In the CIA model, this demand for real balances is independent of the nominal interest rate, equation (15.18) with equality. But for the models where real balances reduce

transaction costs or provide utility directly, this demand may depend on the nominal interest rate, e.g., equations (15.31) and (15.37). Since the environment in our examples is stationary so far we have studied their stationary equilibria for which quantities and prices are constant. In these equilibria real balances are constant and therefore the price of money is inversely proportional to the money stock. This reflects the quantity theory which is usually taken to state that the price level is proportional to the money stock.

**The Friedman rule** In the CIA model we have shown that in general the stationary equilibrium outcome is sub-optimal relative to the economy without frictions, that is, without the CIA constraint. We also saw that the outcome of the frictionless economy can be recovered if the money stock is shrinking at a rate such that the price of money is increasing and the real rate of return on money is equal to the real discount rate. In other words, the price level is shrinking and there is deflation at the real interest rate. Furthermore, from the Fisher equation, it then follows that the nominal interest rate is zero. This is the Friedman Rule: if money is dominated in return but fills a function in the economy, then optimal policy is to reduce the opportunity cost of holding money to reduce distortions. For the CIA model such a policy can eliminate the distortion completely, but for the two other reduced-form models of money the full optimum can only be reached exactly if additional "satiation assumptions" are made. In the first economy, transactions cost need to reach zero for some finite value of real balances; in the second, $u$ needs to have a minimum in it second argument.

### Equilibrium indeterminacy

A feature of monetary economies is that they naturally give rise to multiple equilibria. This is mainly because forsaking real resources and buying money today is less advantageous if it loses value entirely in the next period (i.e. if others are not willing to buy it): the price of money today depends naturally on the price of money in the future. In the overlapping-generations economy of section 15.2, money was only used as a savings vehicle and there is therefore always an equilibrium in that economy where money never has value. As we shall see here, reduced-form models often also deliver indeterminacy.

**Hyperinflation** In the reduced-form models, equilibria where money never has value can occur but only under special assumptions on preferences/transactions costs.[14] There may, however, exist non-stationary monetary equilibria. Consider, for example, the CIA model from section 15.3.2. Combine equations (15.20)-(15.23), assuming that the CIA constraint is binding, and normalize $z = 1$. We obtain

$$u_\ell \left(m_t, 1 - m_t\right) \frac{p_{m,t}}{p_{m,t+1}} = \beta u_c \left(m_{t+1}, 1 - m_{t+1}\right). \tag{15.38}$$

[14] In the CIA model, a non-monetary equilibrium would mean zero consumption, which is only possible if utility is bounded below in consumption. In the model where money reduces transaction costs, these costs would need to be bounded at zero real money balances, and in the model where money enters the utility function, this function must be bounded below in real balances.

Now assume that through lump-sum transfers (negative taxes) the money stock changes at the constant gross rate $\gamma > 1$, $M_{t+1} = \gamma M_t$, and we arrive at a first-order difference equation in real balances, $m_t$,

$$\gamma u_\ell \left(m_t, 1 - m_t\right) \frac{m_t}{m_{t+1}} = \beta u_c \left(m_{t+1}, 1 - m_{t+1}\right). \tag{15.39}$$

with no initial condition for real balances. This makes clear how expectations fundamentally drive equilibrium outcomes. One solution is a steady state, but this may not be the only possibility. To make this point particularly sharp, suppose that leisure and consumption are separable and that the marginal utility of leisure is constant and equal to 1: with a slight abuse of notation, $u(c, \ell) = u(c) + \ell$. This yields

$$\gamma m_t = \beta m_{t+1} u'\left(m_{t+1}\right). \tag{15.40}$$

If in addition $\lim_{m \to 0} mu'(m) = 0$, then this expression defines two steady states: one with positive real balances, $\gamma/\beta = u'(\bar{m})$, and a limiting steady state where money has no value and real balances are zero. Furthermore, for any initial real balances $0 < m_0 < \bar{m}$ the path of real balances defined by equation (15.40) converges to the steady state with zero real balances.[15] Given $M_0$ an initial $m_0$ corresponds to a price $p_{m,0}$, and $p_{m,0} < \bar{p}_{m,0}$ for $m_0 < \bar{m}_0$. Thus, for any initial $p_{m,0} < \bar{p}_{m,0}$ the price of money converges to zero faster than the money stock increases, and the value of real balances vanishes in the limit.[16]

**Local indeterminacy** In a different environment, the same policy of constant money growth through lump-sum transfers can result in a continuum of nonstationary equilibria that all converge to the unique stationary equilibrium. For a version of the reduced-form model with money in the utility function from section 15.3.2, Obstfeld (1984) shows that the local dynamics of perfect foresight paths at the steady state allow for a continuum of paths that all converge to the steady state if consumption is a normal good and the magnitude of the elasticity of marginal utility of consumption with respect to real balances is sufficiently small. For all of these equilibria, the price level relative to the money stock remains bounded.

In Section 15.3.3 we will study the interaction of monetary and fiscal policy and explore how policy modifications can lead to a unique equilibrium.

### Different monetary rules

In the above discussion, there is an implicit assumption about the conduct of monetary policy: the monetary authority selects a path for the money supply, a sequence that will be given to the economy, and for which one can then examine the set of implied equilibrium

[15] By linearization of the dynamic system around $m^*$, we see that the system is locally unstable. Linearization around $m = 0$, however, delivers stability, provided $\gamma/(\beta u'(0)) < 1$. Note also that along paths where $m_t \to 0$, the real return on money is always less than $1/\gamma$, that is, for $\gamma > \beta$ the CIA is binding.

[16] The example in this section relies on specific assumptions on utility. Under other assumptions, there is a unique equilibrium. If $u(c) = \log c$, which violates the condition stated as $\lim_{m \to 0} mu'(m) = 1 > 0$, equation (15.40) allows us to solve uniquely for $m_0$, $m_1$, and so on.

allocations. We learned—for the economies we looked at—that for a money supply path featuring growth at a constant rate, there is a unique steady-state equilibrium in which the rate of inflation equals the inverse of the money growth rate. We also learned that other equilibria may exist under some conditions. For the case in which there are government bonds, the nominal interest rate consistent with the money stock sequence would follow from the Fisher equation. However, in most economies of today, monetary policy is not described this way; rather, the more appropriate description of the conduct of monetary policy is that of a "choice of a sequence of nominal interest rates", possibly associated with a rule such as the Taylor rule as described in 15.3.3. We now discuss how to define equilibria where the central bank directly chooses interest rates (and the money supply path becomes endogenous).

**Price level indeterminacy under pure interest rate rules** First, consider a monetary policy that would just specify a sequence of nominal interest rates. Since long, such a policy has been viewed as perilous, as it leaves the price level indetermined. To see why, consider a pure interest-rate peg in the context of the general-equilibrium cash-in-advance model from section (15.3.2). In particular, assume a path for the nominal interest rate for which $i_t > 0$. As in Section (15.3.3) on hyperinflation, we specialize utility to be linear in leisure; in particular, let period utility be $u(c)+\ell$ and let labor productivity $z$ be equal to one. Then recall that the equilibrium conditions, which we only studied in a steady-state version in section 15.3.2, will read (in the order stated in that section) $c_t \leq p_{m,t}M_t$, $u'(c_t) = \lambda_t + \mu_t$, $1 = \lambda_t$, $q_t\lambda_t = \beta\lambda_{t+1}$, and $\lambda_t p_{m,t}/p_{m,t+1} = \beta(\lambda_{t+1} + \mu_{t+1})$. From this follows that $q_t = \beta$ and that $p_{m,t}/p_{m,t+1} = \beta(1 + i_t)$, using the Fisher equation, for all $t \geq 0$. Thus, the interest-rate peg determines inflation. This also means, using the last equilibrium condition, that $p_{m,t}/p_{m,t+1} = \beta u'(c_{t+1})$, thus, pinning down $c_{t+1}$ for all $t \geq 0$. Thus, conditional on the initial price of money, $p_{m,0}$, the peg delivers a determinate deterministic equilibrium from the second period and on.

However, $p_{m,0}$ is not determined. Suppose that the CIA constraint in the initial period does not bind, that is, the initial CIA multiplier is zero, $\mu_0 = 0$, and consumption is determined by $u'(\bar{c}_0) = 1$. For the given initial money stock $M_0$ any positive initial price of money $p_{m,0} > \bar{p}_{m,0} = \bar{c}_0/M_0$ will then satisfy the CIA and represent an equilibrium. Alternatively, if $p_{m,0} \leq \bar{p}_{m,0}$ then the initial CIA constraint binds, and the initial price together with the CIA determines initial consumption, $c_0$, and the FOC for consumption determines the CIA multiplier, $\mu_0$. So, any initial positive price of money indexes an equilibrium, and there is real, and not just nominal, indeterminacy.

Note that, since the nominal interest rate is always positive the CIA always binds, which then determines a path for the money stock $M_{t+1}$ for all $t \geq 0$. In other words, policy has to adjust the outstanding money stock in order to support the interest rate peg, either through transfers or through open market operations. For example, if $i_t = \widetilde{i}$ then $u'(\widetilde{c}) = 1 + \widetilde{i}$ and $\widetilde{p}_{m,t}/\widetilde{p}_{m,t+1} = \beta(1 + \widetilde{i}) = \widetilde{M}_{t+1}/\widetilde{M}_t$. This looks like the money growth rule that is consistent with hyperinflation.

**Interest rate rules with a target** Now consider amending monetary policy with an *interest rate rule*: a function relating the price level (or, more common in practice, to the inflation rate, as in the Taylor rule) to the set interest rate. The idea is then that, under appropriate restrictions on this function, the basic price level indeterminacy disappears. Let us go through the case of money in the utility function. The Fisher equation reads $1 + i_t = (1 + r_t)(P_{t+1}/P_t)$ and to this we add

$$i_t = \phi\left(\frac{P_t}{\widetilde{P}_t}\right). \tag{15.41}$$

Here, $\widetilde{P}_t$ is a target price level, which in general changes over time. How does the addition of the interest rate rule affect the determination of the price level? Suppose, for a moment, that the real side of the economy is not affected by nominal variables. That would mean that the path for $r_t$ can be taken as given. To make matters even simpler, suppose that $r_t$ is constant: $r_t = r$ for all $t$. Then the Fisher equation, together with the interest rate rule, allows us to write

$$\frac{P_{t+1}}{\widetilde{P}_{t+1}} = \frac{P_t}{\widetilde{P}_t}\frac{\widetilde{P}_t}{\widetilde{P}_{t+1}}\frac{1+\phi\left(\frac{P_t}{\widetilde{P}_t}\right)}{1+r} = \frac{P_t}{\widetilde{P}_t}\frac{1+\phi\left(\frac{P_t}{\widetilde{P}_t}\right)}{1+\widetilde{i}}, \tag{15.42}$$

where $1 + \widetilde{i} = (1 + r)(1 + \widetilde{\pi})$ and $1 + \widetilde{\pi} \equiv \widetilde{P}'/\widetilde{P}$ are the targets for the long run nominal interest rate and inflation, respectively. Equation (15.42) thus defines a first-order difference equation for the nominal price relative to its target, without an initial condition. Its local dynamics around a steady state of 1 are straightforwardly determined by the use of a first-order Taylor approximation: $P_{t+1}/\widetilde{P}_{t+1} \sim \left(P_t/\widetilde{P}_t\right)\left[1 + \phi'(1)\left(1 + \widetilde{i}\right)\right]$. If $\phi'(1) > 0$, the local dynamics are determinate: the solution is a constant, $P_t = \widetilde{P}_t$ for all $t$, or else the system explodes. If, on the other hand, $\phi'(1) < 0$, then there is indeterminacy: a continuum of paths are consistent with convergence to the given steady state.[17] In conclusion, we see that if the central bank uses a rule that raises the nominal interest rate in response to an increase of the price level relative to its target, then the indeterminacy of nominal prices is no longer a problem. The corresponding demand for nominal money then follows from the money demand equation, equation (15.37).

The discussion here has taken as given an exogenous path for $r_t$ that, moreover, is constant. The arguments extend to an exogenous path for $r_t$ that converges to a constant. How restrictive is the assumption that the path for $r_t$ is exogenous? We know that $1 + r_t = \beta\frac{u_c(c_{t+1}, M_{t+2}/P_{t+1})}{u_c(c_t, M_{t+1}/P_t)}$. In general, thus, nominal variables appear here and make the analysis more involved. A full equilibrium treatment would require specification of the production side and the equilibrium conditions would then need to be solved for. One possibility is that we again use the assumption that production is linear in labor, $c = z(1 - \ell)$, and that the government does not consume. Then the consumer's first-order condition for leisure, $zu_c(c_t, M_{t+1}/P_t) = v'(\ell_t)$, can be used directly to solve for time-independent $c_t = c$ and $\ell_t = \ell$

[17] The knife-edge case $\phi'(1) = 0$ leads to indeterminacy as well: a continuum of non-diverging price ratios $P_t/\widetilde{P}_t$ are possible (constant) solutions to the equilibrium equations.

under the special assumption that $u$ is separable in consumption and real money balances. Hence, we obtain $r_t = r$.[18] For the more general case of non-separability, one would need to examine the joint system of prices and real money holdings. It is of course possible to do so and to establish joint conditions on $\phi$ and $u$ such that the equilibrium is determinate. The details are not important, so we omit them here. However, dealing with the general case is well motivated as a utility function where money enters separately from consumption seems hard to motivate: after all, money's role is meant to be tied to the purchase of consumption goods.

**Paying interest on money** Recall Friedman's proposal: to pay interest on money. It is indeed conceivable for the central bank to pay interest on money, and it is even a policy in practical use, because many central banks pay interest on reserves. Letting $i_m$ denote the interest paid on money, if $i_m = i$ at all times, money is equivalent to bonds in financial terms for the household. Hence, with an interest rate rule, no reduced-form approach is necessary for obtaining a value of money. If $i_m < i$, money is still financially dominated in return and a reduced-form demand would need to be added for money to have value in equilibrium. Now, the compact budget constraint (15.12) would have $(i - i_m)/(1+i)$ multiplying $M_{t+1}/P_t$ and this is now the real financial cost of holding each unit of money. Using an $i_m \in (0, i)$ as an additional policy variable would not add conceptually to the rest of the analysis in this chapter, which is why our benchmark maintains $i_m = 0$.

### Digression: interest rate rules

Sims (1980) popularized the use of vector autoregressions (VARs) as a way of studying the effects of exogenous shocks on macroeconomic time series while imposing minimal assumptions to identify the exogenous shocks. Usually, shocks are identified by imposing restrictions on the covariance matrix of the residuals from estimated structural, i.e., theory-based, VARs (SVARs). Early VAR applications that studied monetary policy as a source of economic fluctuations were influenced by QTM arguments. Sims (1972), in a small-scale VAR of the U.S. economy with real GDP and nominal money finds a quantitatively large contribution of money shocks to output fluctuations. But later work considered larger-scale SVARs and once short-term interest rates were added to the list of variables, interest rate shocks replaced money shocks as a source of output fluctuations. Bernanke and Blinder (1992) then argued that monetary policy actions set a particular short-term interest rate, namely the Federal Funds rate, and thus interest rate shocks reflect the impact of monetary policy. The Federal Funds rate is the overnight interest rate in the market for interbank loans. Contemporaneously, Goodfriend (1991), based on a reading of U.S. monetary policy implementation at the Federal System, argues that

> "Except for the period from 1934 to the end of the 1940s when short-term interest rates were near zero or pegged, the Fed has always employed either a direct or

[18] This argument works also if $z$ depends on time but is converging to a constant, in which case $r_t$ converges to a constant.

an indirect Federal funds rate policy instrument." (p.8)

In particular, he observes that the Federal Funds rate is adjusted infrequently and that rate changes usually occur in a sequence of consecutive small steps. Taylor (1993a) then provides the "Taylor rule" as an example of a monetary policy rule with an interest rate instrument that also well represents Fed behavior. According to the Taylor rule, the Federal Funds rate responds to deviations of inflation from a 2% target with a coefficient of 1.5 and to percentage deviations of real GDP from potential real GDP with a coefficient of 0.5. Figure 15.5 from Taylor (1993a) plots the actual and rule-prescribed Fed Funds rate path. Variations of the Taylor rule are now an integral part of most quantitative monetary models. The Federal Reserve Bank of Cleveland and the Federal Reserve Bank of Atlanta provide utilities that calculate Fed Funds rate prescriptions for a collection of Taylor-rules.

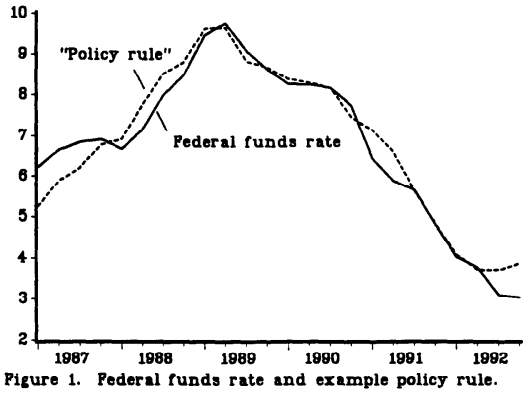



Figure 15.5: The Taylor Rule

### The cashless limit

An important part of the New Keynesian model, to which a chapter of this book is devoted, deals with price determination in the absence of a demand for money: money balances never appear in that model. The price level is still nominal—set in units of currency—and serves a different role: it is the unit in which prices are assumed to be sticky. Price stickiness is not the subject of the present chapter, however; suffice it to say that in the New Keynesian model, a firm's price *in dollar terms* cannot be changed freely. But how is it possible to have nominal variables in the model without having money? We now discuss how Woodford (2003a) motivates his approach.

So consider again the model with money in the utility function. We just saw explicitly that a framework in which $u$ is additively separable in consumption and real money balances is consistent with price-level determinacy under an interest rate rule, which is also the standard assumption about monetary policy made in the New Keynesian model. In that case,

moreover, nominal money balances are simply given *residually* from the money demand relationship $u_{m,t}/u_{c,t} = i_t/(1+i_t)$: money balances do not matter for the determination of either real variables or the price level. This economy is not cashless but it is consistent with using nominal variables without having to mention money balances. However, we also pointed out that additive separability is a very special, and arguably unrealistic, case.

Woodford therefore considers the more general case, where money matters for real allocations and for price-level determination. However, the idea is then that money matters less and less in a limiting, *cashless*, case. For suppose that we imagine that the utility function allows us to represent the "relevance" of real money balances (in comparison to consumption) with a parameter, say, $\omega$, that can be taken to zero. In that limit, then, is it possible that (i) real money balances go to zero while (ii) the price level remains finite and determinate and the remaining real variables can be pinned down as well? I.e., consumers would demand less and less money from the central bank (in nominal terms), since they care less and less about money as $\omega$ approaches zero. Let us look into this possibility here. Thus, consider the Euler equation for nominal bonds,

$$1 + i_t = \beta^{-1} \frac{u_{c,t}(c_t, M_{t+1}/P_t)}{u_{c,t+1}(c_{t+1}, M_{t+2}/P_{t+1})} \frac{P_{t+1}}{P_t},$$

which combines the Euler equation for real bonds with the Fisher equation, together with the money demand equation (15.37),

$$\frac{u_{m,t}(c_t, M_{t+1}/P_t)}{u_{c,t}(c_t, M_{t+1}/P_t)} = \frac{i_t}{1 + i_t}.$$

These two equations are interdependent but let us consider the first equation only in a limiting case. Let us focus on an interest rate rule with $\phi(1) \equiv \widetilde{i}$ along with a price target path such that $\widetilde{P}_{t+1}/\widetilde{P}_t = \widetilde{\pi}$. With constant consumption in steady state, we thus have $1 + \phi(1) = \widetilde{\pi}/\beta$. The idea is now to examine local determinacy around steady states, with successively smaller $\omega$s, i.e., with lower and lower values for real money balances in steady state, with a particular focus on the limiting case of zero. We will define the local dynamics in logarithms; the interpretation is thus one of how percentage changes in variables relate to each other, which is a robust notion also when some variables are close to zero. We thus define

$$\widehat{i}_t \equiv \log \frac{1 + i_t}{1 + \widetilde{i}}, \quad \widehat{c}_t \equiv \log \frac{c_t}{\bar{c}}, \quad \widehat{m}_t \equiv \log \frac{M_{t+1}/P_t}{\bar{m}}, \quad \text{and} \quad \widehat{\pi}_{t+1} \equiv \log \frac{P_{t+1}/P_t}{\bar{\pi}}$$

where bars denote the steady states of variables associated with the policy implied interest rate $\widetilde{i}$. We obtain the first-order Taylor expansion of the Euler equation,

$$\widehat{i}_t = \eta_c^{u_c}[\widehat{c}_{t+1} - \widehat{c}_t] + \eta_m^{u_c}[\widehat{m}_{t+1} - \widehat{m}_t] + \widehat{\pi}_{t+1},$$

where $\eta_c^{u_c} \equiv -\frac{\partial u_c}{\partial c}\frac{c}{u_c}$ and $\eta_m^{u_c} \equiv -\frac{\partial u_c}{\partial m}\frac{m}{u_c}$. The idea is now to parameterize the utility function with an $\omega$ such that, when $\omega$ approaches zero, the associated steady-state $\bar{m}$ value goes to

zero, $\eta_m^{u_c}$ goes to zero, and $\eta_c^{u_c}$ goes to a finite value. If so, the dynamics implied by the linearized Euler equation do not involve money at all, so together with the interest rate rule, it defines a complete system.[19]

A complete characterization of the class of utility functions satisfying these requirements is beyond the scope of the treatment here. An example, however, can be provided. So suppose that $u$ is an increasing, strictly concave function $f$ of $h$, a CRS CES index in the two arguments:

$$h(c,m) = \left[(1-\omega)c^{\rho} + \omega m^{\rho}\right]^{\frac{1}{\rho}}.$$

Here our key parameter appears: $\omega$ is the weight on real balances.[20] With this formulation, the money demand equation $u_m/u_c = h_m/h_c = i/(1+i)$ becomes

$$\frac{M_{t+1}}{P_t} = c_t \left(\frac{\omega}{1-\omega}\frac{i_t}{1+i_t}\right)^{\frac{1}{\rho-1}}. \tag{15.43}$$

We will assume that $\rho < 1$ so that money and consumption are more complementary than Cobb-Douglas. This implies a money demand which has unitary elasticity with respect to consumption and a constant (negative) elasticity with respect to the cost of holding money, $i/(1+i)$. These features hold regardless of the value of $\omega$; when $\omega \to 0$, these elasticities remain, with the only implication being that $M_{t+1}/P_t \to 0$. Thus, if $P_t$ is targeted to equal $\widetilde{P}_t$, with the aid of the interest rate rule, it must be that $M_{t+1} \to 0$.[21] It remains to be shown that when $\omega$ approaches zero, $\eta_m^{u_c}$ goes to zero while $\eta_c^{u_c}$ goes to a positive constant. It is straightforward, but somewhat tedious, to show this to be true under the functional assumptions given.[22]

Notice that the interest-rate rule is critical for the logic of the cashless limit to go through: the interest rate is fixed in a relation to prices such that, as $\omega$ goes to zero, $M/P$ goes to zero through $M$ going to zero. If monetary policy was instead governed by a money supply rule—the simplest form of which is to have $M$ constant over time—then $P$ would go to infinity (and $p_m$ to zero) as $\omega$ and $M/P$ went to zero, and the New Keynesian model could not be built on a core where the price level is infinity.

---

[19] We have not commented here on how $c$ might interact with real money balances but rather assumed it does not.

[20] Notice that we use $m$ as short-hand (without time subscript) here: it stands for real balances purchased; at time $t$ it is $M_{t+1}/P_t = m_{t+1}p_{m,t}/p_{m,t+1}$.

[21] For a more general utility function, note that $(i/(1+i))m/c = (u_m/u_c)m/c$ from the money demand equation. Thus $(i/(1+i))m/c = (u_{mc})m/u_c)/(u_{cm}c/u_c)$. A more general condition that delivers $m/c \to 0$ under a constant interest rate is therefore that $\eta_m^{u_c}/\eta_c^{u_m} \to 0$ as $m/p \to 0$, i.e., that the ratio of cross-elasticities of marginal utility goes to zero.

[22] Note, since $u_c = f'h_c$, that $u_{cc}c/u_c = (f''c/f')h_c + h_{cc}c/h_c$. It is easy to see, by taking successive derivatives of the CES function $h$, that in the limit as $\omega \to 0$, the second term $h_{cc}c/h_c \to 0$. However, the first term becomes $f''(c)c/f'(c)$, which is strictly negative; hence, $\eta_c^{u_c}$ is strictly positive. Turning to $u_{cm}m/u_c$, we obtain $(f''/f')h_m m + h_{cm}m/h_c$. As $\omega \to 0$, both these terms go to zero: the $h$ function has the property that both $h_m m$ and $h_{cm}m/h_c$ go to zero. Hence $\eta_m^{u_c}$ goes to zero.

### Monetary-fiscal interactions

In this section we look more carefully at the government's budget constraint and at how it interacts with monetary policy. This will allow us to touch on a number of conceptual issues. Throughout, we maintain a consolidated view of the government sector, i.e., the fiscal authority plus the central bank. We do not model their behavior but rather discuss the set of policies they could undertake. An alternative would be to formulate a game between two authorities with different objective functions, and a private sector responding to policy, but there is no established approach to that in the literature.

**The financing of a given path of primary deficits** In real terms, the flow budget constraint of the government reads

$$g_t + (1 + i_{t-1})\frac{B_t}{P_t} = \tau_t + \frac{B_{t+1}}{P_t} + \frac{M_{t+1} - M_t}{P_t}. \tag{15.44}$$

Here, the left-hand side is government spending—real purchases of goods plus debt payback, including interest—and the right-hand side describes how spending is financed: via taxes, new borrowing, or money printing, also known as seigniorage. This budget constraint consolidates the budgets of the fiscal and monetary authorities; in particular, $B$ denotes debt to the public, and any debts between the fiscal and monetary authorities are thus netting to zero. In equilibrium the present value of the budget constraint reads

$$\sum_{t=0}^{\infty} q_{0,t}\,(g_t - \tau_t) + (1 + i_{-1})\frac{B_0}{P_0} = \sum_{t=0}^{\infty} q_{0,t}\frac{M_{t+1} - M_t}{P_t}, \tag{15.45}$$

where discounting, $q_{0,t} = 1/\prod_{s=0}^{t}(1 + r_s)$, uses the Fisher equation.[23] The left-hand side of this equation is the present value of the primary deficit, $g - \tau$, plus the initial debt (including interest), and the right-hand side indicates the total remaining financing needs that will have to be covered by money printing. In this obvious sense, the central bank's operations play a role in the consolidated budget.

There is also an alternative description of the consolidated constraint, which we can obtain by defining $D_t \equiv (1+i_{t-1})B_t + M_t$ as the total liabilities of the consolidated government at time $t$. Here, $M$ is not a liability in the usual sense of something having to be paid back, but rather it is an item on the central bank's balance sheet.[24] Using this definition, the

[23] One might think that the government's present value budget constraint represents the forward solution of the flow budget constraint (15.44), subject to a no-Ponzi game condition, analogous to the household's budget constraint in Chapter 4, Section 4.3.1. There is, however, no optimization problem associated with the government, and thus there is no reason to impose a NPG condition on the government. Rather, we derive the government's present value budget constraint from the household's present value budget constraint by imposing market clearing. In this sense, the present value budget constraint (15.45) represents not only a constraint but an equilibrium condition.

[24] In the distant past, when some central banks promised that each unit of money could be exchanged for a fixed number of units of gold, the liability was more apparent.

flow budget constraint can be written (along the lines of the description of the household's constraint, equation (15.12)) as

$$d_t = \tau_t - g_t + \frac{M_{t+1}}{P_t}\frac{i_t}{1+i_t} + \frac{1}{1+r_t}\cdot d_{t+1}. \tag{15.46}$$

Thus, the given initial liabilities $d = D/P$, which can be seen as a present value, equal the current primary surplus, $\tau - g$, plus the seigniorage revenue from the current presence of money in $D$, plus the present value of the liabilities left for next period, $d'$. This notion of seigniorage is more narrow and it represents a form of arbitrage that the government carries out: to the extent the private sector appreciates money for non-financial reasons, the government saves on its financing costs by substituting non-interest-bearing money $M$ for interest-bearing debt $B$ while maintaining the same value for $D$. The interest savings, per unit of money, is $i_t/(1+i_t)$, and the total savings are larger, the larger the real balances share of the liabilities. We can also think of this form of seigniorage as that accruing keeping the path for $D/P$ fixed.

**The seigniorage Laffer curve in equilibrium and in the long run** The previous discussion merely describes government variables, assuming that in equilibrium money and bonds are held by private agents in the amounts specified. The value of money to the private sector has been the subject of the previous sections; in the case of money as a mere storage instrument (the overlapping-generations argument), money would only have value ($P < \infty$) if nominal interest rates are zero, so let us instead adopt the reduced-form liquidity perspective here. There, we could derive a static money demand equation as $M_{t+1}/P_t = c_t f(i_t)$, with $c$ being private-sector consumption and $f$ a strictly decreasing function.[25] Thus, taking $c_t$ as given, the arbitrage-based seigniorage expressed in (15.46) is proportional to $f(i)i/(1+i)$. This function has a decreasing part, $f(i)$, multiplying an increasing part, $i/(1+i)$, that starts at 0 and is bounded above by 1. So, clearly, under rather weak assumptions on $f$, the product expression will have a maximum and describe a "Laffer curve." As the nominal interest rate $i$ increases above zero, money is made, but at some point, further increases in $i$ will lead to declining seigniorage revenues.

Let us now consider a stationary, long-run equilibrium: taxes and spending are constant, $\bar{\tau}$ and $\bar{g}$, as is consumption, $\bar{c}$, inflation, $\bar{\pi} = P'/P$, the real and nominal interest rate, $\bar{q} = 1/(1+\bar{r})$ and $\bar{i}$, and the real liability and money demand levels, $\bar{d}$ and $\bar{m}$. Then, the government budget constraint (15.46) will read

$$\bar{g} + \frac{\bar{r}}{1+\bar{r}}\bar{d} = \bar{\tau} + \bar{c}f(\bar{i})\frac{\bar{i}}{1+\bar{i}}.$$

Many of the models in the literature have exact, or approximate, separation between purely monetary and real phenomena in the long run. With that motivation, let us consider $\bar{g}$, $\bar{r}$, and $\bar{c}$ to be given here. Then an increase in government liabilities (be it through an increase

[25] The unitary elasticity in $c$ here obtained under special functional-form assumptions only but is not key to the argument here.

in money or bonds) must cause an adjustment in the nominal interest rate. If, in addition, the economy is on the "benign" side of the Laffer curve, then an increased $\bar{d}$ will raise $\bar{i}$. Because of the Fisher equation and $\bar{r}$ is given, inflation will need to rise.

The long-run version of the government's budget constraint allows us to touch on the well-known Sargent and Wallace (1985) piece "Some Unpleasant Monetarist Arithmetic". In that paper, the authors argue that if the central bank at any point wishes to fight inflation through an open-market operation where the money stock is decreased and the stock of debt is increased (borrowing to buy and withdraw money from the economy), then because debt pays interest, $\bar{d}$ will rise. So in the long run, $\bar{d}$ will be higher (unless the short-run action is reversed), which will increase $\bar{i}$ and, therefore, inflation. I.e., the arithmetic of the consolidated government's budget makes fighting inflation difficult if the fiscal authority does not collaborate by increasing its primary surplus. Sargent and Wallace framed their argument as one where a "fiscal dominance regime"—where the fiscal authority independently commits to primary surpluses—makes it hard even for a monetarist central bank to fight inflation. Sargent and Wallace's argument can be made without restricting the analysis to steady states, but we omit the details here.

**The irrelevance of open-market operations** Another well-known result in the early literature on money—Wallace (1981)—concerns a situation where open-market operations, i.e., changes in the central bank's balance sheet involving changes in the amount of money in the economy and the amount of outstanding government bonds, do not affect the economy at all. This occurs in a special situation where money functions as a store of value and is not dominated in return by government bonds. The consolidated budget setting here allows us to explain the irrelevance; the original argument in Wallace (1981) instead used the overlapping-generations model of money to derive the result.

Again recall the budget constraint (15.46) and suppose now that $i_t = 0$ for all $t$. Then the budget simply reads

$$\frac{M_t + B_t}{P_t} = \tau_t - g_t + \frac{1}{r_t}\frac{M_{t+1} + B_{t+1}}{P_{t+1}}.$$

It is not clear from this budget constraint whether (and why) money has value, but the maintained assumption is that the private sector regards $M$ and $B$ as fully exchangeable, perfect substitutes. In addition, however, it is also not clear how many equilibria may exist. However, the following can be said: if an equilibrium exists, in the form of a set of sequences $\{M_t, B_t, \tau_t, g_t, r_t\}_{t=0}^{\infty}$ along with other variables (consumption, output, relative prices, etc.), then any $\left\{\hat{M}_t, \hat{B}_t, \tau_t, g_t, r_t\right\}_{t=0}^{\infty}$ with other variables maintaining their values, is also an equilibrium so long as $\hat{M}_t + \hat{B}_t = M_t + B_t$ for all $t$. The proof is trivial: the government's budget constraint is still met and the private sector's situation is entirely unchanged, so if the original sequences constitute an equilibrium, so does the proposed alternative.

All in all, the only truly central assumption behind the irrelevance proposition is that money and bonds be perfect substitutes. This assumption might make you think that the

proposition itself is irrelevant: in reality, bonds give interest payments and money does not! However, the recent long period of zero interest rates again made the proposition relevant: it is really saying that *quantitative easing*, QE, is irrelevant at the zero lower bound. Or, put differently, for QE to have real impact, it would have to be that bonds and money are not identical for some reason. The irrelevance proposition thus forces us to think more about how money and bonds really differ. Moreover, during periods with positive, but low, nominal interest rates, could QE really have quantitative importance, given that money and bonds are very similar quantitatively?

**Indeterminacy issues: fiscal back stops and the fiscal theory of the price level** We have seen, in the various discussions of different models of money, that equilibria may not be unique. Moreover, the number of equilibria may depend on the monetary policy assumed. We now consider two examples. In the first example, a "fiscal backstop" eliminates all equilibria for which money loses value in the limit. In the second example, illustrating the *fiscal theory of the price level*, a policy that fixes the present value of government net-revenues eliminates a continuum of equilibria that start away from the stationary equilibrium but eventually converge to it. In either case, we are left with the unique stationary equilibrium.

So, first, recall the hyperinflation example from section 15.3.3 with a continuum of equilibria indexed by the initial price level $P_0$, where the equilibrium price path increases without bounds for any initial $P_0 > \bar{P}_0$ in such a way that money loses value entirely in the limit. The fiscal-monetary policy specification for this example was constant money growth $M_t = \gamma M_{t-1}$ implemented through lump-sum transfers, that is, negative taxes. A simple fiscal backstop that eliminates the equilibria with price level growth that exceeds the money growth is one where the government promises to buy an unlimited amount of money if, at any time, the price level exceeds a critical value, $\bar{P}_t = \bar{P}_0\gamma^t$, with the purchases financed by lump-sum taxes; see Wallace (1981) and Obstfeld and Rogoff (1983). The critical value grows at the rate $\gamma$, since in the stationary equilibrium with constant real balances the price level grows at the money growth rate. If the market considers this policy credible, the hyperinflationary equilibria cease to exist. Thus fiscal policy is used to support a monetary equilibrium through an off-equilibrium action.[26]

Now consider another example where fiscal policy is used to rule out indeterminacy of the equilibrium path. In particular, consider the pure interest-rate peg in the context of the general-equilibrium cash-in-advance model that we studied in section (15.3.3). For this policy the initial price level, $P_0$, is indeterminate and indexes all perfect foresight equilibria. Furthermore, the interest rate peg is supported by a path for the nominal money stock, $M_{t+1}$. We now describe a policy that eliminates equilibrium indeterminacy. The policy is general in nature: it really only relies on the government's budget constraint as it holds in equilibrium. Hence, it works for a general class of models displaying indeterminacy. The idea is to commit to a fiscal adjustment in every period period *conditional* on the outcomes of $P$ and $M$ (which are those subject to indeterminacy). Such a commitment, then, is constructed

[26]Notice that this fiscal-monetary policy effectively amounts to a promise to give money value in the future. I.e., money would be "backed" (by goods, not gold).

to completely un-do the impact on the government budget of any change in the price level and implied change in the money stock (and hence seigniorage). For our interest rate peg, the simple rule for lump-sum taxes will do

$$\tau_t - \bar{\tau} = -\frac{M_{t+1} - M_t}{P_t} \tag{15.47}$$

for a given $\bar{\tau}$ and fixed spending $g_t = \bar{g}$. Here, lump-sum taxes decline one-for-one with the magnitude of the real money stock change necessary to support the interest rate peg on the equilibrium path. That is, if seigniorage were to decline, so that the right-hand side would rise, the lump-sum tax would go up to compensate for the seigniorage shortfall. Substituting the tax rule into the government's budget constraint (15.45) yields

$$\frac{(1+i_{-1})B_0}{P_0} = \sum_{t=0}^{\infty} \beta^t \left( \tau_t - g_t + \frac{M_{t+1} - M_t}{P_t} \right) = \sum_{t=0}^{\infty} \beta^t \left( \bar{\tau} - \bar{g} \right) = \frac{\bar{\tau} - \bar{g}}{1-\beta} \tag{15.48}$$

Here, the only endogenous variable is the price level $P_0$, and the promise of future fiscal adjustments thus implies a unique solution for the price level. The *fiscal theory of the price level* refers to the notion that the fiscal rule (that is committed to in the future) will imply that the present value of primary surpluses is given, which forces the price level to adjust so that the initial real debt is covered by the present value of future surpluses. The argument of course requires that $B_t > 0$, but the other details of our example are not important. One can view government debt here as an asset: a (forward-looking) claim to future payments in the form of primary surpluses.

Compare this outcome for an interest rate peg with the other case case we studied in section (15.3.3) when the interest rate is determined by a state contingent rule, equation (15.41). There we have shown that with a rule for which the interest rate is sufficiently responsive to deviations of the price level from a target the equilibrium is unique. But if the price level is unique, how can we the sure that it is consistent with the price level as determined by the governments present value budget constraint? Here it becomes important that the present value budget constraint government represents an equilibrium outcome. Therefore if the nominal interest rate rule (15.41) results in a unique equilibrium with corresponding paths for the price level and money stock then fiscal policy, lump sum taxes and spending, has to adjust such that the present value budget constraint holds. Again, the problem is related to the Sargent and Wallace (1985) piece "Some Unpleasant Monetarist Arithmetic": the price level being uniquely determined by the interest rate rule implicitly assumes a "monetary dominance" regime.

The fiscal theory of the price level might seem counterintuitive in that price setters today—imagine a fruit seller—are modeled as needing to set the price so as to make sure that the value of the government's debt becomes equal to a certain value. But note that we are talking about equilibrium prices in a dynamic competitive equilibrium, and in general, we study the properties of an equilibrium and do not ask how an equilibrium comes about. Nevertheless, when indeterminacy is an issue and off-equilibrium path behavior is needed to rule out all equilibria but one, macroeconomics turns into subtle game theory and such analysis is advanced material.

## 15.4 Missing assets

As we have seen, in the overlapping-generations model, money can have value under some restrictions on primitives, primarily involving utility functions and the time profile of endowments. In the dynastic models we then looked at, money will not have value without some reduced-form assumptions on the benefits of holding money. In the present section, we look at a final possibility: the idea that some valuable assets are not available and that money can (at least partially) replace the missing assets. This feature can be introduced in overlapping-generations models too, but we look at a dynastic version and one that is closely related to material earlier in the book: consumer heterogeneity and incomplete insurance against idiosyncratic shocks.

So consider the Huggett model studied in Chapter 9, with the tightest possible borrowing constraint: in a stationary equilibrium agents solve

$$v_\omega(a) = \max_{a' \geq 0} u(a + \omega - qa') + \beta \int_{\omega'} v_{\omega'}(a') F(d\omega'|\omega).$$

We obtain an autarky equilibrium: the agent valuing the bond the most holds it, with everybody else being borrowing-constrained. In the simplest, two-state case with nontrivial probabilities, we obtain

$$qu'(\omega_{hi}) = \beta \left( \pi_{hi|hi} u'(\omega_{hi}) + \pi_{lo|hi} u'(\omega_{lo}) \right)$$

and hence the gross real interest rate is

$$\frac{1}{q} = \frac{1}{\beta} \cdot \frac{1}{\pi_{hi|hi} + \pi_{lo|hi} \frac{u'(\omega_{lo})}{u'(\omega_{hi})}} < \frac{1}{\beta}.$$

Now suppose

$$q = \beta \left( \pi_{hi|hi} + \pi_{lo|hi} \frac{u'(\omega_{lo})}{u'(\omega_{hi})} \right) > 1.$$

Then we have a negative (net) real interest rate, like in the overlapping-generations case. Here as well, people would like to save (rather badly if $\frac{u'(\omega_{lo})}{u'(\omega_{hi})}$ is very high) but are lacking assets for it: intra-personal loans are not allowed because the asset cannot be held in negative amounts. That is, an asset is prevented from existing. Other assets are missing too—insurance claims written contingent on idiosyncratic endowment outcomes—but the key here is that a riskless asset is missing.

Now consider introducing fiat money in the economy: suppose there is a fixed stock of $M$ nominal units. If the price level is constant, $P_t = P$, then money can play the role of a riskless asset, $m = a$, with gross real return one, $q = 1$, that helps consumers with high-endowment realizations save and smooth consumption.[27] What would the steady-state real

[27] As for the overlapping-generations model, one can also imagine non-stationary equilibria here, where money would lose value over time and be worthless in the limit.

value of this money be? The answer can be obtained by solving

$$v_\omega(a) = \max_{a' \geq 0} u(a + \omega - a') + \beta \int_{\omega'} v_{\omega'}(a') F(d\omega'|\omega);$$

with implied decision rule: $a' = g_\omega(a)$. Thus the total value of the money stock is simply the total savings in this economy

$$m = \int_{a \geq 0, \omega} g_\omega(a) \Gamma(\omega, a),$$

where $\Gamma$ is the stationary distribution implied by the decision rule and the distribution of endowment shocks. Given the fixed stock of nominal money, this expression then determines the price level. Of course, the role played by money could also be played by (real) government bonds.

An even simpler missing-asset model is a deterministic version of the model above where endowments alternate between high and low: for half of the population, endowments are high (low) in even (odd) periods, and for the other half of the population they are high (low) in odd (even) periods. If borrowing is not allowed (individual debt assets are ruled out), then the equilibrium would be autarky. But money could be introduced and would have value if the autarky interest rate in the original economy is below zero. This is essentially the environment of Townsend (1980)'s turnpike model of money. Like the overlapping-generations model, the missing-asset model has the attractive feature that it can account for valued money without simply assuming it, but it cannot explain why money continues to have value when being dominated in return.

## 15.5 Multiple currencies

Exchange rate determination is a major topic in international macroeconomics: what determines the rate of exchange between two currencies, such as the dollar and the euro? Undergraduate textbook treatments would refer to certain "parity" conditions—purchasing power parity and/or (covered or uncovered) interest rate parity—as guidance. Here, the purpose is to discuss exchange rate determination from the perspective of the models considered so far. In so doing, we will also more generally touch on implications for determining the value of other money-like assets, such as crypto-currency.

### 15.5.1 Money as a store of value: Kareken-Wallace exchange rate indeterminacy

In models where money's only role is that of a store of value, such as in the overlapping-generations model, what if we introduced more than one type of money (or currency)? The main discussion here will be in the context of one country only. One reason for this is that it simplifies the exposition. Another is that the presence of consumer heterogeneity or

multiple types of goods does not appear central to any arguments. We will also consider the possibility that the stocks of currencies grow at different rates, hence mimicking differences in monetary policies across countries. The discussion follows Kareken and Wallace (1981).

So consider an overlapping-generations model with one good, one (type of) consumer per cohort, and two monies, $a$ and $b$, for the consumer to invest in. The prices of the two monies in terms of the consumption good are $p_a$ and $p_b$. The consumer born at $t$ thus faces the budget constraints

$$c_y + p_{a,t}M_a + p_{b,t}M_b = \omega_y \quad \text{and} \quad c_o = \omega_o + p_{a,t+1}M_a + p_{b,t+1}M_b,$$

and the non-negativity constraints on monies: $M_a \geq 0$ and $M_b \geq 0$. The choice is over $(c_y, M_a, M_b, c_o)$ and let us for simplicity use $u(c_y, c_o) = \log c_y + \log c_o$ as the objective function to maximize. Here, the consumer faces four prices but what matters for decision making are the real returns $p_{a,t+1}/p_{a,t}$ and $p_{b,t+1}/p_{b,t}$. Because short-selling is not possible, the consumer could not conduct arbitrage based on rate of return differences between the currencies. Thus, if at any point in time $t$ one money has a lower return than the other money, no consumer will hold the low-return money: its price will be zero. It would then follow that the price would be zero also before that date since no consumer would buy the money for a positive price and sell it later for a zero price. Similarly, the price of this money after the date $t$ would also have to be zero since otherwise, markets would not clear in the period before the price becomes positive: consumers would demand an infinite amount of money at that point. Hence possible equilibria have the structure that at all times, either (i) both monies are valued; (ii) only one of the monies is valued; or (iii) no money is valued. As the discussion in Section 15.2 should make clear, case (iii) will apply if $\omega_y \leq \omega_o$. If $\omega_y > \omega_o$, however, (i) and (ii) are possible. Clearly, case (ii) is subsumed in our previous analysis: it is possible that, for example, currency $a$ is never valued (if future agents do not value it, present agents will not either, and hence $p_{a,t} = 0$ at all times. So instead consider case (i) and let $e_t$ be the *nominal exchange rate* between the currencies, $e_t \equiv p_{a,t}/p_{b,t}$. The exchange rate measures how many units of money $b$ need to be given up to obtain one unit of money $a$: if $e$ is above 1, one unit of currency $a$ is more valuable than one unit of currency $b$. Clearly, from $p_{a,t+1}/p_{a,t} = p_{b,t+1}/p_{b,t}$ at all times, it follows that $e_t$ must be constant over time: $e_t = e$. Moreover, we can define the total money stock, in currency $b$ units, as $M \equiv eM_a + M_b$ and $p_b M$ as its real value. Thus, the consumer's budget constraints become

$$c_y + p_{b,t}M = \omega_y \quad \text{and} \quad c_o = \omega_o + p_{b,t+1}M,$$

and, just as in Section 15.2, demand for total money by the consumer born at $t$ satisfies the equation $M_{t+1}p_{b,t} = \max\{(\omega_y - \omega_o p_{b,t+1}/p_{b,t})/2, 0\}$. To close the model, we need to specify money supplies. Suppose first that they are constant over time: for all $t$, $M_{a,t} = M_a$ and $M_{b,t} = M_b$. Thus, focusing on equilibria where money has value, $M_{t+1} = eM_a + M_b$ and the equilibrium is determined by

$$p_{b,t}M = \frac{1}{2}\left(\omega_y - \omega_o \frac{p_{b,t}}{p_{b,t+1}}\right)$$

holding at all times. As in the one-money model, this defines a set of equilibrium sequences for $p_{b,t}M$: one is a steady state, where $p_{b,t}M$ is constant and equal to $(\omega_y - \omega_o)/2$; but there is also a continuum of other paths with $p_{b,t}M$ converging to zero over time. The key observation here, however, is that there is no other equilibrium condition, and hence $e$ is not determined. To be more concrete: select an arbitrary $e \in (0, \infty)$. This will define $M = eM_a + M_b$, since $(M_a, M_b)$ are given. Then $p_{b,t}$ follows from knowing $p_{b,t}M$ in the given equilibrium. This allows us to find $p_{a,t}$: it equals $ep_{b,t}$ at all times. Since the consumer's demands for individual currencies are not pinned down—there is complete indifference, since the currencies give identical returns—we can then set $M_{a,t+1} = M_a$ and $M_{bt+1} = M_b$ at all times, since their value sum is then equal to their chosen total amount of saving $(eM_a + M_b)p_{b,t}$.

Suppose now that the money stocks change over time. We can still, for an arbitrary $e$, define a total money supply as $M_t = eM_{a,t} + M_{b,t}$, which will now in general depend on time, and we can look for solutions to the difference equation for $p_{b,t}M_t$

$$p_{b,t}M_{t+1} = \frac{1}{2}\left(\omega_y - \omega_o\gamma_t\frac{p_{b,t}M_{t+1}}{p_{b,t+1}M_{t+2}}\right),$$

where $\gamma_t = M_{t+1}/M_t$. A solution will, then, be nonstationary (due to $\gamma_t$ appearing in the equation). Let us consider a particularly simple case where both money stocks grow at constant gross rates, $\gamma_a$ and $\gamma_b$. Without loss of generality, suppose $\gamma_a \geq \gamma_b$. Then $\gamma_t$ will converge to $1/\gamma_a$, and in an equilibrium where total real money, $p_{m,t}M_{t+1}$, has its value go to a positive constant in the limit, this constant will be $\omega_y - \gamma_a\omega_o$. That is the faster-growing money will dictate the long-run real value of the total money stock, and its total relative value $eM_{a,t}/M_{b,t}$—given that the exchange rate must remain constant—will go to infinity. This can be interpreted as a version of Gresham's law: "bad money drives out good money," where "bad" refers to a faster-growing stock. This statement, of course, is conditioned on an equilibrium of type (iii); another equilibrium is always that where $e = 0$, in which the value of the total money stock will converge to $\omega_y - \gamma_b\omega_o$, which is higher.

### 15.5.2 Dynastic models with a reduced-form liquidity demand

With a reduced-form liquidity demand, a key question immediately arises: how should the two monies appear (in cash-in-advance constraints, in a transactions-cost technology, or in utility)? In an early paper, Lucas (1982) considers a two-country model where there are two traded goods and consumers in both countries value both goods. Country 1 consumers, however, are only endowed with goods of type 1 and country 2 consumers are only endowed with goods of type 2. He, then, assumes that good 1 is subject to a cash-in-advance constraint involving only country 1 money, and similarly for good 2: you need country 2 money to buy it. Therefore, all consumers need both types of currency. The model gives a uniquely determined exchange rate: the value of country 1 money is tied, via the cash-in-advance constraint, to the total demand for good 1, which is real. Thus, the exchange rate is directly tied to the relative demands for the two consumption goods (and to the relative money stocks).

A similar result to Lucas's can be obtained with any of the other models where money has reduced-form liquidity demand. One can, for example, assume that foreign money has some (quantitatively limited) value to one's utility. Monetary policies will matter for exchange rates since the real value of money is pinned down by its reduced-form real role, as well as its financial return. If a country increases its money stock at a slower rate than other countries, then its exchange rate will appreciate over time, everything else equal.

### 15.5.3 Discussion: theory and data

Exchange rates fluctuate significantly over time and are typically described as random walks, implying that their movements are very hard to predict. To what extent can the theories used here be used to understand these facts? The overlapping-generations model we studied above sharply violates these fluctuations: it predicts a constant exchange rate. However, extensions of the Kareken-Wallace insight to allow for extrinsic uncertainty (i.e., non-fundamental random fluctuations, such as "sunspots") imply that the indeterminacy is even greater and allows movements in exchange rates that are unpredictable (martingales). Reduced-form liquidity models admit randomness in exchange rates to the extent there is randomness in fundamentals, such as in money supplies or output.

Relatedly, practitioners (and basic undergraduate textbooks) often refer to *Purchasing Power Parity* (PPP) as a guide for understanding what the value of an exchange rate should be: it should cost the same to buy a given set of (tradable) goods in one currency directly as it would if one swapped into another currency and bought the goods using that currency.[28] Domestic prices do not fluctuate nearly as much as do exchange rates (prices are "sticky"), implying that PPP cannot hold at all points in time, even though it might hold on average over time. So to the extent one could identify a basket of goods that is available in two countries, couldn't this be a way to think about exchange rate determination? Clearly, in the models described above—the overlapping-generations model and the reduced-form liquidity models—PPP holds: there is free trade in goods and currencies. If purchasing power parity appears to be violated in the data, then in a strict sense it is a violation of these theories. But the idea here would be to argue that the theories still hold on average over time, and hence they can be used to predict an upcoming adjustment in exchange rates in the direction of making PPP hold. However, PPP can be restored also by adjustments in the price levels: these are fundamentally endogenous. Prices may move slowly over time, but gradual movements would also allow us to move toward restoring PPP, thus not necessarily involving exchange rate adjustments. A similar condition is *interest rate parity*: the notion that investing in bonds in one country should give the same return as investing in bonds in another country. In particular, a dollar invested in U.S. bonds gives an interest of $i_t$ between $t$ and $t+1$; alternatively, the investor could buy euro at the time $t$ exchange rate, invest in euro bonds paying $\tilde{i}_t$, and convert the proceeds back to dollars at the $t+1$ exchange rate. Again, in all the theories considered above, these two transactions would give the same

[28] The so-called Big Mac index is another example of this idea, though a Big Mac can hardly be regarded as tradable.

return. Can apparent departures from this kind of parity be used to argue that the current exchange rate is too high or too low? No. Parity does not give a prediction for what the exchange rate at $t$ ought to be (conditional on knowing $i_t$ and $\tilde{i}_t$): it gives a prediction for the exchange rate at $t+1$ relative to that at $t$. Hence, it cannot help us understand the level of an exchange rate.

Because of their large fluctuations and violation of parity conditions, exchange rate movements are challenging to understand with our basic theories. The overlapping-generations model and its indeterminacy predictions does suggest a way of thinking about fluctuations. But it does not allow us to understand why dollars are used in most transactions in the United States and euro in euro countries. Reduced-form models, with the stroke of a pen, allow this feature. They also predict that if a country prints money at a faster rate than another country, and it experiences higher inflation as a result, then its exchange rate will depreciate. This feature is broadly in line with data. One interesting example is the Swiss franc, a currency that has experienced decades of consistent appreciation. Is this appreciation because Swiss inflation has been low in an international comparison? Maybe yes, but the appreciation is much stronger than what can be accounted for (using a PPP relationship) by its low inflation. Thus, the obvious fundamentals go some, but far from all of, the way toward understanding the facts.

### 15.5.4 Crypto-currency

Crypto-currency is a form of digital currency aimed to work as a medium of exchange and a store of value. The circulation of alternative currencies, for example in the form of paper money issued by commercial banks, has a long history. Whether the digital nature of crypto-currency makes it special is not clear but it does have certain distinguishable properties: it is costless to "carry" and, typically, has a way of securing privacy in that your holdings are, for example, not directly observable to the government. Thus, for example for criminal activity, crypto-currency may have a comparative advantage, especially as regulation against money laundering has been made more and more potent in many countries. But crypto-currency can be given different formats too; in some versions, it is marketed as an asset simply with above-market return (for some time), and in others as a safe store of value (e.g., the so-called stable-coins, whose value is tied to the dollar).

How can crypto-currency be understood given the above theories? The overlapping-generations model would say that it is just another money that can, based on the expectations of the behavior of future consumers, potentially be used as a store of value. I.e., its value could, when introduced, be positive, but how large is indeterminate. And its exchange rate with standard currencies could move randomly. This theory does not rely on any of the specific properties that crypto-currency has. The reduced-form models of liquidity could also accommodate more monies. A cash-in-advance theory could allow other monies as allowable means of payment for some or all of the goods. A money-in-the-utility function model could introduce another variable in utility, perhaps nested with standard money in a CES form. Of course, none of these approaches seem satisfactory since the results (in the form of the equilibrium value of crypto-currency) appear to follow very directly from the assumptions.

## 15.6 Models of money as a medium of exchange

The cash-in-advance and transaction cost models described above share the feature that money can be described as being used together with consumption and, in that sense, functioning as a medium of exchange. However, these are not models that explain the reason why money may be important in exchange. In undergraduate texts, the informal motivation for money's role as a medium of exchange is the *absence of double coincidence of wants*. I.e., when buyers meet sellers they are rarely in a situation where a direct change of goods or services is mutually beneficial (without access to some form of public record-keeping technology). The model in Kiyotaki and Wright (1989) is the first to carefully spell out frictions that can motivate this idea formally. We briefly describe a setting like theirs here (without equations). A modern version of their model is contained in Lagos and Wright (2005). In all these settings, trade is fundamentally *decentralized*: people don't (always, at least) trade in centralized markets. Also, markets are *anonymous*: there is no record-keeping. In this sense, they are similar to the search/matching model in Diamond (1982).

**Kiyotaki and Wright's (1989) model** We will simply describe the core setting and attempt to explain verbally how it works. There are three kinds of people, all deriving utility from the consumption of a good, but a consumer of type $i \in \{1, 2, 3\}$ only consumes good $i+1$, modulo 3 (i.e., $i+1$ is interpreted as 1 when $i = 3$). We denote the utility benefit of consumption by $u_i$ for consumer $i$. Moreover, there is production: upon consuming, a person of type $i$ produces a good of type $i$ and, hence, is also a producer. Hence, a person's identity $i$ is the good she produces. All goods must be indivisible: production, when it occurs, always delivers one unit of a good. Each period, people meet pairwise. Thus, by construction, when people meet, there is never double coincidence of consumption wants. For example, a person of type 1 can offer her produced good 1, but only a person of type 3 likes that good, and unfortunately that person produces good 3, which person 1 does not consume.

In this economy, storage of goods is possible. In particular, good $i$ can be stored from $t$ to $t+1$ at utility cost $c_i$ that is additive (does not interact with consumption utility). Thus, it is conceivable that a producer of type $i$, when meeting someone, swaps the good just produced for another good with the idea of storing it and possibly swapping this good for good $i+1$, and thereby experience a future utility benefit (produce again and then carry the own-produced good into the period after). The stored good would then function as a *medium of exchange*. Search is assumed to be random and a person of type $i$ is equally likely to meet a person of type $i+1$ and $i+2$, modulo 3.

In each period of the model, a person only has one choice to make: upon meeting another person, whether or not to swap goods, i.e., trade, as a function of what good the other person is carrying. A Nash equilibrium concept is used whereby a swap will only occur voluntarily, i.e., if both people's choices are to trade.

It is relatively straightforward to characterize the set of steady-state equilibria for this model. People's decisions will be made based on an expectation of what other people will do. Hence, deciding to accept a good and store it only makes sense if one's expectation is

that this good will be accepted by others and that such an exchange can (eventually) lead one to be able to consume.[29] Thus, the distribution of goods holdings in the population and the associated exchange strategies of agents is key for decision making today. Existence of equilibria is made easier if randomization is allowed, but pure-strategy equilibria can occur as well. Which good(s) might appear as a medium of exchange is a function, naturally, of the relative storage costs, but also of the relative utility benefits and of consumers' rate of discount. Similarly, whether a given good is a *general* medium of exchange (and thus is accepted in trade against all other goods) depends nontrivially on the parameters of the model. Several steady-state equilibria may also exist, which is not surprising due to the central role played by expectations.

Now *fiat money* would enter this economy as another good that a subset of agents is endowed with at time zero: "fiat" means that it is intrinsically useless in that no consumer likes to consume it, or has any production benefits from it. Like other goods, money is also indivisible and it can be carried from period to period (in no greater quantity than one) at a utility cost $c_\$$. In addition, the storage properties of money are attractive: $c_\$$ is low. There is thus a fixed amount of money in the economy that could, potentially, function as a medium of exchange. Kiyotaki and Wright show that there are assumptions on the primitive parameters such that money, indeed, functions as a medium of exchange, and can be a general medium of exchange. Naturally, there is also another equilibrium where money has no value: it is never accepted in exchange, since it is not expected to be used in exchange in the future.

**Discussion** The Kiyotaki-Wright model of medium exchange satisfies Wallace's dictum: the patterns of exchange emerge endogenously and non-trivially, including the role of fiat money in exchange. In the absence of public record-keeping, whereby people could get "credit" from giving up a good and hence obtain a good without exchange in the future, money at least partially plays this role: if a person comes into a meeting holding money, it must mean that that person gave up a good for money at some point in the past. The model has shortcomings, to be sure, the indivisibility of goods and money being one (relative "prices" in exchange are therefore either 1, 0, or infinity). The full absence of (centralized) markets also makes it difficult to imagine how monetary policy would be conducted. However, the original paper was followed up by many others, gradually relaxing strong assumptions. In particular, Lagos and Wright (2005) relax all the assumptions just mentioned and their model is still tractable due to special assumptions made on utility functions whereby the distribution of money holdings collapses: all agents carry the same amount of money into each period.[30] To go through that framework in detail is second-year material, however. It is also interesting to note that whereas the New Keynesian model, to be described below in Chapter 16, has been generalized in the direction of consumer and firm heterogeneity as

[29] Here, for example, one possibility is that person 1 trades the just-produced good 1 to obtain good 3, stores good 3 and then manages to trade good 3 for good 2.

[30] It is, however, difficult, using their setting, to rule out the use of bonds as a means of payment in the decentralized market. Hence the setting becomes close to a cash-in-advance model where bonds are ruled out as a means of payment by assumption.

well as a labor market with search frictions, it has not been merged with the medium-of-exchange settings. At least in part, this is because the New Keynesian model focuses on (i) cashlessness, with the argument that cash is more and more rarely used, and (ii) sticky prices, which is not the focus of the medium-of-exchange literature.

# Chapter 16

# Nominal frictions and business cycles

*Alisdair McKay and Morten Ravn*

## 16.1 Introduction

The Great Depression was a key event in the development of macroeconomics as a field. To many observers, this episode revealed that market economies can perform very poorly as the mass unemployment of the Depression was a clear sign that something was very wrong. It was then natural to ask what lead to this failure and what could be done about to improve the situation.

The Keynesian school of thought emerged as a result of the Depression and provides a perspective on business cycle fluctuations of all magnitudes not just crises like the Depression. One core idea of Keynesian economics is that the productive factors of the economy may not be used at the optimal level. A second core idea in Keynesian economics is that the level of nominal demand influences the level of production. A central piece of most economic theory is that, in the *long run*, nominal values are unimportant and simply a choice of units. However, going back at least to Hume's 1752 essay *Of Interest*, it has also been recognized that changes in nominal demand such as a change in the money supply may have real effects, at least temporarily, because the process of adjusting prices is neither perfect nor immediate.

This chapter serves several purposes. We will review the evidence that (a) prices adjust only infrequently and do not immediately react to changes in market conditions and (b) changes in nominal variables (here nominal interest rates) affect real outcomes such as real output. We also present the modern incarnation of the Keynesian tradition in the form of the workhorse New Keynesian model. Using this model, we will explain why imperfect price adjustment leads to a role for nominal demand in determining real outcomes, why the market equilibrium can differ from the efficient level of production, and what types of policies can lead to better outcomes.

## 16.2 Empirical evidence on price rigidity

Here we will review the empirical evidence on price rigidity. We will concentrate on nominal rigidities in prices because this is also the focus of most of our theoretical analysis, but many of the same issues that apply to the prices of goods and services also apply to wages. The research we review here examines data on the prices of individual goods as opposed to economy-wide price indices. We discuss aggregate evidence later in the chapter.

The literature on the adjustment of individual prices developed significantly since the early 2000's. In a line of important papers focused on the U.S., Bils and Klenow (2004), Klenow and Kryvtsov (2008) and Nakamura and Steinsson (2008) exploited access to the survey-based micro data underlying the construction of the Consumer Price Index (CPI) to document a series of facts about goods-level price adjustments. Much of this work has now been extended to other countries.[1]

[1]In this respect, an important effort has been made by the European Central Bank in its Inflation Persistence Network which has been summarized in Dhyne, Alvarez, Le Bihan, Veronese, Dias, Hoffmann, Jonker, Lunnemann, Rumler, and Vilmunen (2006).

The data underlying the U.S. CPI is collected by the Bureau of Labor Statistics (BLS). The BLS surveys are carried out monthly or bi-monthly in 75 urban areas collecting price quotes from about 23,000 retail and service establishments for about 85,000 individual items. These individual items are then classified into about 300 goods categories known as entry level items (ELIs).

One feature of these data that has received particular attention is the frequency of price changes. Suppose a price has a probability $1-\theta$ of changing each month. The expected time between price adjustments is then

$$\sum_{t=1}^{\infty}(1-\theta)\theta^t t = \frac{\theta}{1-\theta}.$$

Thus, observing the length of time a price of a good remains at a certain level is informative about the frequency of price changes for this good. In the data, such price change frequencies depend on the category of goods. For instance, magazine prices tend to change very infrequently while the prices of energy and food items tend to be adjusted much more frequently. The mean and median price duration of all items in the CPI are 6.2 months and 3.4 months, respectively (see Klenow and Malin, 2010).[2] The very considerable difference between the mean and the medians indicates significant heterogeneity across categories.

One of the features of these data is that prices have memory—sellers often return the price of an item to a level they have set in the past. One key source of this memory is sales, i.e. temporary discounts. The key problem presented by sales is whether such episodes should be excluded or not when estimating the frequency of price changes. If firms introduce sales as a direct response to a temporary change in market conditions, it would seem appropriate to include sales. In contrast, if the timing and size of the temporary price discount are not affected by economic conditions, price changes associated with sales may not reflect true flexibility in prices and should therefore be excluded. The literature therefore typically report estimates with or without controlling for sales.

Many individual products change over time and are replaced by new versions. Such product replacements imply that for a subset of goods, while their prices have been observed in the past, prices cannot be observed this month (and future months) because they are no longer for sale. In constructing the CPI, the BLS will substitute a new comparable product for the unavailable product. These product substitutions are often associated with price adjustments. It is an open question whether and how one should control for such product substitutions when computing measures of price rigidity. To the extent that product turnover presents an opportunity to adjust prices in the face of shocks, it would seem natural to include such price adjustments in measures meant to inform about the level of price rigidities. At the same time, by its very nature, when there is product substitution, the precise nature of the good changes and so the price change may be unrelated to market conditions.

When temporary sales prices are excluded, the estimates of the mean and median price durations rise to 8.0 months and 6.9 months, respectively (see Klenow and Malin, 2010).

[2]The statistics reported here are computed from the price change frequencies within ELIs weighted using CPI weights. These results relate to survey prices in the three largest cities in the U.S.

Hence, it is clear that whether sales are included or not makes a big difference especially if one concentrates on the median price duration estimates. These estimates treat price quotes as referring to the same item even in the face of product substitutions. Eliminating these as well, the mean and median price durations increase further to 10.1 and 8.3 months, respectively.

In summary, judging from the evidence across goods, there is considerable evidence of infrequent price changes with the range of empirically plausible estimates of price durations for the U.S. ranging from 6 months to 11 months depending on whether one focuses on the mean or the median and depending on the price measure. Estimates for European economies are typically in the upper range of this interval (see Dhyne et al., 2006).

What is not clear from this evidence is whether prices change infrequently because there are infrequent shocks to "market conditions" or because there are frictions that prevent prices from adjusting. Eichenbaum, Jaimovich, and Rebelo (2011) use data from a large retailer to estimate how the prices of products respond to the replacement cost of the goods and find that the price-cost margin varies within a fairly narrow range implying prices respond quickly to changes in market conditions.

Movements in individual prices are large relative the changes in the aggregate price level. For example, the median absolute price change is 11.5 percent (see Klenow and Kryvtsov, 2008). The obvious interpretation of this fact is that most price changes reflect conditions in a specific market as opposed to macroeconomic conditions. It is possible that prices can respond strongly and quickly to sector-specific shocks while adjusting slowly and imperfectly to aggregate shocks.[3] Boivin, Giannoni, and Mihov (2009) find evidence in support of this idea. They study disaggregated data on consumer and producer prices and examine how consumer prices respond to common and sector-specific shocks. They find that most price changes are driven by sector-specific developments rather than aggregate conditions and disaggregated prices are sticky in response to macroeconomic conditions.

## 16.3 The New Keynesian model

Early Keynesian theories were based on simple relationships between aggregate variables that did not have a close connection to the microeconomic decisions made by individuals. Research in the 1980s and 1990s developed the **New Keynesian** model that is based on the same principles of optimization and equilibrium that underlie modern macroeconomics and incorporate frictions that interfere with the immediate and perfect adjustment of prices.

Nominal rigidities can be modeled in several ways. One natural way to model them is to assume prices are fixed for a certain period of time. For example, many employment contracts are reviewed and adjusted at an annual frequency (see Grigsby, Hurst, and Yildirmaz, 2021). Apartment rents are another type of transaction with long-term contracts. The same may

[3]This may occur due to information frictions, e.g. sellers are unaware of all the macroeconomic developments that may impact the optimal price. Price adjustments may also be imperfect if firms want to keep their prices near those of other firms who are not simultaneously updating their prices. This force is known as "real rigidities" and it amplifies the effect of nominal rigidities.

be the case for wholesale prices where prices of goods are subject to longer term contracting between firms. Taylor's (1980) overlapping-contracts model adopts such a view of nominal rigidities and assumes that each price is fixed for some number $n$ periods and therefore each period $1/n$ of the prices in the economy is updated. This assumption requires the modeler to keep track of $n$ different prices. A simpler approach is to assume that a fraction $1-\theta \in [0,1]$ of the prices in the economy is updated each period, but unlike Taylor's model, the prices that are updated each period are randomly drawn from the existing prices. This assumption, which was introduced by Calvo (1983), leads to a much more tractable model because it is no longer necessary to keep track of the previously set prices as we will explain below.

**Environment.** A representative household has preferences for consumption and leisure as represented by

$$U_0 = \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \left[ \frac{C_t^{1-\sigma}}{1-\sigma} - \frac{L_t^{1+\psi}}{1+\psi} \right], \tag{16.1}$$

where $C_t$ denotes consumption and $L_t$ is labor supply and $\beta \in [0,1]$. The household chooses how much to work, how much to save, and how much to consume. The vehicle for savings is a one-period nominal bond that pays interest $i_t$ in period $t+1$ and is in zero net supply.

The household consumes a final good, which is produced by a representative competitive firm out of a continuum of intermediate inputs indexed by $j \in [0,1]$. The final goods producer operates a constant elasticity of substitution production function. Letting $Y_t$ be the quantity of final goods produced, the production function is given as:

$$Y_t = \left( \int_0^1 y_{j,t}^{\frac{\varepsilon-1}{\varepsilon}} dj \right)^{\frac{\varepsilon}{\varepsilon-1}}, \tag{16.2}$$

where $y_{j,t}$ is the quantity of the $j$th intermediate input used and $\varepsilon > 1$ is the elasticity of substitution between varieties. The intermediate inputs are produced according to the production function

$$y_{j,t} = A_t \ell_{j,t}, \tag{16.3}$$

where $\ell_{j,t}$ is the amount of labor used in producing good $j$ and $A_t$ is an aggregate productivity that follows a first-order Markov process. The real marginal cost of producing one unit of intermediate goods is common across firms and equal to $w_t/A_t$, where $w_t$ is the real wage denominated in units of final goods.

To model nominal rigidities we necessarily have to depart from perfect competition. After all, if all firms are price takers, no firm can be said to set its price, and when prices are sticky, firms need to stand ready to satisfy demand at the possible preset price which a competitive firms may not be willing or able to do. We therefore adopt monopolistic competition as a convenient framework in which to model market power. We have seen the production function (16.2) before in Chapter 6. Using the argument that we introduce there, the final goods producer's cost-minimization problem yields a price index

$$P_t = \left( \int_0^1 P_{j,t}^{1-\varepsilon} dj \right)^{1/(1-\varepsilon)},$$

where $P_t$ is the cost of producing one unit of the final good and $P_{j,t}$ is the price of a unit of the $j$th intermediate good.[4] As the final goods producer is competitive, it sells its output at marginal cost, and $P_t$ is therefore also the price of a unit of final goods.

It also follows from the final goods producer's cost-minimization problem that the demand for intermediate good $j$ is given by the following iso-elastic function of the good's relative price:

$$y_{j,t} = \left(\frac{P_{j,t}}{P_t}\right)^{-\varepsilon} Y_t. \tag{16.4}$$

Note that the price elasticity of the demand facing an intermediate goods producer is $\varepsilon$.

Both $P_{j,t}$ and $P_t$ are nominal prices meaning they are set in terms of units of money. Money does not feature in this version of the New Keynesian model except as the unit of account for prices and bonds. Here we are making use of the cashless-limit argument introduced in Chapter 14, in which case households hold zero cash balances yet money can serve as the numeraire. However, extending the model with money demand derived from a money-in-the-utility-function assumption would produce exactly the same equilibrium as long as the marginal rate of substitution between consumption and leisure is independent of real cash balances.[5]

The key assumption that distinguishes New Keynesian models is that they feature nominal rigidities—that is, one or more prices in the economy is sticky. These nominal rigidities can apply to goods prices, to wages, or both; but here we will assume that the prices of intermediate inputs are sticky with the price-setting friction modeled as in Calvo (1983). Specifically, an intermediate goods producer has an i.i.d. probability $1 - \theta \in [0, 1]$ of being able to update its price each period. In the monopolistic competition model, the intermediate goods producers who are allowed to adjust prices in a given period then set prices taking as given the actions of other firms, input prices, and all aggregate variables. Firms serve whatever demand they face given the price they are quoting. Thus, in the short run, output is "demand determined" in the New Keynesian model.

The government is the final participant in this economy. The government sets the nominal interest rate and its choices are often represented through an interest rate rule that specifies the interest rate as an explicit function of macroeconomic variables such as the rate of inflation. We defer making detailed assumptions about the monetary policy rule.

**Decision problems.** The household's decision problem is to maximize

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \left[\frac{C_t^{1-\sigma}}{1-\sigma} - \frac{L_t^{1+\psi}}{1+\psi}\right],$$

subject to the budget constraint

$$B_{t+1} + P_t C_t = P_t w_t L_t + (1 + i_{t-1}) B_t + P_t \int m_{jt} dj \qquad \forall t,$$

[4] Formally, $P_t$ is the derivative of the cost function for the final goods producers with respect to the quantity produced.

[5] Adding money demand would simply determine money supply.

where $B_{t+1}$ is the nominal bonds held from date $t$ to date $t+1$ and $m_{jt}$ is the date-$t$ profit earned by firm $j$. In addition to the budget constraint, the household must also satisfy a no Ponzi game condition of the kind introduced in Chapter 4. It is convenient to rewrite the budget constraint in real terms by dividing both sides by $P_t$

$$b_{t+1} + C_t = w_t L_t + \frac{1+i_{t-1}}{1+\pi_t} b_t + \int m_{jt} dj,$$

where $a_t \equiv B_t/P_{t-1}$ is the real value of savings in period $t-1$ and $1+\pi_{t+1} \equiv P_{t+1}/P_t$ is the gross rate of inflation between periods $t$ and $t+1$.[6]

This decision problem results in two optimality conditions. The Euler equation

$$C_t^{-\sigma} = \beta \mathbb{E}_t \left[ \frac{1+i_t}{1+\pi_{t+1}} C_{t+1}^{-\sigma} \right] \tag{16.5}$$

and the intratemporal labor supply condition

$$C_t^{-\sigma} w_t = L_t^{\psi}. \tag{16.6}$$

The Euler equation in (16.5) implies that the intertemporal path of consumption is determined by the (gross) real interest rate, $(1+i_t)/(1+\pi_{t+1})$. Hence, household's intertemporal choices are no different than in standard flexible price models encountered earlier in this book. Likewise, the labor supply condition in equation (16.6) is also standard and equalizes households' marginal rate of substitution between consumption and work with the real wage. Nominal rigidities affect these conditions only through equilibrium changes in real wages and real interest rates.

An intermediate goods firm chooses how to price its good when it has the opportunity to reset its price. As usual, it is only relative prices that matter so we will define $p_{jt} \equiv P_{jt}/P_t$. Suppose a firm has a relative price $p_{jt}$ at date $t$. If at date $t+1$ the firm does not update its nominal price, its new relative price will be

$$p_{jt+1} = \frac{P_{jt}}{P_{t+1}} = \frac{P_{jt}}{P_t} \frac{P_t}{P_{t+1}} = \frac{p_{jt}}{1+\pi_{t+1}}.$$

If the firm instead does update its price at date $t+1$, it faces the same decision problem as any other firm that is updating its price that period because there are no firm-specific state variables other than the price they set in the past, which is no longer relevant to the firm updating its price. Let $P_{t+1}^R$ be the nominal price set by firms who update their prices in $t+1$, which is known as the **reset price.** Let $p_{t+1}^R = P_{t+1}^R/P_{t+1}$ be the corresponding relative price.

In period $t$, the real profits of firm $j$ are

$$m_{jt} = p_{jt} y_{jt} - w_t \ell_{jt} = (p_{jt} - w_t/A_t) y_{jt} = (p_{jt} - w_t/A_t) p_{jt}^{-\varepsilon} Y_t,$$

[6]Note that in the standard notation used here, $i_t$ applies to savings from $t$ to $t+1$ while $\pi_t$ refers to the inflation between $t-1$ and $t$. Both variables are determined at date $t$.

where the we have used eq. (16.3) to substitute for $\ell_{jt}$ and eq. (16.4) to substitute for $y_{jt}$.

The value function of a firm with relative price $p$ is

$$V(p,\mathcal{S}) = u'(C(\mathcal{S}))\left(p - \frac{w(\mathcal{S})}{A(\mathcal{S})}\right)p^{-\varepsilon}Y(\mathcal{S}) + \beta\mathbb{E}\left[\theta V\left(\frac{p}{1+\pi(\mathcal{S}')},\mathcal{S}'\right) + (1-\theta)V\left(p^R(\mathcal{S}'),\mathcal{S}'\right)\right],$$

where $\mathcal{S}$ is the aggregate state, which evolves independently of an individual firm's decisions because of the monopolistic competition assumption. On the right-hand side of this Bellman equation, the first term is the period payoff, which is the profits earned during the period valued at the marginal utility of consumption of the representative household (the owner of the firm). The second term is the continuation value, which has two components. The first component represents the value if the firm does not update its price next period while the second component is the continuation value if it does update.

A firm that updates its price maximizes $V(p,\mathcal{S})$ therefore finds the price that solves:

$$p^R(\mathcal{S}) = \arg\max_p V(p,\mathcal{S}).$$

As shown in the appendix, the first-order necessary condition for the choice of the optimal reset price can be expressed as:

$$0 = \mathbb{E}_t\sum_{\tau=t}^{\infty}(\beta\theta)^{\tau-t}u'(C_\tau)Y_\tau\left[{p^R_{t,\tau}}^{-\varepsilon} - \varepsilon{p^R_{t,\tau}}^{-\varepsilon-1}\left(p^R_{t,\tau} - \frac{w_\tau}{A_\tau}\right)\right],$$

where $p^R_{t,\tau} \equiv P^R_t/P_\tau = p^R_t/\prod_{s=t+1}^{\tau}(1+\pi_s)$, is the relative price at date $\tau$ of a firm that last updated its price at date $t$. This condition is interesting. First, the term in square brackets times $Y_\tau$ is the marginal change in profit from a change in $p^R_{t,\tau}$. In a flexible price model, this term would equal zero period by period. But, because prices are sticky, firms aim at hitting this condition "on average" which means weighting profits in future periods by the probability $\theta^{\tau-t}$ that price will remain in effect that far in the future (and discounting future marginal profit contributions by the owners' marginal rate of intertemporal substitution). Secondly, firms are more forward looking when prices are stickier. This result is intuitive. In a flexible price setting, firms simply solve static optimization problems that result in equalizing marginal revenue and marginal costs period by period. With sticky prices, they now need to consider how their current choice of price may affect future profits. For example, firms will adjust prices today in response to information about marginal costs in the future.

Rearranging, we obtain

$$p^R_t\mathbb{E}_t\sum_{\tau=t}^{\infty}(\beta\theta)^{\tau-t}u'(C_\tau)Y_\tau\left(\frac{P_\tau}{P_t}\right)^{\varepsilon} = \frac{\varepsilon}{\varepsilon-1}\mathbb{E}_t\sum_{\tau=t}^{\infty}(\beta\theta)^{\tau-t}u'(C_\tau)Y_\tau\left(\frac{P_\tau}{P_t}\right)^{1+\varepsilon}\frac{w_\tau}{A_\tau}.$$

We can then express the condition for the optimal reset price as:

$$p^R_t = \frac{p^N_t}{p^D_t} \tag{16.7}$$

where:

$$\begin{aligned} p_t^N &= \frac{\varepsilon}{\varepsilon - 1} \frac{w_t}{A_t} u'(C_t) Y_t + \theta \beta \mathbb{E}_t (1 + \pi_{t+1})^{1+\varepsilon} p_{t+1}^N \\ p_t^D &= u'(C_t) Y_t + \theta \beta \mathbb{E}_t (1 + \pi_{t+1})^{\varepsilon} p_{t+1}^D. \end{aligned}$$

**Market clearing.** There are three markets at each date: the labor market, the goods market, and the bond market. Labor market clearing requires that the total labor supplied by households equals the total labor used in production

$$L_t = \int \ell_{j,t} dj.$$

As consumption is the only use for final goods, goods market clearing requires that production equals total consumption

$$Y_t = C_t. \tag{16.8}$$

Finally, bond market clearing requires that the net demand for bonds is zero.

**Aggregation.** In the Calvo model, firms are given the opportunity to re-optimize their prices with probability $1 - \theta$ and it is assumed that the process for the arrival rate of this opportunity is memoryless (ie. it follows a Poisson process). Hence, every period, firms are split randomly into a group of firms that cannot adjust their prices and another group that can. Let $\mathcal{J}_t$ be the set of firms that update their prices in period $t$. It follows then that the price level can be expressed as:

$$P_t^{1-\varepsilon} = \int_j P_{j,t}^{1-\varepsilon} dj = \underbrace{\int_{j \notin \mathcal{J}_t} P_{j,t}^{1-\varepsilon} dj}_{\text{non-adjusters}} + \underbrace{\int_{j \in \mathcal{J}_t} P_{j,t}^{1-\varepsilon} dj}_{\text{adjusters}}$$

where each of these two terms are given as:

$$\begin{aligned} \int_{j \notin \mathcal{J}_t} P_{j,t}^{1-\varepsilon} dj &= \theta P_{t-1}^{1-\varepsilon} \\ \int_{j \in \mathcal{J}_t} P_{j,t}^{1-\varepsilon} dj &= (1 - \theta) \left( P_t^R \right)^{1-\varepsilon}. \end{aligned}$$

Here we have used the assumption that firms are chosen randomly to adjust their prices to derive the expression for $\int_{j \notin \mathcal{J}_t} P_{j,t}^{1-\varepsilon} dj$. The expression for $\int_{j \in \mathcal{J}_t} P_{j,t}^{1-\varepsilon} dj$ exploits the fact that all firms that optimize their prices at the same date, choose the same price. It follows from this that:

$$P_t = \left[ \theta P_{t-1}^{1-\varepsilon} + (1 - \theta) \left( P_t^R \right)^{1-\varepsilon} \right]^{1/(1-\varepsilon)}. \tag{16.9}$$

This expression makes it clear that the price level is sticky since it displays inertia. The higher is the probability that the firm cannot adjust prices in any given period, $\theta$, the larger is the backward looking component of the price level. Yet, despite this, inflation is purely

forward looking in this model. To see this, we can also exploit the above result to derive a relationship between inflation and the relative reset price:

$$p_t^R = \frac{\left[\frac{P_t^{1-\varepsilon} - \theta P_{t-1}^{1-\varepsilon}}{1-\theta}\right]^{1/(1-\varepsilon)}}{P_t} = \left[\frac{1-\theta(1+\pi_t)^{\varepsilon-1}}{1-\theta}\right]^{1/(1-\varepsilon)}$$

which we can rearrange to give:

$$1+\pi_t = \left[\frac{1-(1-\theta)\left(p_t^R\right)^{1-\varepsilon}}{\theta}\right]^{1/(\varepsilon-1)}. \tag{16.10}$$

Since firms are purely forward looking when setting $p_t^R$, inflation is purely forward looking, too. As a result, we do not need to keep any record of past prices (or inflation) when determining the current (or future) inflation rates.

Since only a fraction of firms can adjust prices, there can be dispersion of prices across firms. This price dispersion across intermediate input producers is a source of inefficiency as the final goods producers will over-utilize those inputs that have low relative prices and under-utilize those with high relative prices. To see this, integrate equation (16.3) across $j$

$$\int y_{j,t} dj = A_t \int \ell_{j,t} dj = A_t L_t,$$

where the second equality uses the labor market clearing condition. Now substitute (16.4) for $y_{j,t}$ to arrive at

$$Y_t = \frac{A_t}{D_t} L_t, \tag{16.11}$$

where

$$D_t \equiv \int \left(\frac{P_{j,t}}{P_t}\right)^{-\varepsilon} dj \geq 1$$

is a measure of price dispersion. In particular, if $P_{j,t}$ is common across all firms, as would happen in this model if intermediate goods prices were flexible, then there is no price dispersion $P_{j,t} = P_t$ for all $j$ and $D_t = 1$. In this case it follows that $Y_t = A_t L_t$. When there is price dispersion, however, $D_t$ will be greater than one, which reduces the effective productivity of the economy. This loss of efficiency is due to the concave production function for final goods and the misallocation of labor across intermediate varieties. By the logic that the updated prices are randomly selected we have

$$D_t = \int_{j\in\mathcal{J}_t} p_t^{R^{-\varepsilon}} dj + \int_{j\notin\mathcal{J}_t} \left(\frac{P_{j,t-1}}{P_{t-1}}\frac{P_{t-1}}{P_t}\right)^{-\varepsilon} dj = (1-\theta)p_t^{R^{-\varepsilon}} + \theta(1+\pi_t)^{\varepsilon} D_{t-1}. \tag{16.12}$$

**State variables.** There are two state variables of the aggregate economy: the current level of TFP, $A_t$, and the degree of price dispersion $D_{t-1}$. These states evolve according to the exogenous law of motion for $A_t$ and equation (16.12), respectively.

**Equilibrium.** An equilibrium of this economy involves stochastic processes for $C_t$, $L_t$, $Y_t$, $w_t$, $\pi_t$, $i_t$, $A_t$, $D_t$ that satisfy (16.5), (16.6), (16.8), (16.10), (16.11), (16.12), the monetary policy rule, and the exogenous law of motion for $A_t$; in addition, an equilibrium requires a policy rule $p^R$ and a value function $V$ that solve the price-setting problem.

**Flexible-price equilibrium and output gap.** In order to complete the specification of the environment we need a monetary policy rule. An important benchmark for policy in this context is the counterfactual outcome with fully flexible prices. We introduce that now as common monetary policy rules make reference to this benchmark.

The flexible-price economy is a special case where $\theta = 0$ so all prices are updated every period. When an intermediate goods firm sets its price in period $t$, it knows it will for certain be able to adjust its price again in $t+1$ and the only consideration is maximizing profits in period $t$. This static profit maximization problem results in setting a constant markup $\mu \equiv \varepsilon/(\varepsilon - 1)$ over marginal cost as we saw in Chapter 6.[7] So $p_t^R = \mu w_t / A_t$ and the markup is positive as long as $\varepsilon > 1$ which we have imposed earlier. As all firms face the same profit maximization problem when prices are flexible, they all choose to set the same price and the definition of the price index reduces to $P_t^n = P_t^{R,n}$ or $p_t^{R,n} = 1$ where we use the superscript "$n$" to indicate variables in the flexible price, or "natural," equilibrium. It then follows that $w_t^n / A_t = \mu^{-1}$. Moreover, as all firms set the same price, there is no efficiency loss from price dispersion. We can then solve the following system of equations to obtain the equilibrium

$$\begin{aligned} w_t^n &= A_t/\mu \\ Y_t^n &= C_t^n \\ Y_t^n &= A_t L_t^n \\ C_t^{n-\sigma} w_t^n &= (L_t^n)^{\psi}, \end{aligned}$$

where the second, third, and fourth equations are the aggregate resource constraint, the aggregate production function, and the intratemporal labor supply condition. The solution to this system is:

$$Y_t^n = \mu^{-1/(\sigma+\psi)} A_t^{(1+\psi)/(\sigma+\psi)}. \tag{16.13}$$

Let us also define the output gap as

$$x_t \equiv \log Y_t - \log Y_t^n \tag{16.14}$$

i.e. the log of level of output in the sticky-price equilibrium relative to the natural level.

**Interest rate rules.** An interest rate rule describes how monetary policy is set. These can take several forms, but the simplest version is to impose a relationship between nominal interest rates and other variables in the model. For example,

$$i_t = \bar{i} + \phi_\pi(\pi_t - \pi^*) + \phi_x x_t,$$

[7] You can verify this by solving $\max_p \{(p - w_t/A_t)\, p^{-\varepsilon} Y_t\}$.

where $\pi^*$ is the inflation target and $\bar{i} = \beta^{-1}(1+\pi^*) - 1$ is the nominal interest rate that would satisfy the Euler equation (16.5) in a steady state with constant inflation $\pi^*$. The coefficients $\phi_\pi$ and $\phi_x$ determine how strongly nominal interest rates react to deviations of inflation from its target and to the output gap. Interest rate rules are often called **Taylor rules** in light of Taylor's (1993) finding that such a rule with $\phi_\pi = 1.5$ and $\phi_x = 0.5$ provided a good description of the monetary policy choices of the Federal Reserve between 1987 and 1992.[8]

**Monetary policy shocks.** Here we are building the baseline New Keynesian model. More elaborate versions of this model are often used in central banks to analyze potential monetary policy strategies. Researchers and policymakers are therefore often interested in questions of the form "what would happen if we raise interest rates taking the economic conditions as given?" To answer this question, we want to calculate the consequences of a shift in interest rates that is not the result of a shift in the economic fundamentals that appear in the interest rate rule. The solution is to add an exogenous shock to the interest rate rule.

**The three-equation model.** The model we have presented above is often called the **three-equation model** because a first-order approximation of the model around a zero-inflation steady state is summarized by the following three equations. First, we have a log-linearized version of the consumption Euler equation

$$x_t = \mathbb{E}_t x_{t+1} - \frac{1}{\sigma}(\hat{i}_t - \mathbb{E}_t \pi_{t+1} - r_t^n), \tag{16.15}$$

where $x_t$ is the output gap and $r_t^n = -\log\beta + \frac{\sigma(1+\psi)}{\sigma+\psi}(\hat{A}_t - \mathbb{E}_t\hat{A}_{t+1})$ is the real natural rate of interest.[9] This equation represents the demand-side of the economy and is sometimes called the **IS curve** in a reference to older IS-LM models. Notice that output at date $t$ is forward-looking (it depends on $\mathbb{E}_t\hat{Y}_{t+1}$) and is decreasing in the real interest rate where $1/\sigma$ is the elasticity of intertemporal substitution.

Next, as we show in the appendix, we use the first-order condition of the firm's price-setting problem and other equilibrium conditions of the model to arrive at a New Keynesian Phillips curve

$$\pi_t = \kappa x_t + \beta\mathbb{E}_t\left[\pi_{t+1}\right], \tag{16.16}$$

[8] The coefficients in Taylor's specification correspond to annual rates of inflation and annualized interest rates. Interest rates and inflation rates are measured in percentage points per unit of time while the output gap is simply measured in percentage points. Therefore, if you change the length of a unit of time, you need to adjust the coefficients accordingly. For example, in a quarterly model, the equivalent rule would be $\phi_\pi = 1.5$ and $\phi_x = 0.125$.

[9] To derive this equation, define $\hat{i}_t = \log(1+i_t)$. Then take logs of both sides (16.5) noting that in a first-order perturbation solution we assume the shocks are arbitrarily small so we can take logs inside the expectation operator. Finally, we need the definition of the output gap from equations (16.13) and (16.14), which when linearized give $x_t = \hat{Y}_t - \frac{1+\psi}{\sigma+\psi}\hat{A}_t$. Using this to substitute for $\hat{Y}_t$ yields the equation above. Note that $r_t^n$ is equal to the real interest rate that solves the linearized Euler equation when consumption takes its natural level.

where the composite parameter $\kappa \equiv \frac{(1-\theta)(1-\beta\theta)(\sigma+\psi)}{\theta}$ is usually referred to as the slope of the New Keynesian Phillips curve. Here we see that inflation is forward-looking and increasing in the current output gap. The slope of the Phillips curve, $\kappa$, is larger if prices are more flexible (lower $\theta$) or if marginal cost is more sensitive to the level of production.

Lastly, we have an interest rate rule. For our simulations here we will assume

$$\hat{i}_t = \phi_\pi(\pi_t - \pi^*) + \phi_x x_t + \omega_t, \tag{16.17}$$

where $\omega_t$ is an AR(1) monetary policy shock. In addition to these three core equations, we also have exogenous processes for TFP and the monetary policy shock. The three-equation representation of the model, (16.15)-(16.17), is often used as a small-scale model of the economy to explore qualitative features of business cycles and monetary policy.

**Determinacy.** A key consideration for any monetary policy rule is whether it yields a unique equilibrium. Suppose that, for some reason, inflation expectations are elevated despite expectations of a zero output gap. Could that lead to high inflation today? It all depends on the behavior of the nominal interest rate. If the nominal interest rate is unresponsive, high inflation expectations translate to low real interest rates, which through the IS curve lead to a positive output gap. The positive output gap then puts upward pressure on current inflation through the Phillips curve. By this logic, some small expectation of higher inflation in the far future could justify expectations of high inflation (and positive output gaps) all the way back to the present. In the model, nothing pins down the expectations of inflation in the infinitely far future so we indeed have multiple equilibria when nominal interest rates are unresponsive. In order to prevent this outcome, the interest rate rule must respond sufficiently strongly so that the real interest rate *rises* when inflation rises thereby stabilizing the economy. The three-equation model has a unique equilibrium if and only if

$$(1-\beta)\phi_x + \kappa(\phi_\pi - 1) > 0 \tag{16.18}$$

(see the appendix for a derivation). Note that $\phi_\pi > 1$ is sufficient for this condition to hold and necessary if $\phi_x = 0$. This condition is known as the **Taylor principle**.

**The role of nominal demand in determining output.** As we mentioned in the introduction to this chapter, when prices adjust imperfectly, changes in nominal variables can have real effects. Traditionally, this point is often discussed in terms of the money supply—if money is injected, a frictionless model would imply an immediate adjustment of the price level with no effect on any real variables. This outcome is often called the classical dichotomy between real and nominal sides of the economy. With nominal rigidities, this logic breaks down as the increase in nominal demand from the increase in the money supply is not immediately undone by an increase in the price level. Here we will make a similar argument in terms of nominal interest rates. Suppose the government announces a lower path for nominal interest rates going forward (e.g. an expansionary monetary policy shock). As we saw in equation (16.5), the path of aggregate consumption (and therefore aggregate output) is fully

determined by the path of real interest rates. Therefore, if the classical dichotomy were to hold, it would have to be the case that path of expected inflation would immediately fall so as to leave real interest rates unchanged. But if we solve equation (16.16) forward, we see inflation at each date reflects expectations of current and future output gaps, which would be zero under the classical dichotomy. In fact, the model predicts that with lower real interest rates there will be a positive output gap and an increase in inflation that can reinforce the decline in nominal interest rates absent a counter-veiling response of the monetary policy rule.

## 16.4 Monetary policy strategies

The New Keynesian model has an active role for government. After all, a key equation in the model is the monetary policy rule. We now describe what the role for policy is in the New Keynesian model and how this is reflected in some real-world monetary policy strategies.

### 16.4.1 Policy objectives

The equilibrium of the New Keynesian model can be inefficient for two reasons (i) the total level of production can deviate from the efficient level and (ii) labor can be misallocated across intermediate inputs. To see this in more detail, consider the planner's problem

$$\max_{C_t, Y_t, L_t} \frac{C_t^{1-\sigma}}{1-\sigma} - \frac{L_t^{1+\psi}}{1+\psi}$$

subject to

$$C_t = Y_t = A_t L_t.$$

This is a static problem because the planner only has to choose how to allocate labor each period and in the aggregate there is no way to move resources across time. The constraint on the planner uses the fact that when $L_t$ units of labor are used to produce each intermediate good, the quantity of final goods that can be produced is $A_t L_t$. The first-order condition of this problem leads to

$$Y_t^* = A_t^{(1+\psi)/(\sigma+\psi)}, \tag{16.19}$$

where a star denotes the first-best allocation.

To focus on the possibility of an inefficient level of aggregate production, suppose there is no price dispersion so $Y_t = A_t L_t$. Then using (16.19), (16.6), and $C_t = Y_t$ we obtain

$$Y_t = \left(\frac{w_t}{A_t}\right)^{1/(\sigma+\psi)} Y_t^*.$$

In an efficient economy, the real wage would be equal to the marginal product of labor. The equation above shows that if the wage deviates from this efficient level, output will deviate

from the efficient level. When the wage is too low, households will supply too little labor and too little will be produced. Comparing this to the natural level of output (16.13), we see that in the flexible price economy, output is too low because $w_t/A_t = 1/\mu < 1$. So monopoly power is one source of inefficiency. Notice that $w_t/A_t$ is real marginal cost and if firms set prices equal to nominal marginal cost we would have $P_t = (P_t w_t)/A$ or $w_t/A_t = 1$.

Nominal rigidities affect the markups that firms charge and therefore the degree of inefficiency in the economy. To see this, suppose there is a decrease in nominal marginal cost. As some firms do not update their prices, their markups rise. In the aggregate, the increase in average markups is reflected in a lower $w_t/A_t$ ratio and the economy is further from the efficient level of production. Over time, firms will update their prices and return their markups to the desired level $\mu$ and $w_t/A_t$ will return to $1/\mu$.

Now consider the inefficiency due to price dispersion. Using the aggregate production function $Y_t = A_t L_t / D_t$, (16.19), (16.6), and $C_t = Y_t$ we obtain

$$Y_t = D_t^{-\psi/(\sigma+\psi)} \left(\frac{w_t}{A_t}\right)^{1/(\sigma+\psi)} Y_t^*.$$

This equation shows that there is an additional inefficiency, when $D_t \neq 1$. From the concavity of the final goods production function, the efficient use of labor is to produce equal amounts of all the intermediate inputs. The economy will deviate from this outcome if some intermediate goods producers have lower prices than others (see eq. 16.11) . The reason one producer may have a lower price than another is, for example, that one set its price more recently than the other and, in the intervening time, market conditions changed leading to a change in the optimal reset price. As the reset price fluctuates, so too will the inflation rate (see eq. 16.10). One goal for monetary policy then is to stabilize the inflation rate because this minimizes the efficiency loss from relative price dispersion.[10]

Define the *welfare-relevant output gap* as $x_t^w \equiv \log Y_t - \log Y_t^*$. Using equations (16.13) and (16.19) we see that

$$x_t^w = \log Y_t - \log Y_t^n + \log Y_t^n - \log Y_t^* = x_t - \frac{1}{\sigma+\psi} \log \mu.$$

The welfare-relevant output gap differs from the output gap because the natural level of output is distorted by monopoly power.

In the literature that followed Phillips' original contribution, Phillips (1958), it was perceived that there was a stable relationship between (wage) inflation and resource utilization (as measured by unemployment) which therefore presented a trade-off to policymakers. This notion was challenged by Friedman (1968) and Phelps (1967) who posited that the long-run Phillips curve is vertical (at the natural rate of unemployment) as real wages, which should be independent of inflation in the long-run, determine employment. If monetary policy attempts to persistently raise the level of output above the natural level, the results will be

[10]The Calvo model of nominal rigidities can generate large efficiency losses from price dispersion because a firm can be stuck with a very old price that is far out of alignment with the current price level. Models of menu costs in which firms can choose to change their prices subject to a cost tend to generate less price dispersion because firms with very misaligned prices will choose to change their prices.

high inflation and no actual increase in output. For these reasons, a common view is that monetary policy should focus on stabilizing the economy around the flexible-price level of activity even though this may not be the first best. One way of formalizing this view is to give the policymakers another policy tool that can address long-run inefficiencies while leaving monetary policy responsible for responding to aggregate shocks. In the New Keynesian model, a simple extension of the model is to suppose there is lump-sum tax on households that finances a constant production subsidy for intermediate goods producers. In particular, suppose that intermediate goods producers are given a subsidy $\tau^l = 1/\epsilon \in (0,1)$ so that their effective cost of labor is $(1-\tau^l)w_t/A_t$. The production subsidy induces them to produce more and, in the steady-state, the subsidy undoes the distortion from monopoly power.

In summary, in the rest of this chapter we will take the goals for policy to be to (i) bring the aggregate level of production to the natural level and to (ii) minimize the efficiency loss from price dispersion by stabilizing inflation. Locally around a zero-inflation steady state, the welfare of the representative household in the basic New Keynesian model can be expressed as[11]

$$U \approx -\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \left[ \pi_t^2 + \frac{\kappa}{\varepsilon} x_t^2 \right]. \tag{16.20}$$

Welfare decreases with inflation or deflation because it results in price dispersion. Welfare also decreases with a positive or negative output gap because the level of production differs from the optimal level.

### 16.4.2 The divine coincidence

In the baseline New Keynesian model, it is possible to have zero inflation and zero output gaps at all times. The New Keynesian Phillips curve, (16.16), is the key to this argument. From that equation we can see that if there is no output gap at any date, then there will be no inflation today or in any future period. This aspect of the model—that there is no trade off between output gap and inflation stabilization—is known as the **divine coincidence**.

The standard interpretation of the divine coincidence is that the model is somewhat too simple and abstracts from the features of the economy that lead to a meaningful trade-off between inflation stabilization and output stabilization. What might those features be? To break the divine coincidence we need a *time-varying* wedge between the flexible-price level of output and the efficient level of output. Productivity shocks themselves do not create such a wedge because they affect both the flexible-price level of output as well as the efficient level of output in equal proportions. We say the wedge needs to be time-varying because a constant distortion requires a permanent change in the level of activity and is therefore not an issue for which monetary policy is well suited.

An extension of the model, which is often discussed due to its simplicity, involves shocks to the elasticity of substitution between varieties of intermediate goods. The resulting time-varying market power affects $Y_t^n$ but does not change $Y_t^*$ leading to a time-varying gap

[11] Here we use a second order approximation around a steady state in which the the monopoly distortion has been corrected through the labor subsidy $\tau^l = 1/\epsilon$. See Woodford (2003b).

between the socially efficient level of production and the flexible-price level. In the linearized version of the model, these shocks appear as a "cost-push" shock—an exogenous term that is appended to the Phillips curve

$$\pi_t = \kappa x_t + \beta \mathbb{E}_t \left[\pi_{t+1}\right] + \eta_t \tag{16.21}$$

(see Steinsson, 2003, for a derivation). With this added term, it is no longer possible to stabilize inflation and output perfectly because a policy of setting $x_t = 0$ at all dates no longer leads to $\pi_t = 0$. The divine coincidence also breaks if the flexible-price economy does not respond to shocks in the efficient way; for example because there are frictions in the determination of wages (see Blanchard and Galí, 2007).

### 16.4.3 Inflation targets and price level targets

Around the world, most central banks follow a version of a monetary policy strategy called **flexible inflation targeting**. This strategy can be summarized as minimizing deviations of inflation from a target level while also minimizing deviations of output from its natural level. These goals can be formalized in the objective function

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \left[(\pi_t - \pi^*)^2 + \lambda x_t^2\right], \tag{16.22}$$

where $\pi^*$ is the inflation target and $\lambda$ is a parameter that determines the relative weight placed on the two goals. The policymaker then seeks to minimize this loss function, which closely resembles eq. (16.20).

When inflation is unpredictable, it is risky to agree to a long-term nominal contract because the real values stipulated by the contract are unpredictable. This rationale would argue that it is more valuable for monetary policy to minimize the uncertainty over the price level than to target an inflation rate. Under price-level targeting, the central bank seeks to return the price level to a specific path. For example, consider the interest rate rule

$$i_t = \bar{i} + \phi_P(P_t - P_t^*) + \phi_x x_t$$

where $P_t^*$ is the price-level target, which could be the price level in some base year scaled by a constant annual inflation rate. This rule dictates that the central bank raises interest rates whenever the price level is above target, which would, all else equal, put downward pressure on inflation and move the economy back toward the target. In addition to providing more certainty about the price level, price-level targeting is appealing because it makes clear that future policy will undo unwanted changes in the price level—something that is also useful under inflation targeting as we will see next.

### 16.4.4 Expectations, commitment, and time consistency

In the New Keynesian model, the private sector is forward looking: current inflation depends on expectations of future inflation and current demand depends on expectations of future

real interest rates. Therefore, expectations of what monetary policymakers will do in the future affect the macroeconomic outcomes today. If the private sector expects interest rates to be high in the future *ceteris paribus*, output and inflation will be lower today.

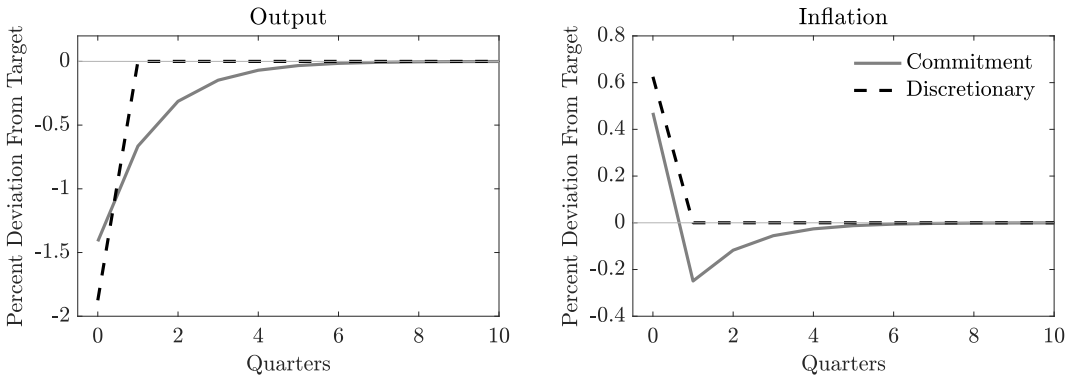



Figure 16.1: Response of output and inflation to a transitory cost push shock under commitment and discretionary policies.

**Notes**: These paths are simulated using the parameters $\beta = 0.99$, $\kappa = 0.2$, $\varepsilon = 3$.

Suppose a central bank pursues the inflation targeting objective (16.22) and the economy is hit by a cost-push shock at date 0. That is, we consider the Phillips curve (16.21) with $\eta_t > 0$ for just one period and zero thereafter. Now consider two possible monetary policy responses. In the first, the central bank raises interest rates at date 0 and then returns them to their long-run level. We will call this the "discretionary" policy for reasons that will become clear. In the second case, the central bank raises interest rates by less, but only gradually normalizes them. We will call this the commitment policy. Figure 16.1 plots the two cases. Under the discretionary policy, inflation rises at date 0 and then returns to zero. The linearized version of the model with the discretionary policy has no endogenous state variables so once the cost-push shock dissipates, the economy immediately returns to steady state.[12] The Phillips curve at date 0 is

$$\pi_0 = \kappa x_0 + \eta_0,$$

where we have used $\pi_1 = 0$ as the economy returns to steady state in the next period. In this case, the policymaker can only use $x_0$ to lean against the cost-push shock. As we see in the figure, output is reduced so as to dampen the inflationary effects of the shock.

Under the commitment policy, inflation is positive at date 0 and then negative for several future periods. The Phillips curve at date 0 is

$$\pi_0 = \kappa x_0 + \beta \pi_1 + \eta_0,$$

[12]In the non-linear model, the degree of price dispersion is an endogenous state variable.

where now $\pi_1$ appears. Under this policy, the central bank reduces $\pi_1$ in order to reduce $\pi_0$ with less impact on $x_0$. As we see in the figure, there is less inflation at date 0 but also a smaller decline in output at date 0. In order to reduce $\pi_1$, the central bank keeps interest rates persistently higher than normal so that output and inflation are persistently lower.

While the outcomes from the commitment policy are better than those from the discretionary policy at date 0, they are worse at future dates. Under the discretionary policy, there are no deviations in output or inflation at any date $t \geq 1$ while, under the commitment policy, output and inflation are too low relative to the targets. This brings us to a time-consistency problem. At date 0, the central bank would like to announce the commitment policy, but it would like to switch to the discretionary policy at date 1. If the private sector anticipates this switch in policy, then the benefits of the commitment policy at date 0 are unattainable because the central bank cannot convince the private sector that $\pi_1$ will be negative. The central bank can only achieve the better outcomes if it is able to commit at date 0 to have tight monetary policy at future dates even though it will not want to do that when those future dates arrive.

Finally, under the discretionary policy, the price-level jumps at date 0 and then remains constant thereafter. Under the commitment policy, the price level rises, but then falls subsequently as inflation is negative for several periods. In fact, if you accumulate the inflation rates in the figure, you find that the long-run price level is unaffected. This result demonstrates that long-run price stability is useful for a central bank even if it is pursuing an inflation targeting framework because it serves to stabilize inflation expectations and therefore helps stabilize inflation rates.

## 16.5 Aggregate evidence of nominal rigidity

We now review the aggregate evidence on price rigidity and the response of real variables to nominal shocks in the form of changes in the nominal interest rate targeted by monetary policy. As we will see, the baseline model is qualitatively consistent with this evidence—in contrast to a flexible-price model—but also has shortcomings.

### 16.5.1 The macroeconomic effects of monetary policy shocks

One of the most studied issues in macroeconomics is the dynamic effects of monetary policy shocks, which lead to exogenous changes in nominal interest rates. This topic attracts so much interest because it speaks to the relevance of the nominal rigidities that underlie the Keynesian perspective. In most models with flexible prices, nominal variables including nominal interest rates have no bearing on real outcomes (e.g. see equation (16.13)). Thus, the impact of changes in nominal interest rates on real variables provides information about the importance of nominal rigidities. Moreover, monetary policy plays a central role in modern macroeconomic policy and understanding the consequences of a change in interest rates is a crucial ingredient to real-world policy decisions.

The main challenge with assessing the effects of a change in interest rates is that monetary policy changes *in response* to developments in the economy. If we simply look at the correlations between variables, we will tend to find that higher nominal interest rates are associated with higher levels of inflation, but this may simply reflect endogeneity of monetary policy decisions (because central banks tend to increase interest rates when inflation rises). We are instead interested in the *causal* effect of interest rates on the economy. That is, we ask what would happen if interest rates increased for reasons unrelated to state of the economy? To answer this question, we need to empirically identify monetary policy shocks as opposed to systematic movements in interest rates in response to economic conditions.

Researchers typically identify monetary policy shocks using information that allows them to estimate the systematic or predictable component of interest rates. Removing this systematic component from the actual change in interest rates yields a residual that is interpreted as a monetary policy shock. One strand of literature, which was pioneered by Kuttner (2001) and Gürkaynak, Sack, and Swanson (2005), is premised on (a) the fact that monetary policy decisions are announced at particular times known to researchers and (b) in a narrow time window around the announcement, changes in interest rates will be dominated by monetary policy as opposed to other economic news. Using financial data, we can make a forecast of the monetary policy decision just minutes before it is announced. This forecast reflects market expectations of how monetary policy will be conducted given current economic conditions—i.e. it is the market's view of the systematic policy response. If the announced decision differs from the forecast, it is due to a deviation of monetary policy from it's usual practice (as judged by financial markets).

Another strand of literature, starting with Romer and Romer (2004), uses central bank forecasts of inflation and other variables to estimate the endogenous component of policy. The motivation for this approach is that central banks often set policy based on their assessment of the economic outlook as reflected in economic forecasts. For example, if the forecast for inflation is elevated or unemployment is expected to be low, policymakers will raise interest rates. By regressing interest rate changes on the central bank forecasts, this approach estimates the systematic component of policy and the residuals from this regression can be interpreted as movements in interest rates that are not due to changes in the economic outlook.

Figure 16.2 plots impulse responses for output, inflation, and nominal interest rates following a monetary policy shock identified using the Romer and Romer method.[13] The left panel of the figure shows that contractionary monetary policy shocks lead to persistently higher nominal interest rates. The center panel of the figure shows that aggregate output declines. The right panel shows that inflation declines so real interest rates rise more than nominal interest rates.[14] So we find that a nominal shock affects real variables.

[13]We use the implementation of this method by Wieland and Yang (2016) and regress the outcome of interest on the estimated monetary policy shocks and the lags of macroeconomic variables.

[14]The increase in inflation at very short horizons is known as the "price puzzle" and is a fairly common feature of empirical estimates of the effects of monetary policy shocks. It likely reflects reverse causation from inflation to interest rates that is not removed by the identification strategy. In some specifications, adding further control variables eliminates the price puzzle (see Sims, 1992).

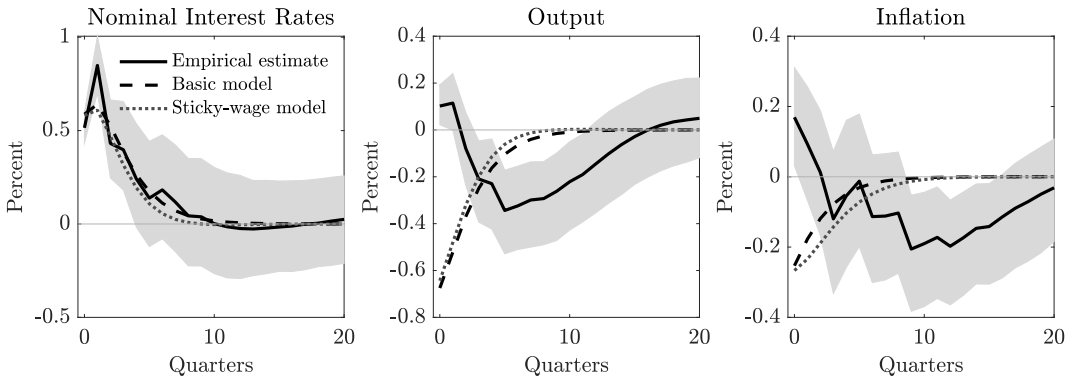



Figure 16.2: Empirical and simulated responses to a monetary policy shock.

**Notes**: The shaded areas around the empirical estimates are 90% confidence bands.

The basic New Keynesian model presented above is qualitatively consistent with the patterns revealed by the empirical estimates: higher nominal interest rates generate increases in real interest rates and reductions in aggregate demand and therefore a decline in output. With lower output, there is lower resource utilization leading to lower marginal costs and a decline in inflation. The timing of the responses, however, is quite different. Figure 16.2 shows model-simulated responses of interest rates, output, and inflation to a monetary policy shock (for now, just focus on the basic model and we will return to the sticky-wage model below). We chose the magnitude of the simulated shock and the monetary policy rule to roughly match the path of nominal interest rates that we estimated. In addition, we chose the other parameters of the model to roughly match the magnitudes of the responses of output and inflation.[15] As shown in the figure, the model-generated responses of output and inflation immediately fall in response to the monetary shock and then gradually return to steady state whereas the empirical responses feature delayed responses of output and inflation.

### 16.5.2 Price rigidity in the aggregate

At the start of this chapter we discussed evidence on the frequency of price changes in micro data. But does this micro-level rigidity translate to rigidity in the aggregate?

One way to answer this question is to rely on a general equilibrium model as discussed above and ask which values of structural parameters allow one best to "match" the data? One might, for example, study the estimated impulse response functions in Figure 16.2 and estimate the slope of the Phillips curve by comparing the magnitude of the inflation response

[15] Here we use equations (16.15) and (16.16) and a policy rule $\hat{i}_t = \gamma \hat{i}_{t-1} + (1-\gamma)(\phi_\pi \pi_t + \phi_x x_t) + \omega_t$ and an AR(1) specification for $\omega_t$ to mimic the estimated path of interest rates. We use parameter values $\gamma = 0.65$, $\phi_\pi = 1.5$, $\phi_x = 0.125$, $\sigma = 5$, $\psi = 1$, $\theta = 1 - \frac{1}{8}$, $\beta = 0.99$, and an autocorrelation of 0.5 for $\omega_t$.

to the magnitude of the output response. One implementation of this idea was carried out by Christiano, Eichenbaum, and Evans (2005). These authors construct a rich New Keynesian model and fit its structural parameters to match the impulse response functions following a monetary policy shock. Their model includes many extensions of the basic New Keynesian framework some of which we describe in the next section. Overall, they find that the data indicate a value of $\theta$ that is consistent with an average contract length of 2.5 quarters which is in the range of values suggested by the literature on the frequency of price changes discussed earlier. Yet, at the same time, they also find that rigidities in wage setting are crucial to explain the dynamics of inflation following a monetary policy shock.

An alternative approach is to estimate the New Keynesian Phillips curve directly in isolation from other parts of the model. Here, the question is how strongly changes in the output gap (or marginal costs of production) translate to changes in prices. Consider the Phillips curve

$$\pi_t = \kappa x_t + \beta \mathbb{E}_t \left[\pi_{t+1}\right] + \eta_t$$

where we have included the shock $\eta_t$ as we would in general not expect the Phillips curve to hold exactly as an empirical relationship. Our interest is in estimating $\kappa = (1-\theta)(1-\theta\beta)(\sigma+\psi)/\theta$ as this speaks to the strength of nominal rigidities.

The direct estimation of the Phillips curve requires one to address several challenges. Any measure of the output gap is bound to be associated with measurement error which induces an attenuation bias in our estimate of $\kappa$ if we were simply to apply ordinary least squares. However, if we have an instrumental variable that is correlated with the true output gap and uncorrelated with the measurement error we can sidestep this source of bias. Inflation expectations are not directly observable either. To measure inflation expectations, the typical practice is to use statistical techniques to construct a forecast of $\pi_{t+1}$ on the basis of information that is available at date $t$. Finally, the presence of the cost-push shock complicates matters. Suppose there is an inflationary cost push shock, $\eta_t > 0$, and that the central bank responds to this by restraining demand and inducing a negative output gap. In this scenario, there is reverse causation from inflation to the output gap.

The literature has adopted a variety of methods to overcome these challenges and it remains an active area of research. Galí and Gertler (1999) is representative of the approach that has been followed by many papers in the literature. This approach rewrites the Phillips curve as

$$\pi_t = \kappa x_t + \beta \pi_{t+1} + \underbrace{\eta_t - \beta\left(\pi_{t+1} - \mathbb{E}_t\left[\pi_{t+1}\right]\right)}_{\equiv \zeta_t},$$

where we have replaced expected inflation with the realization of $\pi_{t+1}$ and included the expectational error $\pi_{t+1} - \mathbb{E}_t[\pi_{t+1}]$ in the error term, which we now denote $\zeta_t$. If we assume $\mathbb{E}_{t-1}[\eta_t] = 0$, any lagged variable known at $t-1$ that predicts $x_t$ or $\pi_{t+1}$ can be a valid instrument.[16] This (strong) assumption allows us to use realized future inflation in place of

[16] Let $z_{t-1}$ be the instrument. The orthogonality condition requires that this instrument be uncorrelated

expected inflation thereby obviating the need for measurement of inflation expectations and also prevents reverse causation as the lagged instruments are not correlated with $\eta_t$.

To measure the output gap, Galí and Gertler make use of the micro-foundations of the New Keynesian model, which say it is real marginal cost that is relevant in price setting. They start by positing a Cobb-Douglas aggregate production function:

$$Y_t = A_t K_t^{\alpha_K} L_t^{\alpha_L}$$

where $K_t$ is the input of capital and $\alpha_K$ and $\alpha_L$ are the output elasticities with respect to the two factor inputs. Assuming that the labor input is flexible while the capital stock is predetermined in period $t$, cost minimization (together with firms being price takers in the input markets) implies that real marginal costs are given as:

$$mc_t = \frac{w_t}{\partial Y_t / \partial L_t} = \frac{s_t^L}{\alpha_L}$$

where $s_t^L = (w_t L_t)/Y_t$ is the labor share of income and $w_t$ is the real wage. Thus, up to a first-order approximation, the log of real marginal costs are given as the log of the labor share.

Galí and Gertler estimate the Phillips curve using the generalized method of moments applied to quarterly U.S. data from 1960 to 1997. Their estimates imply an estimate of the frequency of price adjustment, $1 - \theta$, in the range of 0.085-0.171 per quarter, which indicates much longer average price contract length than the estimates from disaggregated prices discussed earlier.

The New Keynesian Phillips curve implies inflation is purely forward-looking. Before the development of the New Keynesian model, typical specifications of the Phillips curve instead included lagged inflation rather than inflation expectations. This so-called "accelerationist" Phillips curve implies a high degree of inflation persistence. Forward- and backward-looking Phillips curves have very different implications for how inflation can be controlled. In the forward-looking case, a credible central bank can immediately reduce inflation by committing to a low long-run inflation rate and zero output gaps. In the latter case, inflation can only be reduced by imposing negative output gaps on the economy. To explore the issue of inflation persistence, Galí and Gertler also considered a "hybrid" New Keynesian Phillips curve in which they assume that a certain fraction of price setters are purely backward looking and simply set prices by updating the past average reset price (of forward-looking firms) with the past inflation rate. In this extension, they estimate the fraction of firms that are backward looking as opposed to forward looking. Their estimates place much more weight on forward-looking behavior than on backward-looking behavior.

---

with $\zeta_t$, which we can verify as follows

$$\operatorname{cov}(z_{t-1}, \zeta_t) = \mathbb{E}\left[z_{t-1}\zeta_t\right] = \mathbb{E}\left[z_{t-1}\left\{\eta_t - \beta\left(\pi_{t+1} - \mathbb{E}_t\left[\pi_{t+1}\right]\right)\right\}\right] = \mathbb{E}\left[z_{t-1}\eta_t\right] + \beta\mathbb{E}\left[z_{t-1}\left(\pi_{t+1} - \mathbb{E}_t\left[\pi_{t+1}\right]\right)\right] = 0,$$

where $\mathbb{E}\left[z_{t-1}\eta_t\right] = \mathbb{E}\left[z_{t-1}\mathbb{E}_{t-1}\left[\eta_t\right]\right]$ by the law of iterated expectations and $\mathbb{E}_{t-1}\left[\eta_t\right] = 0$ by assumption. Similarly $\mathbb{E}\left[z_{t-1}\left(\pi_{t+1} - \mathbb{E}_t\left[\pi_{t+1}\right]\right)\right] = \mathbb{E}\left[z_{t-1}\left(E_t\left[\pi_{t+1}\right] - \mathbb{E}_t\left[\pi_{t+1}\right]\right)\right] = 0$ by the law of iterated expectations.

The work of Galí and Gertler (1999) has been influential because it establishes a link between inflation and marginal costs as consistent with the price setting condition embedded in the New Keynesian model. However, there remains considerable uncertainty surrounding the Phillips curve parameters as changes in the data series, the sample period, or the econometric specification can result in considerable changes in the parameter estimates, see (see Mavroeidis, Plagborg-Møller, and Stock, 2014). One issue of note is that the correlation between the labor share and inflation appears to have weakened over time. In Figure 16.3 we show the dynamic correlations between the labor share and leads and lags of the inflation rate for U.S. quarterly data, 1960:1-2019:4. We compute cross correlations for both the whole sample and for an early sample (ending in 1997) and a late sample (starting in 1998). Consistent with the results of Galí and Gertler (1999),[17] there is a significant positive relationship between the labor share and inflation in the early sample period. It is this correlation that the estimates of the Phillips curve pick up. In the last part of the sample, however, the sign of the contemporaneous correlation is negative.

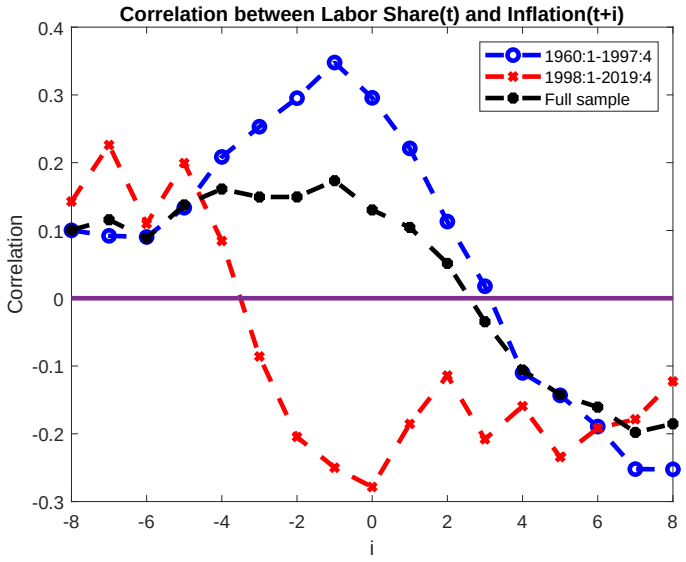



Figure 16.3: Dynamic Correlations

**Notes**: The figure shows the dynamic correlation function between the labor share and lags and leads of the inflation rate. The data have been HP-filtered.

One possible explanation for the falling correlation of inflation and measures of marginal cost is that a smaller share of the variation in inflation now comes from changes in marginal costs (i.e. movements along the Phillips curve) as opposed to other factors (i.e. shifts of the Phillips curve). In this case, estimating $\kappa$ by concentrating on the economy's response to structural shocks that move the economy along the Phillips curve, for example as pursued

[17] Galí and Gertler (1999) measure inflation on the basis of the GDP deflator while we use the PCE deflator.

by Christiano et al. (2005), may be a preferable approach.[18]

## 16.6 Sticky wages and other extensions

The analysis in this chapter has focused on a simple version of the New Keynesian model with sticky prices. This model is useful for illustrating qualitative implications of nominal rigidities but is not a compelling quantitative explanation of the business cycle as we have seen in Figure 16.2. We now describe some of the main extensions of the New Keynesian framework that allow for a richer description of aggregate dynamics. We will pay particular attention to frictions in wage setting as these are a key source of nominal rigidity in addition to price stickiness.

### 16.6.1 Sticky Wages

The New Keynesian model that we have examined so far includes nominal rigidities in goods prices. This focus on the goods market contrasts with the traditional Keynesian literature that, motivated by the high levels of unemployment observed during the Great Depression, focused on sticky nominal wages.[19] The lack of instantaneous adjustment of nominal wages is consistent with the empirical fact that the distribution of wage changes observed in micro data tend to have a spike at zero, see e.g. Kahn (1997). Sticky wages result in periods when labor is under-utilized thereby providing an explanation for labor market "slack" in the sense that there are workers, although willing to work and actively searching for a job, who appear unable to find a job at the going wage. Sticky wages also hold some theoretical appeal as the basic sticky-price New Keynesian makes counterfactual predictions for the cyclicality of firm profits.[20]

The Taylor model of overlapping contracts has considerable empirical and intuitive appeal for many parts of the labor market as many wages are adjusted at an annual frequency (see Grigsby et al., 2021). However, as we have argued earlier, the Calvo model offers much analytical convenience. As mentioned above, there may be asymmetries in nominal wage flexibility in terms of the wage change distribution being asymmetric with few downwards movements in wages (of continuing employee-employer relationships). Here, to keep the analysis simple, we will maintain the symmetric modeling of wage stickiness. Wage stickiness may also be an insurance mechanism that can arise in the absence of nominal rigidities. Firms may, for example, be prepared to offer workers insurance against variations in wages because they are better able and willing to absorb profit fluctuations than workers are at handling

[18] Christiano et al. specify an entire general equilibrium model. An alternative approach is to estimate the Phillips curve in isolation from the rest of the model using the information contained in impulse response functions following identified shocks. See Barnichon and Mesters (2020) and Galí and Gambetti (2020).

[19] A classic reference is Keynes (1936).

[20] The sticky-price model predicts that firm profits decline following a positive demand shock. The resulting negative wealth effect on labor supply is actually central to the increase in labor supply with balanced-growth preferences, see Broer, Hansen, Krusell, and Öberg (2020).

income variations. Such considerations will typically lead to sticky real wages rather than sticky nominal wages and therefore falls somewhat outside the aims of this chapter.

In the sticky-price model we introduced monopolistic competition in the goods market in order to reconcile the assumption of output being demand determined (in the short run) with firms' willingness to supply goods at a possible predetermined price. A similar assumption is required when modeling sticky wages. The standard avenue taken in the literature is that households rent their labor supply to a continuum of labor unions that differentiate labor and rent it to firms setting a nominal wage that is above the cost that they have acquired it for. Households may therefore sometimes be unable to work the number of hours they would have chosen to in the absence of labor unions and nominal rigidities. However, households are assumed to be the ultimate owners of labor unions and are therefore compensated for this by profits received from these institutions.

In this setup, if, for the sake of discussion, we entertain the idea that wages are sticky and adjusted as in the Calvo model while prices are flexible, the model would be almost identical. In the sticky-price model, the marginal cost of production is the wage (adjusted for productivity) and the wage equals the marginal rate of substitution between consumption and leisure because the household is on its labor supply curve (eq. 16.6). In this setup there is a markup between the marginal cost and the price and due to nominal rigidities this markup can vary in response to shocks. We will now sketch out a sticky-wage model in which prices are equal to marginal costs (production is competitive) but there are frictions in the labor market that introduce a wedge between wages and the marginal rate of substitution between consumption and leisure. This approach to modeling sticky-wages has its roots in Blanchard and Kiyotaki (1987) and was developed in more detail by Erceg, Henderson, and Levin (2000).

The union sets its wage subject to Calvo-style nominal rigidities and stands ready to supply any amount of labor that is demanded at that wage. A competitive final goods producer combines these varieties of labor to produce a final good selling its output at marginal cost. Let $W_t$ be the nominal wage-index that reflects the cost of a bundle of labor that allows the final goods producer to produce one more unit of goods. Since goods are sold at marginal cost, we have $P_t = W_t$ in all periods. It then follows that price inflation will be identical to wage inflation. This model yields the exact same Phillips curve as the sticky-price model. Where the model will differ is its implications for real wages, which are now constant as $w_t = W_t/P_t = 1$. In Section 16.4.1, we argued that the distance between the equilibrium level of output and the efficient level depends on the ratio $w_t/A_t$. Such a condition no longer holds. Instead, the relevant issue is how the marginal rate of substitution compares to productivity. In the sticky-wage model, households need not be on their labor supply curves and there can be times when they would like to work more, but wages are high and labor demand is therefore "too low," in the sense of being below its efficient (flexible wage) level.

A model with ***both*** sticky wages and sticky prices is, however, somewhat different from the models with only one nominal rigidity. We describe this model in detail in Appendix 16.A.3. As we show there, inflation is now explained by three equations that replace the

Phillips curve

$$\pi_t = \beta \mathbb{E}_t \left[\pi_{t+1}\right] + \xi^p \left(\hat{w}_t - \hat{A}_t\right) \tag{16.23}$$

$$\pi_t^w = \beta \mathbb{E}_t \left[\pi_{t+1}^w\right] - \xi^w \left(\hat{w}_t - \hat{A}_t\right) + \kappa^w x_t \tag{16.24}$$

$$\hat{w}_t = \hat{w}_{t-1} + \pi_t^w - \pi_t. \tag{16.25}$$

Eq. (16.23) is the price Phillips curve. It is similar to the standard New Keynesian Phillips curve but now depends on the real wage $w_t$ rather than the output gap. Goods price setters choose their prices taking account of current and future marginal costs, which in this model is just the real wage relative to productivity.[21] Eq. (16.24) is the wage Phillips curve where $\pi_t^w$ is wage inflation (i.e. the growth rate of nominal wages). Wage setters will increase nominal wages if the marginal rate of substitution is high relative to real wages. Therefore eq. (16.24) is increasing in the output gap (the marginal rate of substitution rises as households work and consume more) and is decreasing in the real wage.[22] This equation is forward looking for the same reason that the price Phillips curve is. Wage setters know their wage could remain fixed for a number of periods so the look ahead to future market conditions. Finally, eq. (16.25) is the log of the identity $w_t = \frac{W_t}{P_t} = \frac{W_{t-1}}{P_{t-1}} \frac{W_t}{W_{t-1}} \frac{P_{t-1}}{P_t}$. This equation relates the change in the real wage between any two periods to the difference between nominal wage inflation and nominal price inflation.

Figure 16.2 shows results of simulating the model with both sticky wages and sticky prices. There are two things worth noting here. First, we parameterize this model with double the frequency of price and wage adjustments as in the basic model. In the sticky-wage model, prices and wages update once per year on average while in our calibration of the sticky-price model they updated every two years on average. When nominal rigidities layer on top of each other, the pass through from resource utilization to inflation becomes more gradual. Even when they are able to update their prices, intermediate goods firms only raise their prices to the extent their marginal costs rise and the change in their marginal costs is muted by the wage rigidities. Second, note that inflation is more persistent with the two rigidities. Wage inflation initially falls by more than price inflation leading real wages to fall. Thereafter, the low real wages exert a downward force on price inflation and the real wage only returns to its steady state value gradually.

### 16.6.2 Other extensions of the basic New Keynesian model

We showed above that the basic New Keynesian model is qualitatively consistent with empirical estimates of the impact of monetary policy shocks. However, as also made clear, quantitatively, the model does not manage to match the data. The same is the case for other structural shocks often studied in macroeconomics such as total factor productivity

[21] If the model included decreasing returns in production or factors of production other than labor, there could also be an output gap term in eq. (16.23) in addition to the wage term.

[22] Eq. (16.24) is also increasing in productivity because there is a wealth effect on labor supply that raises the marginal rate of substitution between leisure and consumption.

shocks, shocks to investment efficiency, fiscal shocks, uncertainty shocks, etc. Clearly this basic model fails to capture some important features of business cycle fluctuations. For that reason, much work in the area has considered so-called "medium-scale" models that extend the above framework with the hope of improving the model's quantitative performance. Here we will discuss a few of these extensions and the underlying reason for their introduction into this line of work.

**Consumption dynamics**: A main feature of many empirical estimates of the macroeconomic impact of aggregate shocks is gradual adjustment over time. We see this above in Figure 16.2 in the hump-shaped response of output to the identified monetary policy shock, but such dynamics are standard findings in the literature also in response to other shocks. There are many ways in which macroeconomists have attempted to model such dynamics.

Under the permanent-income hypothesis, consumption is determined by permanent income and the path of interest rates. It is impossible then to explain a gradual change in consumption without a very particular path for interest rates. To generate a gradual consumption response, some authors replace the preferences in equation (16.1) with a specification such as

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \left[ \frac{(C_t - \gamma C_{t-1})^{1-\sigma}}{1-\sigma} - \frac{L_t^{1+\psi}}{1+\psi} \right], \tag{16.26}$$

where $\gamma \in [0, 1)$ and $C_{t-1}$ refers to last period's consumption. These preferences are interpreted as reflecting consumption habits. To see why, note that the term involving consumption can be rewritten as

$$C_t - \gamma C_{t-1} = (1-\gamma) C_t + \gamma (C_t - C_{t-1})$$

i.e. as a weighted average of the level and the change in consumption. Thus, in habit formation models, households are concerned about smoothing both the level and the growth rate of consumption. When $\gamma$ is large, households effectively worry about smooth growth rates of consumption and in this case the level of consumption will tend to adjust partially to shocks over time. [23]

An alternative way of generating richer dynamics for aggregate consumption is to abandon the complete markets assumptions underlying the permanent income hypothesis. When households face uninsurable idiosyncratic risks and borrowing constraints, their consumption choices will tend to reflect current labor market conditions. For example, borrowing-constrained agents will spend strongly out of current income so aggregate consumption will respond more strongly to aggregate disposable income. Similarly, an increase in the risk households face, say an increase in the risk of unemployment, will lead them to cut back on consumption for precautionary reasons. These issues are explored in the literature on heterogeneous-agent New Keynesian models.

[23] This specification is usually referred to as an an internal habit model because the household understands that increasing $C_t$ will affect its utility next period. Another approach treats the habit term as referring to past *aggregate* consumption (taken as given). That approach is called an "external habit" or "catching up with Joneses."

**Capital, investment, and adjustment costs**: The model discussed so far has only one factor input: labor. Quantitative models typically also include capital accumulation both because capital is important as a savings vehicle and also because investment demand is an important part of aggregate demand. Typically, one assumes a Cobb-Douglas technology:

$$y_{j,t} = A_t k_{j,t}^{\alpha} \ell_{j,t}^{1-\alpha}. \tag{16.27}$$

Let's assume that capital is owned by households and rented out to firms at the (real) rental rate $r_t^k$. Now assume that capital accumulates over time according to a standard neoclassical specification:

$$K_{t+1} = (1-\delta)K_t + I_t,$$

Unfortunately, this model implies that investment demand becomes extremely sensitive to variations in monetary policy. To see this, note that firms will continuously adjust their capital demand so as to equate the marginal product of capital with the cost of capital $r_t^k + \delta$. In steady state, investment satisfies $I_t = \delta K_t$ and with a small $\delta$, the flow of investment is small relative to the capital stock. Small percentage changes in the desired capital stock then translate to large percentage changes in investment. The model as written would then imply that small changes in interest rates lead to large changes in investment and equilibrium output.

To limit such high investment volatility, it is common in medium-scale models to include sources of adjustment costs. A common specification is:

$$K_{t+1} = (1-\delta)K_t + \xi\left(\frac{I_t}{K_t}\right)K_t$$

where the function $\xi(x_t/k_t)$ is assumed to be increasing but concave and captures adjustment costs.[24] Notice here that, because of concavity, it is costly to vary the investment rate. This specification is therefore able to generate gradual adjustments in the flow of investment and hump-shaped aggregate dynamics.

Empirically, the slope of the Phillips curve is low relative to what we would expect based on the rigidity of prices as measured in micro-data as we discussed in Section 16.5.2. This tension is exacerbated in the New Keynesian model with capital accumulation. In the short-run, the capital stock is predetermined and the only way to produce more in the aggregate is to use more labor. If we use the production function (16.27), the elasticity of output with respect to labor is $1-\alpha$, which means we need to use $1/(1-\alpha) > 1$ units of labor in order to produce one more unit of output. Thus, relative to the specification in (16.3), marginal costs are more sensitive to the quantity produced and the Phillips curve becomes steeper. One way to address this concern is to allow for variable capacity utilization whereby the effective capital services are not predetermined giving the economy an additional opportunity to adjust production beyond changes in labor effort.

[24] Typically one assumes $\xi > 0, \xi' > 0, \xi'' < 0$, and that $\xi(\delta) = \delta$ so that Tobin's Q equals 1 in the deterministic steady state.

# Chapter 17

# Frictional credit markets

*Vincenzo Quadrini*

# Chapter 18

# Frictional labor markets

*Toshihiko Mukoyama and Ayşegül Şahin*[1]

[1]We would like to thank Jifan Wang and Jinxin Wu for their excellent research assistance.

## 18.1 Introduction

Early real business cycle models have assumed a frictionless labor market. In a frictionless labor market, all firms can find workers, and all workers can find jobs with the equilibrium wage that equates labor demand and labor supply. The change in aggregate employment reflects shifts in either the labor demand curve or the labor supply curve.

These models ignore unemployment, that is, the phenomenon that some workers who want to work and look for jobs cannot find jobs. The unemployment rate, defined as the fraction of unemployed workers in the labor force, is an important indicator of the business cycle. The unemployment rate tends to increase when the macroeconomy is in a recession. The elevated unemployment rate is often regarded as one of the most important social costs of recessions. Various government policies, such as unemployment insurance and job training, have been implemented to reduce unemployment and address issues arising from unemployment. To analyze these policies, we need to develop a formal theoretical framework where unemployment arises endogenously.

This chapter introduces labor market frictions in macroeconomic models to analyze unemployment. There are many theories of unemployment in macroeconomics. One simple theory is that there are frictions in wage adjustment. If the wages are too high compared to the level that clears the market, the quantity supplied can exceed the quantity demanded in the labor market. The excess supply of labor can be interpreted as unemployment. Wages can be too high, for example, because of institutional reasons such as minimum wages or unions, or there can be economic reasons. The theory of efficiency wages, for example, postulates that employers want to keep the wage high so that they can induce the workers to exert a high level of effort at work.

In this section, we focus on unemployment arising from search frictions. In the models with search frictions, it takes time, effort, and resources for workers and firms to match with each other. The model describes the element of reality that it takes time for a worker to find a job that is sufficiently good for them, and it takes time for a firm to find a worker they think can perform the task that the job requires. In principle, these frictions can exist in many markets (e.g., it may take time to find the kind of chocolate one wants to buy). Even so, we can easily imagine that this type of friction is particularly severe in the labor market because workers and jobs are heterogeneous in many dimensions. Some search models explicitly deal with the matching of heterogeneous workers and jobs. Some models treat the matching process as a "black box" and use reduced-form functions to formulate it. An example of such a model is the Diamond-Mortensen-Pissarides (DMP) model, described in Section 18.4 below. As we will see, the DMP model has the advantage of being able to fit some of the salient features of the labor market.

## 18.2 Some labor market facts

Figure 18.1 plots the unemployment rate in the postwar U.S. economy, computed from the Current Population Survey (CPS). In the statistics provided below, the entire U.S.

civilian noninstitutional population (16 years old and above) is divided into employment ($E$), unemployment ($U$), and not in the labor force ($N$). The unemployment rate is defined as $U/(E+U)$. In the figure, shaded periods indicate recessions defined by the National Bureau of Economic Research.[2] Unemployment rate describes the ratio of workers who cannot find a job (although they are searching for a job or on temporary layoff) to the entire labor force. The figure clearly indicates that the unemployment rate increases during recessions. There is no apparent long-run trend in the unemployment rate. Whereas the unemployment rate trended up in the 1970s, it has trended down since then, except for the stark increases in the Great Recession and the COVID-19 pandemic.

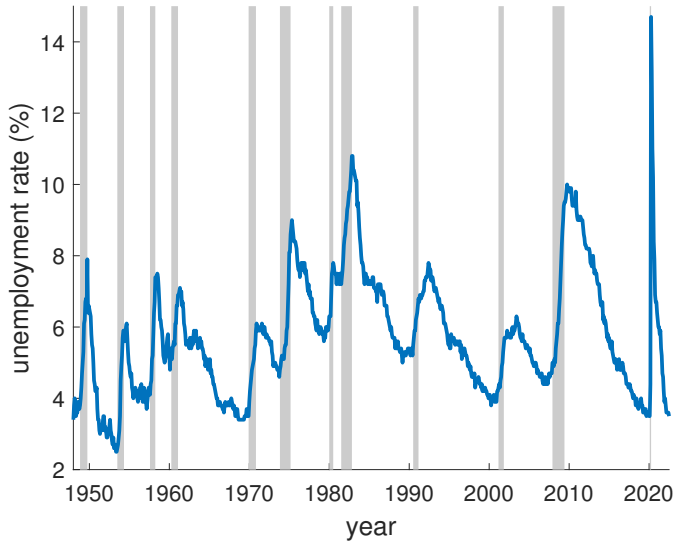



Figure 18.1: Unemployment rate in the United States.

**Source**: CPS.

Figure 18.2 plots the unemployment rate and the vacancy rate ($V$ represents vacancy) in the United States from December 2000 to May 2022. The unemployment rate is identical to Figure 18.1, and the vacancy rate is computed from the Job Openings and Labor Turnover Survey (JOLTS) and the CPS.[3] Here, we simply point out that vacancy and unemployment coexist in the labor market, indicating that there are trivial amounts of friction in the labor market. We will come back to this figure later on.

[2] See `https://www.nber.org/research/business-cycle-dating`.

[3] JOLTS defines the job opening rate as $V/(E+V)$. We transform it to $V/(E+U)$, which is a more relevant object for the theoretical framework below, using the information on the unemployment rate in the CPS.

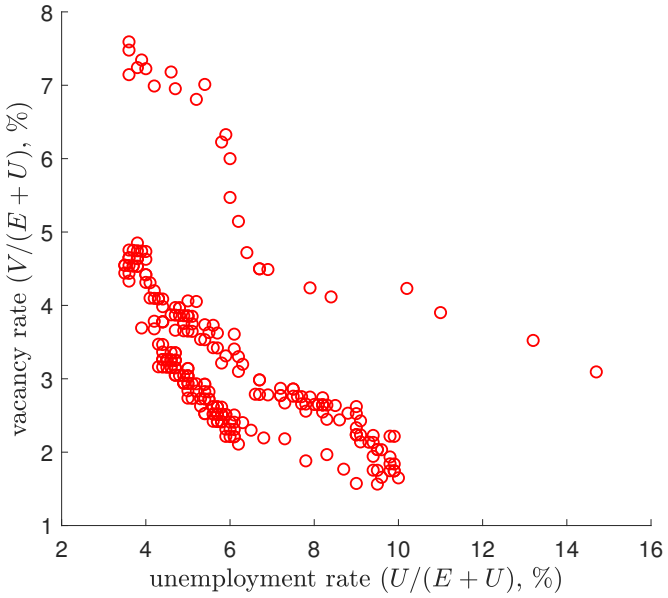



Figure 18.2: Unemployment rate and vacancy rate in the United States.

**Source**: JOLTS and CPS.

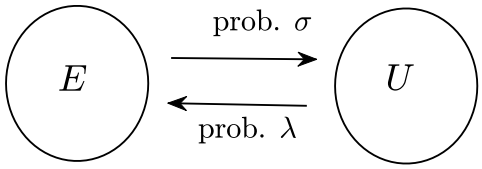



Figure 18.3: The simple model of unemployment.

## 18.3 A simple model of unemployment

As a starting point, consider a simple model of unemployment. The model can be simply described by Figure 18.3. Workers are either employed ($E$) or unemployed ($U$). We normalize the total labor force to 1, and therefore

$$e_t + u_t = 1,$$

where $e_t$ is employment at period $t$ and $u_t$ is unemployment at period $t$. We ignore the movements in and out of the labor force. Because the labor force is 1, $u_t$ is the labor force at period $t$. We assume unemployed workers transit into employment with probability $\lambda$ and employed workers move into unemployment with probability $\sigma$. The probability $\lambda$ is often called the *job finding probability* and the probability $\sigma$ is called the *separation probability*.

Let the unemployment rate at period $t$ be $u_t$. Then

$$u_{t+1} = (1-\lambda)u_t + \sigma(1-u_t) \tag{18.1}$$

holds, where the first term on the right-hand side is the unemployed workers at period $t$ who remain unemployed, and the second term is employed workers who lose their jobs between period $t$ and $t+1$. In the steady state where the unemployment rate is constant at $u_{t+1} = u_t = \bar{u}$,

$$\bar{u} = (1-\lambda)\bar{u} + \sigma(1-\bar{u})$$

holds, and therefore

$$\bar{u} = \frac{\sigma}{\lambda+\sigma}. \tag{18.2}$$

The steady-state unemployment rate is decreasing in the job-finding probability $\lambda$ and increasing in the separation probability $\sigma$. Note the steady state, characterized by the expression (18.2), is unique. Moreover, using (18.1) for period $t$ and $t-1$, we obtain

$$u_{t+1} - u_t = (1-\lambda-\sigma)(u_t - u_{t-1}),$$

and because $|1-\lambda-\sigma| \in (0,1)$, the sequence of the unemployment rate is a Cauchy sequence, and thus it converges to the steady-state value. Similarly

$$u_{t+1} - \bar{u} = (1-\lambda-\sigma)(u_t - \bar{u})$$

holds, and we can see that $|1-\lambda-\sigma|$ determines the speed of convergence. If, for example, monthly $\lambda = 0.45$ and $\sigma = 0.034$ (the numbers we use later in the quantitative analysis), $1-\lambda-\sigma$ is almost half, and $u_t$ converges to the steady-state value very quickly. In this case, $\bar{u}$ can be computed as about 7.0%, and if the economy starts from an unemployment rate of 15%, after six months, the unemployment rate goes down to 7.2%.

The job-finding rate $\lambda$ is also closely linked to the average (or expected) duration of unemployment. As the probability that the unemployment duration is one period is $\lambda$, the probability that it is two periods is $(1-\lambda)\lambda$, the probability that it is three periods is $(1-\lambda)^2\lambda$, and so on, the average duration of unemployment $D$ can be computed as:

$$\begin{aligned} D &= \lambda \cdot 1 + (1-\lambda)\lambda \cdot 2 + (1-\lambda)^2\lambda \cdot 3 + \cdots \\ &= \lambda\{[1 + (1-\lambda) + (1-\lambda)^2 + \cdots] + [(1-\lambda) + (1-\lambda)^2 + \cdots] + [(1-\lambda)^2 + \cdots] + \cdots\} \\ &= \lambda\left[\frac{1}{\lambda} + (1-\lambda)\frac{1}{\lambda} + (1-\lambda)^2\frac{1}{\lambda} + \cdots\right] \\ &= \frac{1}{\lambda}. \end{aligned}$$

Thus the average duration of unemployment is inversely related to $\lambda$. Note that $D$ can also be derived from a recursive formulation. Let the expected duration of unemployment from the viewpoint of time $t$ as $D_t$. Then

$$D_t = \lambda \cdot 1 + (1-\lambda)(1 + D_{t+1}),$$

because if the worker finds a job next period (with probability $\lambda$), the duration is 1, and if she doesn't, it is one period plus the expected duration from the next period. Here, there is no difference in the expected duration forward at period $t$ and period $t+1$, and thus $D_t = D_{t+1} = D$, and we can solve $D = 1/\lambda$.

## 18.4 The Diamond-Mortensen-Pissarides (DMP) model

This section is introduces a model called the search and matching model or the Diamond-Mortensen-Pissarides (DMP) model. The model in this section is a discrete-time version of Pissarides (1985).[4] It is called the "search and matching model" because workers and firms have to engage in search activity (in this model, only firms engage in an active search effort, but both the workers and the firms have to wait until they find the counterpart), and the probability of a successful search is governed by a function called the matching function.

The DMP model features an active search by firms in the form of vacancy postings. Vacancy posting is a form of investment: pay the cost now and receive payoffs later. Firms and workers share the surplus (the difference between market production and home production), and the firm can receive a positive profit. In that sense, we also depart from the competitive labor market. Because the firms are the only party that engage in active search activities, this model is focusing on the demand side of the labor market. As we will see, this assumption is consistent with the behavior of vacancies in the labor market.

### 18.4.1 Matching function and the labor market dynamics

We assume that workers are either employed or unemployed. The total population is normalized to 1. The basic structure of the model is similar to Section 18.3, and we can view the model of this section as endogenizing $\lambda$ of the simple model there.

Firms that look for workers post vacancies to search. The number of vacancies is endogenous—the vacancy posting behavior of firms responds to the costs and benefits of hiring workers. Vacancy posting is a (risky) investment for firms: it is costly to post a vacancy, but if firms successfully hire workers, they can enjoy the profit arising from production in the future. Unemployed workers search for firms to work for. The matching process between firms (vacancies) and unemployed workers is summarized by the *matching function*:

$$\mathcal{M}_{t+1} = M(u_t, v_t),$$

where $\mathcal{M}_{t+1}$ is the number of matches created at the beginning of period $t+1$. The function $M(\cdot, \cdot)$ is increasing in both terms and exhibits constant returns to scale. It also satisfies $M(u_t, v_t) \leq u_t$ and $M(u_t, v_t) \leq v_t$. This function is a "black box" that summarizes the complex process of firms' recruiting activities. In particular, workers and firms are heterogeneous, and it is not an easy task for a firm to find a suitable person for its position. Different firms do not coordinate their recruiting efforts, and they may go after the same person even when there are other people available. The interpretation of the matching function can vary across different models, but in the basic DMP model, the "black box" is interpreted as incorporating all difficulties firms face when recruiting workers.

We assume that the search is random, that is, all vacancies have the same chance of finding workers, and all workers have the same chance of meeting the vacancies. Thus the

[4] The textbook Pissarides (2000) explores various versions of the DMP model in continuous time.

probability of a worker meeting a firm is

$$\frac{M(u_t, v_t)}{u_t} = M\left(1, \frac{v_t}{u_t}\right) = M(1, \theta_t),$$

where $\theta_t$ is defined as $\theta_t \equiv v_t/u_t$ and often referred to as the labor market tightness. Let us denote

$$\lambda_w(\theta_t) \equiv M(1, \theta_t). \tag{18.3}$$

This $\lambda_w(\theta_t)$ corresponds to $\lambda$ in Section 18.3. Note that $\lambda_w(\cdot)$ is increasing in $\theta_t$ from our assumptions about the matching function. The probability of a vacancy meeting a worker is

$$\frac{M(u_t, v_t)}{v_t} = M\left(\frac{u_t}{v_t}, 1\right) = M\left(\frac{1}{\theta_t}, 1\right).$$

Let us define

$$\lambda_f(\theta_t) \equiv M\left(\frac{1}{\theta_t}, 1\right).$$

Note that

$$\lambda_w(\theta_t) = \theta_t \lambda_f(\theta_t)$$

holds.

It turns out that, when $z > b$, all firms and workers accept all matches once they meet. Thus $\lambda_w(\theta_t)$ represents the job-finding probability of an unemployed worker. It also represents the probability that a worker transitions from unemployment to employment, and unlike in Section 18.3, it depends on the labor market tightness. In this section (similarly to Section 18.3), we assume that matches resolve with probability $\sigma \in (0, 1)$.

From the above assumption, the dynamics of unemployment is governed by

$$u_{t+1} = (1 - \lambda_w(\theta_t))u_t + \sigma(1 - u_t). \tag{18.4}$$

The first term on the right-hand side is the unemployed workers at period $t$ who stay unemployed at period $t + 1$. The second term is the employed workers (note that $e_t = 1 - u_t$) who separate from the job between $t$ and $t + 1$.

When the vacancy is constant at $v$, it is straightforward to show that there exists a unique steady state value of $u_t$ (call it $\bar{u}$). To see this, set $u_{t+1} = u_t = \bar{u}$ in (18.4) and obtain

$$\bar{u} = \frac{\sigma}{\lambda_w(v/\bar{u}) + \sigma}. \tag{18.5}$$

This equation can be rewritten as, using (18.3),

$$M(v, \bar{u}) + \sigma\bar{u} = \sigma. \tag{18.6}$$

Because the right-hand side is constant and the left-hand side is increasing in $\bar{u}$, the solution to $\bar{u}$ in (18.5) is unique. Further note that (18.6) describes a negative relationship between $v$ and $\bar{u}$ when $\sigma$ is kept constant.

Now let us go back to Figure 18.2. Figure 18.4 connects the data points of Figure 18.2. It is clear that there is a negative relationship between the unemployment rate and the vacancy rate. This regularity is often called the *Beveridge curve*. The Beveridge curve relationship is consistent with the equation (18.6), and therefore provides support for this component of the DMP model. For this reason, equation (18.6) is often referred to as the Beveridge curve relationship. Moreover, the strong procyclical movement of vacancy indicates that the firm's recruiting activities (i.e., the labor demand movements) play an important role in driving the cyclical movement of the unemployment rate.

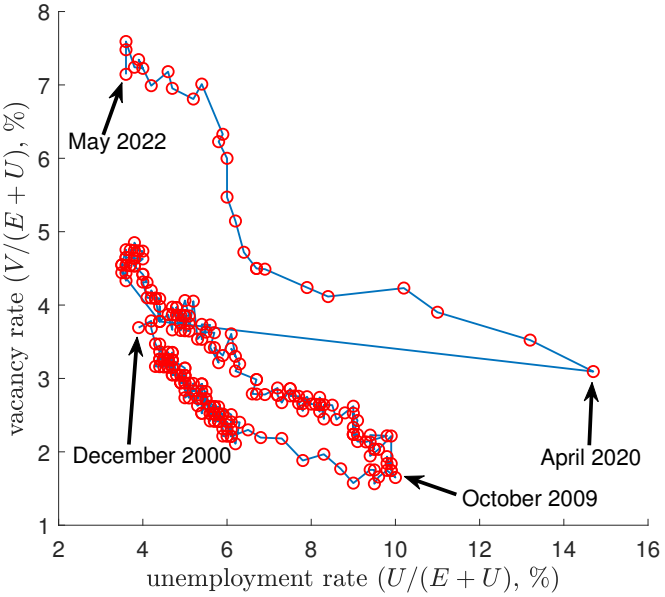



Figure 18.4: Beveridge curve in the United States.

**Sources**: JOLTS and CPS.

### 18.4.2 Market equilibrium with an endogenous vacancy creation

Let us consider how the vacancy level $v_t$ is determined. Here, we assume the production is conducted by a match between one firm (vacancy) and one worker. We abstract from capital stock for the time being and assume that the match between a firm and a worker can produce $z_t$ units of goods. We further assume that, as in the standard real business cycle (RBC) model, there is an aggregate shock to productivity. Therefore, $z_t$ can vary stochastically over time. As in the RBC models, we assume that $z_t$ follows a first-order Markov process.

It turns out that we can formulate the model recursively and the relevant state variable in the general equilibrium is only $z$. Assume that the firm discounts the future profit with

the discount factor $\beta \in (0,1)$. The Bellman equation for a firm that has already matched with a worker is

$$J(z) = z - w(z) + \beta \mathbb{E}[(1-\sigma)J(z') + \sigma V(z')], \tag{18.7}$$

where $J(z)$ is the value of a matched firm. The flow value $z - w(z)$ is profit, where $w(z)$ is the wage paid to the worker when the aggregate productivity is $z$. The parameter $\beta \in (0,1)$ is the discount factor of the firm (which will be identical to the worker's discount factor), and $\mathbb{E}[\cdot]$ indicates the expected value (conditional on the current period information). The prime ($'$) indicates the next period variable. $V(z)$ represents the value of a vacancy.

The Bellman equation for a vacant firm is

$$V(z) = -\kappa + \beta \mathbb{E}[\lambda_f(\theta)J(z') + (1-\lambda_f(\theta))V(z')], \tag{18.8}$$

where $\kappa > 0$ is the cost of posting a vacancy. We also assume that anyone can set up a vacancy and enter the market ("free entry"). Thus, in equilibrium, the value of vacancy is driven down to zero:

$$V(z) = 0. \tag{18.9}$$

(18.8) and (18.9) imply

$$\frac{\kappa}{\lambda_f(\theta)} = \beta \mathbb{E}[J(z')]. \tag{18.10}$$

Intuitively, the cost of vacancy $\kappa$ has to be equal to the expected value of the future filled job $\beta \mathbb{E}[J(z')]$ times the probability of finding a worker $\lambda_f(\theta)$.

To determine the equilibrium wage, we first have to consider the worker side. We assume that a worker is infinitely-lived, consumes what she receives every period, and has linear utility function with discount factor $\beta$ (i.e., the same discount factor as the firm's):[5]

$$\mathbb{E}_0 \left[ \sum_{t=0}^{\infty} \beta^t c_t \right].$$

The Bellman equation for an employed worker is

$$W(z) = w(z) + \beta \mathbb{E}[(1-\sigma)W(z') + \sigma U(z')], \tag{18.11}$$

where $W(z)$ is the value of an employed worker and $U(z)$ is the value of an unemployed worker. We assume that an unemployed worker receives a constant amount of goods $b < z_t$

[5]Because the firms in this economy can earn a profit (in particular, the aggregate profit is positive in the steady-state of the economy), there is a question of who receives the profit (i.e., the ownership of the firms). Here, we implicitly assume that the firms are owned by someone outside the economy who has the same discount factor as the worker. Alternatively, we can assume that the firm is owned by workers. As we will see later in this section, because only the *difference* of income between the employed worker and the unemployed worker matters for the equilibrium dynamics of unemployment, the same results go through if we assume that the workers' initial ownership of the firms (i.e., the stock holdings) is equal across workers. This result follows because, with linear utility, there are no reasons for the workers to trade stocks. In Section 18.8, we assume a closed economy and make the stock holding explicit.

(which has to hold for all possible values of $z_t$). $b$ can be interpreted as home production or unemployment insurance payment. The Bellman equation for an unemployed worker is

$$U(z) = b + \beta\mathbb{E}[\lambda_w(\theta)W(z') + (1 - \lambda_w(\theta))U(z')]. \tag{18.12}$$

Once a firm and a worker match, they are in a *bilateral monopoly* situation: the only possible seller (of the labor service) for the firm is the worker it matched with, and the only possible buyer for the worker is the firm she matched with. In such a situation, we cannot use the marginal principle to determine the wage because there is no competition. The match generates a surplus. In the current period, the match jointly generates $z$—if they separate, the worker can create $b$, and the firm can end up creating nothing. Therefore, it is jointly beneficial for the firm and the worker to be together because of the assumption $z > b$. Unless they are separated, this flow surplus $z - b$ is generated in the future as well. We assume that the firm and the worker split the surplus following the *Generalized Nash Bargaining* rule. The Nash Bargaining rule splits the surplus so that the Nash Product, which is the product of surpluses of each party (in our case, the firm and the worker), is maximized. The Generalized Nash Bargaining rule uses the weighted product instead, where the "weight" is represented as the exponent to each of the surpluses.

In our formulation, the Generalized Nash Bargaining solution solves

$$\max_w (\tilde{W}(w,z) - U(z))^\gamma (\tilde{J}(w,z) - V(z))^{1-\gamma},$$

where $\tilde{W}(w,z)$ is the value of an employed worker when the current wage is $w$. Note that the Bellman equation (18.11) assumes that the wage is the equilibrium value under $z$, $w(z)$. Here, we are allowing $w$ to be any value. Therefore $\tilde{W}(w(z),z) = W(z)$ holds. Similarly, $\tilde{J}(w,z)$ is the value of a job matched with a worker when the wage is $w$. Here, the worker's surplus is $\tilde{W}(w,z) - U(z)$, and the firm's surplus is $\tilde{J}(w,z) - V(z)$. The exponent $\gamma \in (0,1)$ represents the "weight" of the worker's surplus. It is often referred to as the "bargaining power" of the worker. By taking the first-order condition, one can show that $w$ solves

$$(1-\gamma)(\tilde{W}(w,z) - U(z)) = \gamma(\tilde{J}(w,z) - V(z)). \tag{18.13}$$

The detailed derivation of (18.13) is in Appendix 18.A.1.

The six equations (18.7), (18.8), (18.9), (18.11), (18.12), and (18.13) define the equilibrium. These can be rearranged to obtain a difference equation on $\theta_t$:[6]

$$\frac{\kappa}{(1-\gamma)\lambda_f(\theta_t)} = \beta\mathbb{E}\left[z_{t+1} - b + \frac{1-\sigma-\gamma\lambda_w(\theta_{t+1})}{1-\gamma}\frac{\kappa}{\lambda_f(\theta_{t+1})}\right]. \tag{18.14}$$

When $z$ is constant over time, the steady-state value of $\theta_t$ (denote it $\bar{\theta}$) solves

$$\frac{\kappa}{(1-\gamma)\lambda_f(\bar{\theta})} = \beta\left[z - b + \frac{1-\sigma-\gamma\lambda_w(\bar{\theta})}{1-\gamma}\frac{\kappa}{\lambda_f(\bar{\theta})}\right]. \tag{18.15}$$

[6] We can analyze the implications on equilibrium wages using the same set of equations. The analysis of the wages is in Appendix 18.A.2.

This equation (note that the right-hand side is decreasing in $\bar{\theta}$) determines $\bar{\theta} = \bar{v}/\bar{u}$, where $\bar{v}$ is the steady-state value of vacancy. Often this condition is called the job creation condition. The job creation condition, together with the Beveridge curve relationship (18.6) determine $\bar{v}$ and $\bar{u}$.

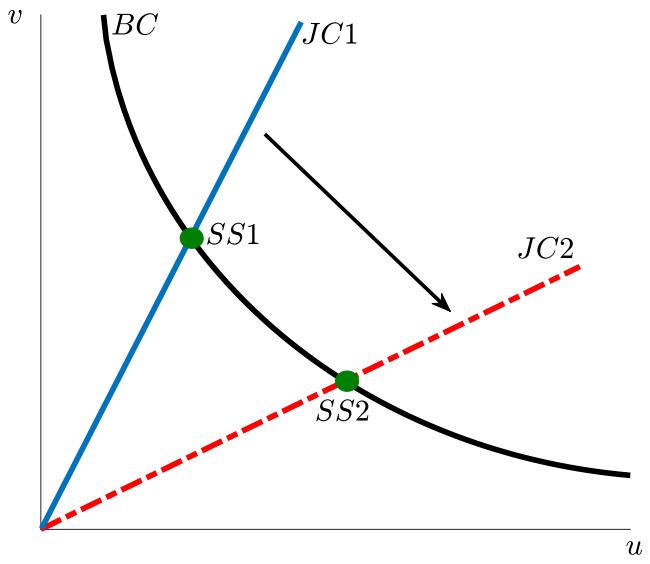



Figure 18.5: The determination of the steady state

Figure 18.5 describes the determination of $v$ and $u$ in the steady state. The $BC$ curve represents (18.6), which describes the steady-state relationship between $u$ and $v$. The straight lines $JC1$ and $JC2$ represent the relationship between $u$ and $v$ that correspond to different values of $\theta$. When $JC1$ represents the value of $\theta$ that satisfy (18.15), the steady-state values of $u$ and $v$ are at $SS1$. When $z$ goes down, from (18.15) we can see $\bar{\theta}$ also goes down. The $JC$ line shifts from $JC1$ to $JC2$. In the new steady-state ($SS2$), $v$ is smaller and $u$ is larger.

The transition dynamics is also easy to analyze. Consider an unanticipated one-time permanent decline in $z$. Because the decline is permanent, the new job creation equation holds with the new steady-state value of $\theta$. In other words, the equation (18.15) holds with the new value of $\bar{\theta}$. In Figure 18.6, the new value of $\bar{\theta}$ is represented by the new line $JC2$. The jump to the new value of $\bar{\theta}$ is immediate. Using equation (18.14), it is possible to show that $\theta_t$ would diverge away from the new steady state (and eventually violate the economy's feasibility) unless $\theta_t$ immediately jumps to the new steady-state value. Because $u$ cannot jump, $v$ immediately drops so that $v/u$ becomes the new value of $\bar{\theta}$. After that, the economy gradually converges to the new steady state ($SS2$) along the $JC2$ line.

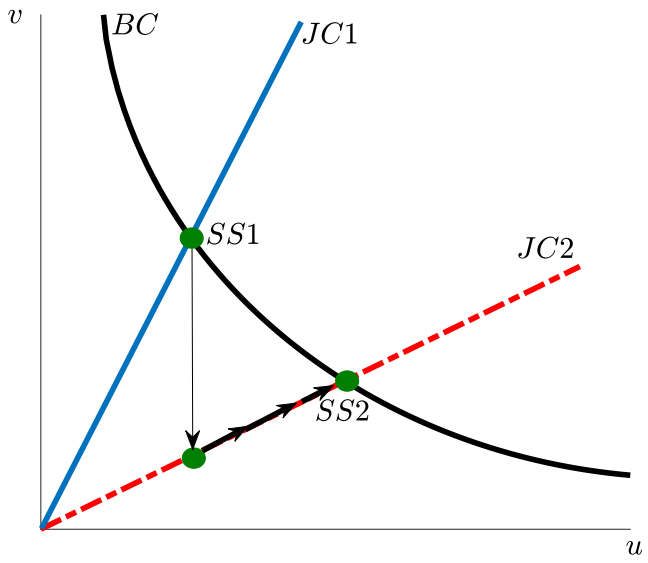



Figure 18.6: The transition dynamics of the DMP model.

### 18.4.3 Efficiency

Unemployment has been considered one of the most important macroeconomic problems. Various policies are proposed and implemented to reduce unemployment. Before considering policies, however, it is critical to know what kind of inefficiencies exist in the economy and whether unemployment in the market equilibrium is too high or too low compared to the social optimum.

One can think of the DMP model as describing the firm's investment problem through vacancy posting. We will show that there are two inefficiencies in making the investment decision.[7] First, because the firm is the only entity that makes the investment, the reward has to be captured solely by the firm. In the DMP model, this is not the case. The surplus from matching is divided by the firms and the workers by the Generalized Nash Bargaining. This fact implies that the firm's reward from investment is "taxed," and the number of vacancies is inefficiently low. Second, a firm's decision imposes externalities on other firms and workers. Here, because the only entities that are making decisions are firms, what matters is the externality imposed on other firms. The externality imposed on the other firms is the number of vacancies times the change in the matching probability for each of the other firms, that is,

$$v \times \frac{\partial}{\partial v}\left(\frac{M(u,v)}{v}\right) = M_2(u,v) - \frac{M(u,v)}{v}. \tag{18.16}$$

The first term on the right-hand side, $M_2(u,v)$, is the number of matches created by the

[7]The exposition below closely follows Fukui and Mukoyama (2024).

marginal vacancy. The second term, $M(u,v)/v$, is the private perception of the likelihood of a new match for a vacancy-posting firm. The difference is the externality. In other words, the externality is the difference between the marginal increase in the number of matches and the average number of matches per vacancy. Because the externality is negative, this inefficiency leads to too many vacancies in the market equilibrium. The balance of the two inefficiencies determines the overall effect.

There are similar inefficiencies in valuing the outcome of investment. The social value of moving a worker from unemployment to employment can be different from the private value. The social value takes the negative externality that an unemployed worker imposes on other unemployed workers. Because of this externality, the market equilibrium overvalues the opportunity cost of employment. At the same time, in the market equilibrium, only the opportunity cost for the worker is taken into account, and thus the opportunity cost is undervalued. The eventual outcome depends on the balance of these two inefficiencies.

To start the formal analysis, let us formulate the social planner's problem when the social planner is subject to the same labor market frictions as in the market equilibrium. This type of problem is often called the problem that solves the "constrained efficient" solution. The social planner maximizes the social welfare

$$\mathbb{E}_0\left[\sum_{t=0}^{\infty}\beta^t(z_t e_t + b(1-e_t) - \kappa v_t)\right],$$

where $\mathbb{E}_0[\cdot]$ is the expected value at time 0. The first term is the production by matched worker-firm, the second is the unemployed workers' home production, and the third is the vacancy-posting cost. Using $v_t = \theta_t u_t = \theta_t(1-e_t)$, let us write the social planner's problem as

$$\max_{\theta_t, e_{t+1}} \mathbb{E}_0\left[\sum_{t=0}^{\infty}\beta^t(z_t e_t + b(1-e_t) - \kappa\theta_t(1-e_t))\right]$$

subject to

$$e_{t+1} = (1-\sigma)e_t + \lambda_w(\theta_t)(1-e_t).$$

Let $\mu_t$ be the Lagrange multiplier of the constraint. Then, the Lagrangian is

$$L = \mathbb{E}_0\left[\sum_{t=0}^{\infty}\beta^t(z_t e_t + b(1-e_t) - \kappa\theta_t(1-e_t)) + \sum_{t=0}^{\infty}\mu_t((1-\sigma)e_t + \lambda_w(\theta_t)(1-e_t) - e_{t+1})\right].$$

The first-order condition on $\theta_t$ is

$$\beta^t\kappa(1-e_t) = \mu_t\lambda_w'(\theta_t)(1-e_t)$$

and therefore

$$\kappa = \lambda_w'(\theta_t)\beta\hat{\mu}_t \tag{18.17}$$

holds, where $\hat{\mu}_t$ is defined as

$$\hat{\mu}_t \equiv \frac{\mu_t}{\beta^{t+1}}. \tag{18.18}$$

Because $\mu_t$ is the time-0 social value of increasing $e_{t+1}$ by one unit (taking expectations at time $t$), $\hat{\mu}_t$ is the concurrent value of one unit of employment at time $t+1$ (taking expectations at time $t$). The left-hand side of (18.17), $\kappa$, is the social cost of increasing $\theta_t$ by one unit. By increasing $\theta_t$ by one unit, employment increases by $\lambda_w'(\theta_t)$ units, so the right-hand side of (18.17) is the social benefit of increasing $\theta_t$ from the viewpoint of the time when the vacancy is created.

The first-order condition on $e_{t+1}$ is

$$\mu_t = \mathbb{E}_t \left[ \beta^{t+1}(z_{t+1} - b + \kappa\theta_{t+1}) + \mu_{t+1}(1 - \sigma - \lambda_w(\theta_{t+1})) \right].$$

Here, $\mathbb{E}_t[\cdot]$ is the expectation taken at period $t$. It can be rewritten as

$$\hat{\mu}_t = \mathbb{E}_t \left[ z_{t+1} - b + \kappa\theta_{t+1} + \beta\hat{\mu}_{t+1}(1 - \sigma - \lambda_w(\theta_{t+1})) \right]. \tag{18.19}$$

The two equations (18.17) and (18.19) characterize the socially optimal outcome, represented by $\theta_t$ and $\hat{\mu}_t$ for $t = 0, 1, \ldots$. The term $\beta\hat{\mu}_{t+1}\lambda_w(\theta_{t+1})$ corresponds to the opportunity cost of employment: a worker has to forgo an opportunity of a possible new match. However, having one more unemployed worker imposes an externality on other workers. The term $\kappa\theta_{t+1}$ corrects for this externality.

Let us derive the market equilibrium equations that correspond to these two. First, define the *expected surplus of a match* as

$$S_t \equiv \mathbb{E}_t[W(z_{t+1}) - U(z_{t+1}) + J(z_{t+1}) - V(z_{t+1})],$$

where the expectation is taken at time $t$. From (18.10) and (18.13) (which implies $(1-\gamma)S_t = \mathbb{E}_t[J(z_{t+1})]$), we obtain

$$\kappa = \beta(1 - \gamma)\lambda_f(\theta_t)S_t. \tag{18.20}$$

From (18.7), (18.11), and (18.12) all for time $t+1$ (and taking expectations at time $t$), and using (18.9) and (18.13) (which implies $\mathbb{E}_{t+1}[W(z_{t+2}) - U(z_{t+2})] = \gamma S_{t+1}$),

$$S_t = \mathbb{E}_t \left[ z_{t+1} - b + \beta S_{t+1}(1 - \sigma - \gamma\lambda_w(\theta_{t+1})) \right]. \tag{18.21}$$

In the final term on the right-hand side, $\beta S_{t+1}\lambda_w(\theta_{t+1})$, is multiplied by $\gamma$ because only the opportunity cost of a match on the worker side is taken into account in the market equilibrium.

To make a comparison between the social planner's solution and the market equilibrium, first, define the elasticity of the firm's worker-finding probability as

$$\eta(\theta) \equiv -\frac{\theta\lambda_f'(\theta)}{\lambda_f(\theta)}.$$

From $\lambda_w(\theta) = \theta\lambda_f(\theta)$, one can derive

$$\lambda_w'(\theta) = \lambda_f(\theta)(1 - \eta(\theta)).$$

This relationship enables us to rewrite the social planner's first-order condition (18.17) as

$$\kappa = \beta(1 - \eta(\theta_t))\lambda_f(\theta_t)\hat{\mu}_t. \tag{18.22}$$

This equation can be used to rewrite (18.19) as

$$\hat{\mu}_t = \mathbb{E}_t \left[ z_{t+1} - b + \beta\hat{\mu}_{t+1}(1 - \sigma - \eta(\theta_{t+1})\lambda_w(\theta_{t+1})) \right]. \tag{18.23}$$

Comparing (18.20) against (18.22) and (18.21) against (18.23), one can see that the social planner's solution $(\theta_t, \hat{\mu}_t)$ and the equilibrium outcome $(\theta_t, S_t)$ are equivalent if

$$\eta(\theta) = \gamma \tag{18.24}$$

for all $\theta$. This condition is often called the *Hosios condition* (Hosios, 1990). For the firm's investment incentive, the market equilibrium "taxes" the vacancy-creation incentive to correct for the negative externality imposed on the other firms. The "tax rate" $\gamma$ has to be larger when $\eta(\theta)$ is larger because, when $\eta(\theta)$ is large, the change of the other firms' matching probability $\lambda_f(\theta)$ due to the firm's vacancy posting is larger. One can also see that

$$\eta(\theta)\lambda_f(\theta) = -\theta\lambda_f'(\theta) = \theta\frac{d}{d\theta}\left(\frac{M(1,\theta)}{\theta}\right) = M_2(u,v) - \frac{M(u,v)}{v},$$

which implies that the term $\eta(\theta)\lambda_f(\theta)$ is indeed the externality imposed to the other firms, derived in (18.16).[8] For the valuation of worker employment, the Hosios condition ensures that the externality that an unemployed worker imposes on other unemployed workers corresponds to the under-valuation of the opportunity cost of employment in the market equilibrium.

Suppose the Hosios condition (18.24) does not hold. For example, suppose that $\eta(\theta) > \gamma$ for all $\theta$. It can be shown that, in this case, the value of $\theta$ in the market equilibrium is too large compared to the constrained-efficient outcome. In this situation, the equilibrium unemployment rate is too low compared to the social optimum. A policy that reduces the unemployment rate, therefore, can lower the social welfare in this situation. In contrast, if $\eta(\theta) < \gamma$, the equilibrium unemployment rate is too high. In this situation, there is room for improving social welfare through social policies targeting lower unemployment.

## 18.5 Labor market facts, once again

The modern macroeconomic study of the labor market considers the gross flows behind the movement of stocks, such as unemployment in Figure 18.1. In fact, the model we presented in Sections 18.3 and 18.4 are the analysis of gross flows between the state $E$ (employment) and the state $U$ (unemployment).

In the data, there are three states of the labor market, $E$ (employment), $U$ (unemployment), and $N$ (not in the labor force). Thus we can consider six flows across these states, as we can see in Figure 18.7.

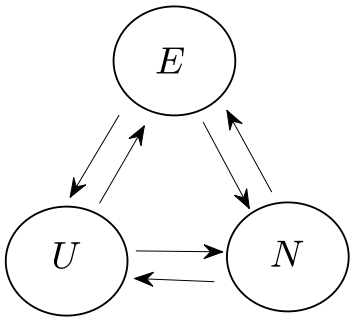



Figure 18.7: Flows among three states.

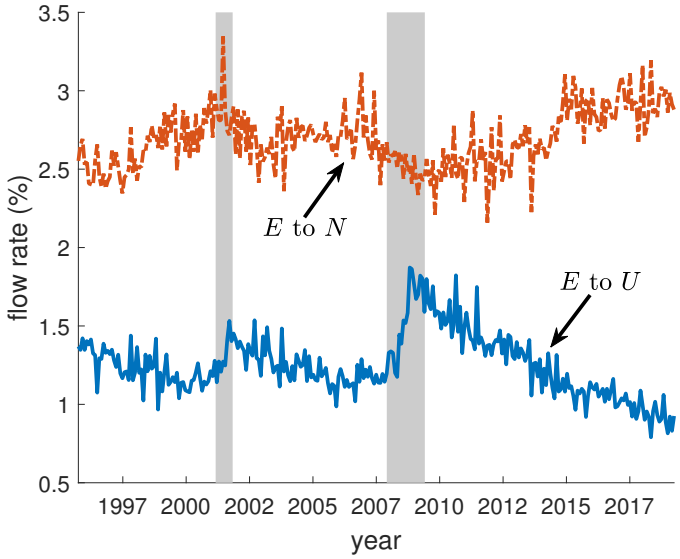



Figure 18.8: Flow rates out of $E$.

**Sources**: CPS (Fallick and Fleischman, 2004).

Figures 18.8, 18.9, and 18.10 plot these flow rates, from Fallick and Fleischman (2004).[9] For example, the $E$ to $U$ flow rate in Figure 18.8 plots the fraction of employed workers that flow into $U$ in the following month. One can see how the gross flows influence the movement of stocks. For example, $U$ increases in recessions because the inflows from $E$ and $N$ go up and the outflows to $E$ and $N$ go down.

[8]The first equation uses the definition of $\eta(\theta)$, and the second equation uses the definition of $\lambda_f(\theta)$. For the third equation, the property $M_2(u,v) = M_2(1,\theta)$, which can be derived by differentiating both sides of $M(u,v) = uM(1,v/u)$ with respect to $v$, is used.

[9]The data is from `https://www.federalreserve.gov/pubs/feds/2004/200434/200434abs.html`. Seasonal adjustment is made using X-13 ARIMA-SEATS from the U.S. Census Bureau.

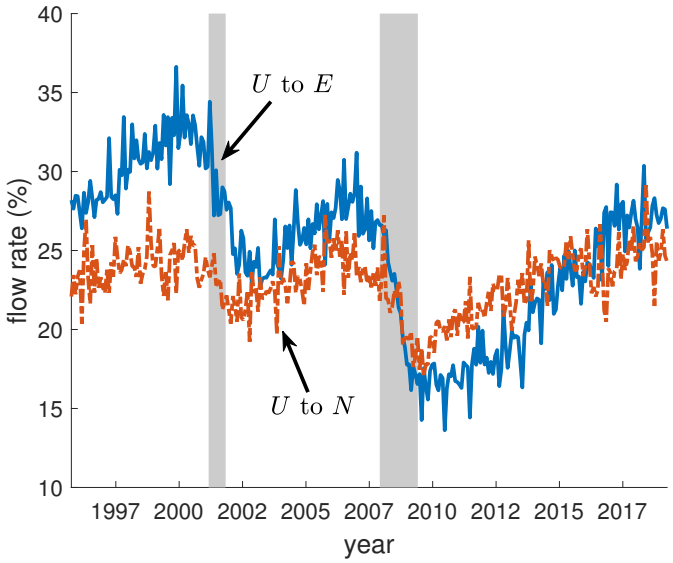



Figure 18.9: Flow rates out of $U$.

**Sources**: CPS (Fallick and Fleischman, 2004).

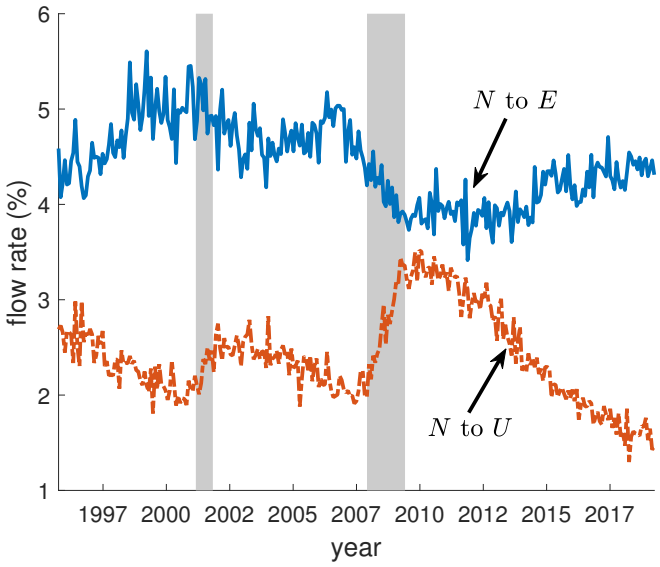



Figure 18.10: Flow rates out of $N$.

**Sources**: CPS (Fallick and Fleischman, 2004).

These stylized facts provide important information for building the models of unemployment. Here, we point out two simple observations. First, the flow rate from $U$ to $E$ is strongly procyclical. This property is consistent with the DMP model in Section 18.4. Second, the flow rate from $E$ to $U$ is strongly countercyclical. This fact implies the constant separation rate in the DMP model above is not consistent with the data. We will come back to this point in Section 18.7.

## 18.6 Unemployment volatility puzzle

We now examine the quantitative performance of the DMP model that we presented in Section 18.4. Let us go back to the difference equation (18.14). We will quantitatively evaluate the model (as in the analysis of the RBC models) by assigning functional forms and parameter values to the model.

Assume that the matching function is specified as the Cobb-Douglas form:

$$M(u, v) = \chi u^{\eta} v^{1-\eta}, \tag{18.25}$$

where $\eta \in (0, 1)$. Then $\lambda_w(\theta) = \chi\theta^{1-\eta}$ and $\lambda_f(\theta) = \chi\theta^{-\eta}$. Note that we are using the same notation $\eta$ as in Section 18.4.3. If we compute the (negative of) elasticity of $\lambda_f(\theta)$ with respect to $\theta$, it is exactly the constant value $\eta$ in this Cobb-Douglas form.

In addition, specify the process of $z_t$ by

$$\hat{z}_{t+1} = \rho\hat{z}_t + \varepsilon_{t+1}, \tag{18.26}$$

where the hat ($\hat{}$) describes the log-deviation: $\hat{z}_t = \ln(z_t) - \ln(\bar{z})$ ($\bar{z}$ is the steady-state value of $z$). The parameter $\rho \in (0, 1)$ represents the persistence of the shock, and the i.i.d. shock $\varepsilon_{t+1} \sim N(0, \sigma_\varepsilon^2)$. Here, $\ln(\bar{z})$ is normalized to zero.

### 18.6.1 Log-linearized solution

We will solve this model by log-linearizing the solution around the steady state. Log-linearizing the equation (18.14) yields

$$\mathcal{A}\hat{\theta}_t = \mathbb{E}[\bar{z}\hat{z}_{t+1} + \mathcal{B}\hat{\theta}_{t+1}], \tag{18.27}$$

where

$$\mathcal{A} = \frac{\kappa\bar{\theta}^{\eta}\eta}{(1-\gamma)\beta\chi}$$

and

$$\mathcal{B} = \frac{1-\sigma}{1-\gamma}\frac{\kappa\bar{\theta}^{\eta}\eta}{\chi} + \frac{\gamma\kappa\bar{\theta}}{1-\gamma}.$$

Appendix 18.A.3 briefly describes the basic method of log-linearization. Further details can be found, for example, in Uhlig (2001).

| Calibrated Parameters | Value |
|---|---|
| $\beta$ | 0.996 |
| $\rho$ | 0.949 |
| $\sigma_\varepsilon$ | 0.0065 |
| $\sigma$ | 0.034 |
| $\chi$ | 0.45 |
| $b$ | 0.4 |
| $\gamma$ | 0.72 |
| $\eta$ | 0.72 |

Table 18.1: Parameter values

To solve the difference equation (18.27) using the method of undetermined coefficients, first guess $\hat{\theta}_t = \mathcal{C}\hat{z}_t$, where $\mathcal{C}$ is the undetermined coefficient. Plugging this guess into (18.27) and using $\mathbb{E}[z_{t+1}] = \rho z_t$, we obtain

$$\mathcal{C} = \frac{\rho}{\mathcal{A} - \rho\mathcal{B}}.$$

Rearranging, we obtain the relationship between $\hat{\theta}_t$ and $\hat{z}_t$ as

$$\hat{\theta}_t = (1-\gamma)\left[\frac{\kappa\bar{\theta}^\eta\eta}{\chi}\left(\frac{1}{\rho\beta} - (1-\sigma)\right) + \kappa\gamma\bar{\theta}\right]^{-1}\hat{z}_t. \tag{18.28}$$

Analytically, this result provides important insights into the model performance against the data. In particular, (18.28) implies (for a given $\bar{\theta}$), a small $\kappa$, a small $\gamma$, a small $\eta$, a large $\chi$, a large $\rho$, a large $\beta$, and a small $\sigma$ makes the response of $\theta_t$ larger.

### 18.6.2 Calibration

Now we set the relevant parameter values. Let one period in the model be one month, so that we can capture the fast labor market dynamics in the U.S. economy. Here, we assume the parameter values in Table 18.1.

The discount factor $\beta$ is set at $0.947^{\frac{1}{12}} = 0.996$. The annual value of 0.947 is taken from the standard real business cycle literature (Cooley and Prescott, 1995).

The parameters for the process for $\hat{z}$, $\rho$, and $\sigma_\varepsilon$, are taken from Hagedorn and Manovskii (2008) and adjusted to a monthly frequency. Here, $\sigma_\varepsilon$ is the standard deviation of $\varepsilon_{t+1}$ in (18.26), assuming that $\varepsilon_{t+1}$ follows a normal distribution.

The values $b = 0.4$, $\eta = 0.72$, and $\gamma = 0.72$ follows Shimer (2005). Furthermore, following Shimer (2005), the parameter value for $\kappa$ is set so that the equation (18.15) is satisfied in the steady state with $\bar{\theta} = 1$. The values of $\chi$ and $\sigma$ also follow Shimer (2005).[10]

[10]These values are larger than the $UE$ flow rates and the $EU$ flow rates in Figures 18.8 and 18.9. This discrepancy is because Shimer (2005) calculates the outflow from and the inflow into $U$ (which include the flows in and out of $N$), instead of $UE$ and $EU$ flow rates.

| | | $u$ | $v$ | $v/u$ | $z$ |
|---|---|---|---|---|---|
| Standard Deviation | | 0.125 | 0.139 | 0.259 | 0.013 |
| Quarterly Autocorrelation | | 0.870 | 0.904 | 0.896 | 0.765 |
| Correlation Matrix | $u$ | 1 | $-0.919$ | $-0.977$ | $-0.732$ |
| | $v$ | — | 1 | 0.982 | 0.460 |
| | $v/u$ | — | — | 1 | 0.967 |
| | $z$ | — | — | — | 1 |

Table 18.2: Summary statistics for quarterly U.S. data

| | | $u$ | $v$ | $v/u$ | $z$ |
|---|---|---|---|---|---|
| Standard Deviation | | 0.005 | 0.016 | 0.020 | 0.013 |
| Quarterly Autocorrelation | | 0.826 | 0.700 | 0.764 | 0.765 |
| Correlation Matrix | $u$ | 1 | $-0.839$ | $-0.904$ | $-0.804$ |
| | $v$ | — | 1 | 0.991 | 0.972 |
| | $v/u$ | — | — | 1 | 0.961 |
| | $z$ | — | — | — | 1 |

Table 18.3: Model statistics

### 18.6.3 Quantitative results

With the given parameter values and the equations (18.4) and (18.28), we can simulate the model by randomly generating the series of $\hat{z}_t$ following (18.26). Table 18.2 is the summary of the U.S. data that we will compare the model against. Here, $z$ measures labor productivity. All are originally from monthly data but averaged to quarterly data, and logged and HP-filtered with the smoothing parameter of 1600. The table is taken from Hagedorn and Manovskii (2008).

Table 18.3 is the statistics from the model-generated data. The model generates the right correlations between variables, but the magnitude of the fluctuations in $u$, $v$, and $\theta$ is too small compared to the data. This discrepancy is often referred to as the *unemployment volatility puzzle* or the *labor market volatility puzzle* (or the "Shimer puzzle," after Shimer (2005)). The model's inability to match the magnitude of labor market fluctuations triggered an extensive body of research in early 2000s.

Intuitively, there are two reasons, corresponding to benefits and costs of hiring, for the quantitatively small response. First, the benefit of hiring a worker is procyclical, but the magnitude is not large. One reason is that the wage increases in booms, and it weakens the response of profit to the productivity shock. Second, the cost of hiring a worker, $\kappa/\lambda_f(\theta)$ moves together with $\theta$. In booms, $\theta$ increases, and this increase in cost dampens the firm's response to a positive productivity shock.

### 18.6.4 Rigid wages

Many possible "solutions" are proposed for the unemployment volatility puzzle. Although there is no clear consensus among researchers in terms of which proposed "solution" is the most plausible one, here we highlight the role of rigid wages. As we explained above, the response of wages to productivity shocks dampens the volatility of profits. Rigid wages would make the benefit of creating a vacancy more volatile.

Empirically, there have been many studies about wage rigidity, both nominal and real.[11] Even without search frictions, rigid wages can generate unemployment by preventing the labor market from clearing. Here, wage rigidity changes the incentive for the firms to demand workers in booms and recessions.[12]

Instead of the Generalized Nash Bargaining, here we assume that wages are rigid at the steady-state value $w = \bar{w}$. Combining (18.7) with (18.10), we obtain

$$\frac{\kappa}{\lambda_f(\theta_t)} = \beta \mathbb{E}\left[ z_{t+1} - \bar{w} + \frac{(1-\sigma)\kappa}{\lambda_f(\theta_{t+1})} \right].$$

Log-linearizing,

$$\hat{\theta}_t = \left[ \frac{\kappa \bar{\theta}^{\eta} \eta}{\chi} \left( \frac{1}{\rho\beta} - (1-\sigma) \right) \right]^{-1} \hat{z}_t. \tag{18.29}$$

With the same calibration, the model outcome can be computed as in Table 18.4. Unemployment fluctuations are of the same magnitude as the data. Comparing (18.28) and (18.29) reveals two factors that the wage rigidity can make the firm's profit more volatile. First, the latter is not multiplied by $(1-\gamma)$, indicating that the additional gain in the surplus is now not shared between the worker and the firm, and firm can receive all the additional gain. Second, the term $\kappa\gamma\bar{\theta}$, which represents the improvement of the worker's bargaining position due to the rise in the future job finding probability, is absent when wages are rigid.

Therefore, one can see that a (real) wage rigidity can address the volatility puzzle. We note, however, the sources and magnitude of the wage rigidity still remains an active area of research.

| | | $u$ | $v$ | $v/u$ | $z$ |
|---|---|---|---|---|---|
| Standard Deviation | | 0.115 | 0.329 | 0.425 | 0.013 |
| Quarterly Autocorrelation | | 0.825 | 0.693 | 0.763 | 0.765 |
| Correlation Matrix | $u$ | 1 | −0.791 | −0.881 | −0.784 |
| | $v$ | — | 1 | 0.986 | 0.969 |
| | $v/u$ | — | — | 1 | 0.961 |
| | $z$ | — | — | — | 1 |

Table 18.4: Model statistics with fixed wages

[11] See, for example, McLaughlin (1994) and Elsby and Solon (2019).

[12] The role of wage rigidity in this context was first explored by Hall (2005) and Shimer (2005).

## 18.7 Endogenous separation

The previous sections focused on the role of the fluctuations in job-finding probability. As we saw in Section 18.5, both job finding and separation rates are both strongly cyclical.[13] In particular, at the onset of recessions, the increase in the separation rate tends to cause a sharp increase in the unemployment rate. In Figure 18.8, the *EU* flow rate increases in recessions, and the magnitude of the increase depends on the severity of the recession, suggesting that the separation rate changes endogenously with the business cycle. In this section, we extend the basic model in Section 18.4 to allow for endogenous separation.

### 18.7.1 Formulation

Instead of facing an exogenous separation shock, the firm has to pay a cost for maintaining the match, $c(\sigma)$. Now $\sigma$ is a choice variable for the firm, but the cost increases if the firm wants to make the separation probability small, that is, $c'(\sigma) < 0$.

The matched firm's Bellman equation is now

$$J(z) = \max_{\sigma} \; z - w(z) - c(\sigma) + \beta \mathbb{E}\left[(1-\sigma)J(z') + \sigma V(z')\right].$$

The rest of the equilibrium conditions ((18.8), (18.9), (18.11), (18.12), and (18.13)) are the same as in the Section 18.4. The optimal value of $\sigma$ is now a function of $z$ (denote it as $\sigma(z)$). The unemployment dynamics is, therefore,

$$u_{t+1} = (1 - \lambda_w(\theta_t(z_t)))u_t + \sigma(z_t)(1-u_t).$$

The first-order condition for $\sigma$ is, using (18.9),

$$-c'(\sigma) = \beta \mathbb{E}[J(z')].$$

From (18.10), this equation implies

$$-c'(\sigma) = \frac{\kappa}{\lambda_f(\theta)}. \tag{18.30}$$

The job creation condition can be derived in the same manner as in the exogenous separation case:

$$\frac{\kappa}{(1-\gamma)\lambda_f(\theta_t)} = \beta \mathbb{E}\left[z_{t+1} - b - c(\sigma(z_{t+1})) + \frac{1-\sigma(z_{t+1})-\gamma\lambda_w(\theta_{t+1})}{1-\gamma}\frac{\kappa}{\lambda_f(\theta_{t+1})}\right]. \tag{18.31}$$

The equation (18.31) determines the dynamics of $\theta_t$. Here, $\sigma(z)$ is determined in equilibrium by (18.30) (because $\theta$ is a function of $z$, $\sigma$ is also a function of $z$).

[13] Fujita and Ramey (2009) is an earlier work that emphasize the importance of the separation margin in unemployment fluctuations.

### 18.7.2 Log-linearized system

Once again, we work with a log-linearized system. First, let the maintenance cost function be

$$c(\sigma) = \phi\sigma^{-\xi},$$

where $\phi > 0$ and $\xi > 0$ are parameters. Assume that the matching function takes the form (18.25). The derivation of the log-linearized system is similar to the exogenous separation case and detailed in Appendix 18.A.4.

First, guess the log-linearized relationship between $\theta_t$ and $z_t$ as

$$\hat{\theta}_t = \mathcal{G}\hat{z}_t. \tag{18.32}$$

Then (18.30) can be log-linearized to

$$\hat{\sigma}(z_t) = -\frac{\eta}{\xi+1}\mathcal{G}\hat{z}_t. \tag{18.33}$$

Thus the log-deviation of the cost is (using the shortened notation of $c(z) = c(\sigma(z))$)

$$\hat{c}(z_t) = \frac{\xi\eta}{\xi+1}\mathcal{G}\hat{z}_t. \tag{18.34}$$

Using (18.33) and (18.34), (18.31) can be The coefficient $\mathcal{G}$ can be solved as

$$\mathcal{G} = \frac{\Theta}{\Gamma},$$

where

$$\Theta \equiv (1-\gamma)\left[\frac{\kappa\bar{\theta}^{\eta}\eta}{\chi}\left(\frac{1}{\rho\beta} - (1-\bar{\sigma})\right) + \kappa\gamma\bar{\theta}\right]^{-1}\bar{z}$$

and

$$\Gamma \equiv 1 + (1-\gamma)\left[\frac{\kappa\bar{\theta}^{\eta}\eta}{\chi}\left(\frac{1}{\rho\beta} - (1-\bar{\sigma})\right) + \kappa\gamma\bar{\theta}\right]^{-1}\bar{c}\frac{\xi\eta}{\xi+1}\left(1 - \frac{1}{1-\gamma}\right).$$

where $\bar{\sigma}$ is the steady-state value of $\sigma$ and $\bar{c} = \phi\bar{\sigma}^{-\xi}$ is the steady-state value of the maintenance cost.

### 18.7.3 Calibration and quantitative results

Parameters $\beta$, $\rho$, $\sigma_\varepsilon$, $\chi$, $b$, $\gamma$, and $\eta$ are set at the same value as in the previous section. For $\sigma$, we set the other parameters so that $\bar{\sigma} = 0.034$ matches the average separation rate in the U.S. data.

The newly-introduced specification is the maintenance cost function, $c(\sigma) = \phi\sigma^{-\xi}$. (18.32), (18.33), and $\lambda_w(\theta) = \chi\theta^{1-\eta}$ implies the ratio of the standard deviations

$$\frac{std(\hat{\lambda}_w)}{std(\hat{\sigma})} = \frac{(1-\eta)(1+\xi)}{\eta}$$

Krusell, Mukoyama, Rogerson, and Şahin (2017, Table 8) indicates the ratio of the standard deviations for the $EU$ flow rate and the $UE$ flow rate is close to 1. Thus, for $\eta = 0.72$, we set $\xi = 1.6$.

As in Section 18.6.2, we set the steady-state value of $\theta$ as 1, and for a given $\kappa$, we can determine the steady-state value of $\phi$ from $\bar{\sigma} = 0.034$. Thus $c(\bar{\sigma})$ can be expressed in $\kappa$, and as in Section 18.6.2, we can set the value of $\kappa$ from the steady-state version of (18.31).

The result is in Table 18.5. The fluctuation of $u$ is still quantitatively very small compared to the data, although it is larger than the constant $\sigma$ case, thanks to the movement in $\sigma$.

| | | $u$ | $v$ | $v/u$ | $z$ |
|---|---|---|---|---|---|
| Standard Deviation | | 0.010 | 0.011 | 0.021 | 0.013 |
| Quarterly Autocorrelation | | 0.862 | 0.623 | 0.764 | 0.765 |
| Correlation Matrix | $u$ | 1 | −0.893 | −0.969 | −0.909 |
| | $v$ | — | 1 | 0.976 | 0.957 |
| | $v/u$ | — | — | 1 | 0.961 |
| | $z$ | — | — | — | 1 |

Table 18.5: Model statistics with endogenous $\sigma$

### 18.7.4 Rigid wages

Once again, we examine the situation where wages are rigid. Following the same steps as those in the Generalized Nash Bargaining case, it can be shown that

$$\frac{\kappa}{\lambda_f(\theta_t)} = \beta \mathbb{E}\left[ z_{t+1} - \bar{w} - c(z_{t+1}) + \frac{(1-\sigma(z_{t+1}))\kappa}{\lambda_f(\theta_{t+1})} \right]$$

characterizes the dynamics of $\theta_t$. Log-linearizing this equation, we obtain

$$\hat{\theta}_t = \left[ \frac{\kappa \bar{\theta}^{\eta} \eta}{\chi} \left( \frac{1}{\rho\beta} - (1-\bar{\sigma}) \right) \right]^{-1} \hat{z}_t.$$

It turns out that the obtained outcome is identical to (18.29). The terms with endogenous $c(z_t)$ and $\sigma(z_t)$, which are new elements here, exactly cancel out. The log-linearized equations for the dynamics of $\sigma(z_t)$ and $c(z_t)$, (18.33) and (18.34), are the same as in the Generalized Nash Bargaining case.

The results are in Table 18.6. The model can replicate the large fluctuations in unemployment. In fact, the fluctuations in $u$ are larger than in the data, suggesting that even a less extreme form of wage rigidity can generate substantial fluctuations in $u$ in this case.

| | | $u$ | $v$ | $v/u$ | $z$ |
|---|---|---|---|---|---|
| Standard Deviation | | 0.217 | 0.232 | 0.433 | 0.013 |
| Quarterly Autocorrelation | | 0.852 | 0.609 | 0.763 | 0.765 |
| Correlation Matrix | $u$ | 1 | $-0.854$ | $-0.960$ | $-0.901$ |
| | $v$ | — | 1 | 0.966 | 0.948 |
| | $v/u$ | — | — | 1 | 0.961 |
| | $z$ | — | — | — | 1 |

Table 18.6: Model statistics with endogenous $\sigma$ and fixed wages

## 18.8 Labor market frictions and the neoclassical growth model

As the final section of this chapter, we connect the DMP model to the main workhorse model of this book: the neoclassical growth model.[14] The important changes are (i) concave utility with an explicit consumption-saving problem and (ii) the use of capital (in addition to labor) in production. In addition, as discussed in the footnote 5 briefly, the earlier sections implicitly assume that all firms are owned by someone outside the economy. In this section, we consider a closed economy, and therefore the profit income from the firm ownership is made explicit.

### 18.8.1 The baseline model with Generalized Nash Bargaining

Imagine that there are consumers on the unit square. The mass of consumers is one. The consumers are indexed by $(i, j)$, where $i \in [0, 1]$ and $j \in [0, 1]$. The index $i$ indicates the family the consumer belongs to. The family $i$, therefore, has members (indexed by $j$) on a unit interval. Families are identical to each other, and therefore, we will consider the *representative family*. Within each family, the members insure each other. That is, although some members are employed and others are unemployed, the income is pooled at the family level, and each member consumes the same amount (here, we do not consider disutility from work). Thus we have families that are homogeneous, and the family members are identical within each family. Below we will consider the decision of the representative family. Because the consumption of each family member is identical, the "family head" only needs to think about the representative member of the family.

Assume that each family member's utility is (because we consider the representative

[14]The model in this section follows Krusell, Mukoyama, and Şahin (2010, Appendix O). The earlier papers incorporating the search and matching framework into the neoclassical growth model include Merz (1995) and Andolfatto (1996).

member of the representative family, we omit the indices $i$ and $j$ below)

$$\mathbb{E}_0\left[\sum_{t=0}^{\infty}\mathbf{U}(c_t)\right],$$

where $\mathbb{E}_0[\cdot]$ is the expectation taken at time 0, $c_t$ is and $\mathbf{U}(\cdot)$ is an increasing and concave period utility function.

The budget constraint for the family is

$$c_t + k_{t+1} = (1 + r_t - \delta)k_t + (1 - u_t)w_t + u_t b + d_t,$$

where $k_t$ is the capital stock holding, $r_t$ is the rental rate of capital, $\delta \in (0, 1]$ is the depreciation rate of capital stock, $u_t$ is the unemployment rate, $w_t$ is the wage per worker, $b$ is the home production of unemployed workers, and $d_t$ is the dividend from the firm.

The labor market setting is the same as in Section 18.4. In this section, we assume that production uses both capital and labor. Normalizing the labor input per match as 1, the output per match is assumed to be $z_t k_t^{\alpha}$, where $\alpha \in (0, 1)$. We assume that the capital is rented by firms from the families every period. Thus, the maximization problem for the choice of capital input by the firm is

$$\max_{k_{f,t}} \; z_t(k_{f,t})^{\alpha} - r_t k_{f,t}.$$

The optimal capital input satisfies

$$\alpha z_t(k_{f,t})^{\alpha-1} = r_t.$$

In equilibrium, there is a mass $(1 - u_t)$ of matches, which equally divides the total capital stock in the economy. Therefore, $r_t$ in the equilibrium is

$$r(z_t, K_t, u_t) = \alpha z_t \left(\frac{K_t}{1 - u_t}\right)^{\alpha-1}.$$

Below, let us call $X_t \equiv (z_t, K_t, u_t)$ for the shorthand. The surplus per match (denoted by $z_t$ in Section 18.4) is now

$$y(X_t) \equiv z_t \left(\frac{K_t}{1 - u_t}\right)^{\alpha} - r(X_t)\left(\frac{K_t}{1 - u_t}\right) = (1 - \alpha)z_t \left(\frac{K_t}{1 - u_t}\right)^{\alpha}.$$

Firms also solve the dynamic problem of vacancy posting, as in the standard DMP model. Because the representative family's utility function is not linear, the discount factor can be different from $\beta$.

For the purpose of exposition, we divide the equilibrium of this model into two blocks: the consumption-saving block and the labor market block. The consumption-saving problem of the representative family can be written as the Bellman equation

$$\mathbf{V}(k, X) = \max_{c,k'} \; \mathbf{U}(c) + \beta\mathbb{E}[\mathbf{V}(k', X')|z] \tag{18.35}$$

subject to

$$c + k' = (1 + r(X) - \delta)k + (1 - u)w(X) + ub + d(X),$$
$$K' = \Omega(X), \tag{18.36}$$

and

$$u' = (1 - \lambda_w(\theta(X)) + \sigma(1 - u), \tag{18.37}$$

where prime ($'$) represents the next period variable. The family takes the rental rate $r(X)$, the wage $w(X)$, the dividend $d(X)$, the law of motion for aggregate capital $\Omega(X)$, and the labor-market tightness $\theta(X)$ as given (as functions of $X$). Later on, we will confirm that $w(X)$, $d(X)$, $\Omega(X)$, and $\theta(X)$ are functions of $X$. From the solution to this Bellman equation, we obtain the decision rules $c(k, X)$ and $k'(k, X)$. Because the families are identical, the equilibrium aggregate consumption is $C(X) = c(K, X)$ and the next period aggregate capital is $\Omega(X) = k'(K, X)$. Note that (18.37) implies we can express $u'$ as a function of $X$, $u'(X)$.

From this information, we can express the state price (i.e., the price of an Arrow security) of the next period state $z'$ when the current state is $X$ as

$$Q(z', X) = \beta f(z'|z)\frac{\mathbf{U}'(C(z', \Omega(X), u'(X)))}{\mathbf{U}'(C(X))}. \tag{18.38}$$

Here, $f(z'|z)$ is the probability density of state $z'$ given the current state $z$ and $u'(X)$ represents the right-hand side of (18.37). The derivation is in Appendix 18.A.5. In summary, once we know the functions $w(X)$, $d(X)$, and $\theta(X)$, we can obtain the state price $Q(z', X)$, in addition to other functions that include $\Omega(X)$. Below, we show that once we have $Q(z', X)$ and $\Omega(X)$, we can obtain $w(X)$, $d(X)$, and $\theta(X)$ in the "labor market block" below. Then these five functions $(Q(z', X), \Omega(X), w(X), d(X), \theta(X))$ can be computed as a fixed point.

Thus suppose we know $Q(z', X)$ and consider the labor market. It works very similarly to the basic DMP model. A firm with a worker has a value $J(X)$, where

$$J(X) = y(X) - w(X) + \int Q(z', X)[(1 - \sigma)J(X') + \sigma V(X')]dz'. \tag{18.39}$$

Here, we discount the future value with $Q(z', X)$ because it represents the price of the next period good (when the state is $z'$) in terms of the current good. The derivation of (18.39) (and the other asset value equations) can be found in Appendix 18.A.6. The value of vacancy is

$$V(X) = -\kappa + \int Q(z', X)[\lambda_f(\theta(X))J(X') + (1 - \lambda_f(\theta(X))V(X')]dz'.$$

Here, the transition equation (18.36) and (18.37) are given, and the functions $w(X)$ and $\theta(X)$ are a part of the unknowns in this block. The free-entry condition, $V(X) = 0$, implies

$$\frac{\kappa}{\lambda_f(\theta(X))} = \int Q(z', X)J(X')dz'. \tag{18.40}$$

This equation is analogous to (18.10) in the basic DMP model.

On the worker side, from the family's viewpoint, a worker brings in a stream of income with a stochastically changing employment state. Thus, we can compute the value of having a worker with specific status for a family using the standard asset pricing theory (the "Lucas tree" model). The value of an employed worker is

$$W(X) = w(X) + \int Q(z', X)[(1-\sigma)W(X') + \sigma U(X')]dz' \tag{18.41}$$

and the value of an unemployed worker is

$$U(X) = b + \int Q(z', X)[\lambda_w(\theta(X))W(X') + (1-\lambda_w(\theta(X)))U(X')]dz'. \tag{18.42}$$

Because $J(X) - V(X)$ and $W(X) - U(X)$ are both linear in $w$, with the same procedure as in Section 18.4, the Generalized Nash Bargaining implies

$$(1-\gamma)(W(X) - U(X)) = \gamma(J(X) - V(X)). \tag{18.43}$$

The Generalized Nash Bargaining here implies that the wage is indeed a function of $X$.

From (18.39), (18.41), (18.42), and $V(X) = 0$,

$$W(X)-U(X)+J(X) = y(X)-b+\int Q(z', X)[(1-\sigma-\lambda_w(\theta(X)))(W(X')-U(X'))+(1-\sigma)J(X')]dz'$$

Using (18.43),

$$\frac{J(X)}{1-\gamma} = y(X) - b + \int Q(z', X)J(X')\frac{1-\sigma-\gamma\lambda_w(\theta(X))}{1-\gamma}dz'. \tag{18.44}$$

Moving one period forward, multiplying $Q(z', X)$ on both sides, integrating over $z'$, and using (18.40) yields the job creation condition:

$$\frac{\kappa}{(1-\gamma)\lambda_f(\theta(X))} = \int Q(z', X)\left[y(X') - b + \frac{1-\sigma-\gamma\lambda_w(\theta(X'))}{1-\gamma}\frac{\kappa}{\lambda_f(\theta(X'))}\right]dz'. \tag{18.45}$$

This condition confirms that the equilibrium $\theta$ can indeed be written as a function of $X$. From (18.39), (18.40), and (18.44), the wage can be solved as

$$w(X) = \gamma(y(X) - b) + b + \gamma\theta(X)\kappa. \tag{18.46}$$

The dividend is all firms' profit minus the vacancy cost. The number of filled jobs is $(1-u)$, and each job creates $y(X) - w(X)$ units of profit. The vacancy cost is $\kappa v = \kappa\theta(X)u$ because $\theta(X) = v/u$. Thus

$$d(X) = (1-u)(y(X) - w(X)) - \kappa\theta(X)u, \tag{18.47}$$

and once again, we confirm that $d(X)$ is a function of $X$.

As with the standard RBC models, there are several alternative methods to compute the equilibrium. The first is, as in the previous sections, to log-linearize the equilibrium conditions and solve for the coefficients.

The second method is to treat the equilibrium conditions as functional equations. For example, one method that can be employed is to first make a guess on $Q(z', X)$, then use the (18.45) to find the function $\theta(X)$ (one can use an iterative method—start from $\theta(X)$ on the right-hand side to obtain $\theta(X)$ in the left-hand side, etc.). Then we can compute $w(X)$ and $d(X)$ from (18.46) and (18.47). Using this information, the representative family's problem (18.35) can be solved using the standard techniques to solve the neoclassical growth models. Finally, $Q(z', X)$ is updated with (18.38). We iterate this process until convergence. The following simulation follows this latter method of computation.

The details of calibration and computation are in Appendix 18.A.7. Calibration is similar to Section 18.6.2. The only difference from Table 18.1 is that, in this model, the steady-state value of $y(X)$ is not 1. Thus we adjust the value of $b$ so that it is 0.4 times the steady-state value of $y(X)$. As in Section 18.6.2, $\kappa$ is endogenously calibrated. The utility is assumed to be a log function $\mathbf{U}(c) = \ln(c)$. The production function parameter $\alpha = 0.4$ as in the standard RBC model (Cooley and Prescott, 1995). The value of $\delta$ in Cooley and Prescott (1995) is 0.012 in quarterly frequency, and thus we set $\delta = 0.004$.

| | | $u$ | $v$ | $v/u$ | $z$ |
|---|---|---|---|---|---|
| Standard Deviation | | 0.005 | 0.017 | 0.022 | 0.015 |
| Quarterly Autocorrelation | | 0.819 | 0.688 | 0.755 | 0.763 |
| Correlation Matrix | $u$ | 1 | −0.831 | −0.899 | 0.089 |
| | $v$ | — | 1 | 0.991 | −0.071 |
| | $v/u$ | — | — | 1 | −0.078 |
| | $z$ | — | — | — | 1 |

Table 18.7: Model statistics with Generalized Nash Bargaining: labor market

Table 18.7 computes the labor market statistics as in the earlier sections. The results are overall in line with Table 18.3. The only noticeable difference is that $corr(z, u)$, $corr(z, v)$, and $corr(z, v/u)$ are close to zero (and the signs are different). The reason is that, in this section's model, the production per worker $y(X)$ is affected not only by $z$, but also by $k$ and $u$. In fact, the correlations of $u$, $v$, and $v/u$ with the labor productivity $y(X)$ are similar to these with $z$ in Table 18.3.

| | $Y$ | $C$ | $I$ | $L$ | $Y/L$ |
|---|---|---|---|---|---|
| Standard Deviation | 0.014 | 0.003 | 0.059 | 0.0004 | 0.014 |
| Correlation with $Y$ | 1 | 0.875 | 0.991 | 0.902 | 0.99992 |

Table 18.8: Model statistics with Generalized Nash Bargaining: business cycles

Table 18.8 computes the standard business cycle statistics that are typically computed in the Real Business Cycle (RBC) literature. Similar to the labor market statistics, all variables are aggregated to quarterly frequency, logged, and HP-detrended (with the smoothing parameter $\lambda = 1,600$). The business cycle properties are overall similar to the standard RBC model: all $C$, $I$, $L$, and $Y/L$ (here, $L$ is computed as $1-u$) are strongly procyclical, $I$ is more volatile than $Y$, and $C$ is less volatile than $Y$. The only significant difference is that $L$ fluctuates much less than $Y$. In this model, this outcome reflects the unemployment volatility puzzle in Section 18.6.

### 18.8.2 Rigid wages

Now consider the case with rigid wages. The consumer's problem is the same as the Generalized Nash Bargaining case, except that the wage is rigid. Different from Section 18.6.4, the output per worker $y(X)$ moves not only with $z$, but also with $k$ and $u$. Because the movement of $y(X)$ is relatively large, it turns out that the flow profit for the firm sometimes becomes negative. To maintain a positive profit, this time we assume the wage to be

$$\tilde{w}(X) = \max\{\bar{w}, y(X)\}.$$

In our simulation, in the majority of the periods, the wage remains $\bar{w}$.

The Bellman equation is

$$\mathbf{V}(k, X) = \max_{c,k'} \ \mathbf{U}(c) + \beta \mathbb{E}[\mathbf{V}(k', X')|z] \tag{18.48}$$

subject to

$$\begin{aligned} c + k' &= (1 + r(X) - \delta)k + (1-u)\tilde{w}(X) + ub + d(X), \\ K' &= \Omega(X), \\ u' &= (1 - \lambda_w(\theta(X)) + \sigma(1-u), \end{aligned}$$

The state price $Q(z', X)$ can, again, be computed as (18.38). The Bellman equation for the matched job is

$$J(X) = y(X) - \tilde{w}(X) + \int Q(z', X)[(1-\sigma)J(X') + \sigma V(X')]dz'. \tag{18.49}$$

The free-entry condition remains the same as (18.40), and thus (18.49) can be rewritten as

$$\frac{\kappa}{\lambda_f(\theta(X))} = \int Q(z', X) \left[ y(X') - \tilde{w}(X) + \frac{(1-\sigma)\kappa}{\lambda_f(\theta(X'))} \right] dz'. \tag{18.50}$$

Similar to (18.47), the dividend is

$$d(X) = (1-u)(y(X) - \tilde{w}(X)) - \kappa\theta(X)u, \tag{18.51}$$

The computation is similar to the Generalized Nash Bargaining case, except that now $\tilde{w}(X)$ does not move as much. First make a guess on $Q(z', X)$. Second, we can solve for $\theta(X)$ from

| | | $u$ | $v$ | $v/u$ | $z$ |
|---|---|---|---|---|---|
| Standard Deviation | | 0.083 | 0.269 | 0.339 | 0.015 |
| Quarterly Autocorrelation | | 0.818 | 0.671 | 0.744 | 0.763 |
| Correlation Matrix | $u$ | 1 | $-0.811$ | $-0.886$ | 0.094 |
| | $v$ | — | 1 | 0.990 | $-0.076$ |
| | $v/u$ | — | — | 1 | $-0.083$ |
| | $z$ | — | — | — | 1 |

Table 18.9: Model statistics with rigid wages: labor market

(18.50). Third, $d(X)$ can be computed from (18.51). Using these information, we can solve (18.48) and update $Q(z', X)$. These steps are repeated until $Q(z', X)$ converges.

The model calibration is the same as in Section 18.8.1. Table 18.9 describes the labor market statistics for the rigid wage case. As in Section 18.6.4, the response of $v$ (and therefore $u$) to the productivity shock is magnified by the rigid wage. Similarly to Table 18.7, the correlations of $u$, $v$, and $v/u$ with $z$ are weak. Once again, this result comes from the fact that $y(X)$ also moves with $k$ and $u$. When we compute the correlation of these variables with $y(X)$, the pattern of correlations is similar to the results in the basic model.

| | $Y$ | $C$ | $I$ | $L$ | $Y/L$ |
|---|---|---|---|---|---|
| Standard Deviation | 0.017 | 0.003 | 0.060 | 0.007 | 0.011 |
| Correlation with $Y$ | 1 | 0.792 | 0.989 | 0.898 | 0.964 |

Table 18.10: Model statistics with rigid wages: business cycles

Table 18.10 lists the business cycle statistics. The results are very similar to Table 18.8 in the previous section. The exception is that the standard deviation of $L$ is one order of magnitude larger, reflecting the larger variability of unemployment.

## 18.9 Heterogeneity of jobs and the frictional job dispersion

So far, we have assumed that jobs are homogeneous, and workers always accept an offered job. In this section, we introduce a model where job offers are heterogeneous. Some jobs pay more than other jobs, and the kind of jobs offered to the worker is stochastic. It is assumed that every period, an unemployed worker can receive only one job offer. After receiving the offer, she decides whether to accept it. The model in this section is called the McCall search model (McCall, 1970) or simply the search model. The search model focuses on the worker's decision, and the demand side is simplified. Because the worker's decisions are operative margin, the labor supply side plays an active role.

The heterogeneity of job offers gives rise to wage dispersion. The labor market friction plays a crucial role; if there are no frictions, all workers will accept only the best (highest-paying) job. Even with labor market frictions, it is not trivial to think of a setting where firms actively offer different wage levels. A well-known example is called the *Diamond paradox*. Diamond (1971) has shown that, if the jobs are homogeneous and the worker has to leave the job to look for another job, all firms offer the workers' *reservation wages* (the lowest wage the worker would accept) even with a small search cost. It is easy to check that this outcome constitutes a Nash equilibrium: workers do not look for another job if they know all other firms are offering their reservation wage, and no firm would want to deviate. The Nash equilibrium is unique because, with any other wage offer distribution, the firm that offers the highest wage has an incentive to lower the wage slightly. In this section, instead of considering firms' wage setting behavior explicitly, we assume away the heterogeneity of wage offers. This type of firm behavior can be justified by relaxing Diamond's assumptions. For example, one can assume that the jobs are heterogeneous or workers can search on the job.

Formally, suppose that the worker is infinitely lived and the utility of the worker is linear:

$$\mathbb{E}_0 \left[ \sum_{t=0}^{\infty} \beta^t c_t \right].$$

An unemployed worker earns $b > 0$ every period. This income can be interpreted as home production, unemployment insurance benefit, or the value of leisure. Each worker receives one job offer every period. Job offers differ in terms of the wage $w$. It is stochastic with distribution function $F(w)$, which has a lower bound of 0 and upper bound $w^u$. The Bellman equation for an unemployed worker is therefore written as

$$U = b + \beta \int_0^{w^u} \max\{W(w), U\} dF(w), \tag{18.52}$$

where $U$ is the value of unemployment and $W(w)$ is the value of employment with wage $w$. The expression (18.52) can be rewritten as

$$(1-\beta)U = b + \beta \int_0^{w^u} \max\{W(w) - U, 0\} dF(w). \tag{18.53}$$

Every period, an employed worker faces the probability $\sigma \in (0,1)$ of losing her job. Therefore, the Bellman equation for an employed worker is

$$W(w) = w + \beta[(1-\sigma)W(w) + \sigma U]. \tag{18.54}$$

Equation (18.54) can be rewritten as

$$W(w) = \frac{w + \beta\sigma U}{1 - \beta(1-\sigma)}. \tag{18.55}$$

From this expression, we can see that $W(w)$ is increasing in $w$ and $W(0) = \beta\sigma U/(1-\beta(1-\sigma)) < U$. We assume that $w^u$ is sufficiently large so that $W(w^u) > U$. Therefore, there exists a threshold $w^*$ where $W(w^*) = U$, $W(w) > U$ for $w > w^*$, and $W(w) < U$ for $w < w^*$. In other words, the wage level $w^*$ is the worker's *reservation wage*. The value of $w^*$ characterizes the worker's choice in this model.

Plugging the expression (18.55) into (18.53) and using that $W(w) > U$ if and only if $w > w^*$,

$$(1-\beta)U = b + \frac{\beta}{1-\beta(1-\sigma)} \int_{w^*}^{w^u} [w - (1-\beta)U] dF(w). \tag{18.56}$$

Considering the expression (18.55) for $w = w^*$ and noting $W(w^*) = U$, we obtain

$$(1-\beta)U = w^*.$$

Thus (18.56) can be rewritten as

$$w^* = b + \frac{\beta}{1-\beta(1-\sigma)} \int_{w^*}^{w^u} [w - w^*] dF(w). \tag{18.57}$$

This equation solves the reservation wage $w^*$.

What can we learn from this model? First, let us consider the frequency of job acceptance. The worker only accepts the jobs that are better than $w^*$. Thus the *job finding probability* $\lambda$ is

$$\lambda = 1 - F(w^*).$$

The job finding probability is decreasing in $w^*$: when the workers are choosier, they find jobs less often.

We can also conduct various comparative statics to analyze how changes in parameters affect the reservation wage $w^*$ (and therefore $\lambda$). Rewrite (18.57) as:

$$\mathbf{G}(w^*, \beta, \sigma, b) = 0,$$

where

$$\mathbf{G}(w^*, \beta, \sigma, b) \equiv w^* - b - \frac{\beta}{1-\beta(1-\sigma)} \int_{w^*}^{w^u} [w - w^*] dF(w).$$

It is straightforward to show that $\partial\mathbf{G}/\partial\beta < 0$, $\partial\mathbf{G}/\partial\sigma > 0$, and $\partial\mathbf{G}/\partial b < 0$. For $w^*$, because (using Leibnitz's rule)

$$\frac{\partial}{\partial w^*} \int_{w^*}^{w^u} [w - w^*] dF(w) = -[1 - F(w^*)],$$

$\mathbf{G}(w^*, \beta, \sigma, b) > 0$. From the implicit function theorem, $w^*$ is increasing in $\beta$ and $b$ and decreasing in $\sigma$. Intuitively, the worker becomes choosier ($w^*$ becomes larger) when $\beta$ increases because the future gain from a better job has a higher weight compared to the opportunity loss from missing the immediate job. An increase in $b$ makes the unemployment state more

attractive and induces workers to wait longer. A higher $\sigma$ implies that even a good job won't last long, and thus it becomes less attractive to wait for a good job offer to arrive.

An interesting comparative-statics exercise with this class of model is to analyze the effect of changes in the wage offer distribution. First, consider the change in the average wage. To analyze the change in average, suppose that the wage offer is $w+\varepsilon$ instead of $w$ above (with the same distribution for $w$), and how the change in $\varepsilon$ changes the reservation wage $w^*+\varepsilon$ when evaluated at $\varepsilon = 0$. Now the $\mathbf{G}$ function is modified to

$$\mathbf{G}(w^*, \varepsilon) \equiv w^* + \varepsilon - b - \frac{\beta}{1-\beta(1-\sigma)} \int_{w^*}^{w^u} [(w+\varepsilon) - (w^*+\varepsilon)] dF(w).$$

$$\frac{\partial}{\partial \varepsilon} \mathbf{G}(w^*, \varepsilon) = 1$$

and (using Leibniz's rule, evaluated at $\varepsilon = 0$)

$$\frac{\partial}{\partial w^*} \mathbf{G}(w^*, \varepsilon) = 1 + \frac{\beta}{1-\beta(1-\sigma)} [1 - F(w^*)].$$

Thus

$$\frac{dw^*}{d\varepsilon} = -\frac{1-\beta(1-\sigma)}{1-\beta(1-\sigma)+\beta[1-F(w^*)]}.$$

The change in the reservation wage is, therefore,

$$\frac{d(w^*+\varepsilon)}{d\varepsilon} = \frac{dw^*}{d\varepsilon} + 1 = \frac{\beta[1-F(w^*)]}{1-\beta(1-\sigma)+\beta[1-F(w^*)]} \in (0,1).$$

Thus the reservation wage goes up, but not one-to-one. When the average wage offer goes up by one dollar, the reservation wage goes up by less than one dollar. This outcome arises because $b$ is kept constant. Because $b$ is the same, the relative attractiveness of the unemployment state (compared to working) goes down. This effect attenuates the effect of the change in the wage offer distribution.

Next, consider the dispersion of the wage offer distribution. To analyze the effect of dispersion, we first have to define the appropriate concept of the dispersion of wage offers in this context. Here, we introduce the concept of the *mean-preserving spread*. For a random variable $x$ with the distribution function $F(x)$, we can construct a random variable $\tilde{x} \equiv x+z$ where $z$ has a distribution function $H_x(z)$ and its mean is zero ($\int z dH_x(z) = 0$). Then, the mean of $x+z$ is $x$, and let us call the new distribution function $G(\tilde{x})$. Then we refer to $G(\cdot)$ as a mean-preserving spread of $F(\cdot)$. It can be shown (see Mas-Colell, Whinston, and Green, 1995, p. 198) that $G(\cdot)$ being a mean-preserving spread of $F(\cdot)$ is equivalent to

$$\int_0^x G(t)dt \geq \int_0^x F(t)dt \text{ for all } x. \tag{18.58}$$

Now let us rewrite the equation (18.57) as[15]

$$\begin{aligned} w^* &= b + \frac{\beta}{1-\beta(1-\sigma)} \left[ \int_0^{w^u} [w - w^*] dF(w) - \int_0^{w^*} [w - w^*] dF(w) \right] \\ &= b + \frac{\beta}{1-\beta(1-\sigma)} \left[ \mu_w - w^* - \int_0^{w^*} [w - w^*] dF(w) \right], \end{aligned} \tag{18.59}$$

where $\mu_w$ is the mean value of $w$. Integration by parts yields

$$\int_0^{w^*} [w - w^*] dF(w) = - \int_0^{w^*} F(w) dw$$

and thus (18.59) can be rewritten as

$$w^* - b - \frac{\beta}{1-\beta(1-\sigma)} \left[ \mu_w - w^* + \int_0^{w^*} F(w) dw \right] = 0. \tag{18.60}$$

It is straightforward to show that the left-hand side is increasing in $w^*$. Suppose that the distribution of $w$, $F(w)$, becomes more dispersed in the sense of the mean-preserving spread. The property (18.58) implies that the left-hand side of (18.60) goes down with this change in the distribution, and thus $w^*$ has to go up. The reservation wage increases with the dispersion of the wage offers. Intuitively, the worker becomes choosier with more dispersed wage offers because the possibility of a good wage offer increases. With higher dispersion, the possibility of a bad offer also increases, but the left tail of the distribution does not matter because these offers are rejected in any case. In other words, the option value of searching increases with the dispersion of the wage offer.

Now, consider the model implications for the realized wage dispersion. The equilibrium wage dispersion in this model is often called the *frictional wage dispersion* because all workers would work at $w = w^u$ if there are no search frictions (i.e., if all jobs are available to the workers). Workers accept a job with $w < w^u$ because it is costly to wait for high-wage job offers. To analyze the frictional wage dispersion, first define the mean (accepted) wage as

$$w^M \equiv \frac{\int_{w^*}^{w^u} w dF(w)}{1 - F(w^*)}.$$

Let us also define

$$\rho \equiv \frac{b}{w^M},$$

that is, the ratio of the unemployed worker's income to the mean wage. When $b$ is interpreted as the income from unemployment insurance, $\rho$ corresponds to the replacement rate. Further, as defined above, let $\lambda \equiv 1 - F(w^*)$ be the job-finding probability, that is, the probability that an unemployed worker transitions into employment.

[15] This derivation follows Ljungqvist and Sargent (2012, pp. 166-167).

Define the *mean-min ratio* of the (accepted) wages, $Mm$, as

$$Mm \equiv \frac{w^M}{w^*}.$$

Note that $w^*$ is the minimum value of the accepted wages. Then, using (18.57) and the above definitions, we can derive

$$Mm = \frac{1 + \beta\lambda/(1 - \beta(1 - \sigma))}{\rho + \beta\lambda/(1 - \beta(1 - \sigma))}.$$

A back-of-the-envelope calculation with (monthly) $\beta = 0.996$, $\sigma = 0.034$, $\lambda = 0.45$, and $nb = 0.4$ yields $Mm = 1.031$. The mean wage is only 3.1% larger than the lowest wage in this economy. In other words, search friction can explain only a tiny part of the observed wage dispersion in this model. This result is often referred to as the *frictional wage dispersion puzzle* in the literature. The reason is that, in the model with this parameterization, it is not very costly to wait for a new job (the offer comes fairly frequently), whereas the benefit of receiving a better wage offer is large. One situation where the frictional wage dispersion is high is, therefore, when the cost of unemployment is high. A large cost of unemployment makes unemployed workers accept low wage offers to avoid unemployment. Another situation is when there are possibilities of on-the-job search. For example, if the offered wage distributions are identical and the offer frequency is also the same between on- and off-the-job search, the worker would accept any job with $w > b$ when unemployed. In this case, the worker does not have to give up the option value of search when accepting a wage offer.

# Chapter 19

# Heterogeneous consumers

*Per Krusell and Víctor Ríos-Rull*

# Chapter 20

# Heterogeneous firms

*Toshihiko Mukoyama*

## 20.1 Introduction

In most macroeconomic models (and earlier in this textbook), it is assumed that there exists an aggregate production function

$$Y = F(K, L),$$

where $Y$ is the aggregate output, $K$ is the aggregate capital, and $L$ is the aggregate labor. This assumption is justified if the production functions for all firms are homogeneous. In reality, firms are heterogeneous in many dimensions. There are large and small firms, young and old firms, growing and contracting firms, and productive and unproductive firms.

Answering many macroeconomic questions requires explicitly taking firm heterogeneity into account. For example, how should we encourage (or discourage) the entry of new firms? How should we support growing firms? Should we worry about the growing prominence of "mega-firms"? What are the causes and consequences of the decline in firm entry and reallocation of resources across firms? Analysis based on the aggregate production function cannot answer these questions. It may also be the case that some economic policies may have different effects on the macroeconomy if firms have different degrees of heterogeneity.

In this chapter, we consider such questions. Looking at the data, we will see some indications that the prominence of large firms in the U.S. economy has been rising in recent years. The reallocation of resources through entry and exit of firms and expansion or contraction of firms seem to be slowing down in recent years (often referred to as the "decline in business dynamism"). These phenomena have potentially important consequences in the macroeconomic context. The rise of big firms may lead to an increase in their market power. The market power in the product and the labor market could be linked to market distortions and changes in labor share. The lack of reallocation of resources from unproductive firms to productive firms may lead to lower aggregate productivity due to "misallocation" (resource allocation that is suboptimal). Misallocation may also lead to a slower rate of innovation and aggregate productivity growth. The dominance of large firms may also have implications for other macroeconomic phenomena, such as business cycle fluctuations.

To answer these questions, we need to break out of the aggregate production function. This chapter covers the basic facts, models, and methods of analyzing an economy with heterogeneous firms.

## 20.2 A simple model

We start by considering a simple example where firm heterogeneity matters for macroeconomic analysis.[1] Suppose that there is a unit mass of firms. The firms produce a homogeneous good under perfect competition. The production function of firm $i$ (where $i$ is the index of firms: $i \in [0, 1]$) is

$$y_i = a_i F(\mathbf{x}_i)^\gamma,$$

[1] A similar framework is used by Hopenhayn (2014a). Some of the results below overlap with his.

where $y_i$ is the output of firm $i$ and $\gamma \in (0,1)$. Firms are heterogeneous in their productivity; $a_i$ is different across firms. $\mathbf{x}_i$ is the input vector for firm $i$. Assume that $F(\mathbf{x}_i)$ exhibits constant returns to scale. Then, because $\gamma < 1$, the overall production of $y_i$ exhibits decreasing returns to scale in inputs $\mathbf{x}_i$. The decreasing returns property is important. With constant returns, the firm(s) with the largest $a_i$ takes over the entire production of the economy, and the outcome is either (i) a monopoly or oligopoly of one or a few firms, which would contradict the perfect-competition assumption; or (ii) only the most efficient firms with common $a_i$ operate as price takers, which would return to the homogeneous-firms scenario. Let $\mathbf{X}$ be the endowment vector of inputs in the economy.

Due to the constant-returns property of $F(\mathbf{x}_i)$, we can solve the firm's problem in two steps: first, solve the cost-minimizing combination of inputs for one unit of $F(\mathbf{x})$. Second, decide the optimal scale of production. The first stage is common across firms:

$$\min_{\mathbf{x}} \ \mathbf{p}\mathbf{x}$$

subject to

$$F(\mathbf{x}) = 1,$$

where $\mathbf{p}$ is the vector of input prices. Let the solution of this problem be $\mathbf{x}^*$ and the minimized unit cost be $c \equiv \mathbf{p}\mathbf{x}^*$.

Let $m_i = F(\mathbf{x}_i)$ be the choice of the firm $i$'s combined inputs. The constant-returns property implies that the optimal input choice is $\mathbf{x}_i = m_i \mathbf{x}^*$ and the cost of production is $cm_i$. The second stage optimization problem is

$$\max_{m_i} \ a_i m_i^{\gamma} - c m_i. \tag{20.1}$$

The first-order condition to this problem is

$$a_i m_i^{\gamma - 1} = \frac{c}{\gamma}. \tag{20.2}$$

Therefore, $y_i = (c/\gamma) m_i$ for all $i$. Adding up for all $i$,

$$Y = \frac{c}{\gamma} M \tag{20.3}$$

holds, where

$$Y = \int y_i di \tag{20.4}$$

is the total output and

$$M = \int m_i di. \tag{20.5}$$

Note that, in equilibrium, $M = \int F(\mathbf{x}_i) di = F(\mathbf{X})$ has to hold. Let us define

$$A \equiv \left( \int a_i^{\frac{1}{1-\gamma}} di \right)^{1-\gamma}. \tag{20.6}$$

From (20.2),

$$A = \frac{c}{\gamma} M^{1-\gamma}$$

holds. Combining with (20.3) and $M = F(\mathbf{X})$,

$$Y = AF(\mathbf{X})^{\gamma}. \tag{20.7}$$

In this environment, this relationship can be viewed as the aggregate production function.[2] The heterogeneity of firms matter through the aggregation (20.6): the aggregate outcome is influenced by the distribution of $a_i$ to the extent that it yields different values of $A$ in (20.6).

To illustrate, suppose that $a_i$ follows a lognormal distribution, where $\ln(a_i) \sim N(\nu - \sigma^2/2, \sigma^2)$. From the property of the lognormal distribution, the average of $a_i$, $\int a_i di$, is $\exp(\nu)$. However, it can be computed that[3]

$$A = \exp\left(\nu + \frac{\gamma}{1-\gamma}\frac{1}{2}\sigma^2\right). \tag{20.8}$$

Therefore, the dispersion parameter $\sigma$ influences the level of $A$ even when the average productivity $\exp(\nu)$ is constant. This result holds because the productive resources are endogenously allocated: a productive firm uses more input than an unproductive firm and therefore has more presence in the aggregate production than merely having higher productivity. When the dispersion parameter $\sigma$ is larger, the economy has more room to allocate resources to the highly productive firms in the right tail. Allocation of inputs is the key to analyzing heterogeneous firms: when the inputs are not allocated optimally, aggregate productivity can be less than what can be achieved optimally. In this chapter, we always keep two questions in mind: (i) how the distribution of $a_i$ is determined, and (ii) how the economy allocates resources to different firms.

## 20.3 Firm heterogeneity in the data

This section describes some facts related to firm heterogeneity. We will focus on the U.S. data. The statistics here are all based on publicly-available data.[4] The first natural question is: how heterogeneous are the U.S. firms? Figure 20.1 shows the firm size distribution as the number of firms in each size category, as a fraction of the total number of firms. Here, the firm size is measured by the number of employees.

[2] An example of this aggregation is when the production function is $y_i = a_i(k_i^{\alpha}\ell_i^{1-\alpha})^{\gamma}$, where $k_i$ is firm $i$'s capital input, $\ell_i$ is the firm $i$'s labor input, and $\alpha \in (0,1)$. In this case, the aggregate production function is

$$Y = A(K^{\alpha}L^{1-\alpha})^{\gamma},$$

where $A$ is given by (20.6). The rental rate of capital in equilibrium is $r = \gamma\alpha A(K^{\alpha}L^{1-\alpha})^{\gamma-1}K^{\alpha-1}L^{1-\alpha}$, the wage rate is $w = \gamma(1-\alpha)A(K^{\alpha}L^{1-\alpha})^{\gamma-1}K^{\alpha}L^{-\alpha}$, and the unit cost of production is $c = (r/\alpha)^{\alpha}(w/(1-\alpha))^{1-\alpha}$.

[3] See Appendix 20.A.1 for derivation.

[4] All figures in this section is drawn from the U.S. Census Bureau's Business Dynamics Statistics. See `https://bds.explorer.ces.census.gov/`.

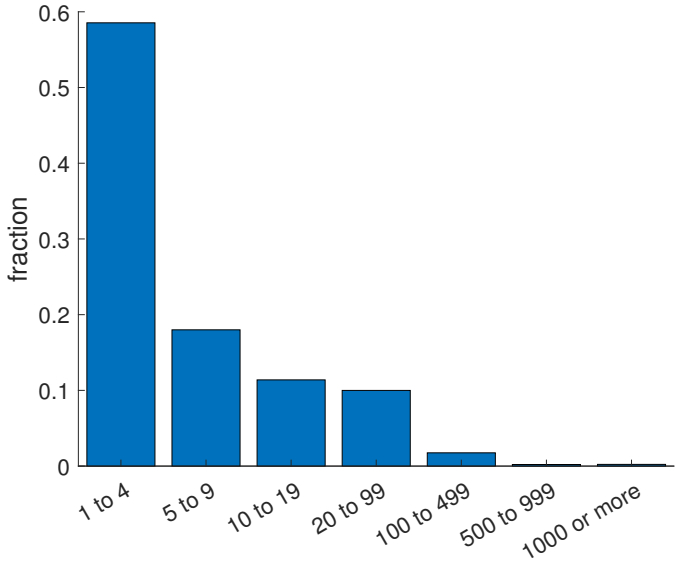



Figure 20.1: Distribution of firm size in 2019.

**Source**: Business Dynamics Statistics.

Figure 20.1 shows that the firm size distribution is quite dispersed. There are over 5 million firms in the U.S., and the majority are very small firms with 1 to 4 employees. At the same time, there are over 10,000 large firms with more than 1,000 employees, and over 1,000 firms with 10,000 employees.

The fact that very small firms account for the majority of firms does not imply that large firms are unimportant. Figure 20.2 plots the employment share of each size category. Approximately half of all employees work at the 1,000+ employment firms. In fact, approximately 30% of workers work at very large firms with 10,000+ employees.

Firm dynamics literature often uses data at the establishment level. An establishment is a fixed physical location where economic activity occurs; it is more straightforward to identify an establishment than a firm. A firm is a collection of establishments under common ownership, and it is often difficult to identify a firm in an administrative dataset. Establishments are also heterogeneous. Figure 20.3 is the establishment size distribution. There are over 7 million establishments in the U.S. economy, and approximately half are very small establishments with 1 to 4 employees.

Figure 20.4 plots the number of establishments that are owned by each firm size category. It shows that many establishments are owned by large firms (approximately 17% of all establishments are owned by the 1,000+ category firms). Whereas almost all "1 to 4" category firms own only one establishment, the firms in the 1,000+ category own 100 establishments on average. Very large firms with 10,000+ employees own about 600 establishments.[5]

[5]See Cao, Hyatt, Mukoyama, and Sager (2022) for a detailed analysis of the number of establishments

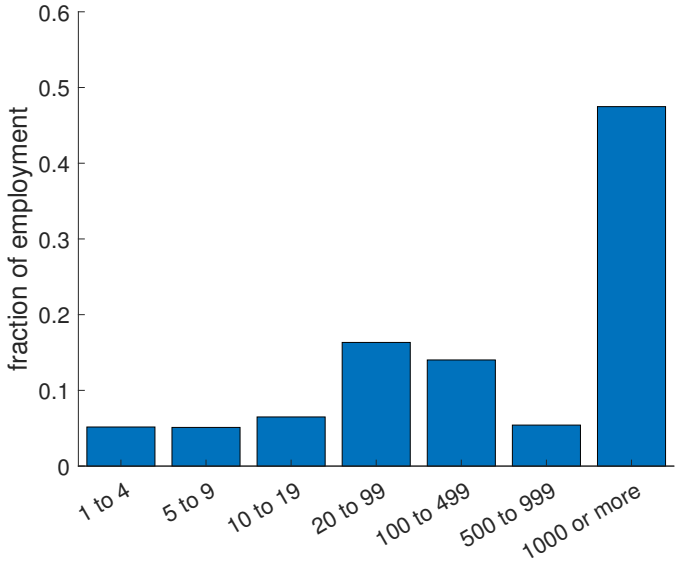



Figure 20.2: Employment share of each size category in 2019.

**Source**: Business Dynamics Statistics.

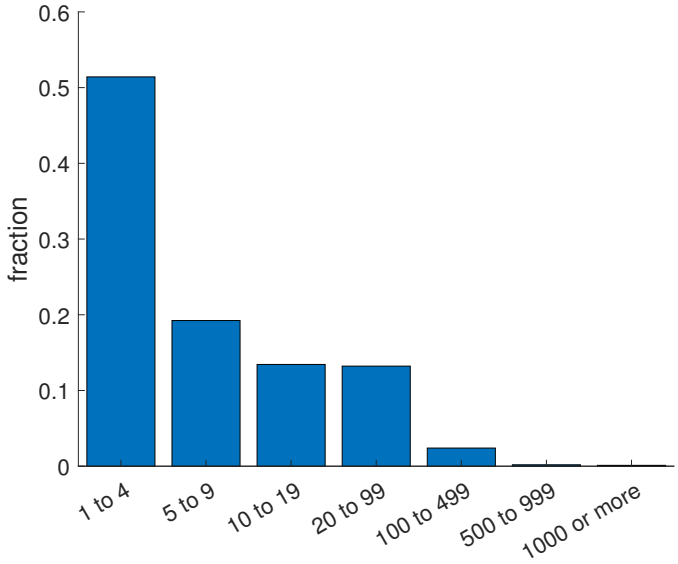



Figure 20.3: Distribution of establishment size in 2019.

**Source**: Business Dynamics Statistics.

per firm and its time-series properties.

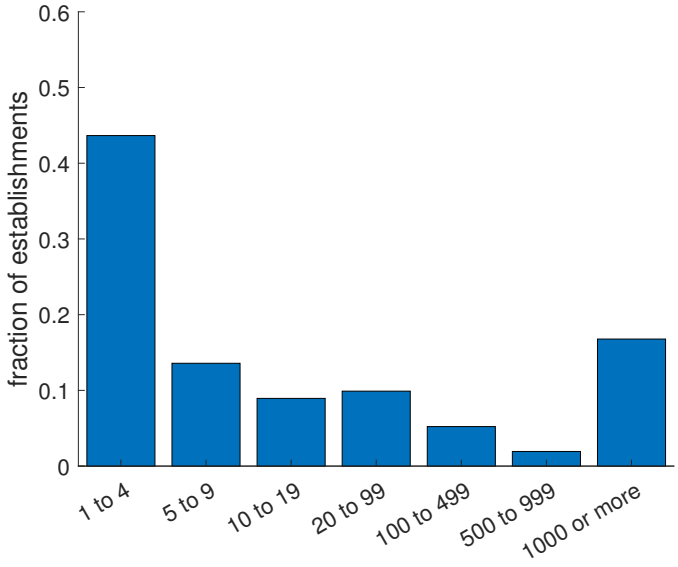



Figure 20.4: Fraction of establishments owned by each firm size category in 2019.

**Source**: Business Dynamics Statistics.

Aside from the cross-sectional heterogeneity, U.S. firms conduct significant adjustments over time. One measure of a firm's size adjustment is job creation and destruction. Job creation (JC) refers to the expansion of firms or establishments, and job destruction (JD) is the contraction of firms or establishments. BDS publishes the establishment-level JC and JD rates. The JC rate is defined as:

$$JC_t \equiv \frac{\sum_{i:\ell_{it}>\ell_{i,t-1}}(\ell_{it}-\ell_{i,t-1})}{\bar{L}_t}, \tag{20.9}$$

where $\ell_{it}$ is the employment of establishment $i$ at year $t$, $L_t$ is the total employment at year $t$ (which is the sum of $\ell_{it}$), $\bar{L}_t \equiv (L_t + L_{t-1})/2$. In words, the JC rate is the sum of employment increases in all expanding establishments, divided by the total employment (the average of time $t$ and $t-1$). The JD rate is similarly defined as:

$$JD_t \equiv \frac{\sum_{i:\ell_{it}<\ell_{i,t-1}}(\ell_{i,t-1}-\ell_{it})}{\bar{L}_t}.$$

The JD rate is the sum of the employment decrease by contracting establishments, divided by the total employment (the average of time $t$ and $t-1$). JC and JD, often called gross job flows, measure the magnitude of labor reallocation across establishments. Figure 20.5 plots the JC and JD rates from the BDS dataset. The shaded area is the recession period defined by the National Bureau of Economic Research (NBER).[6] Three properties are notable. First,

[6]See `https://www.nber.org/research/business-cycle-dating`.

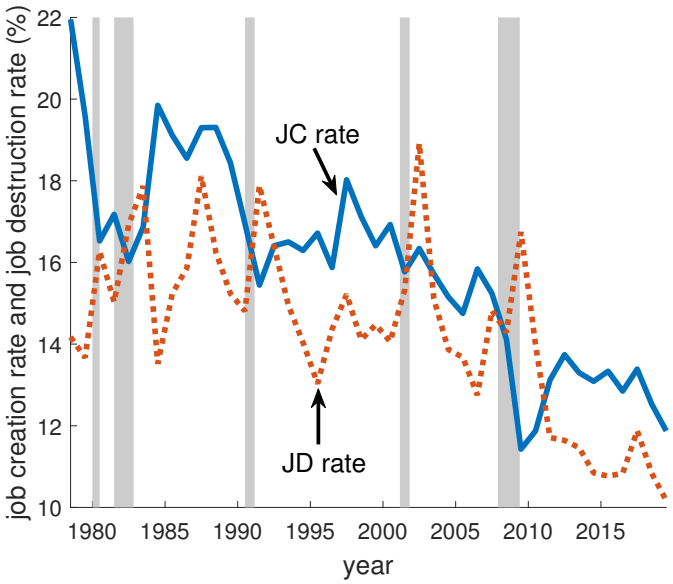



Figure 20.5: Annual job creation and destruction rates (establishment level).

**Source**: Business Dynamics Statistics.

the magnitude of JC and JD is large. Both the JC and JD rates are over 10% in any year. Second, both rates are cyclical. When a recession arrives, the JC rate declines and the JD rate increases. Third, there is a general declining trend in both the JC and JD rates. It is known that a wide range of indicators of reallocation, including the JC and JD rates, have declined in recent years. Some researchers call this trend the "declining business dynamism" of the U.S. economy.

Another measure of reallocation is the rate of entry and exit. Many firms and establishments enter and exit every year. Figure 20.6 plots the entry rate and exit rates of establishments. The entry rate is defined as the number of entering establishments between $t-1$ and $t$ divided by the total number of establishments (the average of time $t-1$ and $t$). The exit rate is defined as the number of exiting establishments between $t-1$ and $t$ divided by the total number of establishments (the average of time $t-1$ and $t$). One can see similar properties here as the JC and JD rates: the entry and exit rates are large, cyclical, and there are overall declining trends.

Over the last few decades, there have been significant changes in the heterogeneity among U.S. firms. In addition to the "declining dynamism" described above, one topic that caught researchers' attention is the dominance of large firms. Figure 20.7 plots the fraction of workers employed by firms in the 10,000+ size category.[7] This fraction has steadily increased since the early 1990s, indicating that large firms are starting to dominate the U.S. economy.

[7] In drawing this figure, the distinction between firms and establishments is very important. See Appendix 20.A.2.

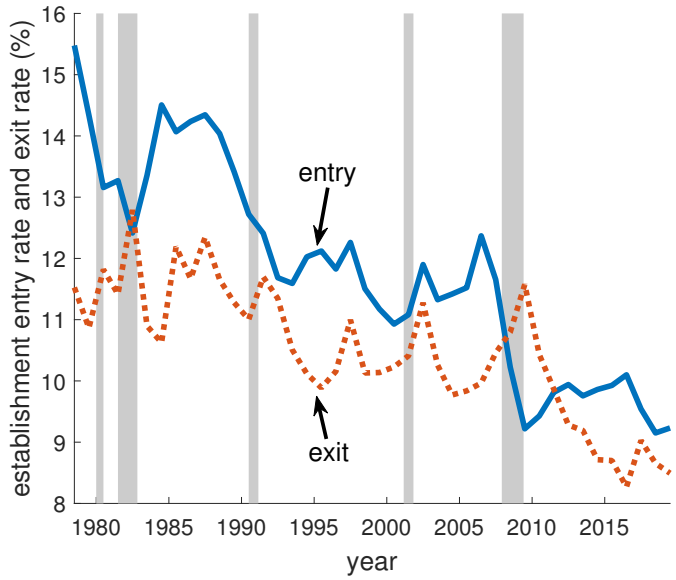



Figure 20.6: Annual establishment entry and exit rates.

**Source**: Business Dynamics Statistics.

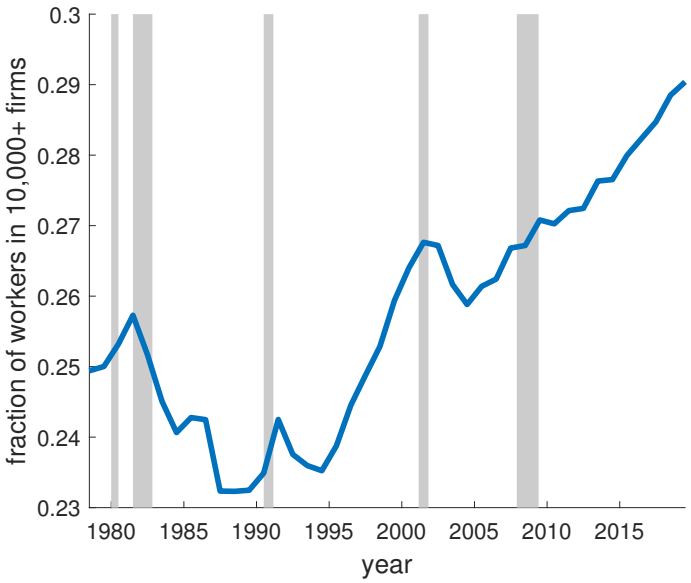



Figure 20.7: Fraction of employees working at 10,000+ employee firms.

**Source**: Business Dynamics Statistics.

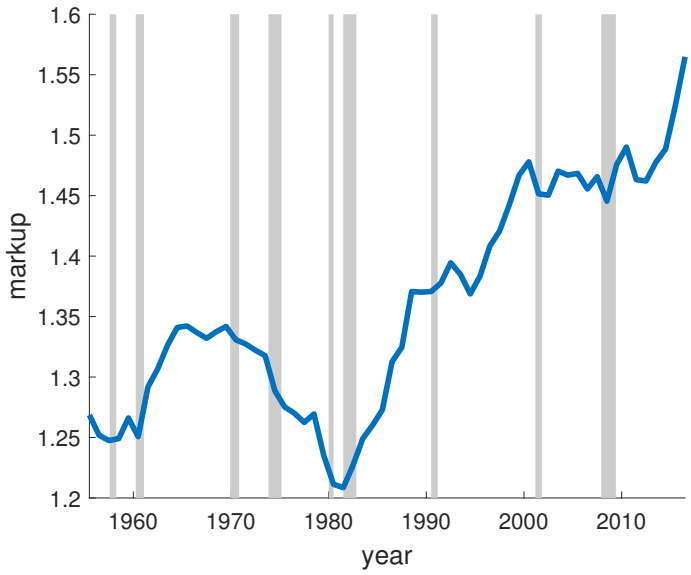



Figure 20.8: Markup of the U.S. public firms.

**Source**: De Loecker, Eeckhout, and Unger (2020).

This dominance raised concerns about the market power of large firms, and is consistent with another strand of research that tries to measure the trend of market power in the U.S. economy. Figure 20.8 reproduces the Figure 1 of De Loecker et al. (2020). It measures the trend of the average markup (i.e., price over the marginal cost) of the U.S. public firms in the Compustat dataset. The markup series exhibits increasing trend since the 1980s.

## 20.4 Reallocation and misallocation

The previous section shows that there is a large degree of reallocation among U.S. firms. How much does the reallocation matter for aggregate productivity? In Section 20.1, we have seen that firm heterogeneity affects the aggregate outcome through the endogenous allocation of inputs. In a dynamic economy where firm productivity changes over time, one can imagine that the constant reallocation of inputs can have an important impact on aggregate productivity.

Foster, Haltiwanger, and Krizan (2001) illustrate the quantitative impact of reallocation through the following simple accounting framework. Let us denote the productivity of establishment $i$ (they use establishment-level data and not firm-level data) at time $t$ be $a_{it}$. The (output-weighted) average productivity $\bar{A}_t$ is defined as

$$\bar{A}_t \equiv s_{it} a_{it},$$

where $s_{it}$ is the output share of establishment $i$. Then, by denoting the $x_t - x_{t-1}$by $\Delta x_t$,

$$\begin{aligned}\Delta \bar{A}_t = & \sum_{i\in C} s_{it-1}\Delta a_{it} + \sum_{i\in C}(a_{it-1} - \bar{A}_{t-1})\Delta s_{it} + \sum_{i\in C}\Delta a_{it}\Delta s_{it} \\ & + \sum_{i\in N} s_{it}(a_{it} - \bar{A}_{t-1}) - \sum_{i\in X} s_{it-1}(a_{it-1} - \bar{A}_{t-1})\end{aligned}$$

holds, where $C$ is the set of continuing establishments (establishments that exist in both time $t-1$ and $t$), $N$ is the set of new establishments (establishments that enter between time $t-1$ and $t$), and $X$ is the set of exiting establishments (establishments that exit between time $t-1$ and $t$). The increase in average productivity can occur for five distinct reasons. First, each of the existing establishments can increase its productivity. Second, an establishment whose productivity is higher than average can increase its share. Third, the first two effects can be magnified if both occur at the same time (i.e., a high-productivity establishment raises the share and its own productivity). Fourth, the entering establishment can be better than the average. Fifth, the exiting establishment can be worse than the average. All factors except for the first one can be interpreted as the contribution of reallocation. That is, if $\Delta s_{it} = 0$ and there are no entry and exit, the only way for the aggregate productivity to increase is for each establishment to increase its productivity. Using the U.S. Manufacturing data from 1977 to 1987, Foster et al. (2001) estimate (see their Table 8.4) that the aggregate change in multifactor productivity (the change in output that is not accounted for by the change in capital, labor, and intermediate goods) is 45% accounted for by the first factor, and the remaining 55% is the contribution of reallocation. This decomposition highlights the importance of reallocation in determining aggregate productivity growth.

Recently, a large body of literature has evaluated the role of various frictions that hinder the optimal allocation of resources. This literature emphasizes the existence of the *misallocation* of productive inputs as the source of low aggregate total factor productivity. A subset of literature, such as Restuccia and Rogerson (2008) and Hsieh and Klenow (2009), emphasizes firm-specific distortions as the sources of misallocation. To see how firm-specific distortions can affect aggregate productivity, consider the model of Section 20.1 and add an assumption that the government taxes the output of firm $i$ at the rate of $\tau_i$. Thus, instead of the problem (20.1), the firm solves

$$\max_{m_i} (1-\tau_i)a_i m_i^{\gamma} - cm_i.$$

The rest of the model is the same, and the GDP is still measured as $Y = \int y_i di$. After going through similar steps as in Section 20.1, one can show that the aggregate production function still takes the form of (20.7), but $A$ is modified to

$$A = \frac{\int a_i^{\frac{1}{1-\gamma}}(1-\tau_i)^{\frac{\gamma}{1-\gamma}} di}{\left(\int a_i^{\frac{1}{1-\gamma}}(1-\tau_i)^{\frac{1}{1-\gamma}} di\right)^{\gamma}}. \tag{20.10}$$

One can easily see that this $A$ is identical to (20.6) when $\tau_i = 0$ for all $i$. Now, suppose that $a_i$ and $(1-\tau_i)$ follow a bivariate lognormal distribution. In particular, $(\ln(a_i), \ln(1-\tau_i)) \sim$

$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = (\nu_a - \sigma_a^2/2, \nu_\tau - \sigma_\tau^2/2)$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_a^2 & \rho\sigma_a\sigma_\tau \\ \rho\sigma_a\sigma_\tau & \sigma_\tau^2 \end{bmatrix}.$$

Plugging this formulation into (20.10), we obtain[8]

$$A = \exp\left(\nu_a + \frac{\gamma}{1-\gamma}\frac{1}{2}(\sigma_a^2 - \sigma_\tau^2)\right). \tag{20.11}$$

One can easily see that this expression is identical to (20.8) when $\sigma_\tau = 0$. A large dispersion in $(1-\tau_i)$ is detrimental to aggregate productivity. This result is because, when $(1-\tau_i)$ is dispersed, some productive firms do not expand, because $(1-\tau_i)a_i$ is low even when $a_i$ is large. At the same time, some unproductive firms employ a large amount of input because $(1-\tau_i)a_i$ is large even though $a_i$ is small. Note that, in this setting, $\nu_\tau$ does not influence aggregate productivity because it does not distort the allocation of input across firms, and the total supply of inputs is fixed.

Another subset of literature examines various specific policies and institutions that cause misallocation. Examples are size-specific taxes and regulations, entry regulations, and firing and hiring regulations.

## 20.5 Firm heterogeneity in general equilibrium

When firms are forward-looking, frictions for reallocation have further effects through the firms' behavior. Hopenhayn and Rogerson (1993) highlight this mechanism in the context of firing taxes and quantify the outcome in a general equilibrium framework. In the following, we introduce the Hopenhayn and Rogerson (1993) framework with slightly different notations. In addition to the experiments on firing taxes (replicating the Hopenhayn and Rogerson (1993) exercises), this section conducts experiments with entry barriers.[9]

### 20.5.1 Setup

There is a continuum of firms in the economy. We focus on the steady-state where the prices and aggregate quantities (employment, output, and the number of firms) are constant over time. In this section, we omit the firm's index $i$ when there is no risk of confusion. Each firm uses only labor $\ell_t$ as an input. Firms behave competitively and maximize the

[8]See Appendix 20.A.3 for derivation. Hsieh and Klenow (2009) obtain a similar expression. This case is special in that $\rho$ does not appear in the expression for aggregate productivity. In general, correlation between $(1-\tau_i)$ and $a_i$ matter for the aggregate productivity, and Restuccia and Rogerson (2008) emphasize the importance of such correlation. See Hopenhayn (2014b) for related discussions.

[9]The analysis of entry barriers is not in Hopenhayn and Rogerson (1993) but is subsequently conducted by, for example, Moscoso Boedo and Mukoyama (2012).

profit with wage $w_t$. The firm's production function is $y_t = a_t \ell_t^\gamma$, where $a_t$ is (exogenous) idiosyncratic productivity. In addition to wages, firms have to pay $c_f$ units of goods as the fixed operation cost every period. The firing taxes imposed by the government take the form of $\tau \max(0, \ell_{t-1} - \ell_t)$, where $\tau > 0$ is the firing tax for dismissing one worker. The government transfers all firing taxes back to the representative consumer. Therefore, the firm's flow profit is

$$\pi(\ell_{t-1}, \ell_t, a_t) = a_t \ell_t^\gamma - w_t \ell_t - c_f - \tau \max(0, \ell_{t-1} - \ell_t).$$

Note that the output price is normalized to 1, and the only endogenous price in each period is $w_t$.

The timing for the firms within a period is as follows. At the beginning of each period, the incumbent firm from the last period decides whether to exit. If the firm exits, it pays the firing cost $\tau \ell_{t-1}$. If it stays, it receives the current period value of $a_t$ from the stochastic process

$$\ln(a_t) = \alpha + \rho \ln(a_{t-1}) + \varepsilon_t,$$

where $\alpha$ and $\rho \in [0, 1)$ are parameters and $\varepsilon_t \sim N(0, \sigma^2)$. After observing $a_t$, the firm decides the employment and produce.

Note that unlike the model in Section 20.1, the firm's employment decision is dynamic in the presence of a positive firing tax. When the firm decides to hire a worker, it foresees that it has to pay the firing cost when it wants to shed workers in the future due to negative productivity shock. This effect makes the firm reluctant to hire a worker when it receives a positive productivity shock. The dynamic programming problem for the firm is

$$W(a, \ell_{-1}) = \max_{\ell} \pi(\ell_{-1}, \ell, a) + \beta \max(\mathbb{E}[W(a', \ell)|a], -\tau \ell), \tag{20.12}$$

where the subscript $-1$ represents the previous period value and prime ($'$) represents the next period value. $\beta \in (0, 1)$ is the consumer's discount factor (which is equal to the firm's discount factor) and $\mathbb{E}[\cdot|a]$ represents the expected value given $a$.

We assume free entry; that is, anyone can enter as long as the entry cost is paid. After the entry cost is paid, the firm draws productivity, employs workers, and produces. The entry cost is assumed to be $c_e + \kappa$, where $c_e$ is the technological entry cost, including the investment required when enter, and $\kappa$ is additional (wasteful) policy-related cost that we interpret as "entry barriers." Free entry implies

$$W^e = c_e + \kappa, \tag{20.13}$$

where $W^e$ is the value of the entry that satisfies

$$W^e = \int (W(a, 0) + c_f) d\nu(a),$$

where $\nu(a)$ is the exogenous distribution of $a$ for a new entrant. Note that we assume entrants do not have to pay the fixed operation cost $c_f$.

The representative consumer owns the firms, works, and consumes. The utility is

$$\sum_{t=0}^{\infty} \beta^t [u(C_t) - \chi L_t^s],$$

where $u(C_t)$ is increasing and concave utility from consumption $C_t$, $\chi > 0$ is a parameter, and $L_t^s$ is the labor supply. In the steady-state, the consumer's problem is static:

$$\max_{C, L^s} \; u(C) - \chi L^s$$

subject to

$$C \leq wL^s + \Pi + R,$$

where $\Pi$ is the total profit of the firms and $R$ is the total transfer. The first-order condition is

$$wu'(wL^s + \Pi + R) = \chi. \tag{20.14}$$

Therefore, the labor supply is a function of $w$ and $\Pi + R$.

As Kaas (2021) points out, the competitive equilibrium of this model has a structure that is often called "block recursive." That is, the equilibrium price (wage in this model) can be computed without the information on the distribution of state variables across incumbent firms. To see this, note that $W(a, \ell_{-1})$ can be computed from (20.12) once the value of $w$ is known. Thus, $W^e$ can be computed as a function of $w$. The free-entry condition (20.13) can be used to pin down the equilibrium $w$. For a given mass of entry $M$, the decision rule of (20.12) can be used to compute the stationary distribution of firms across different state variables $a$ and $\ell_{-1}$. With the stationary distribution, one can compute the total labor demand $L^d$, the total profit $\Pi$, and the total firing tax $R$. All $L^d$, $\Pi$, and $R$ are functions of the entry mass $M$. Thus, the labor supply equation (20.14) can be used to determine the level of entry mass $M$ that is consistent with the labor market equilibrium $L^d = L^s$.

Hopenhayn and Rogerson (1993) calibrate the model with $\tau = 0$ to the U.S. economy and examine the effect of $\tau$ quantitatively. The model here is identical to theirs except that (i) some notations are different, and (ii) they normalize the wage as 1 and the market equilibrium determines the product price $p$, which corresponds to $1/w$ in our notation. We also set the baseline $\kappa = 0$.

The calibration procedure follows Hopenhayn and Rogerson (1993). One period is set at five years. First, the functional form of the utility function for consumption is assumed to be natural log: $u(c) = \ln(c)$. Some parameters are set ex-ante. The discount factor is set at $\beta = 0.8$, corresponding to the value of 4% per year. The production function parameter $\gamma$ is 0.64, corresponding to the labor share.

To set other parameters, we assume that the $(\tau, \kappa) = (0, 0)$ case corresponds to the U.S. economy and find the parameter values so that various statistics from the model-generated data match the corresponding data moments. The parameters for the productivity process are set using the property of the model that the property of the productivity shock is directly reflected in the firm's employment decision. By using the plant-level data from the U.S.

manufacturing, $\alpha = 0.076$ so that the average size of the firms is 61.7 (the actual model moment is 62.4), $\rho = 0.93$ so that the autocorrelation of $\ln(\ell)$ is 0.93, and $\sigma = 0.253$ so that the variance of the growth rate for $\ell$ is 0.53. The operation cost $c_f = 18.0$ so that the exit rate is 37% (the actual model moment is 34%). The entrant's productivity distribution $\nu$ is set so that the size distribution of young firms matches the U.S. data. The entry cost $c_e = 9.04$ so that the free-entry condition holds with $w = 1$. The disutility of working $\chi$ is set so that the steady-state labor supply $L$ is 0.6.

### 20.5.2 The effects of firing taxes

| | $\tau = 0$ | $\tau = 0.1$ | $\tau = 0.2$ |
|---|---|---|---|
| Wage | 1.000 | 0.977 | 0.957 |
| Total output | 100 | 97.7 | 95.7 |
| Total employment | 100 | 98.3 | 97.4 |
| Labor productivity | 100 | 99.4 | 98.3 |
| $JC(= JD)$ rate | 0.28 | 0.25 | 0.21 |

Table 20.1: Model results with firing taxes

Table 20.1 summarizes the steady-state outcomes of the model with $\tau = 0$, $\tau = 0.1$, and $\tau = 0.2$, keeping $\kappa = 0$. Because one period is assumed to be five years and the period wage (earnings per worker) with $\tau = 0$ is 1, $\tau = 0.1$ corresponds to six months' salary of a worker.[10] For total output, total employment, and labor productivity, $\tau = 0$ case is normalized to 100.[11]

There are several important points to note. First, it is not a priori obvious whether the equilibrium employment $L$ goes up or down when $\tau$ increases. The reason is that the effect on firing (the firms fire less because of the taxes) brings $L$ up, whereas the effect on hiring (the firms do not hire much even with a positive $a$ shock, given that, in the future, the firm may have to fire these extra workers) brings $L$ down. Which one dominates is a quantitative question; here, the latter effect dominates, and $L$ decreases when $\tau$ increases to $\tau = 0.1$ and $\tau = 0.2$.

Second, the productivity $Y/L$ declines with $\tau$. The reason is the misallocation mentioned in Section 20.4. Because a firm with a good $a$ shock does not expand as much as the first best, and a firm with a bad $a$ does not fire as much with the firing tax, labor is not allocated properly across firms. These incentives imply that the marginal product of labor is dispersed (in the first best, the marginal product of labor is equalized). The difference from Section

[10] Moscoso Boedo and Mukoyama (2012) computes the costs of business regulations corresponding to $\tau$ that explicitly shows up in the World Bank's Doing Business dataset. The cross-country median of $\tau$ is about eight months of annual wages, and the average of low-income countries is 1.2 times the annual wages.

[11] Although the calibration is the same as Hopenhayn and Rogerson (1993), the numbers are not exactly the same. The reason for the discrepancy is likely from detailed differences in computation.

20.4 is that the misallocation stems from the firm's dynamic decisions, especially for hiring. Firing and exiting behaviors are also affected by dynamic considerations. In the general equilibrium, misallocation also affects the wage level and entry of firms.

Third, the job creation ($JC$) rate, defined as (20.9), decreases with $\tau$.[12] The reallocation of labor across firms is reduced because of the reluctance to hire and fire described above. The lack of reallocation is, therefore, closely linked to productivity loss due to misallocation.

### 20.5.3 The effects of entry barriers

| | $\tau = 0$ | $\kappa = 0.5$ | $\kappa = 5.0$ |
|---|---|---|---|
| Wage | 1.000 | 0.986 | 0.879 |
| Total output | 100 | 98.6 | 87.9 |
| Total employment | 100 | 99.5 | 96.2 |
| Labor productivity | 100 | 99.1 | 91.4 |
| $JC(= JD)$ rate | 0.28 | 0.28 | 0.28 |

Table 20.2: Model results with entry barriers

Table 20.2 describes the model outcome for $\kappa = 0$, $\kappa = 0.5$, and $\kappa = 5.0$, keeping $\tau = 0$.[13] As in Table 20.1, for total output, total employment, and labor productivity, $\kappa = 0$ case is normalized to 100.

Entry barriers have a substantial effect on the model outcome. A higher cost of entry implies that the value of firms has to be higher in equilibrium (see equation (20.13)), and a high value of firms implies that the equilibrium wage has to be low. A low wage affects labor productivity through three channels. First, a low wage implies that low-productivity firms are less likely to exit. This effect pushes down aggregate productivity. Second, a low exit rate means that the entry rate is also low in the steady state. Because the entrants are of lower productivity than incumbents, this effect increases aggregate productivity. Finally, the size of incumbents is larger with lower wages, and because of the decreasing returns to scale, a large scale implies lower productivity. The first and third effect pushes down the aggregate productivity, and the second effect pushes up; the lowering force dominates quantitatively.

The outcome of this exercise also highlights a heterogeneity in policy effects across different firms. Whereas the entry barriers harm entry, it increases the value of incumbent firms.

[12] Here, because the economy is in the steady state, the job creation rate is equal to the job destruction rate.

[13] Moscoso Boedo and Mukoyama (2012) measures the costs of entry regulations corresponding to $\kappa$ in the World Bank's Doing Business dataset. The cross-country median value of $\kappa$ is 3.4 times the annual wages (about 0.7 times the five-year wages), and the average of the low-income countries is 29.9 times the annual wages (corresponding to 6 times the five-year wages. Note that although $\kappa$ in the current model is not in terms of annual wages, the baseline annual wage is set at 1.0, and thus the units are comparable.

When thinking about the policies on entry, there is an important conflict of interest between incumbent firms and potential entrants.[14]

## 20.6 Alternative market arrangements

The above discussions have assumed that all markets are perfectly competitive. We have seen in earlier chapters that many macro models consider alternative market arrangements. This section introduces two alternative market arrangements with market power in the context of firm heterogeneity. There are two takeaways from this section. First, the insights on misallocation in Section 20.4 goes through with minor modifications. Second, the inclusion of market power allows us to analyze how firm heterogeneity interacts with other macro variables of interest, such as the aggregate level of markups.

### 20.6.1 Monopolistic competition

A popular alternative formulation in the macroeconomic context is monopolistic competition. In this setting, firms produce differentiated goods, and only one firm produces each good.

A popular setting considers two types of goods, the *final good* and *intermediate goods*. The final good is produced in a perfectly-competitive sector with constant returns to scale technology:

$$Y = \left[ \int y_i^{\frac{\sigma-1}{\sigma}} di \right]^{\frac{\sigma}{\sigma-1}}. \tag{20.15}$$

where $\sigma$ is the elasticity of substitution parameter. We assume $\sigma > 1$ so that the monopoly problem of each intermediate-good producer is well-defined. The aggregate $Y$ in (20.15) can alternatively be considered as the utility by a consumer. This aggregation (20.15) is sometimes called the Dixit-Stiglitz utility function.

Consider the setting in Section 20.1, except that (20.15) replaces (20.4) and that each good $i$ is now monopolistically produced by an intermediate-good producer $i$. The intermediate-good producer's production structure is the same as in Section 20.1. The same inputs are used by the intermediate-good producers, and the input market is perfectly competitive. The aggregation of input (20.5) remains the same.

First, consider the cost-minimization problem of the final good producer:

$$\min_{\{y_i\}} \int p_i y_i di$$

subject to (20.15) for a given $Y$. Letting the Lagrange multiplier of the constraint be $\lambda$, the first-order condition is

$$p_i = \lambda y_i^{-\frac{1}{\sigma}} Y^{\frac{1}{\sigma}}, \tag{20.16}$$

which implies

$$\int p_i y_i di = \lambda Y$$

[14] See Mukoyama and Popov (2014) for a politico-economic analysis of policies on firm entry.

and thus $\lambda$ represents the (minimum) cost of producing one unit of the final good. Because the final-good market is perfectly competitive, $\lambda$ is also the price of the final good. Let us call this price $P$. From (20.16),

$$P = \left[ \int p_i^{1-\sigma} di \right]^{\frac{1}{1-\sigma}}$$

holds. As in Section 20.1, normalize the final-good price to be 1. Therefore, $P = \lambda = 1$ and the inverse demand function for good $i$ is, from (20.16),

$$p_i = y_i^{-\frac{1}{\sigma}} Y^{\frac{1}{\sigma}}. \tag{20.17}$$

Keep in mind here that $Y$ here is a shorthand that represents the production of all other goods in (20.15).

The monopolistic producer $i$ maximizes profit given the inverse demand function (20.17) and its production function. The problem, which corresponds to (20.1) in Section 20.1, is

$$\max_{m_i} \ (a_i m_i^{\gamma})^{-\frac{1}{\sigma}} Y^{\frac{1}{\sigma}} a_i m_i^{\gamma} - c m_i. \tag{20.18}$$

In the Nash equilibrium, each intermediate good firm $i$ has to solve this problem given the other firms' input choices: $m_j$ for all $j \neq i$. The important assumption in the monopolistic competition is that each firm is small compared to the aggregate so that it ignores the effect of $m_i$ (therefore $y_i$) on $Y$. Therefore, each firm takes $Y$ as given. With a similar step as in Section 20.1, we can obtain the relationship

$$Y = \frac{c}{\gamma} \frac{\sigma}{\sigma - 1} M$$

instead of (20.3), and the same expression for the aggregate production function (20.7) where $A$ is now modified to

$$A \equiv \left( \int a_i^{\frac{1}{\frac{\sigma}{\sigma-1} - \gamma}} di \right)^{\frac{\sigma}{\sigma-1} - \gamma}$$

instead of (20.6). Note that we have $\sigma/(\sigma - 1)$ instead of 1, reflecting that each firm faces another factor (in addition to the decreasing returns to scale) that limits firm size. One important difference between this formulation from the basic model in Section 20.1 is that we can now accommodate constant returns to scale (or even some increasing returns to scale).

### 20.6.2 Oligopoly and endogenous markups

In the model of monopolistic competition with the constant elasticity of substitution aggregation (20.15), intermediate-good producers set a constant markup. To see this, first take a look at the first-order condition of the problem (20.18):

$$\left(1 - \frac{1}{\sigma}\right) \gamma a_i^{1-\frac{1}{\sigma}} m_i^{\gamma - \frac{\gamma}{\sigma} - 1} Y^{\frac{1}{\sigma}} = c. \tag{20.19}$$

Using (20.17), $y_i = a_i m_i^{\gamma}$, and the fact that the marginal cost $\mathcal{M} \equiv \partial(cm_i)/\partial y_i$ can be expressed as

$$\mathcal{M} = \frac{cm_i^{1-\gamma}}{\gamma a_i}, \tag{20.20}$$

(20.19) can be rewritten as

$$p_i = \frac{\sigma}{\sigma - 1}\mathcal{M}. \tag{20.21}$$

Therefore, $\sigma/(\sigma - 1)$ is the markup and is constant as long as $\sigma$ is constant.

In many contexts, this constant markup property is a convenient model feature. However, this feature also imposes some limitations: the model cannot be used to analyze the endogenous changes in markups when the economic environment or policies change. The question of markup determination is particularly relevant in the recent U.S. economy. As mentioned in Section 20.3, De Loecker et al. (2020) observe that the level of markups has increased since the 1980s in the U.S. economy. There are three different paths explored by researchers: (i) departing from the monopolistic competition assumption; (ii) departing from the CES assumption; and (iii) considering the endogenous difference in productivity across firms. This section briefly introduces the first approach, following the formulation based on Atkeson and Burstein (2008).

Consider the same setting as in Section 20.6.1, but each intermediate good itself is the combination of several products. Following Atkeson and Burstein (2008), call the collection of $J$ firms that produces inputs for the intermediate good $i$ as the *sector* $i$. Within each sector, let us index each firm by $j$ and call a particular firm's product a *brand*. The production of intermediate good $i$ is dictated by the function

$$y_i = \left[\sum_{j=1}^{J} q_{ij}^{\frac{\eta-1}{\eta}}\right]^{\frac{\eta}{\eta-1}}. \tag{20.22}$$

We assume that $\eta < \infty$, that is, brands are imperfect substitutes. We also assume that $\eta > \sigma > 1$, that is, the brands are more substitutable within the sector than across sectors. The final-good production function is (20.15). We keep assuming that each intermediate good is small compared to the entire economy so that each firm does not consider the influence of its production decision on the final good production $Y$ (or the general price level). However, it is sufficiently large within each sector so that it is aware of the influence on $y_i$ (and the price of intermediate good $i$).

In this setting, the final good producer has to solve two layers of the cost-minimization problem: (i) find the best combination of $y_i$ for a given $Y$, and (ii) find the best combination of $q_{ij}$ for a given $y_i$. The first cost-minimization problem is identical to the one in Section 20.6.1. The inverse demand function of $y_i$ is given by (20.17). The second-stage cost-minimization problem can be solved similarly, and the result is

$$\frac{\hat{p}_{ij}}{p_i} = q_{ij}^{-\frac{1}{\eta}} y_i^{\frac{1}{\eta}}, \tag{20.23}$$

where $\hat{p}_{ij}$ is the price of the brand $j$ in sector $i$ and $p_i$ is now the price of the sector-$i$ good:

$$p_i = \left[ \sum_{j=1}^{J} \hat{p}_{ij}^{1-\eta} \right]^{\frac{1}{1-\eta}}.$$

Note that (20.22) and (20.23) imply

$$p_i y_i = \sum_{j=1}^{J} \hat{p}_{ij} q_{ij}. \tag{20.24}$$

which is the consequence of (20.22) being a constant returns to scale function. Let the production function for firm $j$ in sector $i$ be

$$q_{ij} = a_{ij} m_{ij}^{\gamma}, \tag{20.25}$$

where $m_{ij}$ is the "combined input" as before. The firm maximizes profit $\hat{p}_{ij} q_{ij} - c m_{ij}$. Using the inverse demand function, the problem the firm solves is

$$\max_{q_{ij}, m_{ij}} \; q_{ij}^{-\frac{1}{\eta}} y_i^{\frac{1}{\eta}} y_i^{-\frac{1}{\sigma}} Y^{\frac{1}{\sigma}} q_{ij} - c m_{ij}.$$

subject to (20.22) (with $q_{ij}$ for the other firms as given) and (20.25). As in the monopolistic competition case, the firm takes $Y$ as given. From the first-order condition, noting the relationship (20.17) and (20.23), the pricing rule can be derived as

$$\hat{p}_{ij} = \frac{\varepsilon(s_{ij})}{\varepsilon(s_{ij}) - 1} \mathcal{M}, \tag{20.26}$$

where $\mathcal{M}$ is the marginal cost (analogous to (20.20))

$$\mathcal{M} = \frac{c m_{ij}^{1-\gamma}}{\gamma a_{ij}}$$

and

$$\varepsilon(s_{ij}) = \left[ \frac{1}{\eta}(1 - s_{ij}) + \frac{1}{\sigma} s_{ij} \right]^{-1}. \tag{20.27}$$

Here, $s_{ij}$ is

$$s_{ij} = \frac{\hat{p}_{ij} q_{ij}}{p_i y_i} = \frac{\hat{p}_{ij} q_{ij}}{\sum_{h=1}^{J} \hat{p}_{ih} q_{ih}}, \tag{20.28}$$

where the second equation utilizes the relationship (20.24). Thus, $s_{ij}$ is the sales share of the firm $j$ in sector $i$. Because $s_{ij} \in [0, 1]$ and $\sigma < \eta$, $\varepsilon(s_{ij})$ takes the value between $\sigma$ and $\eta$. In particular, $\varepsilon(s_{ij})$ is $\eta$ when $s_{ij} = 0$, monotonically decreases with $s_{ij}$, and reaches $\varepsilon(s_{ij}) = \sigma$ when $s_{ij} = 1$.

Comparing (20.21) and (20.26), the difference is that $\sigma$ is replaced by $\varepsilon(s_{ij})$. In the current model, the markup of firm $j$ can vary depending on the sales share. When firms are symmetric,

$$\varepsilon(s_{ij}) = \left[\frac{1}{\eta}\frac{J-1}{J} + \frac{1}{\sigma}\frac{1}{J}\right]^{-1}.$$

It is intuitive that the markup is the highest with $\varepsilon(s_{ij}) = \sigma$ when $J = 1$ (monopoly within the sector), which is exactly the monopolistic competition case (20.21). The markup decreases as $J$ increases and $\varepsilon(s_{ij}) \to \eta$ as $J \to \infty$. This framework allows the monopoly power (represented by the number of firms $J$) to be linked to the markup. When firms are not symmetric (for example, the values of $a_{ij}$ are different across firms), firm heterogeneity can feed into heterogeneity in markups.

In the above derivation, we made the Cournot competition assumption: each producer chooses its quantity given the quantities of the other producers in the same sector. Alternatively, we can make the Bertrand competition assumption: each producer chooses its price given the prices of the other producers in the same sector. It can be shown that the formula (20.26) still holds, with different $\varepsilon(s_{ij})$:

$$\varepsilon(s_{ij}) = \eta(1 - s_{ij}) + \sigma s_{ij}$$

instead of (20.27), where $s_{ij}$ is still defined by (20.28). The intuition is similar to the Cournot competition case. See Appendix 20.A.4 for the details of the derivation.

## 20.7 Business cycles and heterogeneous firms

As we saw in Section 20.3, many statistics on firm behavior, such as job creation and destruction rates and entry and exit rates, have clear cyclicality. It is natural, therefore, to think of the causes and consequences of such cyclicality in firm dynamics.

Note that in the model of Section 20.5, firms face idiosyncratic shocks, but the aggregate economy is stationary. The basic logic is simple: firm-level shocks are smoothed out by being summed up across firms to create GDP. To illustrate, suppose that the GDP $Y_t$ is the sum of firm-level output $y_{it}$, $i = 0, 1, ..., N$.[15] The growth rate of $y_{it}$ is identically and independently distributed with mean 0 and variance $\sigma^2$. That is,

$$\frac{y_{i,t+1} - y_{it}}{y_{it}} = \sigma\varepsilon_{i,t+1},$$

where $\varepsilon_{i,t+1}$ is a random variable with mean zero and variance one. Then, the growth rate of $Y_t$ is

$$\frac{Y_{t+1} - Y_t}{Y_t} = \frac{1}{Y_t}\sum_{i=1}^{N}\Delta y_{i,t+1} = \sum_{i=1}^{N}\frac{y_{it}}{Y_t}\sigma\varepsilon_{i,t+1}.$$

[15] This illustration is based on the exposition of Gabaix (2011).

Thus, the standard deviation of GDP growth rate, $\sigma_Y$, is

$$\sigma_Y = \sigma \sqrt{\sum_{i=1}^{N} \left( \frac{y_{it}}{Y_t} \right)^2}. \tag{20.29}$$

When the firms are of equal size, that is, $y_{it}/Y_t = 1/N$, $\sigma_Y = \sigma/\sqrt{N}$. When there are one million firms in the economy (there are over 5 million firms in the U.S. data), $1/\sqrt{N} = 0.001$. A typical standard deviation of the firm-level volatility is between 10% to 20%;[16] thus, the effect of idiosyncratic shocks on the aggregate volatility is about two orders of magnitude smaller than the GDP volatility in the U.S. data. In other words, it is negligible compared to the actual business cycle fluctuations.

### 20.7.1 Aggregate shocks and firm dynamics

A strand of literature takes the law of large numbers as given and adds aggregate shocks in analyzing aggregate fluctuations. In that case, the model in Section 20.5 can be modified to have a production function for the firm $i$:

$$y_{it} = z_t s_{it} \ell_{it}^{\gamma}.$$

Here, the variable $z_t$ is added. As in the standard real business cycle model, $z_t$ can be interpreted as the aggregate productivity shock. The model can then be calibrated and computed. As discussed earlier, the Hopenhayn and Rogerson (1993) model has a *block-recursive* structure. Therefore, computing this type of model is often substantially easier than the computation of standard heterogeneous-agent models.

It is known that the modified Hopenhayn and Rogerson (1993) model does well in replicating the aggregate fluctuations in statistics such as job creation and destruction rates and entry and exit rates. There are some firm-level statistics that are difficult to replicate by a simple modification of the Hopenhayn and Rogerson (1993) model.[17]

### 20.7.2 Can idiosyncratic shocks generate aggregate fluctuations?

Despite the law of large numbers result, some researchers believe that micro-level shocks can have an important role in generating aggregate fluctuations. The calculation at the beginning of this section made two assumptions: (i) the distribution of idiosyncratic shock has a finite variance, and (ii) there are no input-output networks. This subsection introduces the analysis of the economic environment where one of these two assumptions does not hold.

This research agenda is attractive because, since the outset of the real business cycle research agenda, the fluctuations in aggregate shocks are often criticized for being a "black box." If we know that the cycles stem from idiosyncratic shocks, one can imagine that an effective stabilization policy would target the firms whose idiosyncratic shocks matter for the aggregate fluctuations.

[16] Gabaix (2011) estimates it to be 12% in the U.S, data.

[17] See Lee and Mukoyama (2018) for a detailed analysis.

### Hulten's theorem

Before discussing how micro shocks can matter for macroeconomic fluctuations, it is useful to introduce a simple theorem by Hulten (1978). Let us think of a setting where there are $N$ different sectors indexed by $i$. Here, we use the terminology "sectors" because we would like to think of a competitive equilibrium. Using "firms" would yield the same result as long as firms behave as price-takers. Sector $i$'s production is $y_i$. The production function is

$$y_i = a_i F(k_i, \ell_i, x_{i1}, x_{i2}, ..., x_{iN}),$$

where $a_i$ is the TFP, $x_{ij}$ is the sector-$j$ product used in sector $i$, $k_i$ is capital used in sector $i$, and $\ell_i$ is labor used in sector $i$. Note that the total sales $\sum_i p_i y_i$ is different from the total value added (i.e., the GDP), which is equal to $\sum_i p_i c_i$, because some of the output is used as the intermediate goods. Let $Y = \sum_i p_i c_i$.

Hulten's (1978) theorem states that the output effect of

$$\frac{dY}{Y} = \sum_i D_i \frac{da_i}{a_i},$$

where $D_i$ is the weight called the Domar weight:

$$D_i = \frac{p_i y_i}{\sum_i p_i c_i}. \tag{20.30}$$

$D_i$'s denominator is the total value added whereas the numerator is the sales of sector $i$. The proof of the theorem can be found in Appendix 20.A.5.

Two pieces of intuition are key to understanding Hulten's theorem. First, why does only $a_i$ matter and not anything about inputs? The intuition is the envelope theorem. Because the economy achieves the first best, the input allocation is already optimized. Therefore, to the first-order approximation, the adjustment of inputs due to shocks to $a_i$ does not have an impact on welfare. With the homothetic utility, the welfare result can be mapped to GDP. Second, why is the numerator of the weight measured as sales? This question seems natural, especially because the denominator is in value added, which implies that the Domar weight does not necessarily sum up to 1. The intuition is that, when there are input-output networks, the improvement in the TFP in a downstream firm also raises the value of intermediate inputs.

To see the second point more clearly, consider the following simple example. Suppose that there are two sectors, sectors 1 and 2. Sector 1 produces the consumption good, whose price is normalized to 1. The production function is $y_1 = a_1 x_1^{1-\gamma} \ell^{\gamma}$, where $y_1$ is the output, $a_1$ is the TFP, $x_1$ is the intermediate input, and $\ell$ is the labor input. The parameter $\gamma \in (0, 1)$. Sector 2 produces the intermediate good using capital: $y_2 = a_2 k$. The capital supply is fixed at $K$ and the labor supply is fixed at $L$. In the competitive equilibrium, because $Y = a_1 (a_2 K)^{1-\gamma} L^{\gamma}$,

$$\frac{dY}{Y} = \frac{da_1}{a_1} + (1-\gamma)\frac{da_2}{a_2} \tag{20.31}$$

holds. Note that the weights in front of both TFP growths do not sum up to 1. To confirm Hulten's Theorem, let us compute the Domar weight of each sector. Let the price of the intermediate good be $p$. In the competitive equilibrium, $p = (1-\gamma)a_1 x_1^{-\gamma} L^{\gamma}$ holds, where $x_1 = a_2 K$. Thus, the value added of sector 2 is

$$V_2 = (1-\gamma)a_1 (a_2 k)^{1-\gamma} L^{\gamma}.$$

The value added of sector 1 is

$$V_1 = Y - px_1 = \gamma a_1 (a_2 k)^{1-\gamma} L^{\gamma}.$$

Thus, the Domar weight of sector 2 (because the sales are equal to the value added in sector 2) is $V_2/Y = (1-\gamma)$. The Domar weight of sector 1 (because the sales are $Y$) is $Y/Y = 1$. Both correspond to the coefficients on the right-hand side of (20.31), confirming Hulten's theorem. Note that if we consider the value added instead of sales in sector 1, the coefficient is computed as $V_1/Y = \gamma$.

**Large firms**

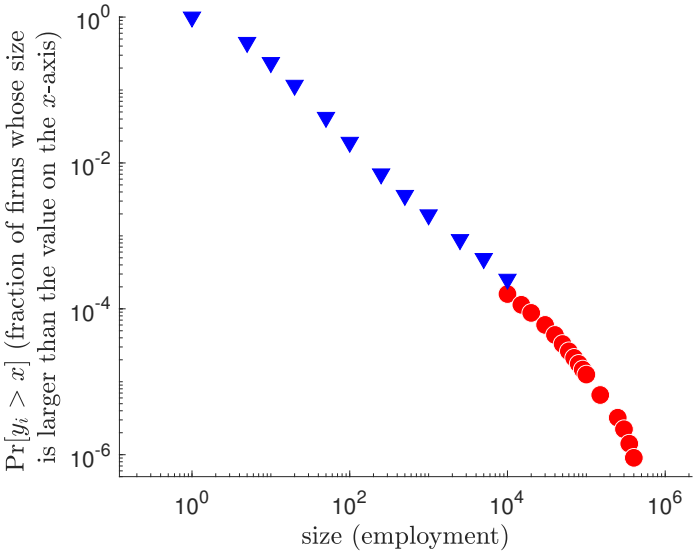



Figure 20.9: Distribution of firm size in log-log axes.

**Source**: Business Dynamics Statistics and Compustat. Reproduced from Carvalho and Grassi (2019).

As we discussed in Section 20.3, there are many large firms in the U.S. economy. In fact, it is known that the U.S. firm size distribution is "fat-tailed": it can be closely approximated by a Pareto distribution at the right tail. Figure 20.9 reproduces the data plot of Figure

2A in Carvalho and Grassi (2019). It plots the size of a firm, measured by employment, against the firm size "percentile" (the fraction of firms whose size is larger than the value on the $x$-axis). The blue triangles use the same information as in Figure 20.1.[18] Because the publicly-available BDS table does not contain finer information on large firms, Carvalho and Grassi (2019) utilizes the Compustat data, which includes only publicly-traded firms (red circles). To the extent that large privately-held firms are rare, the red circles approximate the distribution of large firms in the U.S. economy. One can see that the plot with log-log axes is close to a straight line, indicating that the distribution is well-approximated by the Pareto distribution.

Gabaix (2011) argues that when the distribution of firm size is fat-tailed, that is, there is a considerable presence of large firms (as in the Pareto distribution), the formula (20.29) implies a significant impact of $\sigma$ on $\sigma_Y$. In particular, he considers the case where the firm size distribution is Pareto

$$\Pr[y_i > x] = \chi x^{-\zeta}, \tag{20.32}$$

where $\chi$ and $\zeta$ are constants. The variable $y_i$ here is employment and $\Pr[y_i > x]$ is the fraction of firms whose sizes are larger than $x$. It is known that the U.S. firm size distribution has $\zeta$ close to 1. (Because (20.32) implies $\ln(\Pr[y_i > x]) = -\zeta \ln(x) + \ln(\chi)$, this can be seen from the slope of Figure 20.9 being close to $-1$.) Therefore, the empirically relevant case is $\zeta = 1$. In this case, he derives

$$\sigma_Y \sim \frac{v_\zeta}{\ln(N)}\sigma$$

as $N \to \infty$, that is, $\sigma_Y \ln(N)$ converges to $v_\zeta \sigma$ in distribution, where $v_\zeta$ is a random variable that does not depend on $N$ or $\sigma$. This formula implies that the GDP volatility is proportional to $1/\ln(N)$ instead of $1/\sqrt{N}$. The function $1/\ln(N)$ declines slower than $1/\sqrt{N}$ as $N$ becomes large (for example, $1/\ln(1,000,000) \approx 0.072$ whereas $1/\sqrt{1,000,000} = 0.001$). Gabaix (2011) calculates that the coefficient on $\sigma$ in the above formula would be approximately 0.12 instead of 0.001. The effect of idiosyncratic shocks is, therefore, two magnitudes larger than the identical-firm case; thus the volatility induced by the idiosyncratic shock can have an effect of a similar magnitude to observed business cycles. In a recent work, Carvalho and Grassi (2019) build a quantitative business cycle model similar to the one in Section 20.5 and analyze the business cycle dynamics driven by idiosyncratic shocks to large firms.

### Effects of production networks

An additional important factor that can magnify the effect of idiosyncratic shocks on the aggregate economy is that the Domar weight (20.30) divides the firm sales by the aggregate value added. Many large firms, such as Walmart, Amazon, and GM, have significantly larger sales than their value added, magnifying their contribution to the aggregate fluctuations. These firms are at the downstream of the production network, and the impact of their productivity shocks on the aggregate GDP is larger than their value added.

[18] The time period for Figure 20.9 is 1977–2012 whereas the time period for Figure 20.1 is 2019. Figure 20.9 plots in finer categories than Figure 20.1 does.

Some researchers believe that there are further important implications of the production network. Note that Hulten's theorem builds on two assumptions: (i) the economy is efficient, and (ii) the first-order approximation is sufficiently accurate. When these two assumptions are not appropriate, network structures can play a role in considering the aggregate effects of idiosyncratic shocks. For example, in a recent paper, Baqaee and Farhi (2019) argue that there are situations where the second-order effect is quantitatively important. In another paper, Baqaee and Farhi (2020) consider economies with distortions, and there the network structure also plays a role.

## 20.8 Endogenous productivity

Given the importance of the idiosyncratic shocks in generating heterogeneity across firms and their dynamics, it is natural to wonder what factors influence idiosyncratic productivity. One natural framework that can be used to analyze this issue is the models of endogenous productivity change, introduced in the economic growth chapter. Klette and Kortum (2004) introduced a framework that can generate the entry, exit, expansion, and contraction of firms due to endogenous innovation. Below, we explain the Klette and Kortum (2004) model with discrete time.[19]

Consider an economy with a continuum of products on the unit interval $[0, 1]$. We will focus on the balanced-growth path of the economy, where all aggregate variables grow at the same rate. The representative consumer's utility function is

$$\sum_{t=0}^{\infty} \beta^t \ln(C_t),$$

where $\beta \in (0, 1)$ is the discount factor and $C_t$ is defined as

$$C_t = \exp\left( \int_0^1 \ln\left( \sum_{k=-1}^{J_t(j)} q_t(j,k) c_t(j,k) \right) dj \right). \tag{20.33}$$

Here, $j \in [0, 1]$ is the index of the product. The aggregation with natural log implies that the elasticity of substitution across goods is 1. The index $k$ is an integer which runs from $-1$ to $J_t(j)$. The value of $k$ represents the generation of the product: a newer generation (i.e., a larger $k$) product is of higher quality. $J_t(j)$ is the cutting-edge (state-of-the-art) generation of good $j$. Generation $k$ of product $j$ (call it product $(j, k)$ for short) contributes to consumption in the form of $q_t(j, k)c_t(j, k)$. Here, $q_t(j, k)$ is the *quality* of the product $(j, k)$ and $c_t(j, k)$ is the *quantity* of the product $(j, k)$. The fact that the aggregation is additive across different generations implies that different generations are perfect substitutes. If generation $k'$ has twice the quality of generation $k$, that is, $q_t(j, k')/q_t(j, k) = 2$, consuming one unit of product $(j, k')$ is equivalent to consuming two units of product $(j, k)$.

[19] Ates and Saffie (2021) also develops a discrete-time version of the Klette and Kortum (2004) model.

The consumer faces two layers of problems: intratemporal and intertemporal. The intratemporal problem is to think of how to allocate the expenditure across different goods at each point in time. The intertemporal problem is to decide how much to spend across time.

Let us start by thinking about the intratemporal problem. Let $E_t$ be the expenditure at period $t$ and $p_t(j,k)$ is the price of product $(j,k)$. First note that within product $j$, it is optimal to only purchase the generation whose "quality-adjusted price" $p_t(j,k)/q_t(j,k)$ is the lowest. Therefore, let us only consider such generation $k$ for each $j$. Then, the intratemporal problem is

$$\max_{c_t(j,k)} \int_0^1 \ln\left(q_t(j,k)c_t(j,k)\right) dj$$

subject to

$$\int_0^1 p_t(j,k)c_t(j,k)dj \leq E_t. \tag{20.34}$$

The solution is

$$c_t(j,k) = \frac{E_t}{p_t(j,k)}. \tag{20.35}$$

Given the solution to the intratemporal problem, the consumption of each period can be rewritten as

$$C_t = E_t \exp\left(\int_0^1 [\ln(q_t(j,k)) - \ln(p_t(j,k))]dj\right).$$

This relationship can be rewritten as

$$P_t C_t = E_t,$$

where the price index for consumption is

$$P_t \equiv \exp\left(\int_0^1 [\ln(p_t(j,k)) - \ln(q_t(j,k))]dj\right).$$

As in Section 20.6.1, normalize $P_t = 1$. This normalization implies

$$\int_0^1 \ln(p_t(j,k))dj = \int_0^1 \ln(q_t(j,k))dj \tag{20.36}$$

at any $t$.

The intertemporal problem is now

$$\max_{C_t} \sum_{t=0}^{\infty} \beta^t \ln(C_t)$$

subject to

$$\sum_{t=0}^{\infty} \left(\frac{1}{1+r}\right)^t C_t \leq \mathcal{A}_0,$$

where $\mathcal{A}_0$ is the present discounted value of all future labor and asset incomes. The asset income in this economy comes from the claim to the profit of the firms. Here, we are imposing that the interest rate $r > 0$ is constant, given that we will focus on the balanced growth path. The optimization results in the Euler equation:

$$\frac{1}{C_t} = \beta(1+r)\frac{1}{C_{t+1}}.$$

Along the balanced-growth path, $C_{t+1}/C_t$ grows at a constant rate. Let us define $\gamma \equiv (C_{t+1} - C_t)/C_t$. Then,

$$\frac{1}{1+r} = \frac{\beta}{1+\gamma}. \tag{20.37}$$

Each product $(j,k)$ is produced by one firm. One firm can own several product lines. A *firm* here is indeed defined as a collection of product lines. A small firm owns only a few product lines, and a large firm owns many product lines. Although firms are heterogeneous in this dimension, the analysis of the firm decision in the Klette and Kortum (2004) model is relatively simple because the model has a structure that allows each of the firm's product lines to independently make the decision. In the following, we exploit that property and analyze the firm's decisions at the product-line level.

First, consider the production decision. Producing one unit of a product takes one unit of labor.[20] Thus, the unit production cost is $w_t$. Given that the production cost is the same, the cutting-edge producer (the "leader") has an advantage over other producers (with the older generation of product $j$). Because the demand elasticity is 1, the optimal pricing is to set the price as high as possible. Here, the cutting-edge producer cannot increase the price in an unlimited manner. Once the price is set sufficiently high, the $J_t(j)-1$ generation producer can enter profitably. The highest price the cutting-edge producer can charge is

$$p_t(j, J_t(j)) = \lambda w_t, \tag{20.38}$$

where $\lambda > 1$ represents the technology step $q_t(j,k+1)/q_t(j,k)$. Here $\lambda$ coincides with the markup rate. This pricing behavior is called *limit pricing*.

Given the price, the period profit from one product line for the leader is

$$\pi_t \equiv (p_t(j, J_t(j)) - w_t)\frac{C_t}{p_t(j, J_t(j))} = \left(1 - \frac{1}{\lambda}\right) C_t. \tag{20.39}$$

From (20.36) and (20.38),

$$\ln(\lambda w_0) = \int_0^1 \ln(q_0(j, J_0(j))dj.$$

By normalizing $q_0(j, J_0(j)) = \lambda$ for all $j$, this relationship implies $w_0 = 1$.

Second, consider the innovation decision. We assume that innovation is also conducted in each product line. In each product line, the firm decides the intensity of innovation $\eta$

[20] As we saw in Section 20.6.1, the production scale remains finite even with constant returns to scale because of the monopoly power.

with the required labor input $R(\eta)$; therefore, the innovation cost is $w_t R(\eta)$, where $R(\cdot)$ is an increasing and convex function. The intensity $\eta$ represents the probability that the firm gains another product line. When innovation is successful, in addition to the current product line, the firm can start producing another product line. This newly-added product is $\lambda$ times better than the current cutting-edge product. This newly-added product is randomly chosen from $[0, 1]$. Each firm is infinitesimally small compared to the entire economy; thus the probability of innovating over its own product is zero, and the new innovation always takes the market from another firm. Because other firms also innovate, each firm can also lose its market for the product line. Let $\mu$ be the probability that other firms innovate and take over the current product line. Along the balanced-growth path, $\mu$ is constant over time.

Denoting the value of a leader product line by $V_t$, the Bellman equation for the firm is

$$V_t = \max_{\eta} \ \pi_t - w_t c(\eta) + \frac{1}{1+r}(1+\eta-\mu)V_{t+1}. \tag{20.40}$$

The final term is the expected future value of the product line. There are four possible scenarios for the current leader of the product $j$: (i) it innovates and is not taken over by another firm; (ii) it fails to innovate and is taken over by another firm; (iii) both occurs; and (iv) neither occurs. The probability of (i) is $\eta(1-\mu)$ and the future value is $2V_{t+1}$. The probability of (ii) is $\mu(1-\eta)$ and the future value is 0. The probability of (iii) is $\mu\eta$ and the future value is $V_{t+1}$. The probability of (iv) is $(1-\mu)(1-\eta)$ and the future value is $V_{t+1}$. Therefore, the expected future value is computed as $(1+\eta-\mu)V_{t+1}$, which can be seen in the final term.

Along the balanced-growth path, $V_t$, $\pi_t$, and $w_t$ all grow at the common rate $(1+\gamma)$. Dividing both sides of (20.40) by $(1+\gamma)^t$ and using (20.37), (20.39), and $w_0 = 1$,

$$v = \max_{\eta} \ \left(1 - \frac{1}{\lambda}\right) C_0 - R(\eta) + \beta(1+\eta-\mu)v, \tag{20.41}$$

where $v \equiv V_t/(1+\gamma)^t$. The first-order condition is

$$R'(\eta) = \beta v. \tag{20.42}$$

There are many potential entrants in the economy. As in Section 20.5, we assume free entry. Entrant (with probability 1) can hire $c_e$ units of labor and enter. The free-entry condition is

$$V_t = w_t c_e.$$

Normalizing,

$$v = c_e. \tag{20.43}$$

Let the entry rate (the amount of entry at each period) be $\nu$. We assume that each product line receives only (up to) one innovation per period. Thus, the fraction of the product lines that receive innovation is $\mu$. Because the innovation is done by either entrants or incumbents,

$$\mu = \eta + \nu \tag{20.44}$$

holds.

There are three types of labor demand: (i) production, (ii) innovation by incumbents, and (iii) entry. From (20.35) and $E_t = C_t$, production at time 0 is $c_0(j,k) = C_0/\lambda$ and thus the labor demand is $C_0/\lambda$. For innovation by incumbents, $R(\eta)$ units of labor are used. For entry, $\nu$ units are demanded. The labor supply is fixed at $L$. Thus, the labor-market equilibrium condition (using (20.44)) is

$$\frac{C_0}{\lambda} + R(\eta) + \nu = L. \tag{20.45}$$

In sum, the general equilibrium of the model solves four unknowns ($v$, $\eta$, $C_0$, and $\mu$) with four equations:

$$v = \left(1 - \frac{1}{\lambda}\right) C_0 - R(\eta) + \beta(1 + \eta - \mu)v,$$

which is from (20.41), the first-order condition (20.42), the free-entry condition (20.43), and the labor-market equilibrium condition (20.45).

Finally, we can calculate the economy's growth rate, $\gamma$. In this economy, the consumption $C_t$ in (20.33) is equal to $E_t$, which is aggregate expenditure (see (20.34)). Along the balanced-growth path, the growth rate of $E_t$ is also equal to the growth rate of $w_t$. From (20.36) and (20.38),

$$\begin{aligned}
(\gamma \approx) \ln(w_{t+1}) - \ln(w_t) &= \int_0^1 \ln(p_{t+1}(j, J_{t+1}(j))) - \ln(p_t(j, J_t(j)))dt \\
&= \int_0^1 \ln(q_{t+1}(j, J_{t+1}(j))) - \ln(q_t(j, J_t(j)))dt \\
&= \int_0^1 (\ln(\lambda^{J_{t+1}(j)}) - \ln(\lambda^{J_t(j)}))dt \\
&= \mathbb{E}[\ln(\lambda^{J_{t+1}})] - \mathbb{E}[\ln(\lambda^{J_t})] \\
&= \mathbb{E}[J_{t+1}]\ln(\lambda) - \mathbb{E}[J_t]\ln(\lambda) \\
&= \mu(t+1)\ln(\lambda) - \mu t \ln(\lambda) \\
&= \mu \ln(\lambda).
\end{aligned}$$

The first equality follows from (20.38), the second is from (20.36), and the third is from the definition that $J_t(j)$ is the cutting-edge generation at industry $j$. In the fourth equality, we utilize the law of large numbers. Because each industry is subject to the i.i.d. shock and there is a continuum of industries, we can replace the cross-sectional average with the expected value when we interpret $J_t$ as a random variable. The next inequality uses the fact that $J_t$ is viewed as a sum of Bernoulli trials with winning probability $\mu$ (i.e., every period, the probability that a product *receives* an innovation is $\mu$). The growth rate of the economy depends on $\mu$, which is driven by the innovation intensity by incumbents ($\eta$) and by entrants ($\nu$), and the innovation step $\lambda$.

The major strength of Klette and Kortum (2004) model over the traditional endogenous growth models is that the definition of a firm is clear, and one can analyze the dynamics of firms and the firm-size distribution. The analysis of firm-size distribution is somewhat

more complex than the original Klette-Kortum model, given that many events can happen to a firm in the same period with the discrete-time formulation. The details are described in Appendix 20.A.6. Although the firm-size distribution is not analytically straightforward in the discrete-time version, it is straightforward to compute it on a computer. The advantage of this model over the models with exogenous idiosyncratic shocks, such as the one in Section 20.5, is that it can analyze how the policies and changes to the economic environment can affect the productivity process itself.

One simple statistic that we can compute is the average growth rate of the firm size. First, note that because each product line produces the same quantity and employs the same number of workers, the firm size distribution coincides with the distribution of product lines across firms. As can be seen in the discussion of (20.41), the average number of product lines in the next period per each line this period is $(1+\eta-\mu)$. Therefore, the average growth rate of the firm size is $\eta - \mu$. From (20.44), $\eta - \mu = -\nu < 0$. Thus, the average growth rate of a firm is negative, and a large firm almost always contracts (due to the law of large numbers). The property that a large and small firm have a common average growth rate is called Gibrat's Law, although it is usually stated in the context of the positive average growth rate.

There are several counterfactual predictions in Klette and Kortum (2004) model. Recent literature made progress in modifying the model so that the model can replicate the salient features of the data.

First, the firm with a higher quality product does not earn a higher profit on that product line. This feature stems from two assumptions: (i) the elasticity of substitution across goods is 1 and (ii) the technology is not cumulative. For the first point, the natural log specification implies that the revenue of the leader is the same regardless of prices and quality level. With higher substitutability, both can matter for the size and profit of the firm. For the second point, any outside firm can innovate over the state-of-the-art product at the same cost as the incumbent. If the model is extended so that the incumbent firm improves its own product quality in equilibrium, there is an additional reason for the size and profit difference (therefore, the idiosyncratic productivity shock in Section 20.5) across firms.

Second, the firm-size distribution does not feature a Pareto tail. Intuitively, it is very difficult to create many large firms in this economy because the firm size contracts on average. One alternative example is that, instead of a negative growth rate, a large firm has a positive constant growth rate $g$. Suppose, in addition, all firms receive an exit shock with the probability $\delta \in (0,1)$. Consider a very large firm so that we can ignore the integer constraint of product lines. Let us start from the firms between size $n$ and $n+\Delta$, where $\Delta$ is a small number relative to $n$. When the stationary density at $n$ is $h(n)$, the mass of firms between these sizes is approximated by $h(n)\Delta$. In the next period, the surviving mass is $(1-\delta)h(n)\Delta$. After one period, size $n$ will grow to $(1+g)n$ and size $n+\Delta$ will grow to $(1+g)(n+\Delta)$. Thus, the mass between these new sizes will be $(1+g)h((1+g)n)\Delta$. Therefore, in the stationary distribution,

$$(1+g)h((1+g)n)\Delta = (1-\delta)h(n)\Delta$$

has to hold. Guess that the distribution is Pareto: $h(n) = Fn^{-(\zeta+1)}$, where $F > 0$ and $\zeta > 0$

are parameters. In particular, $\zeta$ is the tail index that showed up in Section 20.7.2. The above equation can then be rewritten as

$$(1+g)F((1+g)n)^{-(\zeta+1)}\Delta = (1-\delta)Fn^{-(\zeta+1)}\Delta.$$

This equality holds for any $n$ and $\Delta$ when

$$\zeta = -\frac{\ln(1-\delta)}{\ln(1+g)} > 0.$$

Thus, we verified that the firm dynamics with positive (and constant) growth, combined with a constant exit rate, can be consistent with the stationary distribution that is Pareto. The tail index $\zeta$ is small (i.e., a thick tail) when $\delta$ is small or $g$ is large. One question is how we can modify the Klette and Kortum (2004) model to have a positive firm growth rate at the right tail. One possibility is to break equation (20.44).[21] For example, if some innovation *creates* new products, there can be firm expansion without contributing to $\mu$. Suppose that the new product creation among the total innovation is $\xi$ (i.e., among the total $\eta+\nu$ innovation, $\xi$ creates new products, and $\mu = \eta+\nu-\xi$ replaces existing products). Then, the average growth rate of a firm, which is still $\eta-\mu$, is now equal to $\xi-\nu$ (instead of just $-\nu$). If $\xi$ is sufficiently large, $\xi-\nu$ can be positive.

[21] Luttmer (2011) discusses related insights.

# Chapter 21

# International macro

*Giancarlo Corsetti and Luca Dedola*

# Chapter 22

# Emerging markets

*Juan Carlos Hatchondo and Leo Martinez*

# Chapter 23

# Sustainability

***John Hassler, Per Krusell, and Conny Olovsson***

# Bibliography


Abowd, J. and D. Card (1989). On the covariance structure of earnings and hours changes. *Econometrica 57*(2), 411–445.

Achdou, Y., J. Han, J.-M. Lasry, P.-L. Lions, and B. Moll (2022). Income and wealth distribution in macroeconomics: A continuous-time approach. *The Review of Economic Studies 89*(1), 45–86.

Aiyagari, S. R. (1994). Uninsured idiosyncratic risk and aggregate saving. *The Quarterly Journal of Economics 109*(3), 659–684.

Aiyagari, S. R. and E. R. McGrattan (1998). The optimum quantity of debt. *Journal of Monetary Economics 42*(3), 447–469.

Allen, F. (1985). Repeated principal-agent relationships with lending and borrowing. *Economics Letters 17*(1-2), 27–31.

Altug, S. and R. A. Miller (1990). Household choices in equilibrium. *Econometrica*, 543–570.

Andolfatto, D. (1996). Business Cycles and Labor-Market Search. *American Economic Review 86*, 112–132.

Aschauer, D. A. (1989). Is public expenditure productive? *Journal of Monetary Economics 23*(2), 177–200.

Ates, S. T. and F. Saffie (2021). Fewer but Better: Sudden Stops, Firm Entry, and Financial Selection. *American Economic Journal: Macroeconomics 13*, 304–356.

Atkeson, A. and A. Burstein (2008). Pricing-to-Market, Trade Costs, and International Relative Prices. *American Economic Review 98*, 1998–2031.

Atkeson, A., V. Chari, and P. Kehoe (1999). Taxing capital income: A bad idea. Technical Report 2331, Federal Reserve Bank of Minneapolis.

Attanasio, O. and S. J. Davis (1996). Relative wage movements and the distribution of consumption. *Journal of Political Economy 104*(6), 1227–1262.

Attanasio, O. P. (1999). Consumption. *Handbook of Macroeconomics 1*, 741–812.

Attanasio, O. P. and N. Pavoni (2011). Risk sharing in private information models with asset accumulation: Explaining the excess smoothness of consumption. *Econometrica 79*(4), 1027–1068.

Attanasio, O. P. and G. Weber (2010). Consumption and saving: models of intertemporal allocation and their implications for public policy. *Journal of Economic Literature 48*(3), 693–751.

Baqaee, D. R. and E. Farhi (2019). The Macroeconomic Impact of Microeconomic Shocks: Beyond Hulten's Theorem. *Econometrica 87*, 1155–1203.

Baqaee, D. R. and E. Farhi (2020). Productivity and Misallocation in General Equilibrium. *Quarterly Journal of Economics 135*, 105–163.

Barnett, W. A. (1980). Economic monetary aggregates an application of index number and aggregation theory. *Journal of Econometrics 14*(1), 11–48.

Barnichon, R. and G. Mesters (2020). Identifying modern macro equations with old shocks. *The Quarterly Journal of Economics 135*(4), 2255–2298.

Barro, R. J. (1974). Are government bonds net wealth? *Journal of Political Economy 82*(6), 1095–1117.

Barro, R. J. (1979). On the determination of public debt. *Journal of Political Economy 82*(6), 1095–1117.

Benhabib, J., A. Bisin, and M. Luo (2019). Wealth distribution and social mobility in the us: A quantitative approach. *American Economic Review 109*(5), 1623–1647.

Bernanke, B. S. and A. S. Blinder (1992). The federal funds rate and the channels of monetary transmission. *The American Economic Review 82*(4), 901–921.

Bewley, T. (1980). The optimum quantity of money. In J. Kareken and N. Wallace (Eds.), *Models of Monetary Economies*. Federal Reserve Bank of Minneapolis.

Bewley, T. (1983). A difficulty with the optimum quantity of money. *Econometrica*, 1485–1504.

Bils, M. and P. J. Klenow (2004). Some evidence on the importance of sticky prices. *Journal of political economy 112*(5), 947–985.

Blanchard, J. O. and N. G. Mankiw (1988). Consumption: Beyond certainty equivalence. *American Economic Review 78*(2), 173–177.

Blanchard, O. (2023). *Fiscal Policy under Low Interest Rates*. The MIT Press.

Blanchard, O. and J. Galí (2007). Real wage rigidities and the new keynesian model. *Journal of Money, Credit and Banking 39*(s1), 35–65.

Blanchard, O. J. and N. Kiyotaki (1987). Monopolistic competition and the effects of aggregate demand. *American Economic Review 77*(4), 647–666.

Blundell, R. and I. Preston (1998). Consumption inequality and income uncertainty. *The Quarterly Journal of Economics 113*(2), 603–640.

Boivin, J., M. P. Giannoni, and I. Mihov (2009). Sticky prices and monetary policy: Evidence from disaggregated us data. *American economic review 99*(1), 350–84.

Broer, T. (2013). The wrong shape of insurance? what cross-sectional distributions tell us about models of consumption smoothing. *American Economic Journal: Macroeconomics 5*(4), 107–140.

Broer, T., N.-J. H. Hansen, P. Krusell, and E. Öberg (2020). The new keynesian transmission mechanism: A heterogeneous-agent perspective. *The Review of Economic Studies 87*, 77–101.

Broer, T., M. Kapička, and P. Klein (2017). Consumption risk sharing with private information and limited enforcement. *Review of Economic Dynamics 23*, 170–190.

Browning, M. and T. F. Crossley (2001). The life-cycle model of consumption and saving. *Journal of Economic Perspectives 15*(3), 3–22.

Browning, M. and A. Lusardi (1996). Household saving: Micro theories and micro facts. *Journal of Economic Literature 34*(4), 1797–1855.

Bullard, J. and K. Mitra (2002). Learning about monetary policy rules. *Journal of monetary economics 49*(6), 1105–1129.

Calvo, G. A. (1983). Staggered prices in a utility-maximizing framework. *Journal of monetary Economics 12*(3), 383–398.

Campbell, J. Y. (1987). Does saving anticipate declining labor income? an alternative test of the permanent income hypothesis. *Econometrica*, 1249–1273.

Campbell, J. Y. (2003). Consumption-based asset pricing. *Handbook of the Economics of Finance 1*, 803–887.

Campbell, J. Y. and N. G. Mankiw (1989). Consumption, income, and interest rates: Reinterpreting the time series evidence. *NBER Macroeconomics Annual 4*, 185–216.

Cantor, R. (1985). The consumption function and the precautionary demand for savings. *Economics Letters 17*(3), 207–210.

Cao, D., H. R. Hyatt, T. Mukoyama, and E. Sager (2022). Firm Growth with New Establishments. mimeo. Georgetown University, U.S. Census Bureau, and Federal Reserve Board.

Carroll, C. D. (2001). A theory of the consumption function, with and without liquidity constraints. *Journal of Economic Perspectives 15*(3), 23–45.

Carroll, C. D., R. E. Hall, and S. P. Zeldes (1992). The buffer-stock theory of saving: Some macroeconomic evidence. *Brookings Papers on Economic Activity 1992*(2), 61–156.

Carroll, C. D. and M. S. Kimball (1996). On the concavity of the consumption function. *Econometrica*, 981–992.

Carvalho, V. M. and B. Grassi (2019). Large Firm Dynamics and the Business Cycle. *American Economic Review 109*, 1375–1425.

Castaneda, A., J. Diaz-Gimenez, and J.-V. Rios-Rull (2003). Accounting for the us earnings and wealth inequality. *Journal of Political Economy 111*(4), 818–857.

Chamley, C. (1986). Optimal taxation of capital income in general equilibrium with infinite lives. *Econometrica 54*(3), 607–622.

Christiano, L. J., M. Eichenbaum, and C. L. Evans (2005). Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of Political Economy 113*(1), 1–45.

Clarida, R. H. (1987). Consumption, liquidity constraints and asset accumulation in the presence of random income fluctuations. *International Economic Review*, 339–351.

Clower, R. (1967). A reconsideration of the microfoundations of monetary theory. *Economic Inquiry 6*(1), 1–8.

Cochrane, J. H. (1991). A simple test of consumption insurance. *Journal of Political Economy 99*(5), 957–976.

Cole, H. L. and N. R. Kocherlakota (2001). Efficient allocations with hidden income and hidden storage. *The Review of Economic Studies 68*(3), 523–542.

Cooley, T. F. and E. C. Prescott (1995). Economic Growth and Business Cycles. In T. F. Cooley (Ed.), *Frontiers of Business Cycle Research*. Princeton University Press.

De Loecker, J., J. Eeckhout, and G. Unger (2020). The Rise of Market Power and the Macroeconomic Implications. *Quarterly Journal of Economics 135*, 561–644.

De Nardi, M. and G. Fella (2017). Saving and wealth inequality. *Review of Economic Dynamics 26*, 280–300.

Deaton, A. (1992). *Understanding consumption*. Oxford University Press.

Dhyne, E., L. J. Alvarez, H. Le Bihan, G. Veronese, D. Dias, J. Hoffmann, N. Jonker, P. Lunnemann, F. Rumler, and J. Vilmunen (2006). Price changes in the euro area and the united states: Some facts from individual consumer price data. *Journal of Economic Perspectives 20*(2), 171–192.

Diamond, P. (1982). Aggregate demand management in search equilibrium. *Journal of Political Economy 90*(5), 881–94.

Diamond, P. A. (1971). A Model of Price Adjustment. *Journal of Economic Theory 3*, 156–168.

Doepke, M. and R. M. Townsend (2006). Dynamic mechanism design with hidden income and hidden actions. *Journal of Economic Theory 126*(1), 235–285.

Domeij, D. and M. Floden (2006). The labor-supply elasticity and borrowing constraints: Why estimates are biased. *Review of Economic Dynamics 9*(2), 242–262.

Domeij, D. and J. Heathcote (2004). On the distributional effects of reducing capital taxes. *International Economic Review 45*(2), 523–554.

Eichenbaum, M., N. Jaimovich, and S. Rebelo (2011). Reference prices, costs, and nominal rigidities. *American Economic Review 101*(1), 234–62.

Elsby, M. W. and G. Solon (2019). How Prevalent Is Downward Rigidity in Nominal Wages? International Evidence from Payroll Records and Pay Slips. *Journal of Economic Perspectives 33*, 185–201.

Erceg, C. J., D. W. Henderson, and A. T. Levin (2000). Optimal monetary policy with staggered wage and price contracts. *Journal of monetary Economics 46*(2), 281–313.

Fallick, B. and C. A. Fleischman (2004). Employer-to-Employer Flows in the U.S. Labor Market: The Complete Picture of Gross Worker Flows. FEDS Working Papers 2004-34.

Fisher, J. D. and D. S. Johnson (2006). Consumption mobility in the united states: Evidence from two panel data sets. *Topics in Economic Analysis & Policy 6*(1).

Flavin, M. A. (1981). The adjustment of consumption to changing expectations about future income. *Journal of Political Economy 89*(5), 974–1009.

Foster, L., J. Haltiwanger, and C. J. Krizan (2001). Aggregate Productivity Growth: Lessons from Microeconomic Evidence. In C. R. Hulten, E. R. Dean, , and M. J. Harper (Eds.), *New Developments in Productivity Analysis*. NBER.

Friedman, M. (1957). *Theory of the consumption function*. Princeton University Press.

Friedman, M. (1968). The role of monetary policy. *The American Economic Review 58*(1).

Fujita, S. and G. Ramey (2009). The Cyclicality of Separation and Job Finding Rates. *International Economic Review 50*, 415–430.

Fukui, M. and T. Mukoyama (2024). Efficiency in Job-Ladder Models. Boston University and Georgetown University.

Gabaix, X. (2011). The Granular Origins of Aggregate Fluctuations. *Econometrica 79*, 733–772.

Galí, J. and L. Gambetti (2020). Has the us wage phillips curve flattened? a semi-structural exploration. In G. Castex, J. Galí, and D. Saravia (Eds.), *Changing Inflation Dynamics, Evolving Monetary Policy*, Volume 27 of *Series on Central Banking, Analysis, and Economic Policies*. Central Bank of Chile.

Galí, J. and M. Gertler (1999). Inflation dynamics: A structural econometric analysis. *Journal of monetary Economics 44*(2), 195–222.

Goodfriend, M. (1991). Interest rates and the conduct of monetary policy. *Carnegie-Rochester Conference Series on Public Policy 34*, 7–30.

Grigsby, J., E. Hurst, and A. Yildirmaz (2021). Aggregate nominal wage adjustments: New evidence from administrative payroll data. *American Economic Review 111*(2), 428–71.

Gürkaynak, R., B. Sack, and E. Swanson (2005). Do actions speak louder than words? the response of asset prices to monetary policy actions and statements. *International Journal of Central Banking 1*(1).

Hagedorn, M. and I. Manovskii (2008). The Cyclical Behavior of Equilibrium Unemployment and Vacancies Revisited. *American Economic Review 98*, 1692–1706.

Hall, G. J. and T. J. Sargent (2020, May). Debt and taxes in eight U.S. wars and two insurrections. Working Paper 27115, National Bureau of Economic Research.

Hall, R. E. (1978). Stochastic implications of the life cycle-permanent income hypothesis: theory and evidence. *Journal of Political Economy 86*(6), 971–987.

Hall, R. E. (1988). Intertemporal substitution in consumption. *Journal of Political Economy 96*(2), 339–357.

Hall, R. E. (2005). Employment Fluctuations with Equilibrium Wage Stickiness. *American Economic Review 95*, 50–65.

Hayashi, F., J. Altonji, and L. Kotlikoff (1996). Risk-sharing between and within families. *Econometrica*, 261–294.

Heathcote, J. (2005a). Fiscal policy with heterogeneous agents and incomplete markets. *The Review of Economic Studies 72*(1), 161–188.

Heathcote, J. (2005b). Fiscal policy with heterogeneous agents and incomplete markets. *Review of Economic Studies 72*, 161–188.

Heathcote, J., K. Storesletten, and G. L. Violante (2009). Quantitative macroeconomics with heterogeneous households. *Annual Review of Economics 1*(1), 319–354.

Heathcote, J., K. Storesletten, and G. L. Violante (2017). Optimal tax progressivity: An analytical framework. *Quarterly Journal of Economics 132*(4), 1693–1754.

Hopenhayn, H. (2014a). Firms, Misallocation, and Aggregate Productivity: A Review. *Annual Review of Economics 6*, 735–770.

Hopenhayn, H. and R. Rogerson (1993). Job Turnover and Policy Evaluation: A General Equilibrium Analysis. *Journal of Political Economy 101*, 915–938.

Hopenhayn, H. A. (2014b). On the Measure of Distortions. NBER Working Paper 20404.

Hosios, A. J. (1990). On the Efficiency of Matching and Related Models of Search and Unemployment . *Review of Economic Studies 57*, 279–298.

Hsieh, C.-T. and P. J. Klenow (2009). Misallocation and Manufacturing TFP in China and India. *Quarterly Journal of Economics 124*, 1403–1448.

Huggett, M. (1993). The risk-free rate in heterogeneous-agent incomplete-insurance economies. *Journal of Economic Dynamics and Control 17*(5-6), 953–969.

Huggett, M. (1997). The one-sector growth model with idiosyncratic shocks: Steady states and dynamics. *Journal of Monetary Economics 39*(3), 385–403.

Hulten, C. R. (1978). Growth Accounting with Intermediate Inputs. *Review of Economic Studies 45*, 511–518.

Jappelli, T. and L. Pistaferri (2006). Intertemporal choice and consumption mobility. *Journal of the European Economic Association 4*(1), 75–115.

Jappelli, T. and L. Pistaferri (2017). *The economics of consumption: theory and evidence*. Oxford University Press.

Jappelli, T. and L. Pistaferri (2020). Reported MPC and unobserved heterogeneity. *American Economic Journal: Economic Policy 12*(4), 275–297.

Jensen, M. K. (2018). Distributional comparative statics. *The Review of Economic Studies 85*(1), 581–610.

Judd, K. L. (1985). Redistributive taxation in a simple perfect foresight model. *Journal of Public Economics 28*(1), 59–83.

Kaas, L. (2021). Block-Recursive Equilibria in Heterogeneous-Agent Models. mimeo. Goethe University Frankfurt.

Kahn, S. (1997). Evidence of nominal wage stickiness from microdata. *The American Economic Review 87*(5), 993–1008.

Kaplan, G. and G. L. Violante (2014). A model of the consumption response to fiscal stimulus payments. *Econometrica 82*(4), 1199–1239.

Kaplan, G. and G. L. Violante (2018). Microeconomic heterogeneity and macroeconomic shocks. *Journal of Economic Perspectives 32*(3), 167–194.

Kaplan, G. and G. L. Violante (2022). The marginal propensity to consume in heterogeneous agent models. *Annual Review of Economics 14*, 747–775.

Kaplan, G., G. L. Violante, and J. Weidner (2014). The wealthy hand-to-mouth. *Brookings Papers on Economic Activity*, 121–154.

Kareken, J. and N. Wallace (1981). On the indeterminacy of equilibrium exchange rates. *The Quarterly Journal of Economics 96*(2), 207–222.

Kehoe, T. J. and D. K. Levine (1993). Debt-constrained asset markets. *The Review of Economic Studies 60*(4), 865–888.

Kehoe, T. J. and D. K. Levine (2001). Liquidity constrained markets versus debt constrained markets. *Econometrica 69*(3), 575–598.

Keynes, J. M. (1936). *The General Theory of Interest, Employment and Money*. New York: Harcourt, Brace & World.

Kiyotaki, N. and R. Wright (1989). On money as a medium of exchange. *Journal of Political Economy 97*(4), 927–54.

Klein, P., P. Krusell, and J.-V. Ríos-Rull (2008, 07). Time-Consistent Public Policy. *The Review of Economic Studies 75*(3), 789–808.

Klenow, P. J. and O. Kryvtsov (2008). State-dependent or time-dependent pricing: Does it matter for recent us inflation? *The Quarterly Journal of Economics 123*(3), 863–904.

Klenow, P. J. and B. A. Malin (2010). Microeconomic evidence on price-setting. In *Handbook of monetary economics*, Volume 3, pp. 231–284. Elsevier.

Klette, T. J. and S. Kortum (2004). Innovating Firms and Aggregate Innovation. *Journal of Political Economy 112*, 986–1018.

Kocherlakota, N. R. (1996). Implications of efficient risk sharing without commitment. *The Review of Economic Studies 63*(4), 595–609.

Krueger, D. and H. Lustig (2010). When is market incompleteness irrelevant for the price of aggregate risk (and when is it not)? *Journal of Economic Theory 145*(1), 1–41.

Krueger, D. and F. Perri (2006). Does income inequality lead to consumption inequality? Evidence and theory. *The Review of Economic Studies 73*(1), 163–193.

Krusell, P., T. Mukoyama, and A. Şahin (2010). Labour-Market Matching with Precautionary Savings and Aggregate Fluctuations. *Review of Economic Studies 77*, 1477–1507.

Krusell, P., T. Mukoyama, R. Rogerson, and A. Şahin (2017). Gross Worker Flows over the Business Cycle. *American Economic Review 107*, 3447–3476.

Krusell, P., T. Mukoyama, A. Şahin, and A. A. Smith Jr (2009). Revisiting the welfare effects of eliminating business cycles. *Review of Economic Dynamics 12*(3), 393–404.

Krusell, P. and A. A. Smith, Jr (1998). Income and wealth heterogeneity in the macroeconomy. *Journal of Political Economy 106*(5), 867–896.

Kuttner, K. N. (2001). Monetary policy surprises and interest rates: Evidence from the fed funds futures market. *Journal of monetary economics 47*(3), 523–544.

Kydland, F. E. and E. C. Prescott (1977). Rules rather than discretion: The inconsistency of optimal plans. *Journal of Political Economy 85*(3), 473–491.

Lagos, R. and R. Wright (2005). A unified framework for monetary theory and policy analysis. *Journal of Political Economy 113*(3), 463–484.

LaSalle, J. P. (1986). *The Stability and Control of Discrete Processes*. New York: Springer-Verlag.

Lee, Y. and T. Mukoyama (2018). A Model of Entry, Exit, and Plant-level Dynamics over the Business Cycle. *Journal of Economic Dynamics and Control 96*, 1–25.

Leeper, E. M., T. B. Walker, and S.-C. Yang (2010). Government investment and fiscal stimulus. *Journal of Monetary Economics 57*, 253–92.

Leland, H. E. (1968). Saving and uncertainty: The precautionary demand for saving. *The Quarterly Journal of Economics 82*(3), 465–473.

Ljungqvist, L. and T. J. Sargent (2012). *Recursive Macroeconomic Theory, 3rd Ed.* Cambridge, Massachusetts: MIT Press.

Ljungqvist, L. and T. J. Sargent (2018). *Recursive macroeconomic theory*. MIT press.

Lucas, R. (1982). Interest rates and currency prices in a two-country world. *Journal of Monetary Economics 10*(3), 335–359.

Lucas, R. E. and J. P. Nicolini (2015). On the stability of money demand. *Journal of Monetary Economics 73*, 48–65. Carnegie-Rochester-NYU Conference Series on Public Policy "Monetary Policy: An Unprecedented Predicament" held at the Tepper School of Business, Carnegie Mellon University, November 14-15, 2014.

Lucas, R. E. and N. L. Stokey (1983). Optimal fiscal and monetary policy in an economy without capital. *Journal of Monetary Economics 12*(1), 55–93.

Luttmer, E. G. (2011). On the Mechanics of Firm Growth. *Review of Economic Studies 78*, 1042–1068.

Mace, B. J. (1991). Full insurance in the presence of aggregate uncertainty. *Journal of Political Economy 99*(5), 928–956.

Mankiw, N. G. and R. Reis (2002, 11). Sticky Information versus Sticky Prices: A Proposal to Replace the New Keynesian Phillips Curve*. *The Quarterly Journal of Economics 117*(4), 1295–1328.

Mas-Colell, A., M. D. Whinston, and J. R. Green (1995). *Microeconomic Theory*. New York, New York: Oxford University Press.

Mavroeidis, S., M. Plagborg-Møller, and J. H. Stock (2014, March). Empirical evidence on inflation expectations in the new keynesian phillips curve. *Journal of Economic Literature 52*(1), 124–88.

Mazzocco, M. and S. Saini (2012). Testing efficient risk sharing with heterogeneous risk preferences. *American Economic Review 102*(1), 428–468.

McCall, J. J. (1970). Economics of Information and Job Search. *Quarterly Journal of Economics 84*, 113–126.

McLaughlin, K. J. (1994). Rigid Wages? *Journal of Monetary Economics 34*, 383–414.

Meghir, C. and L. Pistaferri (2011). Earnings, consumption and life cycle choices. In *Handbook of Labor Economics*, Volume 4, pp. 773–854. Elsevier.

Mehra, R. and E. C. Prescott (1985). The equity premium: A puzzle. *Journal of Monetary Economics 15*(2), 145–161.

Merz, M. (1995). Search in the Labor Market and the Real Business Cycle. *Journal of Monetary Economics 36*, 269–300.

Miller, B. L. (1974). Optimal consumption with a stochastic income stream. *Econometrica*, 253–266.

Modigliani, F. and R. Brumberg (1954). Utility analysis and the consumption function: An interpretation of cross-section data. *Franco Modigliani 1*(1), 388–436.

Moscoso Boedo, H. J. and T. Mukoyama (2012). Evaluating the Effects of Entry Regulations and Firing Costs on International Income Differences. *Journal of Economic Growth 17*, 143–170.

Muellbauer, J. (1994). The assessment: consumer expenditure. *Oxford Review of Economic Policy 10*(2), 1–41.

Mukoyama, T. and L. Popov (2014). The Political Economy of Entry Barriers. *Review of Economic Dynamics 17*, 383–416.

Nakamura, E. and J. Steinsson (2008). Five facts about prices: A reevaluation of menu cost models. *The Quarterly Journal of Economics 123*(4), 1415–1464.

Nelson, J. A. (1994). On testing for full insurance using consumer expenditure survey data. *Journal of Political Economy 102*(2), 384–394.

Obstfeld, M. (1984). Multiple stable equilibria in an optimizing perfect-foresight model. *Econometrica 52*(1), 223–228.

Obstfeld, M. and K. Rogoff (1983). Speculative hyperinflations in maximizing models: Can we rule them out?. *Journal of Political Economy 91*(4), 675 – 687.

Parker, J. A., N. S. Souleles, D. S. Johnson, and R. McClelland (2013). Consumer spending and the economic stimulus payments of 2008. *American Economic Review 103*(6), 2530–2553.

Phelps, E. S. (1967). Phillips curves, expectations of inflation and optimal unemployment over time. *Economica 34*(135), 254–281.

Phillips, A. W. (1958). The relation between unemployment and the rate of change of money wage rates in the united kingdom, 1861-1957. *economica 25*(100), 283–299.

Piazzesi, M. and M. Schneider (2016). Housing and macroeconomics. *Handbook of Macroeconomics 2*, 1547–1640.

Pissarides, C. A. (1985). Short-Run Equilibrium Dynamics of Unemployment, Vacancies, and Real Wages . *American Economic Review 75*, 676–690.

Pissarides, C. A. (2000). *Equilibrium Unemployment Theory, 2nd ed.* Cambridge, Massachusetts: MIT Press.

Quadrini, V. and J.-V. Ríos-Rull (2015). Inequality in macroeconomics. In *Handbook of Income Distribution*, Volume 2, pp. 1229–1302. Elsevier.

Restuccia, D. and R. Rogerson (2008). Policy Distortions and Aggregate Productivity with Heterogeneous Establishments. *Review of Economic Dynamics 11*, 707–720.

Romer, C. D. and D. H. Romer (2004). A new measure of monetary shocks: Derivation and implications. *American Economic Review 94*(4), 1055–1084.

Samuelson, P. (1958). An exact consumption-loan model of interest with or without the social contrivance of money. *Journal of Political Economy 66*.

Sandmo, A. (1970). The effect of uncertainty on saving decisions. *The Review of Economic Studies 37*(3), 353–360.

Sargent, T. J. and P. Surico (2011, February). Two illustrations of the quantity theory of money: Breakdowns and revivals. *American Economic Review 101*(1), 109–28.

Sargent, T. J. and N. Wallace (1985). Some unpleasant monetarist arithmetic. *Federal Reserve Bank of Minneapolis Quarterly Review 9*(1), 15 – 31.

Schechtman, J. and V. L. Escudero (1977). Some results on "an income fluctuation problem". *Journal of Economic Theory 16*(2), 151–166.

Schulhofer-Wohl, S. (2011). Heterogeneity and tests of risk sharing. *Journal of Political Economy 119*(5), 925–958.

Shimer, R. (2005). The Cyclical Behavior of Equilibrium Unemployment and Vacancies. *American Economic Review 95*, 25–49.

Sibley, D. S. (1975). Permanent and transitory income effects in a model of optimal consumption with wage income uncertainty. *Journal of Economic Theory 11*(1), 68–82.

Sims, C. A. (1972). Money, income, and causality. *American Economic Review 62*(4), 540 – 552.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica 48*(1), 1 – 48.

Sims, C. A. (1992). Interpreting the macroeconomic time series facts: The effects of monetary policy. *European economic review 36*(5), 975–1000.

Steinsson, J. (2003). Optimal monetary policy in an economy with inflation persistence. *Journal of Monetary Economics 50*(7), 1425–1456.

Stiglitz, J. E. and M. Rothschild (1970). Increasing risk: I. A definition. *Journal of Economic Theory 2*(3), 225–243.

Storesletten, K., C. I. Telmer, and A. Yaron (2001). The welfare cost of business cycles revisited: Finite lives and cyclical variation in idiosyncratic risk. *European Economic Review 45*(7), 1311–1339.

Straub, L. and I. Werning (2020, January). Positive long-run capital taxation: Chamley-judd revisited. *American Economic Review 110*(1), 86–119.

Taylor, J. B. (1980). Aggregate dynamics and staggered contracts. *Journal of political economy 88*(1), 1–23.

Taylor, J. B. (1993a). Discretion versus policy rules in practice. *Carnegie-Rochester Conference Series on Public Policy 39*, 195–214.

Taylor, J. B. (1993b). Discretion versus policy rules in practice. In *Carnegie-Rochester conference series on public policy*, Volume 39, pp. 195–214. Elsevier.

Townsend, R. M. (1980). *Models of money with spatially separated agents*, pp. 265–303. Federal Reserve Bank of Minneapolis Minneapolis.

Townsend, R. M. (1994). Risk and insurance in village india. *Econometrica*, 539–591.

Uhlig, H. (1996). A law of large numbers for large economies. *Economic Theory 8*, 41–50.

Uhlig, H. (2001). A Toolkit for Analysing Nonlinear Dynamic Stochastic Models Easily. In R. Mariom and A. Scott (Eds.), *Computational Methods for the Study of Dynamic Economies*. Oxford University Press.

Wallace, N. (1981). A hybrid fiat-commodity monetary system. *Journal of Economic Theory 25*(3), 421 – 430.

Wallace, N. (1998). A dictum for monetary theory. *Federal Reserve Bank of Minneapolis Quarterly Review 22*(1), 20 – 26.

Wieland, J. F. and M.-J. Yang (2016, March). Financial dampening. Working Paper 22141, National Bureau of Economic Research.

Woodford, M. (2003a). *Interest and Prices–Foundations of a Theory of Monetary Policy*. Princeton and Oxford: Princeton University Press.

Woodford, M. (2003b). *Interest and prices: Foundations of a theory of monetary policy*. Princeton, N.J.: Princeton University Press.

Zeldes, S. P. (1989). Optimal consumption with stochastic income: Deviations from certainty equivalence. *The Quarterly Journal of Economics 104*(2), 275–298.

# Appendices

# 3.A Appendix to Chapter 3

This appendix proves a discrete-time version of the Uzawa (1961) Theorem, following the continuous-time proof by Schlicht (2006). Schlicht's continuous-time proof was also discussed in Jones and Scrimgeour (2008). We will separate the Theorem into two parts.

First, consider the following aggregate production function

$$Y_t = \tilde{F}_t(K_t, L_t), \tag{3.A.1}$$

where $Y_t$ is output, $K_t$ is capital, and $L_t$ is labor input. The sequence of functions $\tilde{F}_t$ is defined for all $K$ and $L$ in the positive orthant and each $\tilde{F}_t$ exhibits constant returns to scale. Assume that the population (labor) grows at the constant rate $n$:

$$L_t = L_0(1+n)^t.$$

The first theorem asserts that, if there is a balanced growth path where $Y_t$ and $K_t$ grow at the same rate $\gamma$, the sequence of functions $\tilde{F}_t$ has to take a particular form, under which technological progress is labor-augmenting (Harrod neutral).

**Theorem 3.A.1 (Uzawa Theorem, Part I)** *Suppose that for all $Y$, $K$, and $L$ such that $Y = F_0(K, L)$,*

$$Y_t = \tilde{F}_t(K_t, L_t), \text{ where } Y_t = Y(1+\gamma)^t, K_t = K(1+\gamma)^t, \text{ and } L_t = L(1+n)^t$$

*holds for all $t$. That is, the sequence of functions $\tilde{F}_t$ is consistent with balanced growth, regardless of the initial conditions and factor combinations. Then there exists a function $F(K, L)$, defined for all $K$ and $L$, such that for all $K_t$, $L_t$, and $t$*

$$\tilde{F}_t(K_t, L_t) = F(K_t, A_t L_t).$$

*In particular, the function $F$ is identical to $\tilde{F}_0$. The variable $A_t$ grows at a constant rate $(1+\gamma)(1+n) - 1$.*

**Proof.** Take any $K_t$, $L_t$, and $t$. Define $K$ by $K = K_t/(1+\gamma)^t$, $L$ by $L = L_t/(1+n)^t$, and $Y$ by $Y = \tilde{F}_t(K_t, L_t)/(1+\gamma)^t$. Define $F(k, l) = \tilde{F}_0(k, l)$ for any $(k, l)$.

We know that $Y = \tilde{F}_0(K, L)$. So $Y = F(K, L)$. Because $\tilde{F}_0$ exhibits constant returns to scale, we can multiply both sides by $(1+\gamma)^t$ and obtain $Y(1+\gamma)^t = F(K(1+\gamma)^t, L(1+\gamma)^t)$. Transferring back using the definitions of $Y$, $K$, and $L$, we obtain

$$\tilde{F}_t(K_t, L_t) = F(K_t, A_t L_t),$$

where

$$A_t = \left(\frac{1+\gamma}{1+n}\right)^t.$$

■

The proof makes it clear that $F(K_t, A_t L_t)$ completely characterizes $\tilde{F}_t$ over its domain.

Now, in Part II of the Theorem, we specify a setting that is common in macroeconomic modeling and show that, indeed, output and capital have to grow at the same rate when the growth rates of capital, output, consumption, investment are constant over time. The capital accumulation equation is specified as

$$K_{t+1} - K_t = I_t - \delta K_t. \tag{3.A.2}$$

$I_t$ is investment and $\delta > 0$ is depreciation rate. In the goods market, output equals consumption $C_t$ plus investment $I_t$:

$$Y_t = C_t + I_t. \tag{3.A.3}$$

Note that the setting is more general than the Solow model covered in the main text: here, we don't impose any assumption on investment.

**Theorem 3.A.2 (Uzawa Theorem, PartII)** *Suppose that, under the assumptions* (3.A.2) *and* (3.A.3)*, there exists a growth path in which the investment is strictly positive* $I_t > 0$ *and the growth rates of* $Y_t$*,* $C_t$*,* $I_t$*, and* $K_t$ *are constant over time. Call these growth rates* $\gamma_Y$*,* $\gamma_C$*,* $\gamma_I$*, and* $\gamma_K$*. Then* $\gamma_Y = \gamma_K$*.*

**Proof.** From (3.A.2),

$$1 + \gamma_K = \frac{I_t}{K_t} - \delta$$

holds. Thus, $I_t/K_t$ has to remain constant, which implies $\gamma_I = \gamma_K$. Therefore, it suffices to show $\gamma_Y = \gamma_I$. Subtracting (3.A.3) for time $t$ from (3.A.3) for time $t+1$:

$$C_{t+1} - C_t + I_{t+1} - I_t = Y_{t+1} - Y_t.$$

Then

$$\frac{C_{t+1} - C_t}{C_t} C_t + \frac{I_{t+1} - I_t}{I_t} I_t = \frac{Y_{t+1} - Y_t}{Y_t} Y_t$$

and thus

$$\gamma_C C_t + \gamma_I I_t = \gamma_Y Y_t$$

holds. Again using (3.A.3),

$$(\gamma_C - \gamma_Y) C_t = (\gamma_Y - \gamma_I) I_t. \tag{3.A.4}$$

Suppose, by contradiction, $\gamma_Y \neq \gamma_I$. Because $I_t > 0$, (3.A.4) implies $C_t \neq 0$ and (3.A.4) can be rewritten as

$$\frac{I_t}{C_t} = \frac{\gamma_C - \gamma_Y}{\gamma_Y - \gamma_I}$$

and therefore $I_t/C_t$ has to remain constant. This fact means $\gamma_I = \gamma_C$, but because of (3.A.3) $\gamma_I = \gamma_C$ implies $\gamma_I = \gamma_C = \gamma_Y$. Contradiction. ■

Note that, in the Solow model, the latter half of the proof is unnecessary because $\gamma_Y = \gamma_I$ by the assumption of the constant saving rate.

# 4.A Appendix to Chapter 4

This Appendix includes the proofs of theorems and propositions in Chapter 4, and analyzes the NGM with a phase diagram.

## 4.A.1 Constraints in the consumption-saving problem

Consider the infinite horizon version of the consumption-saving problem P2, where the consumer maximizes

$$\max_{\{c_t,a_{t+1}\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t u(c_t),$$

subject to the sequential budget constraint

$$a_{t+1} = (1+r)a_t + w_t - c_t. \tag{4.A.5}$$

As we saw in the main text, this constraint is not sufficient in guaranteeing a well-behaved maximization problem, as Ponzi schemes are possible. In the finite-horizon model, we imposed the terminal condition $a_{T+1} \geq 0$. Would it make sense to impose a similar condition in the limit of this environment, $\lim_{T\to\infty} a_{T+1} \geq 0$? The answer is no, because such constraint would be "too restrictive." To see this, consider a situation where $w_t = \bar{w} > 0$ for all $t$. The consumer could borrow a small amount, for example $\varepsilon < \bar{w}/(1+r)$, which can be repaid with her income $\bar{w}$ in the following period. The outcome of this, $a_{t+1} = -\varepsilon$ for all $t$, violates the condition $\lim_{T\to\infty} a_{T+1} \geq 0$ but it would not constitute a Ponzi-scheme. In other words, such behavior is completely feasible in this environment. What is then the "natural borrowing limit" for this economy?

To find a reasonable alternative, let us start by assuming the flow budget constraint (4.A.5) always has to hold. Combining (4.A.5) up to time $T$, we can obtain

$$\sum_{t=0}^{T} \frac{c_t}{(1+r)^t} + \frac{a_{T+1}}{(1+r)^T} = (1+r)a_0 + \sum_{t=0}^{T} \frac{w_t}{(1+r)^t}. \tag{4.A.6}$$

It seems reasonable to impose

$$\lim_{T\to\infty} \frac{a_{T+1}}{(1+r)^T} \geq 0 \tag{4.A.7}$$

to ensure that

$$\sum_{t=0}^{\infty} \frac{c_t}{(1+r)^t} \leq (1+r)a_0 + \sum_{t=0}^{\infty} \frac{w_t}{(1+r)^t} \tag{4.A.8}$$

holds as $T \to \infty$. Here, we are assuming $\sum_{t=0}^{T} w_t/(1+r)^t$ is finite. Constraint (4.A.8) corresponds to the lifetime budget constraint in the main text. Condition (4.A.7) is the no Ponzi game (nPg) condition. It is straightforward to check that imposing (4.A.7) would prevent the Ponzi scheme described in the main text.

The natural borrowing limit in this setting turns out to be

$$a_{t+1} \geq -\sum_{s=t+1}^{\infty} \frac{w_s}{(1+r)^{s-t}}. \tag{4.A.9}$$

The next Theorem shows the equivalence of different ways of imposing the constraints.

**Theorem 4.A.1** *The following three constraint sets are equivalent.*

**(i)** *The flow budget constraint (4.A.5) for $t = 0, 1, ....$ and no Ponzi game condition (4.A.7)*

**(ii)** *The flow budget constraint (4.A.5) for $t = 0, 1, ....$ and the natural borrowing limit (4.A.9) for $t = 0, 1, ....$*

**(iii)** *The lifetime budget constraint (4.A.8)*

**Proof.** To show the equivalence, we start from showing that (i)⇒(ii), then (ii)⇒(iii), and then (iii)⇒(i).

**(i)⇒(ii):** Using (4.A.5) from time $t+1$ to $T$,

$$\frac{a_{T+1}}{(1+r)^T} = \frac{a_{t+1}}{(1+r)^t} + \sum_{s=t+1}^{T} \frac{w_s - c_s}{(1+r)^s} \tag{4.A.10}$$

holds. Rewriting (4.A.10):

$$a_{t+1} = (1+r)^t \frac{a_{T+1}}{(1+r)^T} - \sum_{s=t+1}^{T} \frac{w_s - c_s}{(1+r)^{s-t}}.$$

Taking $T \to \infty$,

$$a_{t+1} = (1+r)^t \lim_{T\to\infty} \frac{a_{T+1}}{(1+r)^T} - \sum_{s=t+1}^{\infty} \frac{w_s - c_s}{(1+r)^{s-t}}.$$

Using (4.A.7), this equation implies

$$a_{t+1} \geq -\sum_{s=t+1}^{\infty} \frac{w_s - c_s}{(1+r)^{s-t}}.$$

Because $c_s \geq 0$ for all $s$,

$$a_{t+1} \geq -\sum_{s=t+1}^{\infty} \frac{w_s}{(1+r)^{s-t}}$$

holds, which is (4.A.9). Because $t$ was arbitrary, we are done.

**(ii)⇒(iii):** To show that (4.A.9) for all $t$ implies (4.A.8), first note (4.A.9) implies

$$\frac{a_{t+1}}{(1+r)^t} \geq -\sum_{s=t+1}^{\infty} \frac{w_s}{(1+r)^s}. \tag{4.A.11}$$

The flow budget constraint (4.A.5) implies (4.A.6) holds for any $T$:

$$\frac{a_{T+1}}{(1+r)^T} = (1+r)a_0 + \sum_{t=0}^{T} \frac{w_t - c_t}{(1+r)^t}.$$

Thus, combining with (4.A.11),

$$(1+r)a_0 + \sum_{t=0}^{T} \frac{w_t - c_t}{(1+r)^t} \geq -\sum_{t=T+1}^{\infty} \frac{w_t}{(1+r)^t}.$$

holds. Rearranging and taking $T \to \infty$ (the right-hand side converges to zero[1]) results in (4.A.8).

**(iii)⇒(i):** Note that because (4.A.8) doesn't specify $a_1, a_2, ...$, one can create (4.A.6) by appropriately defining $a_T$. In turn, one can also create (4.A.5) that corresponds these $a_1, a_2, ...$, because (4.A.6) and (4.A.5) are equivalent.

Take a limit of (4.A.6) for $T \to \infty$,

$$\sum_{t=0}^{\infty} \frac{c_t}{(1+r)^t} + \lim_{T\to\infty} \frac{a_{T+1}}{(1+r)^T} = (1+r)a_0 + \sum_{t=0}^{\infty} \frac{w_t}{(1+r)^t}.$$

Rewriting,

$$\lim_{T\to\infty} \frac{a_{T+1}}{(1+r)^T} = (1+r)a_0 + \sum_{t=0}^{\infty} \frac{w_t}{(1+r)^t} - \sum_{t=0}^{\infty} \frac{c_t}{(1+r)^t}.$$

From (4.A.8), this equation implies (4.A.7).

■

[1]This result follows from $\sum_{t=0}^{\infty} \frac{w_t}{(1+r)^t}$ being finite. The right-hand side can be rewritten as

$$-\sum_{t=0}^{\infty} \frac{w_t}{(1+r)^t} + \sum_{t=0}^{T} \frac{w_t}{(1+r)^t},$$

whose limit with $T \to \infty$ equals zero.

### 4.A.2 Balanced growth and CRRA utility

The following theorem shows that CRRA utility is the only form of utility function consistent with balanced growth. In Appendix 3.A, we have seen that capital, output, and consumption have to grow at a same rate along the balanced growth. This fact also implies that the rental rate of capital $r = f'(k)$ is constant, because $K/Y = k/f(k)$ is constant (implying that $k$ is constant).

**Theorem 4.A.2** *For a twice differentiable utility function $u(c)$ to be consistent with balanced growth with any arbitrary growth rate $\gamma$, $u(c)$ has to be a CRRA utility function:*

$$u(c) = \log(c)$$

*or*

$$u(c) = \frac{c^{1-\sigma}}{1-\sigma},$$

*where $\sigma > 0$ and $\sigma \neq 1$.*

**Proof.** We require, for consistency with balanced growth, that

$$\frac{u'((1+\gamma)c)}{u'(c)} = \frac{1+r}{\beta} \equiv \#$$

holds for all $c$. The requirement must work for any $\gamma$—in which case $r$ may end up depending on $\gamma$. Therefore, we write $\#(\gamma)$.

So we have, for each $c$ and $\gamma$, that

$$u'((1+\gamma)c) = u'(c)\#(\gamma).$$

Because the expression holds for all $c$, we obtain

$$(1+\gamma)u''((1+\gamma)c) = u''(c)\#(\gamma)$$

so we can write, by dividing the second equation times $c$ by the first,

$$\frac{(1+\gamma)cu''((1+\gamma)c)}{u'((1+\gamma)c)} = \frac{u''(c)c}{u'(c)}.$$

Given that we require the condition to hold for all $\gamma$, we know that the function $u$ needs to be on the CRRA form: $u''(c)c/u'(c)$ must equal a constant. We label it $-\sigma$.

To see what functional form is implied, we write

$$-\frac{\sigma}{c} = \frac{u''(c)}{u'(c)}$$

and integrate so that we obtain

$$\log(c^{-\sigma}) = \log u'(c).$$

This immediate implies

$$u'(c) = c^{-\sigma} \quad \Rightarrow \quad u(c) = \frac{c^{1-\sigma}}{1-\sigma}$$

whenever $\sigma \neq 1$; when $\sigma = 1$, we obtain $u(c) = \log c$.[2] ■

### 4.A.3 Proof to Proposition 4.4

Consider any alternative feasible and interior sequence $\mathbf{x} \equiv \{x_{t+1}\}_{t=0}^{\infty}$, i.e., a sequence in the interior of $\Gamma(x_t)$ $\forall t$. We want to show that for any such sequence,

$$\lim_{T\to\infty} \sum_{t=0}^{T} \beta^t \left[\mathcal{F}\left(x_t^*,\ x_{t+1}^*\right) - \mathcal{F}\left(x_t,\ x_{t+1}\right)\right] \geq 0.$$

Define

$$A_T(\mathbf{x}) \equiv \sum_{t=0}^{T} \beta^t \left[\mathcal{F}\left(x_t^*,\ x_{t+1}^*\right) - \mathcal{F}\left(x_t,\ x_{t+1}\right)\right].$$

We will show that, as $T$ goes to infinity, $A_T(\mathbf{x})$ is bounded below by zero.

By concavity of $\mathcal{F}$,

$$A_T\left(\mathbf{x}\right) \geq \sum_{t=0}^{T} \beta^t \left[\mathcal{F}_1\left(x_t^*,\ x_{t+1}^*\right)\left(x_t^* - x_t\right) + \mathcal{F}_2\left(x_t^*,\ x_{t+1}^*\right)\left(x_{t+1}^* - x_{t+1}\right)\right].$$

Now notice that for each $t$, $x_{t+1}$ shows up twice in the summation. Hence, we can rearrange the expression to read

$$\begin{aligned} A_T\left(\mathbf{x}\right) \quad \geq \quad & \sum_{t=0}^{T-1} \beta^t \left\{\left(x_{t+1}^* - x_{t+1}\right)\left[\mathcal{F}_2\left(x_t^*,\ x_{t+1}^*\right) + \beta\mathcal{F}_1\left(x_{t+1}^*,\ x_{t+2}^*\right)\right]\right\} + \\ & + \mathcal{F}_1\left(x_0^*,\ x_1^*\right)\left(x_0^* - x_0\right) + \beta^T \mathcal{F}_2\left(x_T^*,\ x_{T+1}^*\right)\left(x_{T+1}^* - x_{T+1}\right). \end{aligned}$$

Some information contained in the first-order conditions will now be useful:

$$\mathcal{F}_2\left(x_t^*,\ x_{t+1}^*\right) + \beta\mathcal{F}_1\left(x_{t+1}^*,\ x_{t+2}^*\right) = 0,$$

together with $x_0^* - x_0 = 0$ ($x_0$ can only take on one feasible value), allows us to derive

$$A_T\left(\mathbf{x}\right) \geq \beta^T \mathcal{F}_2\left(x_T^*,\ x_{T+1}^*\right)\left(x_{T+1}^* - x_{T+1}\right).$$

[2] Because of l'Hôpital's rule, we can summarize as

$$u(c) = \lim_{s\to\sigma} \frac{c^{1-s} - 1}{1-s}$$

for all $\sigma \geq 0$.

In addition, $\mathcal{F}_2\left(x_T^*,\ x_{T+1}^*\right) = -\beta\mathcal{F}_1\left(x_{T+1}^*,\ x_{T+2}^*\right)$, so we obtain

$$A_T\left(\mathbf{x}\right) \geq \beta^{T+1}\mathcal{F}_1\left(x_{T+1}^*,\ x_{T+2}^*\right)\left(x_{T+1} - x_{T+1}^*\right) \geq -\beta^{T+1}\mathcal{F}_1\left(x_{T+1}^*,\ x_{T+2}^*\right)x_{T+1}^*.$$

In the finite horizon case, $x_{T+1}^*$ would have been the level of capital left out for the day after the (perfectly foreseen) end of the world; a requirement for an optimum in that case is clearly $x_{T+1}^* = 0$.

As $T$ goes to infinity, the right-hand side of the last inequality goes to zero by the transversality condition. That is, we have shown that the utility implied by the candidate path must be higher than that implied by the alternative.

### 4.A.4 Analyzing the NGM using the phase diagram

The dynamics of the social planner's solution to the NGM can be summarized by the following two difference equations: the resource constraint

$$k_{t+1} - k_t = f(k_t) - \delta k_t - c_t \tag{4.A.12}$$

and the Euler equation

$$u'(c_t) = \beta u'(c_{t+1})(f'(k_{t+1}) + 1 - \delta). \tag{4.A.13}$$

The steady-state values of $(k_t, c_t)$, denoted by $(\bar{k}, \bar{c})$, satisfy (from $k_{t+1} = k_t$ and $c_{t+1} = c_t$)

$$0 = f(\bar{k}) - \delta\bar{k} - \bar{c} \tag{4.A.14}$$

and

$$1 = \beta(f'(\bar{k}) + 1 - \delta). \tag{4.A.15}$$

From (4.A.12), we can see that $k_{t+1} > k_t$ if and only if $f(k_t) - \delta k_t - c_t > 0$, or

$$c_t < f(k_t) - \delta k_t.$$

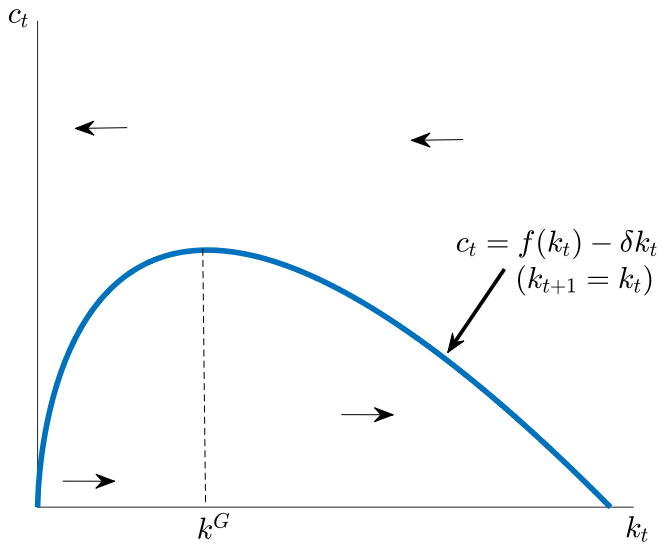
Similarly, $k_{t+1} < k_t$ if and only if $c_t > f(k_t) - \delta k_t$ and $k_{t+1} = k_t$ if and only if

$$c_t = f(k_t) - \delta k_t. \tag{4.A.16}$$

Therefore, in the $(k_t, c_t)$-plane, the curve (4.A.16) divides the entire plane (in particular the positive orthant that we care) into two regions: below the curve where $k_t$ increases over time and above the curve where $k_t$ decreases over time. Figure 1 draws this relationship. The arrows describe how $k_t$ moves over time in that region.

Note that it is straightforward to draw (4.A.16). The first term, $f(k_t)$, is an increasing and concave production function and the second term, $\delta k_t$, is a straight line through the origin. The vertical difference is $f(k_t) - \delta k_t$, which is hump-shaped. Note that $f(k_t) - \delta k_t$ is maximized at the point $k^G$ where $f'(k^G) = \delta$ is satisfied. The capital stock $k^G$ is often called the *golden rule capital stock*: it was considered as a "desirable" capital stock in earlier economic growth literature, because, when the steady-state resource constraint (4.A.14) is satisfied, this value of capital maximizes the consumption $c = f(k) - \delta k$.

Figure 1: Drawing $k_{t+1} > k_t$, $k_{t+1} < k_t$, and $k_{t+1} = k_t$




From (4.A.13), $c_{t+1} > c_t$ if and only if $\beta(f'(k_{t+1}) + 1 - \delta) > 1$. Note that from (4.A.15) and because $f'(k)$ is strictly decreasing in $k$, this condition is equivalent to $k_{t+1} < \bar{k}$. Using (4.A.12), the condition can be rewritten as

$$c_t > f(k_t) - \delta k_t + (k_t - \bar{k}).$$

Similarly, $c_{t+1} < c_t$ if and only if $c_t < f(k_t) - \delta k_t + (k_t - \bar{k})$ and $c_{t+1} = c_t$ if and only if

$$c_t = f(k_t) - \delta k_t + (k_t - \bar{k}). \tag{4.A.17}$$

In the $(k_t, c_t)$ plane, the curve (4.A.17) divides the plane into two regions: above the curve, $c_t$ increases over time, and below the curve, $c_t$ decreases over time. Figure 2 draws this relationship. Once again, the arrows describe the movement over time. In this figure, arrows are vertical because they signify the movement of $c_t$.

Drawing the curve (4.A.17) is also straightforward. Because the first two terms on the right-hand side are the same as the right-hand side in (4.A.16), we simply need to add $(k_t - \bar{k})$ to the hump-shaped curve we have already drawn.

Figure 3 puts two curves together. This diagram is called the *phase diagram*. Note here that the steady state $(\bar{k}, \bar{c})$, which corresponds to the crossing point of two curves (because it is the point where $k_{t+1} = k_t$ and $c_{t+1} = c_t$ both hold), is placed to the left of the largest point of the hump-shaped $k_{t+1} = k_t$ curve. This comparison follows from the facts that (i) the largest point of the hump-shaped curve, $k^G$, satisfies $f'(k^G) = \delta$, (ii) from (4.A.15) the steady state satisfies $f'(\bar{k}) = \delta + 1 - 1/\beta$ (and the right-hand side is larger than $\delta$), and (iii) $f'(k)$ is decreasing in $k$.

Figure 2: Drawing $c_{t+1} > c_t$, $c_{t+1} < c_t$, and $c_{t+1} = c_t$

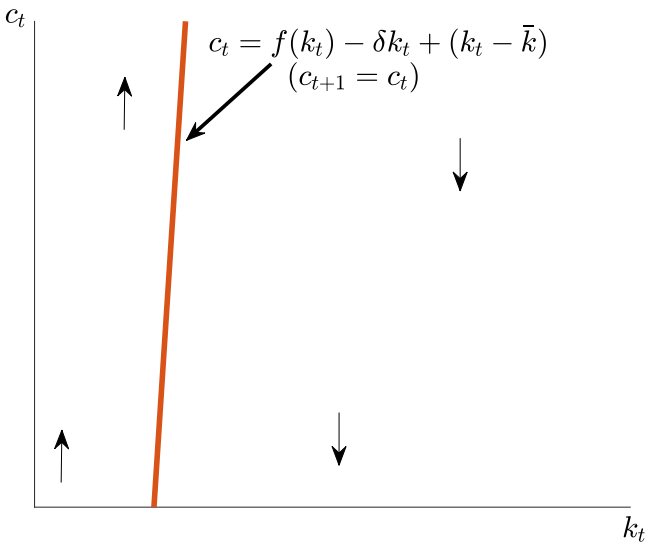
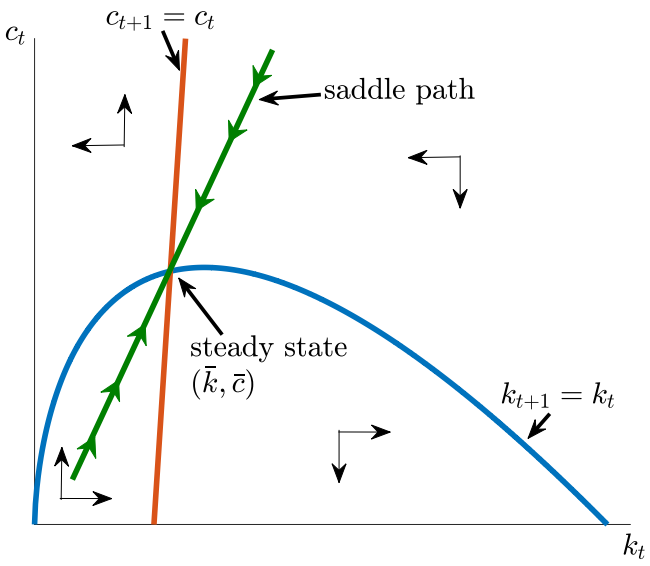



Figure 3: The phase diagram




It turns out that the dynamics of $(k_t, c_t)$ exhibit a *saddle-path* dynamics. The line with arrows in the figure is called the "saddle path" or the "stable arm." The $(k_t, c_t)$ sequence

converges to the steady state if and only if it starts on the saddle path. The steady state, in this case, is called the "saddle point," and the property of the dynamics is also referred to as being *saddle-path stable* (or *saddle-point stable*). If $(k_t, c_t)$ is not on the saddle path, it will eventually diverge from the steady state.

To analyze the dynamics, recall how the economy evolves. First, for a given $k_0$, the social planner chooses $c_0$. Once $c_0$ is chosen, the conditions (4.A.12) and (4.A.13) determines $(k_1, c_1)$, and the sequence of $(k_t, c_t)$ can be determined by these difference equations. The only question is: which value of $c_0$ should the social planner choose? It turns out that the social planner has to choose $c_0$ so that $(k_0, c_0)$ is on the saddle path. Thus, the path of $(k_t, c_t)$ follows the saddle path and converges to the steady state.

To see why $c_0$ on the saddle path has to be chosen for a given $k_0$, first (counterfactually) suppose that $c_0$ is chosen above the saddle path. From the diagram, one can see that $(k_t, c_t)$ eventually goes into the region where $k_t$ keeps decreasing and $c_t$ keeps increasing. In a finite time, $k_t$ becomes zero, and at that point, it becomes impossible to follow the differential equations (4.A.12) and (4.A.13) while satisfying the constraints $k_t \geq 0$ and $c_t \geq 0$. Thus, this choice of $c_0$ is not consistent with the optimal conditions.

If $c_0$ is chosen from below the saddle path, eventually $(k_t, c_t)$ goes into the region where both $k_t$ and $c_t$ keep decreasing. In particular, at a finite $T$, $k_t > k^G$ (and therefore $f'(k_t) < \delta$) for all $t > T$. One can show that this sequence of $(k_t, c_t)$ violates the TVC. First note that, using (4.A.13) repeatedly,

$$\beta^t u'(c_t)(f'(k_t) + 1 - \delta)k_t = \beta^{t-1} u'(c_{t-1})k_t = u'(c_0)\frac{1}{\prod_{s=0}^{t-1}(f'(k_s) + 1 - \delta)}k_t$$

holds. The TVC requires

$$\lim_{t\to\infty} \beta^t u'(c_t)(f'(k_t) + 1 - \delta)k_t = \lim_{t\to\infty} u'(c_0)\frac{1}{\prod_{s=0}^{t-1}(f'(k_s) + 1 - \delta)}k_t$$

to be zero. However, both $u'(c_0) > 0$ and $k_t$ is bounded from below by a strictly positive value (in particular, $k_t > k^G$ for all large $t$), and $\prod_{s=0}^{t-1}(f'(k_s)+1-\delta) \to 0$ because $f'(k_t)+1-\delta < 1$ for all $t > T$. Thus, $u'(c_t)(f'(k_t) + 1 - \delta)k_t$ diverges to $+\infty$, violating the TVC. And this result is intuitive: along this path, $c_t$ is consistently low (the consumers are eating very little), despite that $k_t$ keeps growing: a clear sign of oversaving. Eventually, the value of $k_t$ becomes inefficiently high, to the extent that the net return from the capital at the margin $f'(k) - \delta$ is negative (this situation is often called as *dynamic inefficiency*). We discuss the issues of dynamic inefficiency further in Chapter 6 (and its Appendix).

It is also straightforward to check that the saddle path satisfies the TVC. Because the steady state $\bar{k} < k^G$, $f'(\bar{k}) > \delta$ and therefore $\prod_{s=0}^{t-1}(f'(k_s) + 1 - \delta) \to \infty$ as $t \to \infty$, implying $u'(c_t)(f'(k_t) + 1 - \delta)k_t \to 0$ as $t \to \infty$.

In sum, the optimal path of $(k_t, c_t)$ converges to the steady state $(\bar{k}, \bar{c})$. This convergence property is qualitatively the same as in the Solow model. The difference here is that the saving rate is chosen by the social planner, and it can vary over time. The eventual steady state is derived from the social planner's optimizing behavior. One might wonder why the

social planner eventually chooses a $\bar{k}$ below the golden rule $k^G$, which maximizes the steady-state consumption. The reason is that the consumer (and thus the social planner) discounts the future. Even though the consumption level is higher under $k^G$ than under $\bar{k}$ after reaching the steady state, the consumer has to save extra to reach from $\bar{k}$ to $k^G$, and she can consume less during the transition. Because the consumer discounts the future, the cost from the short-run reduction of consumption is higher than the benefit of the eventual increase in consumption in the long run.

# 5.A Appendix to Chapter 5

To show that there is *aggregation* when $u$ is a power function (with power $1-\sigma$), use the functional form to obtain

$$((1-\delta+r(K))k+w(K)-g(k,K))^{-\sigma}=$$

$$\beta\left(g(k,K)(1-\delta+r(G(K)))+w(G(K))-g(g(k,K),G(K))\right)^{-\sigma}\left[1-\delta+r(G(K))\right].$$

Raising both sides to $-(1/\sigma)$ we obtain

$$(1-\delta+r(K))k+w(K)-g(k,K)=$$

$$\left(\beta\left[1-\delta+r(G(K))\right]\right)^{-\frac{1}{\sigma}}\left(g(k,K)(1-\delta+r(G(K)))+w(G(K))-g(g(k,K),G(K))\right).$$

By collecting terms and defining $A(K)$, $B(K)$, $C(K)$, and $D(K)$ appropriately, we see that we need that, for all $(k,K)$,

$$A(K)+B(K)k+C(K)g(k,K)+D(K)g(g(k,K),G(K))=0.$$

This makes clear that a solution that is linear in $k$ given $K$, is available, i.e.,

$$g(k,K)=\mu(K)+\lambda(K)k,$$

when, for all $K$, $\mu(K)$ and $\lambda(K)$ solve

$$A(K)+C(K)\mu(K)+D(K)\left(\mu(G(K))+\lambda(G(K))\mu(K)\right)=0$$

$$B(K)+C(K)\lambda(K)+D(K)\lambda(G(K))\lambda(K)=0.$$

The first of these equations makes sure the object multiplying $k$ is zero; the second makes sure the remainder is zero too. Thus, the Euler equation is satisfied for all $(k,K)$.

These last two functional equations in $\lambda$ and $\mu$ need to be solved and the solution (their functional forms) clearly depends on the functions taken as given here—$r$, $w$, and $G$—as well as on the primitive constants. For the consumer's problem, a solution can be sought independently of the shape of $G$. An equilibrium furthermore involves the consistency requirement that $G(K)=\mu(K)+\lambda(K)K$ for all $K$.

Aggregation obtains not only in this case. It also works for exponential utility and quadratic utility. Moreover, in all these three functional-form cases, it works also when one replaces consumption as an argument with an affine function of consumption, e.g., $\log(4c-3)$ or $-e^{-9c+7}$. You can see how applying these three functions, with their affine extensions, will deliver the same kind of equation in $g(k,K)$ as above. This class of preferences, in its entirety, is sometimes referred to as HARA preferences.[3]

Let us finally consider our typical closed-form example: utility is logarithmic, the production function is Cobb-Douglas, and there is full depreciation: $\delta=1$. Then $r(K)=\alpha AK^{\alpha-1}$

[3]HARA stands for *hyperbolic absolute risk aversion.*

and $w(K) = (1-\alpha)AK^{\alpha}$. We know from before that the equilibrium law of motion will be $K' = G(K) = \alpha\beta AK^{\alpha}$. In this case, it is straightforward—but somewhat tedious—to verify that

$$k' = g(k, K) = \beta r(K)k,$$

i.e., $\mu(K) = 0$ and $\lambda(K) = \beta r(K)$. How does one find this solution? The best approach is typically to solve a 2-period economy, starting in period 2 and working backwards. Then functional forms appear and one can guess on decision rules of this sort. After substitution of the guess into the Euler equation one can then verify that the guess works, which involves finding the specific parameters of the adopted functional form as a final fixed-point problem.

# 6.A Appendix to Chapter 6

Let us denote $F(k, \omega_y + \omega_o) + 1 - \delta$ by $f(k)$.

**Theorem 6.A.1** *A steady state $k^*$ is efficient if and only if $R^* \equiv f'(k^*) \geq 1$.*

Intuitively, the steady state consumption is $c^* = f(k^*) - k^*$. Figure 4 shows the attainable levels of steady state capital stock and consumption $(k^*,\ c^*)$, given the assumptions on $f$. The $\left(k^G,\ c^G\right)$ locus corresponds to the "golden rule" level of steady state capital and consumption, that maximize $c^G$.

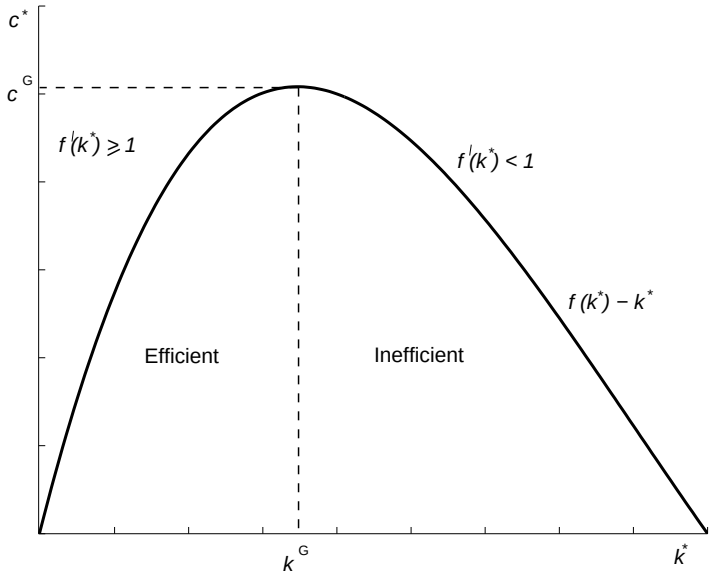



Figure 4: Efficiency of the steady state

**Proof.**

(i) $R^* < 1$: $k^*$ is inefficient.

Assume that $k^*$ is such that $f'(k^*) < 1$. Let $c^*$ denote the corresponding level of steady state consumption, let $c_0 = c^*$. Now consider a change in the consumption path, whereby $k_1$ is set to $k_1 = k^* - \varepsilon$ instead of $k_1 = k^*$. Notice this implies an increase in $c_0$. Let $k_t = k_1\ \forall t \geq 1$. We have that

$$\begin{aligned} c_1 - c^* &= f(k_1) - k_1 - f(k^*) + k^* \\ &\equiv f(k^* - \varepsilon) - (k^* - \varepsilon) - f(k^*) + k^*. \end{aligned}$$

Notice that strict concavity of $f$ implies that

$$f(k^*) < f(k^* - \varepsilon) + [k^* - (k^* - \varepsilon)]\, f'(k^* - \varepsilon)$$

for $\varepsilon \in \left(0,\, k^* - k^G\right)$, and we have that $f'\left(k^* - \varepsilon\right) < 1$. Therefore,

$$f(k^*) < f(k^* - \varepsilon) + k^* - (k^* - \varepsilon)\,.$$

This implies that

$$c_1 - c^* > 0,$$

which shows that a permanent increase in consumption is feasible.

(ii) $R^* \geq 1$: $k^*$ is efficient.

Suppose not, then we could decrease the capital stock at some point in time and achieve a permanent increase in consumption (or at least increase consumption at some date without decreasing consumption in the future). Let the initial situation be a steady state level of capital $k_0 = k^*$ such that $f'(k^*) \geq 1$. Let the initial $c_0$ be the corresponding steady state consumption: $c_0 = c^* = f(k^*) - k^*$. Since we suppose that $k^*$ is inefficient, consider a decrease of capital accumulation at time 0: $k_1 = k^* - \varepsilon_1$, thereby increasing $c_0$. We need to maintain the previous consumption profile $c^*$ for all $t \geq 1$: $c_t \geq c^*$. This requires that

$$\begin{aligned} c_1 = f(k_1) - k_2 &\geq f(k^*) - k^* = c^*, \\ k_2 &\leq f(k_1) - f(k^*) + k^*, \\ \underbrace{k_2 - k^*}_{\varepsilon_2} &\leq f(k_1) - f(k^*). \end{aligned}$$

Concavity of $f$ implies that

$$f(k_1) - f(k^*) < f'(k^*)\underbrace{[k_1 - k^*]}_{-\varepsilon_1}.$$

Notice that $\varepsilon_2 \equiv k_2 - k^* < 0$. Therefore, since $f'(k^*) \geq 1$ by assumption, we have that

$$|\varepsilon_2| > |\varepsilon_1|\,.$$

The size of the decrease in capital accumulation is increasing. By induction, $\{\varepsilon_t\}_{t=0}^{\infty}$ is a decreasing sequence (of negative terms). Since it is bounded below by $-k^*$, we know from real analysis that it must have a limit point $\varepsilon_\infty \in [-k^*,\, 0)$. Consequently, the consumption sequence converges as well:

$$c_\infty = f\left(k^* - \varepsilon_\infty\right) - \left(k^* - \varepsilon_\infty\right).$$

It is straightforward to show, using concavity of $f$, that

$$c_\infty < c^*.$$

Then the initial increase in consumption is not feasible if the restriction is to maintain at least $c^*$ as the consumption level for all the remaining periods of time.

■

# 7.A Appendix to Chapter 7

## 7.A.1 Recursive equilibrium for the stochastic growth model

When defining a recursive equilibrium, the first challenge is to identify the aggregate state variable(s). In the neoclassical growth model without uncertainty, we saw that the aggregate state was the stock of capital. If there are stochastic shocks to productivity, and we consider such shocks here, some aspect of the productivity process will need to be added to the aggregate state. What precise aspect depends on the nature of the productivity process. If we assume that it is first-order Markov in nature, so that nothing beyond the current value of $\omega$ will matter either for productivity today or for the probabilities of different outcomes of $\omega$ in the future, we can simply add $\omega$; hence the aggregate state variable is $(K, \omega)$.

**Definition 16** *A **recursive competitive equilibrium** consists of functions $r(K,\omega)$, $w(K,\omega)$, $G^\star(K,\omega)$, $V^\star(k,K,\omega)$, and $g^\star(k,K,\omega)$ such that*

1. *$V^\star(k,K,\omega)$ solves, for all $(k,K,\omega)$,*

   $$V(k,K,\omega) = \max_{k'} u((1-\delta+r(K,\omega))k + w(K,\omega) - k') + \beta V(k', G^\star(K,\omega), \omega') \quad \forall (k,K).$$

   *and $k' = g^\star(k,K,\omega)$ attains the maximum in this problem*

2. *for all $K$, $r(K,\omega) = A(\omega)F_K(K,1)$ and $w(K,\omega) = A(\omega)F_L(K,1)$; and*

3. *$G^\star(K,\omega) = g^\star(K,K,\omega)$ for all $(K,\omega)$.*

## 7.A.2 Proof of the law of iterated expectations

For any $\tau \geq t \geq 0$

$$\mathbb{E}_0\left\{\mathbb{E}_t\left[x_\tau(\omega^t)\right]\right\} = \mathbb{E}_0\left[x_\tau(\omega^\tau)\right].$$

This is an application law of iterated expectations, which is more general in that it does not just apply to stochastic processes. To prove this, note

$$\mathbb{E}_0\left\{\mathbb{E}_t\left[x_\tau(\omega^\tau)\right]\right\} = \sum_{\omega^t} \pi\left(\omega^t\right) \mathbb{E}_\tau\left[x_\tau(\omega^\tau)\right].$$

Now replace the conditional expectation $\mathbb{E}_\tau$

$$\mathbb{E}_0\left\{\mathbb{E}_t\left[x_\tau(\omega^\tau)\right]\right\} = \sum_{\omega^t} \pi\left(\omega^t\right) \sum_{\omega^\tau} x_\tau(\omega^\tau)\pi(\omega^\tau|\omega^t).$$

Now notice that $\pi\left(\omega^t\right)\pi(\omega^\tau|\omega^t) = \pi(\omega^t \cup \omega^\tau) = \pi(\omega^t|\omega^\tau)\pi(\omega^\tau)$. Because $t < \tau$, once we know $\omega^\tau$ we already have observed $\omega^t$. This means that the probability distribution $\pi(\omega^t|\omega^\tau)$ is

degenerate—it places all the probability mass on a single $\omega^t$. Continuing, we have

$$\begin{aligned}\mathbb{E}_0\left\{\mathbb{E}_t\left[x_\tau(\omega^\tau)\right]\right\} &= \sum_{\omega^t}\sum_{\omega^\tau} x_\tau(\omega^\tau)\pi\left(\omega^\tau\right)\pi(\omega^t|\omega^\tau) \\ &= \sum_{\omega^\tau} x_\tau(\omega^\tau)\pi\left(\omega^\tau\right)\sum_{\omega^t}\pi(\omega^t|\omega^\tau),\end{aligned}$$

where the second line follows from changing the order of summation and noting that $x_\tau(\omega^\tau)\pi\left(\omega^\tau\right)$ does not depend on $t$. Finally, we use the fact that the probability distribution $\pi(\omega^t|\omega^\tau)$ sums to one so we have

$$\mathbb{E}_0\left\{\mathbb{E}_t\left[x_\tau(\omega^\tau)\right]\right\} = \sum_{\omega^\tau} x_\tau(\omega^\tau)\pi\left(\omega^\tau\right) = \mathbb{E}_0\left[x_\tau(\omega^\tau)\right].$$

# 13.A Appendix to Chapter 13

## 13.A.1 Data Appendix

The data on revenues, expenditures, and deficits is obtained from the Bureau of Economic Statistics (BEA), Table 3.1 "Government Current Receipts and Expenditures," which is part of the NIPA tables[4]. The series, expressed in current billion dollars, span the interval 1929-2021. They include Federal, State, and Local government budget measures (sometimes referred to as General Government or National Government statistics). In the plots, the series are expressed as percentages of "GDP," corresponding to *Gross Domestic Product* (line 1 of Table 1.5.5).

**Figure 13.1:** Data is obtained from the OECD Dataset: National Accounts at a Glance, and the variable corresponds to *Total expenditure of general government, percentage of GDP*.

**Figure 13.2:** "Revenues" correspond to *Total Receipts* (line 34 of Table 3.1) and "Expenditures" to *Total Expenditures* (line 37 of Table 3.1). We use total rather than current measures because these include public investment.

**Figure 13.3:** All series are obtained from Table 3.1 "Government Current Receipts and Expenditures," constructed by the BEA. "Income Taxes" corresponds to *Personal current taxes* (line 3), "Sales and Import Taxes" to *Taxes on production and imports* (line 4), "Corporate Taxes" to *Taxes on corporate income* (line 12), and "Social Ins. Taxes" to *Contributions for government social insurance* (line 7).

**Figure 13.4 - Left Panel:** "Total Deficit or Surplus," in Figure 13.4, is constructed as the difference between Revenues and Expenditures (defined above),

$$\text{Total Deficit}=\text{Revenues-Expenditures}.$$

"Net Interest" is the difference between *Interest and Miscellaneous Receipts* (line 11 of Table 3.1) and *Interest Payments* (line 27 of Table 3.1). The government simultaneously owns assets that yield interest and owes debt for which it has to pay interest. In the figure, we plot net interest payments. The "Primary Deficit" is defined as

$$\text{Primary Deficit}= \text{Total Deficit - Net Interest}.$$

**Figure 13.4 - Right Panel:** FRED provides debt series for Federal and State governments between 1946 and 2021. "Federal Debt" corresponds to *Federal Government; Debt Securities and Loans; Liability, Level* (or FGSDODNS), while "State and Local Debt" corresponds to *State and Local Governments; Debt Securities and Loans; Liability, Level* (or SLGSDODNS).

[4]See https://apps.bea.gov/iTable

Both series are obtained from the Flow of Funds tables constructed by the Board of Governors of the Federal Reserve Bank System. It is worth noticing that the Federal Debt series does not correspond exactly to the one provided by the White House historical series due to differences in accounting methods (i.e. which items are included and timing in which certain transactions are incorporated when computing the flow of funds). The series between 1916 and 1945 are obtained from the Survey of Current Business, September 1946 page 13, Table 5. They correspond to Net Public Debt, end of calendar year.

**Figure 13.5 - Left Panel:** All series are obtained from Table 3.1 (described above). "Govt Consumption (+ Investment)" is the sum of *Consumption expenditures* (line 20) and *Gross government investment* (line 39). "Transfers" is the sum of *Current transfer payments* (line 22) and *Subsidies* (line 30). "Interest Payments" are gross, obtained from line 27 (i.e. we are not including interest receipts). The sum of these is equal to Expenditures, defined above.

Expenditures= Govt Consumption (+ Investment) + Transfers + Interest Payments.

**Figure 13.5 - Right Panel:** All series are obtained from Table 3.16 "Government Current Expenditures by Function" constructed by the BEA. "Defense" corresponds to *National Defense* (line 7), "Healthcare" corresponds to *Health* (line 28), and "Education" is obtained directly form line 30. "Income Security" is obtained from line 36. It includes *Disability* (line 37), *Welfare and social services* (line 39), *Unemployment* (line 40), *Retirement* (line 38) and other income insurance programs (line 41). "Other" is constructed as the sum of *General public service* (line 2), *Public order and safety* (line 8), *Economic affairs* (line 13), *Housing and community services* (line 27), *Recreation and culture* (line 29), minus *Interest payments* (line 5).

## 13.A.2 Tax reform with wealth effects

In this section, we re-compute the tax reform from Section 13.3.3, but assuming that utility takes the form

$$u(c,\ell) = \ln c - \frac{\ell^{1+\frac{1}{\phi}}}{1+\frac{1}{\phi}}.$$

The first order condition with respect to labor implies

$$\ell^{1/\phi} = \left(\frac{1-\tau_t^l}{1+\tau_t^c}\right) w \frac{1}{c_w} \quad \text{with} \quad c_w = \left(\frac{1-\tau_t^l}{1+\tau_t^c}\right) w.$$

The labor supply is *independent* of taxes in this case, $\ell = 1$. This happens because the substitution effect, that would make $\ell$ decline when after-tax labor income goes down is exactly offset by the income effect, caused by a decline in $c_w$ that results in lower after-tax income.

The main difference between Figure 13.7 and Figure 5 is that now labor supply remains constant when labor taxes are increased. As a result, we do not observe a decline in output,

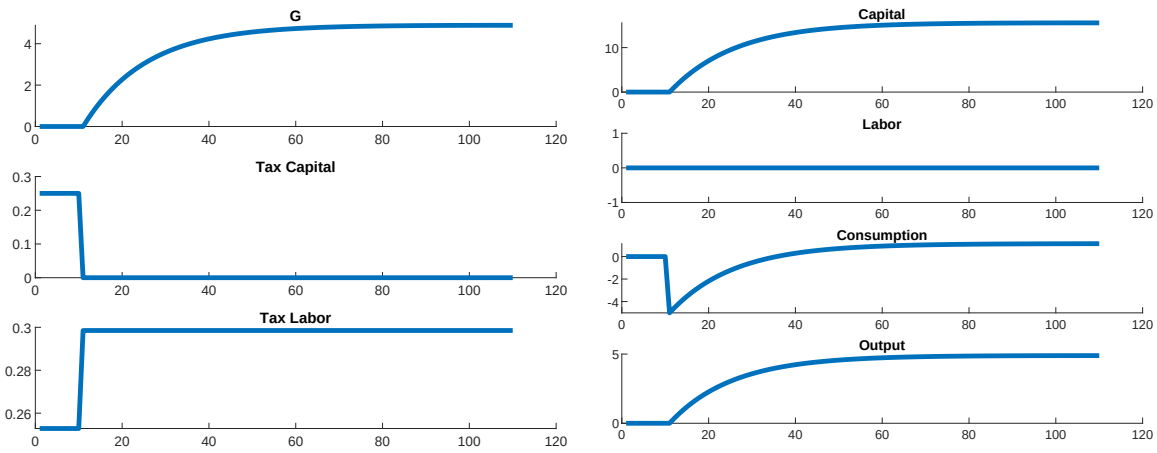



Figure 5: Eliminating capital income taxes (with wealth effects)

which allows the government to have higher G (recall that the exercise is constructed such that $G/Y = 0.2$ throughout the simulation). In the long-run, because labor does not go down, there is higher GDP and aggregate consumption is slightly higher. The tax reform is more effective in this scenario because the costs of replacing capital taxes with labor taxes are smaller when wealth effects are present.

# 16.A Appendix to Chapter 16

## 16.A.1 Derivation of the New Keynesian Phillips curve

The envelope condition of the intermediate goods producer's Bellman equation is

$$V_p(p,\mathcal{S}) = u'(C(\mathcal{S}))Y(\mathcal{S})\left[p^{-\varepsilon} - \varepsilon p^{-\varepsilon-1}\left(p - \frac{w(\mathcal{S})}{A(\mathcal{S})}\right)\right] + \beta\mathbb{E}\left[\theta V_p\left(\frac{p}{1+\pi(\mathcal{S}')}, \mathcal{S}'\right)\frac{1}{1+\pi(\mathcal{S}')}\right]$$

and the first-order condition of the price-setting problem is

$$V_p(p,\mathcal{S}) = 0.$$

$V_p$ gives the benefit to the firm of having a higher price. Notice that the first term on the right-hand side of the envelope condition is the marginal increase in profit this period from setting a higher price (multiplying by $u'(C)$ values this change in profit in terms of utility). The envelope condition then has the form of an expected discounted sum of marginal changes in profits today and, if prices are sticky ($\theta > 0$), in the future.

Combining the envelope condition and first order condition, the solution to the price-setting problem at date $t$, $p_t^R$, must satisfy

$$0 = \mathbb{E}_t \sum_{\tau=t}^{\infty} (\beta\theta)^{\tau-t} u'(C_\tau) Y_\tau \left[{p_{t,\tau}^R}^{-\varepsilon} - \varepsilon {p_{t,\tau}^R}^{-\varepsilon-1}\left(p_{t,\tau}^R - \frac{w_\tau}{A_\tau}\right)\right],$$

where $p_{t,\tau}^R \equiv P_t^R/P_\tau = p_t^R/\prod_{s=t+1}^{\tau}(1+\pi_s)$, is the relative price at date $\tau$ of a firm that last updated its price at date $t$. Rearranging we obtain

$$\mathbb{E}_t \sum_{\tau=t}^{\infty} (\beta\theta)^{\tau-t} u'(C_\tau) Y_\tau {p_{t,\tau}^R}^{-\varepsilon} = \frac{\varepsilon}{\varepsilon-1}\mathbb{E}_t \sum_{\tau=t}^{\infty} (\beta\theta)^{\tau-t} u'(C_\tau) Y_\tau {p_{t,\tau}^R}^{-\varepsilon-1}\frac{w_\tau}{A_\tau}.$$

We now log-linearize both sides of this equation around a zero-inflation steady state in which $p_{t,\tau}^R = 1$ for all $\tau$ to obtain

$$\mathbb{E}_t \sum_{\tau=t}^{\infty} (\beta\theta)^{\tau-t} \hat{p}_{t,\tau}^R = \mathbb{E}_t \sum_{\tau=t}^{\infty} (\beta\theta)^{\tau-t}\left(\hat{w}_\tau - \hat{A}_\tau\right), \tag{16.A.18}$$

where hats denote log deviations from steady state. From the definition of $p_{t,\tau}^R$ we have[5]

$$\hat{p}_{t,\tau}^R = \hat{P}_t^R - \hat{P}_\tau = \hat{P}_t^R - \hat{P}_t - (\hat{P}_\tau - \hat{P}_t) = \hat{p}_t^R - (\hat{P}_\tau - \hat{P}_t).$$

Eq. (16.A.18) then becomes

$$\begin{aligned}
\hat{p}_t^R &= (1-\beta\theta)\mathbb{E}_t \sum_{\tau=t}^{\infty} (\beta\theta)^{\tau-t}\left[\hat{w}_\tau - \hat{A}_\tau + \hat{P}_\tau - \hat{P}_t\right], \\
&= (1-\beta\theta)\left(\hat{w}_t - \hat{A}_t\right) + (1-\beta\theta)\mathbb{E}_t \sum_{\tau=t+1}^{\infty} (\beta\theta)^{\tau-t}\left[\hat{w}_\tau - \hat{A}_\tau + \hat{P}_\tau - \hat{P}_t\right] \\
&= (1-\beta\theta)\left(\hat{w}_t - \hat{A}_t\right) + \beta\theta\mathbb{E}_t\left(\hat{p}_{t+1}^R + \hat{P}_{t+1} - \hat{P}_t\right).
\end{aligned} \tag{16.A.19}$$

[5] Here we use $\widehat{(1+\pi_s)} = \pi_s$.

Note that $\pi_{t+1} \approx \log(1+\pi_{t+1}) = \hat{P}_{t+1} - \hat{P}_t$. Log-linearizing eq. (16.10) yields

$$\pi_t = \frac{1-\theta}{\theta}\hat{p}_t^R \tag{16.A.20}$$

and combining this with eq. (16.A.19) yields

$$\pi_t = \frac{(1-\theta)(1-\beta\theta)}{\theta}\left(\hat{w}_t - \hat{A}_t\right) + \beta\mathbb{E}_t\pi_{t+1}. \tag{16.A.21}$$

This equation has the form of a New Keynesian Phillips curve in which the term $\hat{w}_t - \hat{A}_t$ is the (log-linearized) real marginal cost of producing goods. The last step is to express this marginal cost in terms of the output gap. We do that using the production function, the aggregate resource constraint, and the household's labor supply condition.

Log-linearizing the aggregate production function, eq. (16.11), yields

$$\hat{Y}_t = \hat{A}_t - \hat{D}_t + \hat{L}_t$$

and log-linearizing eq. (16.12) around the steady state values $\bar{D} = \bar{p}^R = 1 + \bar{\pi} = 1$ yields

$$\hat{D}_t = -\varepsilon(1-\theta)\hat{p}_t^R + \theta\varepsilon\pi_t + \theta\hat{D}_{t-1}$$

and using (16.A.20) this simplifies to $\hat{D}_t = \theta\hat{D}_{t-1}$. As there is no price dispersion in steady state, $\hat{D}_t = \theta\hat{D}_{t-1}$ implies that $\hat{D}_t = 0$ for all $t$. That is, price dispersion does not affect the first-order approximation to the dynamics of the economy.

Log-linearizing the labor supply condition,(16.6), we obtain

$$-\sigma\hat{C}_t + \hat{w}_t = \psi\hat{L}_t.$$

Now using the resource constraint $Y_t = C_t$ and the log-linearized production function we have

$$\hat{w}_t = (\sigma+\psi)\hat{Y}_t - \psi A_t. \tag{16.A.22}$$

Log-linearizing (16.13) we obtain

$$Y_t^n = \frac{1+\psi}{\psi+\sigma}\hat{A}_t. \tag{16.A.23}$$

Finally, we can combine (16.A.22) and (16.A.23) to obtain

$$\hat{w}_t - \hat{A}_t = (\sigma+\psi)(\hat{Y}_t - \hat{Y}_t^n).$$

We then plug this into (16.A.21) to obtain equation (16.16).

### 16.A.2 Taylor Principle

This explanations draws on Bullard and Mitra (2002). We are interested in the stability of the dynamic system defined by the IS curve (16.15), the Phillips curve (16.16), and the interest rate rule (16.17). Here we substitute out for the nominal interest rate using the interest rate rule. We will study a deterministic economy, which is without loss of generality given the certainty equivalent property of first-order accurate economies. The system can then be expressed as

$$\begin{aligned}\beta\sigma\hat{Y}_{t+1} &= \beta\sigma\hat{Y}_t + \beta\phi_\pi\pi_t + \beta\phi_x\hat{Y}_t - \pi_t + \kappa\hat{Y}_t \\ \beta\pi_{t+1} &= \pi_t - \kappa\hat{Y}_t,\end{aligned}$$

where the first equation is the IS curve with $\pi_{t+1}$ substituted out using the Phillips curve. For simplicity, we are assuming $\pi^* = 0$. We can then write this system as

$$\begin{pmatrix}\beta\sigma & 0 \\ 0 & \beta\end{pmatrix}\begin{pmatrix}\hat{Y}_{t+1} \\ \pi_{t+1}\end{pmatrix} = \begin{pmatrix}\beta\sigma + \beta\phi_x + \kappa & \beta\phi_\pi - 1 \\ -\kappa & 1\end{pmatrix}\begin{pmatrix}\hat{Y}_t \\ \pi_t\end{pmatrix}.$$

Inverting the coefficient matrix on the right-hand side we have

$$\frac{1}{\sigma + \phi_x + \kappa\phi_\pi}\begin{pmatrix}\sigma & 1 - \beta\phi_\pi \\ \kappa\sigma & \beta\sigma + \beta\phi_x + \kappa\end{pmatrix}\begin{pmatrix}\hat{Y}_{t+1} \\ \pi_{t+1}\end{pmatrix} = \begin{pmatrix}\hat{Y}_t \\ \pi_t\end{pmatrix}.$$

This model has no state variables and two forward-looking variables. For a unique equilibrium, we need both eigenvalues of the coefficient matrix that multiplies $(\hat{Y}_{t+1}, \pi_{t+1})$ to be inside the unit circle. The characteristic polynomial of this matrix is $p(\lambda) \equiv \lambda^2 + a_1\lambda + a_0$ where

$$\begin{aligned}a_1 &= -\left(\frac{\sigma + \beta\sigma + \beta\phi_x + \kappa}{\sigma + \phi_x + \kappa\phi_\pi}\right) \\ a_0 &= \frac{\beta\sigma}{\sigma + \phi_x + \kappa\phi_\pi}.\end{aligned}$$

When are the roots of this polynomial inside the unit circle? We can answer this using the Jury Stability Criterion,[6] which in this specific case (a quadratic polynomial with a unit coefficient on $\lambda^2$) requires that $|a_0| < 1$ and $|a_1| < 1 + a_0$. The former condition can be expressed as $-(1-\beta)\sigma < \phi_x + \kappa\phi_\pi$, which holds as we assume $\beta \in [0, 1]$, all the parameters are weakly positive and $\sigma$ is strictly positive. Turning to the condition $|a_1| < 1 + a_0$, this can be expressed as condition (16.18).

### 16.A.3 A model with nominal wage and price rigidities

The representative household has the same preferences as in the sticky-price model except we now denote hours worked by $N_t$ while $L_t$ will refer to effective labor as we explain below.

[6] See the discussion of the related Schur-Cohn Criterion in on p. 27 of LaSalle (1986).

So preferences are

$$U_0 = \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \left[ \frac{C_t^{1-\sigma}}{1-\sigma} - \frac{N_t^{1+\psi}}{1+\psi} \right]. \tag{16.A.24}$$

This labor is supplied to a continuum of labor unions that differentiate it. Let $n_{i,t}$ be the amount of labor used by union $i \in [0,1]$.

The technology for producing goods is the same as in eqs. (16.2) and (16.3). In particular, $\ell_{j,t}$ remains the amount of labor used in producing good $j$. The labor used in producing the final good is a composite of the various types supplied by the unions. Specifically, the total effective labor supply is

$$L_t = \left( \int_0^1 n_{i,t}^{\frac{\varepsilon-1}{\varepsilon}} di \right)^{\frac{\varepsilon}{\varepsilon-1}}, \tag{16.A.25}$$

and market clearing requires $\int_0^1 \ell_{j,t} dj = L_t$. Here we have assumed the elasticity of substitution between labor varieties is the same as that between intermediate goods. This assumption is not necessary and we make it only for the sake of highlighting the similarity between the sticky-wage and sticky-price models.

Let $W_{i,t}$ be the nominal wage of type-$i$ labor—that is, the labor supplied by union $i$. The cost-minimization problem of intermediate firm $j$ implies firm $j$'s demand for type-$i$ labor is $(W_{i,t}/W_t)^{-\varepsilon} \ell_{j,t}$, where $W_t$ is the cost of producing one unit of labor aggregate. This aggregate wage index is

$$W_t = \left( \int_0^1 W_{i,t}^{1-\varepsilon} di \right)^{1/(1-\varepsilon)}, \tag{16.A.26}$$

Total labor income across all unions is

$$\int_0^1 \int_0^1 W_{i,t} (W_{i,t}/W_t)^{-\varepsilon} \ell_{j,t} dj di = \int_0^1 W_{i,t} (W_{i,t}/W_t)^{-\varepsilon} di L_t = W_t L_t.$$

Effective labor supply, $L_t$, is related to hours worked as follows. Total hours worked must equal the total usage by the unions $N_t = \int_0^1 n_{i,t} di$. In turn, total labor usage by union $i$ must equal the total supplied to each firm

$$n_{i,t} = \int_0^1 (W_{it}/W_t)^{-\varepsilon} \ell_{jt} dj = (W_{it}/W_t)^{-\varepsilon} L_t.$$

Putting the two together we have

$$N_t = \underbrace{\int_0^1 (W_{it}/W_t)^{-\varepsilon} di}_{\equiv D_t^W} L_t.$$

The term $D_t^W \geq$ reflects wage dispersion, which results in hours worked exceeding the effective labor supply as work effort is inefficiently allocated across types of labor.

We introduce a nominal rigidity for wages that is analogous to the one we assumed for prices. Each period, union $i$ is able to update the wage for $i$-type labor with probability $1-\theta_w \in [0,1]$. As in the intermediate goods firm's problem, the union must supply whatever quantity of labor is demanded at the prevailing price. By participating in the union, the household agrees to work the amount of hours the union needs in exchange for the wage the union sets.

The household takes aggregate labor income and hours worked as given as determined by the labor union. Their Euler equation is unchanged from the sticky-price model. By the envelope theorem, the marginal value of funds at date $t$ is $u'(C_t)$ and the marginal disutility of labor supply is $N_t^{\psi}$.

We now turn to the union's wage-setting problem, which is quite similar to the firm's price-setting problem. When raising the wage, the firm raises labor income but this comes at a disutility cost. As the wage is sticky, the union makes a forward-looking choice. The objective of the union that sets its wage at date $t$ is

$$\mathbb{E}_t \sum_{\tau=t}^{\infty} (\beta\theta^w)^{\tau-t} \left[ u'(C_\tau) \frac{W_t^R}{P_\tau} - N_\tau^{\psi} \right] \left( \frac{W_t^R}{W_\tau} \right)^{-\varepsilon} L_\tau,$$

where $W_t^R$ is the wage chosen at date $t$. In this objective, we assume that real labor income is valued at $\beta^{\tau-t}u'(C_\tau)$. Each union takes this marginal utility as given because each union contributes an infinitesimal part of the total income of the household. Similarly, the union takes the disutility of $\beta^{\tau-t}N_\tau^{\psi}$ of work effort as given. The first order condition w.r.t. $W_t^R$ is

$$\mathbb{E}_t \sum_{\tau=t}^{\infty} (\beta\theta^w)^{\tau-t} u'(C_\tau) w_\tau \left( w_{t,\tau}^R \right)^{-(\varepsilon-1)} L_\tau = \frac{\varepsilon}{\varepsilon-1} \mathbb{E}_t \sum_{\tau=t}^{\infty} (\beta\theta^w)^{\tau-t} N_\tau^{\psi} \left( w_{t,\tau}^R \right)^{-\varepsilon} L_\tau,$$

where $w_t \equiv W_t/P_t$ and $w_{t,\tau}^R \equiv W_t^R/W_\tau$. Log-linearizing around a zeros inflation steady state we obtain

$$\mathbb{E}_t \sum_{\tau=t}^{\infty} (\beta\theta^w)^{\tau-t} \hat{w}_{t,\tau}^R = \mathbb{E}_t \sum_{\tau=t}^{\infty} (\beta\theta^w)^{\tau-t} \left( \sigma\hat{C}_\tau + \psi\hat{N}_\tau - \hat{w}_\tau \right). \tag{16.A.27}$$

Notice this is analogous to (16.A.18) with the main difference being that log marginal cost, $\hat{w}_\tau - \hat{A}_\tau$, is replaced with the difference between the log marginal rate of substitution between consumption and leisure, $\sigma\hat{C}_\tau + \psi\hat{N}_\tau$ and the log real wage, $\hat{w}_\tau$. The rest of the derivation of the wage Phillips curve follows steps similar to those for the price Phillips curve so we will omit many of the details here. Defining $\hat{w}_t^R \equiv \hat{W}_t^R - \hat{W}_t$ we have

$$\hat{w}_{t,\tau}^R = \hat{W}_t^R - \hat{W}_t - (\hat{W}_\tau - \hat{W}_t) = \hat{w}_t^R - (\hat{W}_\tau - \hat{W}_t)$$

and using this, (16.A.27) becomes

$$w_t^R = (1-\beta\theta^w) \left( \sigma\hat{C}_t + \psi\hat{N}_t - \hat{w}_t \right) + \beta\theta^w \mathbb{E}_t \left( \hat{w}_{t+1}^R + \pi_{t+1}^w \right).$$

As in the sticky-price model we have

$$\pi_t^w \equiv \frac{W_t}{W_{t-1}} - 1 \approx \frac{1-\theta^w}{\theta^w}\hat{w}_t^R$$

so we have

$$\pi_t^w = \xi^w \left(\sigma \hat{C}_t + \psi \hat{N}_t - \hat{w}_t\right) + \beta \mathbb{E}_t \left[\pi_{t+1}^w\right],$$

where

$$\xi^w \equiv \frac{(1-\beta\theta^w)(1-\theta^w)}{\theta^w}.$$

The log-linearized aggregate production function is now $\hat{Y}_t = \hat{A}_t + \hat{L}_t = \hat{A}_t + \hat{N}_t$ and the log-linearized aggregate resource constraint remains $\hat{Y}_t = \hat{C}_t$. These lead to

$$\sigma \hat{C}_t + \psi \hat{N}_t = (\sigma + \psi)\hat{Y}_t - \psi \hat{A}_t.$$

As the flexible-price economy is unaffected relative to the sticky-price model, we still have $\hat{Y}_t^n = (1+\psi)/(\sigma+\psi)\hat{A}_t$. The wage Phillips curve then becomes

$$\pi_t^w = \kappa^w x_t + \xi^w \hat{A}_t - \xi^w \hat{w}_t + \beta \mathbb{E}_t \left[\pi_{t+1}^w\right],$$

where $\kappa^w \equiv \xi^w(\sigma+\psi)$.

Turning to the price Phillips curve, the derivation is the same as in Appendix 16.A.1 up to eq. (16.A.21). We therefore have

$$\pi_t = \xi^p \hat{w}_t - \xi^p \hat{A}_t + \beta \mathbb{E}_t \left[\pi_{t+1}\right],$$

where $\xi^p \equiv (1-\beta\theta)(1-\theta)/\theta$.

# 18.A Appendix to Chapter 18

## 18.A.1 Detailed derivation of the Generalized Nash Bargaining solution

Consider the problem

$$\max_w (\tilde{W}(w,z) - U(z))^\gamma (\tilde{J}(w,z) - V(z))^{1-\gamma},$$

where

$$\tilde{W}(w,z) = w + \beta \mathbb{E}[(1-\sigma)W(z') + \sigma U(z')],$$
$$\tilde{J}(w,z) = z - w + \beta \mathbb{E}[(1-\sigma)J(z') + \sigma V(z')],$$

and $U(z)$ and $V(z)$ are defined in (18.12) and (18.8). Note that these definitions imply $\partial \tilde{W}(w,z)/\partial w = 1$ and $\partial \tilde{J}(w,z)/\partial w = -1$.

The first-order condition for the maximization problem above is

$$\frac{\partial \tilde{W}(w,z)}{\partial w} \gamma (\tilde{W}(w,z) - U(z))^{\gamma-1} (\tilde{J}(w,z) - V(z))^{1-\gamma} = \frac{\partial \tilde{J}(w,z)}{\partial w} (1-\gamma)(\tilde{W}(w,z) - U(z))^{\gamma} (\tilde{J}(w,z) - V(z))^{-\gamma}.$$

Using $\partial \tilde{W}(w,z)/\partial w = 1$ and $\partial \tilde{J}(w,z)/\partial w = -1$ and reorganizing, we obtain (18.13).

## 18.A.2 Analysis of wages in the basic DMP model in Section 18.4

From the six equations (18.7), (18.8), (18.9), (18.11), (18.12), and (18.13), we can derive the equilibrium wages. In the DMP model, the worker's marginal product $z$ and the opportunity cost of working $b$ are different, and the surplus $z - b$ is divided between firms and workers. A simplistic wage rule could then be giving the workers $b + \gamma(z-b)$, where $\gamma \in (0,1)$ is the parameter that governs the workers' bargaining power. The actual solution in the basic DMP model is more complex because the Nash bargaining solution is over the present value of surpluses. Even though the ultimate source of the surplus is $z - b$, how it is divided between $w - b$ (workers) and $z - w$ (firms) is affected by the future possibilities.

Using (18.9) on (18.7) and multiplying $\gamma$,

$$\gamma J(z) = \gamma(z - w) + \beta \mathbb{E}[\gamma(1-\sigma)J(z')]$$

holds, and subtracting (18.12) from (18.11) and multiplying $(1-\gamma)$,

$$(1-\gamma)(W(z) - U(z)) = (1-\gamma)(w-b) + \beta \mathbb{E}[(1-\gamma)(1-\sigma-\lambda_w(\theta))(W(z') - U(z'))]$$

holds. Combining these two and using (18.13),

$$w = b + \gamma(z-b) + \beta \mathbb{E}[(1-\gamma)\lambda_w(\theta)(W(z') - U(z'))].$$

Thus, the wage is equal to the static share of the surplus $b + \gamma(z-b)$ plus the term that involves $\lambda_w(\theta)$. The latter term arises because the worker loses the opportunity of searching for a new job by being matched, and the worker is compensated for this loss. Using (18.13) and (18.10), the expression for the wage can alternatively be rewritten as

$$w = b + \gamma(z-b) + \gamma\theta\kappa.$$

## 18.A.3 Method of log-linearization

One method of analyzing macroeconomic models (especially business cycle models) is to first approximate the model by a system of linear equations. One popular method is to log-linearize the equilibrium condition. The advantage of the log-linearization is that the outcome is in terms of percentage deviations and therefore easy to interpret.

There are many methods of log-linearization. One method of carrying out the log-linearization of equations is following the next steps:

**Step 1.** Rewrite variable $X_t$ as $X_t = \bar{X}e^{x_t}$, where $x_t \equiv \log(X_t/\bar{X})$ and $\bar{X}$ is the steady-state value of $X$. $x_t$ is the percent deviation of $X_t$ from its steady-state level.

**Step 2.** After simplifying, use the approximation $\bar{X}e^{x_t} \approx \bar{X}(1+x_t)$ to (log-)linearize the equation.

**Step 3.** Further simplify the equation using the steady-state relationship. Note that expected value for the future can easily be dealt with because the expectation operator is linear.

As an example, consider the evolution of capital stock in the neoclassical growth model: $K_{t+1} = K_t^{\alpha}N_t^{1-\alpha} + (1-\delta)K_t - C_t$. The steady-state relationship is $\bar{K} = \bar{K}^{\alpha}\bar{N}^{1-\alpha} + (1-\delta)\bar{K} - \bar{C}$. The first step is to rewrite the equation as $\bar{K}e^{k_{t+1}} = (\bar{K}e^{k_t})^{\alpha}(\bar{N}e^{n_t})^{1-\alpha} + (1-\delta)\bar{K}e^{k_t} - \bar{C}e^{c_t}$. Simplifying, $\bar{K}e^{k_{t+1}} = \bar{K}^{\alpha}\bar{N}^{1-\alpha}e^{\alpha k_t + (1-\alpha)n_t} + (1-\delta)\bar{K}e^{k_t} - \bar{C}e^{c_t}$. Following the step 2 yields, $\bar{K}(1+k_{t+1}) = \bar{K}^{\alpha}\bar{N}^{1-\alpha}(1+\alpha k_t + (1-\alpha)n_t) + (1-\delta)\bar{K}(1+k_t) - \bar{C}(1+c_t)$. In step 3, we obtain $\bar{K}k_{t+1} = \bar{K}^{\alpha}\bar{N}^{1-\alpha}(\alpha k_t + (1-\alpha)n_t) + (1-\delta)\bar{K}k_t - \bar{C}c_t$.

As we saw above, if the equilibrium conditions are power functions, log-linearization is fairly straightforward. When they involve more complex functions, it may be useful to first Taylor approximate the original function $f(X_t)$ by

$$f(X_t) \approx f(\bar{X}) + f'(\bar{X})(X_t - \bar{X})$$

and then apply the above method. Because the approximated function is linear in $X_t$, it is straightforward to apply the above method.

## 18.A.4 Log-linarization of Section 18.7.2

This Appendix derives the log-linearized system in Section 18.7.2. With the Cobb-Douglas matching function, we can rewrite (18.31) as

$$\frac{\kappa}{(1-\gamma)\beta\chi}\theta_t^{\eta} = \beta\mathbb{E}\left[z_{t+1} - c(z_{t+1}) - b + \frac{\kappa\theta_{t+1}^{\eta}}{(1-\gamma)\chi} - \frac{\kappa\sigma(z_{t+1})\theta_{t+1}^{\eta}}{(1-\gamma)\chi} - \frac{\gamma\kappa\theta_{t+1}}{1-\gamma}\right].$$

In (18.33) and (18.34), we saw that $\hat{\sigma}(z_t)$ and $\hat{c}(z_t)$ can be approximated as

$$\hat{\sigma}(z_t) = \mathcal{E}\hat{z}_t$$

and

$$\hat{c}(z_t) = \mathcal{F}\hat{z}_t.$$

Then, the above equation can further be rewritten as

$$\frac{\kappa}{(1-\gamma)\beta\chi}\bar{\theta}^{\eta}e^{\eta\hat{\theta}_t} = \beta\mathbb{E}\left[\bar{z}e^{\hat{z}_{t+1}} - \bar{c}e^{\mathcal{F}\hat{z}_{t+1}} - b + \frac{\kappa\bar{\theta}^{\eta}e^{\eta\hat{\theta}_{t+1}}}{(1-\gamma)\chi} - \frac{\kappa\bar{\sigma}e^{\mathcal{E}\hat{z}_{t+1}}\bar{\theta}^{\eta}e^{\eta\hat{\theta}_{t+1}}}{(1-\gamma)\chi} - \frac{\gamma\kappa\bar{\theta}e^{\hat{\theta}_{t+1}}}{1-\gamma}\right].$$

Using the approximation $1 + x \approx e^x$ and the steady-state relationship,

$$\frac{\kappa}{(1-\gamma)\beta\chi}\bar{\theta}^{\eta}\eta\hat{\theta}_t = \beta\mathbb{E}\left[\bar{z}\hat{z}_{t+1} - \bar{c}\mathcal{F}\hat{z}_{t+1} + \frac{\kappa\bar{\theta}^{\eta}\eta\hat{\theta}_{t+1}}{(1-\gamma)\chi} - \frac{\kappa\bar{\sigma}\bar{\theta}^{\eta}(\mathcal{E}\hat{z}_{t+1} + \eta\hat{\theta}_{t+1})}{(1-\gamma)\chi} - \frac{\gamma\kappa\bar{\theta}\hat{\theta}_{t+1}}{1-\gamma}\right].$$

With the same procedure as in the exogenous separation case, we obtain

$$\hat{\theta}_t = \mathcal{G}\hat{z}_t.$$

where

$$\mathcal{G} = (1-\gamma)\left[\frac{\kappa\bar{\theta}^{\eta}\eta}{\chi}\left(\frac{1}{\rho\beta} - (1-\bar{\sigma})\right) + \kappa\gamma\bar{\theta}\right]^{-1}\left(\bar{z} - \bar{c}\mathcal{F} - \frac{\kappa\bar{\sigma}\bar{\theta}^{\eta}\mathcal{E}}{(1-\gamma)\chi}\right). \tag{18.A.28}$$

We can solve for $\mathcal{G}$ from (18.A.28) after plugging in $\mathcal{E}$ and $\mathcal{F}$ from (18.33) and (18.34).

The solution is

$$\mathcal{G} = \frac{\Theta}{\Gamma},$$

where

$$\Theta \equiv (1-\gamma)\left[\frac{\kappa\bar{\theta}^{\eta}\eta}{\chi}\left(\frac{1}{\rho\beta} - (1-\bar{\sigma})\right) + \kappa\gamma\bar{\theta}\right]^{-1}\bar{z}$$

and

$$\Gamma \equiv 1 + (1-\gamma)\left[\frac{\kappa\bar{\theta}^{\eta}\eta}{\chi}\left(\frac{1}{\rho\beta} - (1-\bar{\sigma})\right) + \kappa\gamma\bar{\theta}\right]^{-1}\bar{c}\frac{\xi\eta}{\xi+1}\left(1 - \frac{1}{1-\gamma}\right).$$

## 18.A.5 Derivation of equation (18.38)

First, modify the model so that the families can issue and sell/buy Arrow securities (contingency claims) for the next period state. Because the Arrow securities' net supply is zero and the families are identical (similar to the Lucas "tree" model), the existence of the Arrow securities does not affect the equilibrium allocation. Consider the problem

$$\max_{\{c_t, k_{t+1}, a_{t+1}(z_{t+1})\}_{t=0}^{\infty}} \mathbb{E}_0\left[\sum_{t=0}^{\infty}\mathbf{U}(c_t)\right],$$

$$c_t + k_{t+1} + \int Q_t(z_{t+1})a_{t+1}(z_{t+1})dz_{t+1} = (1 + r_t - \delta)k_t + (1-u_t)w_t + u_t b + d_t + a_t(z_t),$$

where $Q_t(z_{t+1})$ is the price of an Arrow security that is issued (and traded) in period $t$ and pays one unit of consumption goods if the next period state turns out to be $z_{t+1}$. $a_{t+1}(z_{t+1})$ is the quantity the family purchases at period $t$. Note that because the sum of $a_t$ is zero and because the families are homogeneous, the equilibrium values of $a_t$ are going to be zero. Therefore, existence of the Arrow securities do not alter the equilibrium allocation of goods.

With the recursive formulation,

$$\mathbf{V}(k, a(z), X) = \max_{c,k',\{a'(z')\}} \mathbf{U}(c) + \beta \int \mathbf{V}(k', a(z'), X') f(z'|z) dz'$$

subject to

$$c + k' + \int Q(z', X) a'(z') dz' = (1 + r(X) - \delta)k + (1 - u)w(X) + ub + d(X) + a(z),$$

$$K' = \Omega(X),$$

and

$$u' = (1 - \lambda_w(\theta(X))) + \sigma(1 - u).$$

Letting the Lagrange multiplier on the budget constraint when the stochastic state is $z$ be $\mu(z)$, the first-order conditions for $c$, $k'$, and $a'(z')$ are

$$\mathbf{U}'(c) = \mu(z), \tag{18.A.29}$$

$$\beta \int \mathbf{V}_1(k', a(z'), X') f(z'|z) dz' = \mu(z), \tag{18.A.30}$$

and

$$\beta \mathbf{V}_2(k', a(z'), X') f(z'|z) = \mu(z) Q(z', X). \tag{18.A.31}$$

The envelope conditions are

$$\mathbf{V}_1(k, a(z), X) = \mu(z)(1 + r(X) - \delta) \tag{18.A.32}$$

and

$$\mathbf{V}_2(k, a(z), X) = \mu(z). \tag{18.A.33}$$

Combining (18.A.29), (18.A.31), and (18.A.33), we obtain

$$Q(z', X) = \beta f(z'|z) \frac{\mathbf{U}'(c')}{\mathbf{U}'(c)}, \tag{18.A.34}$$

which is (18.38).

### 18.A.6 Derivation of $J(X)$, $V(X)$, $W(X)$, and $U(X)$ equations in Section 18.8.1

To derive the values of jobs and workers, we explicitly allow the trade of the claims to profits and wages. Let the quantity of the claim for the profit at the beginning of period $t$ be $q_t^J$. In the beginning of the period $t$, the asset (claims) trading occurs with the price $J(X)$, and the new level of asset is $\hat{q}_t^J$. The profit is distributed proportional to $\hat{q}_t^J$. In the next period, some jobs stay matched and some jobs become vacant. Similar notations are used for $V(X)$, $W(X)$, and $U(X)$. Note that $q_t^J$, $\hat{q}_t^J$, $q_t^W$, and $\hat{q}_t^W$ each sum up to $(1-u_t)$, $q_t^U$ and $\hat{q}_t^U$ each sum up to $u_t$, and $q_t^U$ and $\hat{q}_t^U$ each sum up to $v_t = \theta_t u_t$. Because the consumers (families) are homogeneous and total number of families is one, the equilibrium values of asset holdings are equal to the corresponding sum.

The budget constraint now becomes (also incorporating the Arrow securities, as in Appendix 18.A.5)

$$\begin{aligned} c_t + k_{t+1} + \int Q_t(z_{t+1}) a_{t+1}(z_{t+1}) dz_{t+1} + (\hat{q}_t^J - q_t^J) J_t + (\hat{q}_t^V - q_t^V) V_t + (\hat{q}_t^W - q_t^W) W_t + (\hat{q}_t^U - q_t^U) U_t \\ = (1 + r_t - \delta) k_t + \hat{q}_t^W w_t + \hat{q}_t^U b + \hat{q}_t^J (y_t - w_t) - \hat{q}_t^V \kappa + a_t(z_t). \end{aligned}$$

Considering the definition of $d_t$ (see (18.47)), we can show that the budget constraint in equilibrium is identical to the baseline model and therefore the equilibrium allocation is not altered by the possibility of trading claims.

The transition equations for the asset holdings are

$$\begin{aligned} q_{t+1}^J &= \lambda_f(\theta_t) \hat{q}_t^V + (1-\sigma) \hat{q}_t^J, \\ q_{t+1}^V &= (1 - \lambda_f(\theta_t)) \hat{q}_t^V + \sigma \hat{q}_t^J, \\ q_{t+1}^W &= \lambda_w(\theta_t) \hat{q}_t^U + (1-\sigma) \hat{q}_t^W, \end{aligned}$$

and

$$q_{t+1}^U = (1 - \lambda_w(\theta_t)) \hat{q}_t^U + \sigma \hat{q}_t^W.$$

As in Appendix 18.A.5, we can write down the dynamic programming problem, now with new state variables (which includes the claim holdings) and new constraints above. The value function is now $\mathbf{V}(q^J, q^V, q^W, q^U, k, a(z), X)$. The problem is

$$\mathbf{V}(q^J, q^V, q^W, q^U, k, a(z), X) = \max_{c, k', \{a'(z')\}, \{q^{i\prime}, \hat{q}^i\}^{i=J,V,W,U}} \mathbf{U}(c) + \beta \int \mathbf{V}(q^{J\prime}, q^{V\prime}, q^{W\prime}, q^{U\prime}, k', a(z'), X') f(z'|z) dz'$$

subject to

$$\begin{aligned} c + k' + \int Q(z', X) a'(z') dz' + (\hat{q}^J - q^J) J(X) + (\hat{q}^V - q^V) V(X) + (\hat{q}^W - q^W) W(X) + (\hat{q}^U - q^U) U(X) \\ = (1 + r(X) - \delta) k + \hat{q}^W w(X) + \hat{q}^U b + \hat{q}^J (y(X) - w(X)) - \hat{q}^V \kappa + a(z), \end{aligned}$$

$$q^{J\prime} = \lambda_f(\theta(X))\hat{q}^V + (1-\sigma)\hat{q}^J,$$
$$q^{V\prime} = (1-\lambda_f(\theta(X)))\hat{q}^V + \sigma\hat{q}^J,$$
$$q^{W\prime} = \lambda_w(\theta(X)))\hat{q}^U + (1-\sigma)\hat{q}^W,$$
$$q^{U\prime} = (1-\lambda_w(\theta(X)))\hat{q}^U + \sigma\hat{q}^W.$$
$$K' = \Omega(X),$$

and

$$u' = (1-\lambda_w(\theta(X)) + \sigma(1-u).$$

Let the Lagrange multiplier of the budget constraint be $\mu$ and the transition equations be $\nu^J$, $\nu^V$, $\nu^W$, and $\nu^U$. The first-order conditions on $c$, $k'$, $a'(z')$, and the envelope conditions on $k$ and $a(z)$ are the same as in Appendix 18.A.5 (i.e., (18.A.29) to (18.A.33)). As a result, we obtain (18.A.34).

The first-order conditions for $\hat{q}^J$, $\hat{q}^V$, $\hat{q}^W$, and $\hat{q}^U$ are:

$$\mu J(X) = \mu(y(X) - w(X)) + \nu^J(1-\sigma) + \nu^V\sigma, \tag{18.A.35}$$

$$\mu V(X) = -\mu\kappa + \nu^J\lambda_f(\theta(X)) + \nu^V(1-\lambda_f(\theta(X))), \tag{18.A.36}$$

$$\mu W(X) = \mu w(X) + \nu^W(1-\sigma) + \nu^U\sigma, \tag{18.A.37}$$

and

$$\mu U(X) = \mu b + \nu^W\lambda_w(\theta(X)) + \nu^U(1-\lambda_w(\theta(X))). \tag{18.A.38}$$

The first-order conditions for $q^{J\prime}$, $q^{V\prime}$, $q^{W\prime}$, and $q^{U\prime}$ are:

$$\nu^J = \beta\int \mathbf{V}_1(q^{J\prime}, q^{V\prime}, q^{W\prime}, q^{U\prime}, k', a(z'), X')f(z'|z)dz'. \tag{18.A.39}$$

$$\nu^V = \beta\int \mathbf{V}_2(q^{J\prime}, q^{V\prime}, q^{W\prime}, q^{U\prime}, k', a(z'), X')f(z'|z)dz'. \tag{18.A.40}$$

$$\nu^W = \beta\int \mathbf{V}_3(q^{J\prime}, q^{V\prime}, q^{W\prime}, q^{U\prime}, k', a(z'), X')f(z'|z)dz'. \tag{18.A.41}$$

$$\nu^U = \beta\int \mathbf{V}_4(q^{J\prime}, q^{V\prime}, q^{W\prime}, q^{U\prime}, k', a(z'), X')f(z'|z)dz'. \tag{18.A.42}$$

Envelope conditions for $q^J$, $q^V$, $q^W$, and $q^U$ are:

$$\mathbf{V}_1(q^J, q^V, q^W, q^U, k, a(z), X) = \mu J(X) \tag{18.A.43}$$

$$\mathbf{V}_2(q^J, q^V, q^W, q^U, k, a(z), X) = \mu V(X) \tag{18.A.44}$$

$$\mathbf{V}_3(q^J, q^V, q^W, q^U, k, a(z), X) = \mu W(X) \tag{18.A.45}$$

$$\mathbf{V}_4(q^J, q^V, q^W, q^U, k, a(z), X) = \mu U(X) \tag{18.A.46}$$

Combining (18.A.35), (18.A.39), (18.A.40), (18.A.43), and (18.A.44) and utilizing (18.A.29) and (18.A.34), we can obtain

$$J(X) = y(X) - w(X) + \int Q(z', X)[(1-\sigma)J(X') + \sigma V(X')]dz',$$

which is (18.39) in the main text. We can similarly obtain (using equations (18.A.35) to (18.A.46))

$$V(X) = -\kappa + \int Q(z', X)[\lambda_f(\theta(X))J(X') + (1-\lambda_f(\theta(X))V(X')]dz',$$

$$W(X) = w(X) + \int Q(z', X)[(1-\sigma)W(X') + \sigma U(X')]dz',$$

and

$$U(X) = b + \int Q(z', X)[\lambda_w(\theta(X))W(X') + (1-\lambda_w(\theta(X)))U(X')]dz'.$$

## 18.A.7 Calibration and computation of Section 18.8

The parameter values in Table 18.1 apply in this model, except for $b$. Below, parameters $b$ and $\kappa$ are endogenously calibrated. First, with $\bar{\theta} = 1$, the steady-state unemployment rate

$$\bar{u} = \frac{\sigma}{\chi + \sigma}.$$

From the Euler equation of the family's consumption-saving problem,

$$\bar{r} = \frac{1}{\beta} - 1 + \delta.$$

Because

$$\bar{r} = \alpha \left(\frac{\bar{K}}{1-\bar{u}}\right)^{\alpha-1}$$

can be solved for $\bar{K}$:

$$\bar{K} = \left(\frac{r}{\alpha}\right)^{\frac{1}{1-\alpha}} (1-\bar{u}).$$

Once we know $\bar{K}$, we can compute

$$\bar{y} = (1-\alpha)\left(\frac{K}{1-u}\right)^{\alpha}.$$

The parameter $b$ is set by

$$b = 0.4\bar{y}.$$

From the job creation condition, $\kappa$ is calibrated as

$$\kappa = \beta\chi(\bar{y} - b)\left(1 - \beta\frac{1-\sigma-\gamma\chi}{1-\gamma}\right)^{-1}.$$

Now we can compute the steady-state values of wage and dividend:

$$\bar{w} = \gamma(\bar{y} - b) + b + \gamma\kappa$$

$$\bar{d} = (1 - \bar{u})(y - \bar{w}) - \kappa\bar{u}.$$

# 20.A Appendix to Chapter 20

## 20.A.1 Derivation of Equation (20.8)

As a preparation, note two facts about the normal distribution and the lognormal distribution. First, when $X$ is normally distributed $X \sim N(\mu, \sigma^2)$, The random variable $\alpha X$ also follows a normal distribution $N(\alpha\mu, \alpha^2\sigma^2)$. Second, when $X$ is normally distributed $X \sim N(\mu, \sigma^2)$, $\exp(X)$ is lognormally distributed, and $\mathbb{E}[\exp(X)] = \exp(\mu + \sigma^2/2)$.

From (20.6),

$$\ln(A) = (1-\gamma)\ln\left(\int a_i^{\frac{1}{1-\gamma}} di\right).$$

Because $\ln(a_i)$ is normally distributed with $N(\nu - \sigma^2/2, \sigma^2)$, $\ln(a_i)/(1-\gamma)$ is also normally distributed (with mean $(\nu - \sigma^2/2)/(1-\gamma)$ and variance $\sigma^2/(1-\gamma)^2$) and

$$a_i^{\frac{1}{1-\gamma}} = \exp\left(\frac{1}{1-\gamma}\ln(a_i)\right)$$

is lognormally distributed, with mean

$$\exp\left(\frac{1}{1-\gamma}\left(\nu - \frac{\sigma^2}{2}\right) + \frac{1}{2}\frac{\sigma^2}{(1-\gamma)^2}\right) = \exp\left(\frac{1}{1-\gamma}\left(\nu + \frac{\gamma}{1-\gamma}\frac{1}{2}\sigma^2\right)\right).$$

Because $a_i$ is i.i.d., from the law of large numbers,

$$\int a_i^{\frac{1}{1-\gamma}} di = \mathbb{E}\left[a_i^{\frac{1}{1-\gamma}}\right] = \exp\left(\frac{1}{1-\gamma}\left(\nu + \frac{\gamma}{1-\gamma}\frac{1}{2}\sigma^2\right)\right),$$

where $\mathbb{E}[\cdot]$ represents expected value. Going back to the first equation, this outcome implies

$$\ln(A) = \nu + \frac{\gamma}{1-\gamma}\frac{1}{2}\sigma^2,$$

and thus (20.8) follows.

## 20.A.2 Firms versus establishments in the size statistics

Figure 20.7 shows that the fraction of workers working in large firms has increased since the 1990s. The same pattern does not hold for large establishments. Therefore, making a distinction between a firm and an establishment is important in this context.

The BDS dataset does not provide the 10,000+ category for establishments. First, we confirm that the same pattern as Figure 20.7 holds when the threshold moves to 1,000. Figure 6 plots the fraction of employees working at 1,000+ employee firms. One can see that the graph is (aside from the shift in level) almost identical to Figure 20.7.

Figure 7 computes the same statistics for 1,000+ employee establishments. Although we see some increase after the mid-2000s, overall the profile has been relatively flat. Thus,

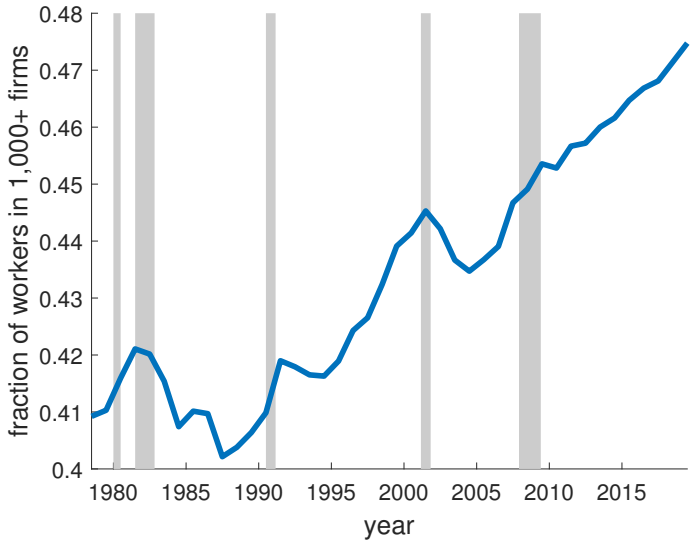



Figure 6: Fraction of employees working at 1,000+ employee firms. Source: Business Dynamics Statistics

the concentration of employment at the top is a firm phenomenon and not an establishment phenomenon.

The contrast between Figures 6 and 7 leads us to suspect that it is the number of establishments that is contributing the concentration of employees at very large firms. Figure 8 confirms this to be the case. It shows that the fraction of establishments that belong to very big firms (10,000+ employees) has steadily increased.

### 20.A.3 Derivation of Equation (20.11)

As in Appendix 20.A.1, let us first prepare with a basic property of the normal distribution. When $X$ and $Y$ are jointly normally distributed with $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = (\mu_x, \mu_y)$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}.$$

Then $\alpha X + \beta Y$ also follows a normal distribution with mean $\alpha\mu_x + \beta\mu_y$ and variance $\alpha^2\sigma_x^2 + \beta^2\sigma_y^2 + 2\alpha\beta\rho\sigma_x\sigma_y$.

From (20.10),

$$\ln(A) = \ln\left(\int a_i^{\frac{1}{1-\gamma}}(1-\tau_i)^{\frac{\gamma}{1-\gamma}}di\right) - \gamma\ln\left(\int a_i^{\frac{1}{1-\gamma}}(1-\tau_i)^{\frac{1}{1-\gamma}}di\right).$$

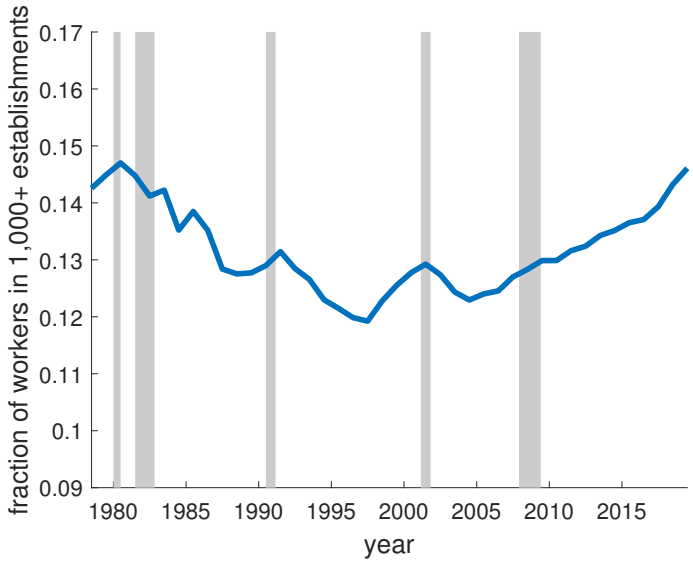



Figure 7: Fraction of employees working at 1,000+ employee establishments. Source: Business Dynamics Statistics

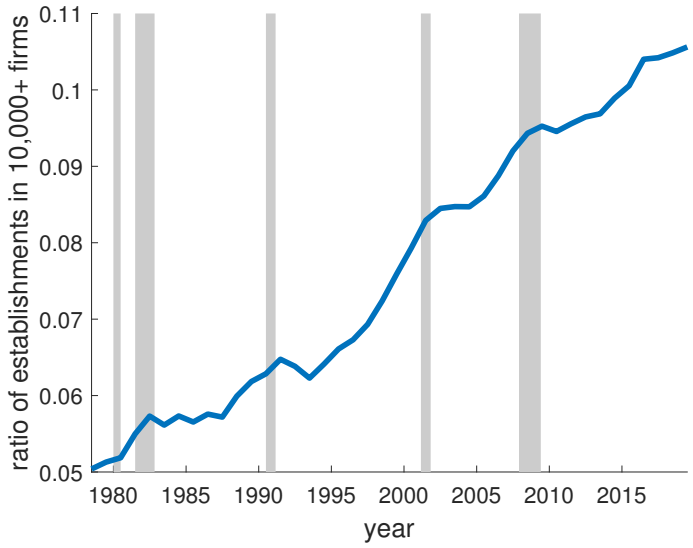



Figure 8: Fraction of establishments that belong to 10,000+ employee firms. Source: Business Dynamics Statistics

Because $\ln(a_i)$ and $\ln(1-\tau_i)$ follow a bivariate normal distribution,

$$\frac{1}{1-\gamma}\ln(a_i)+\frac{\gamma}{1-\gamma}\ln(1-\tau_i)$$

follow a normal distribution with mean

$$\frac{1}{1-\gamma}\left(\nu_a-\frac{\sigma_a^2}{2}\right)+\frac{\gamma}{1-\gamma}\left(\nu_\tau-\frac{\sigma_\tau^2}{2}\right)$$

and variance

$$\frac{1}{(1-\gamma)^2}\sigma_a^2+\frac{\gamma^2}{(1-\gamma)^2}\sigma_\tau^2+2\rho\frac{\gamma}{(1-\gamma)^2}\sigma_a\sigma_\tau.$$

From the law of large numbers,

$$\int a_i^{\frac{1}{1-\gamma}}(1-\tau_i)^{\frac{\gamma}{1-\gamma}}di=\mathbb{E}\left[a_i^{\frac{1}{1-\gamma}}(1-\tau_i)^{\frac{\gamma}{1-\gamma}}\right]=\mathbb{E}\left[\exp\left(\frac{1}{1-\gamma}\ln(a_i)+\frac{\gamma}{1-\gamma}\ln(1-\tau_i)\right)\right]$$

holds, and it can be computed as

$$\exp\left(\frac{1}{1-\gamma}\left(\nu_a-\frac{\sigma_a^2}{2}\right)+\frac{\gamma}{1-\gamma}\left(\nu_\tau-\frac{\sigma_\tau^2}{2}\right)+\frac{1}{2}\left(\frac{1}{(1-\gamma)^2}\sigma_a^2+\frac{\gamma^2}{(1-\gamma)^2}\sigma_\tau^2+2\rho\frac{\gamma}{(1-\gamma)^2}\sigma_a\sigma_\tau\right)\right).$$

We can similarly compute

$$\int a_i^{\frac{1}{1-\gamma}}(1-\tau_i)^{\frac{1}{1-\gamma}}di$$

and after some algebra, we obtain

$$\ln(A)=\nu_a+\frac{\gamma}{1-\gamma}\frac{1}{2}(\sigma_a^2-\sigma_\tau^2),$$

implying (20.11).

### 20.A.4 Derivation of the Bertrand competition result in Section 20.6.2

First, as in the monopolistic competition case, consider the cost-minimization problem for the final-good producer. Within a sector, it has to solve

$$\min_{\{q_{ij}\}}\sum_{j=1}^{J}\hat{p}_{ij}q_{ij}$$

subject to

$$y_i=\left[\sum_{j=1}^{J}q_{ij}^{\frac{\eta-1}{\eta}}\right]^{\frac{\eta}{\eta-1}}$$

for given $y_i$. The first-order condition is

$$\hat{p}_{ij} = \lambda_i q_{ij}^{-\frac{1}{\eta}} y_{ij}^{\frac{1}{\eta}}.$$

Similarly to the monopolistic competition case, we can think of $\lambda_i$ as the price of the combined good for sector $i$ (call it $p_i$) and

$$p_i = \left[ \sum_{j=1}^{J} \hat{p}_{ij}^{1-\eta} \right]^{\frac{1}{1-\eta}}.$$

An intermediate-good producer with Bertrand competition therefore solves

$$\max_{\hat{p}_{ij}} \ \hat{p}_{ij} q_{ij} - cm_{ij},$$

given

$$q_{ij} = a_{ij} m_{ij}^{\gamma} \tag{20.A.47}$$

and

$$\hat{p}_{ij} = p_i q_{ij}^{-\frac{1}{\eta}} y_{ij}^{\frac{1}{\eta}}. \tag{20.A.48}$$

Now, in addition to these, the producer is aware that its price $\hat{p}_{ij}$ affects the sectoral price $p_i$ and thus the sectoral demand. Thus

$$p_i = y_i^{-\frac{1}{\sigma}} Y^{\frac{1}{\sigma}} \tag{20.A.49}$$

(which is the solution of the cost-minimization problem across sectors, with the price index normalized to one) and

$$p_i = \left[ \sum_{j=1}^{J} \hat{p}_{ij}^{1-\eta} \right]^{\frac{1}{1-\eta}} \tag{20.A.50}$$

are also given to the producer (in addition to $Y$).

The first-order condition is

$$\hat{p}_{ij} \frac{\partial q_{ij}}{\partial \hat{p}_{ij}} + q_{ij} = \frac{\partial (cm_{ij})}{\partial q_{ij}} \frac{\partial q_{ij}}{\partial \hat{p}_{ij}}$$

and thus

$$\left( 1 - \frac{q_{ij}/\hat{p}_{ij}}{\partial q_{ij}/\partial \hat{p}_{ij}} \right) \hat{p}_{ij} = \mathcal{M}.$$

Here, the marginal cost is

$$\mathcal{M} = \frac{\partial (cm_{ij})}{\partial q_{ij}} = \frac{cm_{ij}^{1-\gamma}}{\gamma a_{ij}},$$

as in the case with the Cournot competition, computed using (20.A.47). Using (20.A.48), (20.A.49), and (20.A.50),

$$\frac{q_{ij}/\hat{p}_{ij}}{\partial q_{ij}/\partial \hat{p}_{ij}} = \frac{1}{\varepsilon(s_{ij})},$$

where

$$\varepsilon(s_{ij}) = \eta(1 - s_{ij}) + \sigma s_{ij}.$$

Thus

$$\hat{p}_{ij} = \frac{\varepsilon(s_{ij})}{\varepsilon(s_{ij}) - 1}\mathcal{M}.$$

### 20.A.5 Proof of Hulten's theorem

This Appendix outlines the proof of Hulten's (1978) theorem, partly following Baqaee and Farhi (2019). The theorem to prove is

$$\frac{dY}{Y} = \sum_i D_i \frac{da_i}{a_i},$$

where $D_i$ is the Domar weight:

$$D_i = \frac{p_i y_i}{\sum_i p_i c_i}.$$

Below we show the theorem in a relatively simple static economy. The only input is labor, and labor is inelastically supplied. Suppose that there are $N$ goods. Assume that all markets are competitive. The representative consumer's utility takes the form

$$U(c_1, ..., c_N)$$

and the budget constraint is

$$\sum_{i=1}^{N} p_i c_i = w\bar{\ell} + \sum_{i=1}^{N} \pi_i,$$

where $w$ is the wage rate and $\bar{\ell}$ is the fixed amount of labor supply, which is the only input for production. $c_i$ is the consumption of good $i$, and $\pi_i$ is profit from sector $i$. Assume that the preferences are homothetic so that $U(c_1, ..., c_N)$ is linearly homogeneous.

The production function for sector $i$ is

$$y_i = a_i F_i(\ell_i, x_{i1}, x_{i2}, ..., x_{iN}),$$

where $\ell_i$ is labor input at sector $i$ and $x_{ij}$ is the quantity of product $j$ used in sector $i$. The profit is

$$\pi_i = p_i y_i - w\ell_i - \sum_{j=1}^{N} p_j x_{ij}.$$

The market-clearing conditions are

$$y_i = \sum_{j=1}^{N} x_{ji} + c_i$$

for all $i$ and

$$\bar{\ell} = \sum_{i=1}^{N} \ell_i.$$

First, consider the consumer's expenditure minimization problem for a given utility level

$$\min_{c_1,...c_N} \sum_{i=1}^{N} p_i c_i$$

subject to

$$U(c_1, ..., c_N) = u.$$

The Lagrangian is

$$L = \sum_{i=1}^{N} p_i c_i - \lambda(U(c_1, ..., c_N) - u).$$

The first-order condition for this problem is

$$p_i = \lambda \frac{\partial U(c_1, ..., c_N)}{\partial c_i}. \tag{20.A.51}$$

Let us normalize the prices (i.e., choose the numeraire) $p_i$ so that $\lambda = 1$ in equilibrium. Let

$$Y \equiv \sum_{i=1}^{N} p_i c_i$$

be the GDP (and TFP) of this economy. From (20.A.51) and linear homogeneity, it can be rewritten as

$$Y = \sum_{i=1}^{N} \frac{\partial U(c_1, ..., c_N)}{\partial c_i} c_i = U(c_1, ..., c_N). \tag{20.A.52}$$

Therefore, in equilibrium, the level of utility also represents GDP.

Because the first welfare theorem holds, the competitive equilibrium is Pareto optimal, and solves the social planner's problem

$$\max_{c_i, x_{ij}, \ell_i} U(c_1, ..., c_N)$$

subject to

$$c_i + \sum_{j=1}^{N} x_{ji} = a_i F_i(\ell_i, x_{i1}, x_{i2}, ..., x_{iN})$$

and

$$\sum_{i=1}^{N} \ell_i = \bar{\ell}.$$

The Lagrangian for the social planner is

$$L = U(c_1, ..., c_N) + \sum_{i=1}^{N} \mu_i \left( a_i F_i(\ell_i, x_{i1}, x_{i2}, ..., x_{iN}) - c_i - \sum_{j=1}^{N} x_{ji} \right) + \nu \left( \bar{\ell} - \sum_{i=1}^{N} \ell_i \right).$$

From the first-order condition,

$$\frac{\partial U(c_1, ..., c_N)}{\partial c_i} = \mu_i. \tag{20.A.53}$$

The envelope theorem implies

$$\frac{dU}{da_i} = \mu_i F_i(\ell_i, x_{i1}, x_{i2}, ..., x_{iN}) = \mu_i y_i \frac{1}{a_i}.$$

From (20.A.52), this equation implies

$$\frac{dY}{da_i} = \mu_i y_i \frac{1}{a_i}. \tag{20.A.54}$$

From (20.A.51) and (20.A.53), together with our normalization of $\lambda = 1$,

$$\mu_i = p_i$$

holds. Using this relationship and dividing both sides of (20.A.54) by $Y = \sum_{i=1}^{N} p_i c_i$ yields

$$\frac{dY}{Y} = \frac{p_i y_i}{\sum_{i=1}^{N} p_i c_i} \frac{da_i}{a_i}.$$

Repeating the same procedure for all $a_i$, we obtain the theorem.

## 20.A.6 Firm size distribution in Section 20.8

Let the mass of firms with $k$ product lines at time $t$ be $M_{kt}$. The transition equations for $M_{kt}$ are as follows.

First, for $M_{1t}$, there are three kinds of firms in $M_{1,t+1}$ next period. First is the entrants, second is the one-product firms that remain with one product, and third is the multi-product firms who had more than one product but lost some product lines and became one-product firms.

$$M_{1,t+1} = \nu_t + M_{1,t} \sum_{i=0}^{1} \mathbf{P}_\mu(1, i) \mathbf{P}_\eta(1, i) + \sum_{h=1}^{\infty} \left( M_{1+h,t} \sum_{i=h}^{1+h} \mathbf{P}_\mu(1 + h, i) \mathbf{P}_\eta(1 + h, i - h) \right),$$

where

$$\mathbf{P}_{\mu}(a,b) = \left( \begin{array}{c} a \\ b \end{array} \right) \mu^{b}(1-\mu)^{a-b}$$

is the probability of losing $i$ product lines when starting from $k$ lines, and

$$\mathbf{P}_{\eta}(a,b) = \left( \begin{array}{c} a \\ b \end{array} \right) \eta^{b}(1-\eta)^{a-b}$$

is the probability of gaining $i$ product lines when starting from $k$ lines.

Second, for $M_{kt}$ for $k > 1$, there are three different types in $M_{k,t+1}$. First is the firms starting from $k-h$ products, with a net gain of $h$ products and getting to $k$. The second is the $k$-product firms that remain with $k$ products. The third is firms with $k+h$ products, where $h > 0$ and that have a net loss of $h$ products and get to $k$ products.

$$\begin{aligned} M_{k,t+1} = & \sum_{h=1}^{\lfloor k/2 \rfloor} \left( M_{k-h,t} \sum_{i=0}^{\lfloor k/2-h \rfloor} \mathbf{P}_{\mu}(k-h,i)\mathbf{P}_{\eta}(k-h,i+h) \right) \\ & M_{k,t} \sum_{i=0}^{k} \mathbf{P}_{\mu}(k,i)\mathbf{P}_{\eta}(k,i) + \sum_{h=1}^{\infty} \left( M_{k+h,t} \sum_{i=h}^{k+h} \mathbf{P}_{\mu}(k+h,i)\mathbf{P}_{\eta}(k+h,i-h) \right). \end{aligned}$$

Here, $\lfloor x \rfloor$ represents the integer not exceeding $x$. To find the stationary distribution of firm size, we can look for $\{M_1, M_2, ...\}$ such that $M_{1,t+1} = M_{1,t}$, $M_{2,t+1} = M_{2,t}$, ... hold.