中级宏观经济学

一份自足的课程讲义

颜色教授主讲
北京大学光华管理学院　应用经济学系
周睿整理、翻译并用 LaTeX 排版

2022 年秋
（2026 年重排、统一并译为中文）

如何阅读本书

这是一门一学期**中级宏观经济学**课程的讲义，以新古典传统讲授。它写来是供从头读到尾的，像一本简短的教科书，而不是供临时翻查的公式表。

宏观经济学问两个问题。第一个关乎*长期*：为什么有的国家富、有的国家穷？是什么让一个经济体在几十年里持续增长？第二个关乎*短期*：产出为什么会围绕其趋势波动？政策对此能有所作为吗？本书循此分野展开。我们先学习如何*度量*总量经济——产出、价格和劳动力市场——因为后续每一个论证归根结底都是关于这些数字的命题。随后我们从索洛模型出发，一路构建长期增长理论，直至其有微观基础的后继模型，以及现代增长之前的马尔萨斯世界。接着我们研究货币与银行，以及把货币与价格联系起来的货币数量论。最后才转向短期：凯恩斯交叉、IS–LM 模型、总需求与总供给、财政政策及李嘉图等价所设的界限，以及菲利普斯曲线、预期，和“规则对相机抉择”之争。

关于这门学科的一句提醒：宏观经济学不像微观经济学那样整洁。同一个事实可以用不止一个模型来解读，数据很少能干净利落地了结一场争论，而一位好的讲授者会旁逸斜出，谈到一些制度性的细节——中国的中央银行实际如何运作、某项政策为何在某一年失灵——这些并不能整齐地纳入任何单一的定理。凡能照亮理论之处，我们都把这些旁枝保留了下来，置于**注**框之中，而非把每一处观察都硬塞进一个并不真正契合的模型。请把这些框读作一门鲜活学科的肌理，而非未尽之处。

各章的组织方式

每一章都以一段平实的中文*引子*开篇，说明我们要回答什么问题、为何即将引入某个特定的工具。随后我们将这一想法形式化，推演其代数，并从图中读出经济学含义。定义和关键结论都用彩框标出，便于复习时迅速查找。本书每一幅图都是重新绘制的，没有任何扫描件。

彩框颜色约定

- **绿色：定义**

  ——我们赖以构建的对象（GDP、自然失业率、货币乘数，…）。

- **蓝色：定理 / 结论**

  ——各模型的核心结果。

- **粉色：假设**

  ——一个模型所维持的前提，集中列于一处。

- 全书还穿插着**例**（worked examples）、**注**（remarks）以及强调框。

记号约定

全书中，Y 表示总产出（实际 GDP），C 表示消费，I 表示投资，G 表示政府购买，$\mathrm{NX} = X - M$ 表示净出口。K 表示资本存量，L（或 N）表示劳动；小写的 $k = K/L$ 与 $y = Y/L$ 是其人均（劳均）对应量。储蓄率记为 s，折旧率记为 δ，人口增长率记为 n，劳动增强型技术进步率记为 g；稳态值加星号，如 k^*。价格与通胀方面：P 是价格水平，π 是其变化率，上标 e（如 P^e、π^e）表示预期值。利率方面：r 是实际利率，$i = r + \pi^e$ 是名义利率。在跨期问题中，β 是贴现因子，$\mathcal{L}$ 是拉格朗日函数。货币层次为 $M_0 \subset M_1 \subset M_2$。

目录

第一章　导论：宏观经济学的思维方式

学完中级微观经济学，一个学生已经能颇为得心应手地分析单个家庭、单个企业或单个市场。消费者把价格和收入当作给定，去挑选自己最偏好的消费组合；企业把工资率和资本租金率当作给定，去决定生产多少。全部的分析功夫都花在*选择*本身上，而进入选择问题的那些数字——这个价格、那笔收入——都来自外部，已经被人测量好了。宏观经济学恰恰从这种便利结束的地方开始。它研究的对象不是某个家庭或某家企业，而是*作为整体*的经济：一个国家的总产出、所有价格的平均水平、愿意工作却找不到工作的人所占的比例、一代人的生活水准提高的速度。没有人会把这些数字递到我们手上。在建立任何一个总量经济模型之前，我们必须先学会如何*度量*它。

这是宏观经济学的第一个特殊之处，也是本书在任何模型出场之前先用三章讨论度量的原因。但还有第二个、更深层的差别，它塑造了整个学科。整体的行为并不是各个部分行为的简单加总。一个家庭决定多储蓄，会提高这个家庭的财富；可如果所有家庭同时都想多储蓄，总支出就会下降，企业卖得更少，收入随之减少，到头来整个经济储蓄的可能并不比从前多。*加总不等于相加*。要对整体说出任何靠得住的话，我们就需要理论——而正如我们将看到的，宏观经济学家用两种截然不同的风格构建了这套理论，一种看重逻辑的纯粹，另一种看重实用。本章先把全书要回答的问题、要用的方法以及它所依托的两大思想传统摆出来。它是后续一切内容的地图。

1.1　宏观经济学研究什么

宏观经济学研究的是作为整体的经济，而不是它的各个组成部分。研究对象一变，我们首先要解决的问题也跟着变了。在微观经济学里，度量相对容易：消费者把收入和价格当作外生给定，企业要核算自己的成本，但牵涉到的那些数量是具体的、有边界的。一旦我们上升到整个国家的层面，连最基本的数量也得被*构造*出来。我们并不像一个购物者那样直接观察到“价格”；我们观察到的是数以百万计的单个价格，必须设法把它们合成一个单一的*价格水平*（第 三 章）。我们也观察不到“产出”；我们观察到的是无数企业以彼此不可通约的单位进行的生产，必须把它们加总成像国内生产总值这样的一个度量（第 二 章）。

宏观经济学：把经济作为整体来研究

微观经济学分析单个的经济主体和市场，把价格与收入当作*给定*。宏观经济学分析的是总量——总产出、价格水平、失业率、长期增长率——这些量并非给定，而是必须先被*度量并加总*出来。这正是本书先讲国民收入核算（第 二 章）、价格

指数（第 三 章）和劳动力市场（第 四 章），然后才转向任何理论的原因。

正因为这些总量是构造出来的、而不是直接读取来的，它们在某种程度上也是*被构造的数字*——而被构造的数字会受到构造时所做选择的影响。这与凭空捏造不是一回事。像 GDP 这样的数据，是依照一套确定的统计程序、从真实的底层数据组装出来的，但这套程序包含了主观判断，而这些判断是可以被拿捏的。因此，认真的宏观经济数据读者会拿头条总量去对照那些更难做手脚的物理替代指标，也去对照其它经济体的同类序列。

注（与官方数字相互印证）.

> 由于头条总量建立在构造选择之上，分析者常常用难以粉饰的物理指标来做交叉验证。一个有名的例子是所谓的克强指数，它不单看上报的产出数字，而是用电力消费量、铁路货运量和银行贷款来衡量中国的经济活动；其前提是，一家工厂烧掉的电量，比它声称创造的产值更难造假。把同一序列拿到不同经济体之间去比较——比如中国对美国——是又一道理智的核验。这一切并不意味着官方数字是错的；它意味着一个严肃的宏观经济学家会把每一个总量都当作一项背后有方法的度量，并追问那套方法可能漏掉了什么。

1.2　宏观经济学家怎样工作

面对总量经济，人究竟是怎样做宏观经济学的？这门学科有它特有的推进方式。它先是收集*程式化事实*（stylized facts）——数据中那些宽泛而稳健的规律性，比如观察到消费与收入紧密同向变动，或者产出与就业在经济周期中同涨同落——然后试图解释它们。解释可以通过几种不同的方法来进行，值得在一开始就把它们区分清楚，因为本书后面会把它们全用上。

- **简约式分析（reduced-form analysis）**。人们跑回归，把各总量变量联系起来，报告数据中发现的相关关系。这是证据，而且是有用的证据，但它有两个众所周知的局限。一个回归系数未必揭示背后的*机制*；它顶多记录下一种关系。而且宏观经济是一个*动态、环环相扣的系统*，变量之间彼此反馈，所以单个回归看似讲出的那种干净的单向因果故事，往往是站不住脚的。

- **建模**。人们写下一个明确的模型——关于经济主体、技术和市场的若干假设——并推导出它的含义。回报是得到一个由既定前提合乎逻辑地推出的*预测结果*，因此任何预测的来源都可以被追溯，模型也可以被追问。

- **模拟**。人们把历史数据喂进一个模型，看它的行为与实际发生的情况吻合得有多好，借此检验理论与历史记录的拟合程度。

- **校准（calibration）**。人们不是把每个参数都从头重新估计，而是依靠积累起来的经验估计——成堆的既有研究和数据——来确定一个模型的参数，然后再问这样一个被既有经验证据约束住的模型意味着什么。

这些方法是互补的，而不是相互竞争的。一个程式化事实激发出一个模型；模型用既有的估计来校准；校准过的模型拿去与历史做模拟；简约式回归既提供了最初的那个事实，又对模型的预测做了核验。读者几乎会在每一章里都看到这个循环以缩微的形式重演。

注（关于资料与教材）.

> 本课程以新古典传统讲授，英文方面的天然搭档是 Williamson 的《宏观经济学》（*Macroeconomics*），它从明确的微观基础出发展开整个学科。在制度层面——也就是宏观经济政策在中国究竟是怎样运作的——讲授参考了把宏观经济理论与中国政策相搭配的材料。在数据方面，标准的源头是中国国家统计局、CEIC 和 Wind 这类商业宏观数据库，以及在国际和美国序列上常用的美联储经济数据（FRED）服务。知道数字从哪里来，是知道它们意味着什么的一部分。

1.3　两大传统：新古典与凯恩斯

宏观经济学里有一个根本性的方法论岔路口，理解了它，就能在很大程度上解释这门学科为什么是今天这个样子。两个分支在一个问题上分道扬镳：人应当从个体出发往上构建，还是直接从总量入手？

1.3.1　新古典传统：微观基础

新古典的观点在微观经济学和宏观经济学之间不划出截然的界线。在这幅图景里，原则上根本就没有什么独有的“宏观”现象在发生：经济就是它那些做最优化的消费者和企业的加总，因此宏观经济结果应当可以从个体的微观经济行为中推导出来，并与之相一致。现代新古典宏观经济学于是坚持要有*微观基础*（microfoundations）——每一个总量关系都必须扎根于在约束下最大化效用或利润的经济主体的选择之中。

这种纪律的吸引力是实实在在的。一个有微观基础的模型在*逻辑上是严丝合缝的*：它的结论由关于偏好、技术和约束的明确假设严格推出，因此人们总能确切地说出某个结果为什么成立，并把它一直追溯到一个可供争辩的前提。代价同样是实实在在的。人类社会包含着太多互相牵动的部分，而一个干净、内部自洽的逻辑体系往往很难把它们全部容纳进来。坚持要把每一个宏观经济现象都从第一性原理推出，会让最终得到的模型对于我们实际观察到的那个杂乱经济*解释力*弱得出人意料——任何曾经试图把一个漂亮模型套到顽固数据上的人，都熟悉这种沮丧。

1.3.2　凯恩斯传统：让数据说话

凯恩斯式的回应是改变出发点。与其从个体最优化往上构建总量，不如*直接*研究总量，从总量数据中读出整体的行为，让数据说话——哪怕代价是把微观基础悬置不谈。

最典型的例子是总量消费函数。人们不去从每个家庭的效用最大化推导一个社会消费多少，而是看总量消费数据，提出一个简单的关系，

$$C = C_0 + \mathrm{MPC} \cdot (Y - T),$$

其中 $Y-T$ 是总量可支配收入（产出扣去税收），C_0 是自发消费，MPC 是*边际消费倾向*（marginal propensity to consume）——可支配收入每增加一元中被花掉的那一部分。值得注意的一点是，MPC 没有任何微观经济学上的来历。它不是从任何偏好结构推导出来的；它只是一个拟合数据的参数，因为数据大致就是那样表现的。同样的精神后面会出现在索罗模型（第 五 章）里，那里直接*假定*一个不变的储蓄率，而不是从一个跨期最优化问题推导储蓄。

凯恩斯式的取舍

凯恩斯式的进路用*科学的纯粹性换取实用性*。通过从总量出发、把 $C = C_0 + \text{MPC}\cdot(Y-T)$ 这样的关系直接拟合到数据上，它放弃了对这些关系的微观经济推导。它换来的，是一套可操作、有经验依据的论述，目标直指*理解数据本身*——这样的模型也许能很好地解释世界，尽管按其构造方式，它讲不出每个经济主体为什么会做出总量所暗示的那种行为。

两个传统都不是简单地对或错。新古典模型更严谨，将占据我们增长理论的大部分篇幅；凯恩斯模型对短期政策更立竿见影地有用，将组织起本书的后半部分。对宏观经济学成熟的解读，是把两者同时放在心里。

注（远在凯恩斯之前，人们就在研究总量）.

> 若以为宏观经济问题始于凯恩斯，那就错了。古典经济学家早在他之前就把国民财富作为整体来研究——亚当·斯密的《国富论》在真切的意义上就是一部关于总量经济的著作——他们用的是那个时代的古典方法。凯恩斯改变的不是研究的主题，而是研究的方法：他让人们可以正大光明地按总量自身的逻辑去给它建模，而不必先把它还原为个体。

1.4　两个大问题

尽管细节散落各处，宏观经济学其实围绕两个问题来组织，二者以时间跨度相区分。

- **长期经济增长**。为什么有些国家富、有些国家穷，又是什么决定了一个经济的产出在数十年里的长期*趋势*？这是增长理论的领地——索罗模型及其有微观基础的后继者（第 五–七 章）。
- **短期经济周期**。产出为什么不是沿着趋势平滑增长，而是绕着趋势*波动*，有繁荣也有衰退，政策又能不能把这些波动熨平？这是短期宏观经济学的领地——凯恩斯交叉、IS–LM 以及总需求与总供给（第 十–十二 章）。

把这两个问题放在一起把握，一个简洁的办法是把产出的路径看成趋势加上偏离：

$$\underbrace{\text{观察到的产出}}_{\text{我们看到的}} = \underbrace{\text{长期增长趋势}}_{\text{第 五–七 章}} + \underbrace{\text{绕趋势的周期性冲击}}_{\text{第 十–十二 章}}.$$

增长理论研究趋势；经济周期理论研究对趋势的偏离。本书正是按这个划分推进的：先度量各总量，再解释趋势，最后解释波动。

1.5　宏观经济政策的目标

这一切对政策又为什么重要？宏观经济政策按惯例瞄准四个目标，而本书余下的部分都可以读作一项研究：研究是什么挡在了实现这些目标的路上。

1. **经济增长**——在长期内提高产出和生活水准。
2. **充分就业**——确保每一个主动想工作的人都能找到工作。
3. **控制通胀**——把价格水平稳定到足以让货币保持为一把可靠的标尺。
4. **国际收支平衡**——保持国际收支的平衡，这是经济一旦对贸易和资本流动开放就会浮现的关切。

有三点澄清值得现在就标出来，每一点后面都有专门的章节展开。

第一，“充分就业”并不意味着失业率为零。它意味着每一个主动找工作的人都能找到工作，从而失业的*周期性*成分为零，测得的失业率只反映*自然失业率*（natural rate）——也就是那种即便在景气时期也会持续存在的摩擦性和结构性失业，源于人们在不同工作和行业之间流动。我们将在劳动力市场一章（第 四 章）把这一点讲精确。

第二，这些目标无法同时达成。当波动由需求驱动时，就业和通胀往往同向变动，于是朝着更充分就业去推，往往会抬高通胀，反之亦然：两者不能同时被最优化。这种张力是菲利普斯曲线（Phillips curve，第 十四 章）的主题，也是稳定政策的核心两难。

注（中国式的国际收支平衡）.

> 国际收支平衡这个目标，会因一国在世界经济中所处的位置不同而呈现出不同的样貌。中国凭借强劲的出口长期保持着巨额的*贸易顺差*，央行对汇率握有近乎绝对的掌控权。讲授中反复提出的一个问题是：由此积累起来的对外债权，是否真的像它表面看上去那么值钱？堆积对外顺差，在某种意义上就是把你自己的产出借给世界其余部分，而这笔“储蓄”的价值，取决于它最终被换成了什么。本书只在边角处才回到开放经济的问题，但值得早早指出：“国际收支平衡”是一个鲜活而有争议的政策目标，而不是一条已成定论的会计恒等式。

1.6　关于本学科性情的一句话

一开始就坦诚地说清这是一门什么样的学科，是值得的。微观经济学是齐整的：一个提法得当的问题，核心通常是一个干净的最优化，并有一个确定的答案。宏观经济学则刻意地不那么齐整。同一个事实往往可以透过不止一个模型来解读；数据很少能干净利落地了结一场争论；而一堂课上最富启发的话，有时恰恰是一句一行字的制度性旁白——一家央行实际上是怎样进行某项操作的，某项政策为什么偏偏在某一年失败了，某个国家是怎样度量某个特定序列的——而这样的话哪一章都安放得不算妥帖。我们不会假装并非如此，也不会硬把每一条这样的观察塞进一个定理里。哪里有模型，我们就

对它一丝不苟；哪里是课堂岔进了真实经济的肌理，我们就把这段岔话保留下来，单独装进一个注记框里。请把那些框子读作本学科的一部分，而不是没收拾干净的线头。如今，宏观经济学的问题、方法与性情都已在视野之内，下一章我们就转向其中任何一项都首先需要的那项工作：度量总量经济。

第二章　国民收入核算

在解释经济为何繁荣与萧条、增长与停滞之前，我们首先需要一套度量经济产出的办法。宏观经济学研究的是总量——总产出、总收入、总支出——而一个总量的价值，全在于代表它的那个数字。国民收入核算正是这样一套定义与恒等式的体系，它把千百万家庭与厂商纷繁的活动，浓缩成寥寥几个可以度量的数量。它之于宏观经济学家，犹如财务报表之于企业：单看枯燥乏味，却是后续一切分析不可或缺的根基。

核算的核心对象是国内生产总值（GDP），即一个经济体在一段时期内所生产的一切产品的市场价值。本章的大部分篇幅，都是在细致拆解这一句话，因为其中几乎每一个限定词都是经过深思熟虑的选择，并各有其后果。随后我们会看到，同一个总量可以由三条不同的路径得到——按生产计、按收入计、或按支出计——而这三条路径殊途同归，并不是有待证明的定理，而是核算体系本身内建的恒等关系。从支出路径，我们读出对需求的著名分解：消费、投资、政府购买与净出口；从收入路径，读出从 GDP 一路细化到个人可支配收入的链条；从金融市场，读出把家庭储蓄、政府赤字与贸易差额捆在一起的储蓄—投资恒等式。本章最后讨论两个实务问题：产出如何按产业归类，以及在官方数据发布的间隙，分析者如何对 GDP 进行旁敲侧击的估计。

有一句告诫贯穿始终：没有任何单一指标能概括整个国民经济，而 GDP 尤其是为*可度量*而设计的，并不是衡量福利的完整标尺。分清它计入了什么、又遗漏了什么，这场仗就赢了一半。

2.1　国内生产总值

定义 2.1: 国内生产总值

国内生产总值（*GDP*）是指在给定时期内，一个经济体在其地理疆界之内所生产的全部最终产品和服务的市场价值。

慢慢读，这条定义包含了五个限定词，每一个都排除了某些东西。我们逐一来看。

2.1.1　五个限定词

（1）给定时期——GDP 是一个流量。 GDP 总是针对一段时间报告，通常是一个季度或一年；全球的任何国家都有年度和季度 GDP。正因为它是在*一段时间内*度量的，所以 GDP 是一个*流量*（flow），而不是*存量*（stock）。流量是指在一段时间内测度的数量，存量是指在某个时点上测度的数量。这一区分绝非咬文嚼字——它是宏观经济学中

最常见的混淆来源，我们应当养成习惯：对每一个宏观变量都先问一句，它是存量还是流量。

举例来说，财富水平就是一个存量：在这一时刻，经济中总的财富水平是可以测度的。而 GDP 是一个流量，二者并不是一回事。2010 年前后中国 GDP 超过了日本，只能说明从那一年起，中国每一年比日本生产了更多的产品和服务；它并不意味着中国累积的财富水平超过了日本——后者很可能尚未发生。

注（收入、财富，以及金融为何相互抵消）.

> 把三组相关的流量与存量分开来看会很有帮助。收入（流量）等于劳动收入——工资、奖金——加上资本收入——金融资产的收益，例如股权分红、利息，但并不是资本存量本身。财富（存量）则是实物资产与金融资产之和。宏观经济学主要关心收入；财富更多是个人才关心的问题。还要注意，对整个社会而言，金融资产并不是净财富：每一项金融资产都是某个主体对另一个主体的权益要求凭证，在全社会加总时这些要求相互抵消。正因如此，宏观核算可以将其忽略而账目仍然平衡。

注（何时会出现月度 GDP）.

> 通常没有月度 GDP，因为数据来不及处理。但在特殊情形下，当局也会估算它：2022 年疫情扰动期间，中国密切跟踪月度数据，以制定增长目标、稳定预期——2022 年 4 月约为 -2.2%，5 月约为 -0.1%，6 月约为 $+4\%$，合计起来二季度约为 $+0.5\%$。在中国，每月 16—20 日左右公布按收入法初步核算的 GDP；月度的实体经济数据基本上都是名义值，因为来不及做价格平减。完整的支出法年度核算——其中包含居民消费和政府消费的比重——则要等到次年的五六月份才公布。

（2）市场价值。由于 GDP 要把苹果、理发与软件统统加在一起，它就必须用一个共同的单位来给一切定价，而这个单位就是市场价格。依赖市场价格有一个直接的推论：那些从不在市场上交易的东西没有可观察的价格，因而原则上难以核算——自给自足的自然经济成分、家庭家务劳动以及非法交易都在其列。关键在于，产品或服务能且会以市场价格易手，而不是说每一单位都必须真的发生过一次交易（下文将讨论的存货，就是先被计入、后才售出的产出）。

（3）地理范围——“国内”。 GDP 是按疆界定义的。中国的 GDP 涵盖在中国境内进行的生产，不论生产者是中国国民还是外国人。在全球化时代，这是自然的选择，也与“地方政策须与当地条件相适配”的逻辑一致；我们将在下文把它与按国民计的口径（GNP）作对比。

（4）最终品，而非中间品。一件产品究竟是最终品还是中间品，取决于它的用途。被用作进一步生产之投入而消耗掉的，是中间品；为消费或投资而购买的，是最终品。GDP 只核算最终品。这样做是为了避免重复计算：如果把面粉和蛋糕都算上，面粉就被计了两次。注意，同一件实物产品可以是其一，也可以是其二，全看用途——家庭买来在家做蛋糕的面粉是最终品，而烘焙店买来做蛋糕出售的面粉则是中间品。

（5）市场销售或推算的产出。 GDP 核算的是被生产并在市场上销售的东西，外加少数几样并非真的售出、却理应按市场价值估算的东西，即通过推算（imputation）计入。最典型的例子是自有住房的居民“向自己购买”的住房服务：核算时会加上一笔推算租金。（把买房的价格计入 GDP，与把它日后产生的住房服务流计入 GDP，二者并不

矛盾——它们是不同的东西，下文讨论投资时会说明。）

2.1.2　何者计入、何者不计

五个限定词已能处理大多数情形，但边界仍足够微妙，值得把那些标准的处理口径集中列在一处。其组织原则很简单：*GDP 只把新创造的价值计入一次*。

注（GDP 计入与否的边界情形）.

- *没有新创造，就不计入*。政府税收和转移支付不计入 GDP——它们只是把既有资金重新分配。（政府雇员的*工资*要计入 GDP，因为那是对其所提供服务的报酬。）股票交易不计入 GDP，因为它只是所有权的转移、并无新增产出；交易的印花税不计入 GDP，因为那只是把一部分资金转移给政府；但*佣金*要计入 GDP，因为经纪是一项被生产出来的服务。
- *只计新创造的那一部分*。新房成交价并非全部计入 GDP，因为其中一部分是地价。只有土地上新建建筑的价值才计入。购房时缴给政府的税款不计入。
- *存货作为投资计入*。今年生产、尚未售出的产品仍是今年的产出，无论它最终何时成交。从支出口径看，当存货品日后被卖出时，消费（或投资）支出上升，而存货同时等额下降，二者相抵。
- *推算价值*。未在市场上销售、但概念上可市场化的产出，会经推算后计入——自有住房服务是标准情形。
- *研发如今算作投资*。研发支出过去被视为生产过程的中间投入，如今计为投资。
- *服务按价格而非质量计价*。一项服务按其价格计入，与它质量好坏无关。
- *防御性支出会虚增 GDP*。仅仅为抵消某种负外部性而发生的活动同样被计入，从而高估了真实产出与福利。倘若某企业在生产时造成污染，它的产出本身计入了 GDP，事后的治理清污又再次计入 GDP——GDP 于是被二次虚增。

最后一点是最尖锐的提醒：GDP 是一笔产出的清点，而不是福利的指数。这条定义的选取，一半是为了它的含义，一半是为了它的可度量性，结果便是它有意地把大量价值留在了账外。

2.1.3　从国内到国民：GNP 与 GNI

地理这一限定词有一位天然的“近亲”。GDP 把边界划在一片*疆土*周围，国民生产总值则把边界划在一群*国民*周围。

定义 2.2: 国民生产总值

国民生产总值（*GNP*）是指在给定时期内，一个国家的*国民*所生产的全部最终产品和服务的市场价值，无论他们在世界何处进行生产。

两个口径只在“计入哪些人”上有所不同，并由一个干净的调整项相联系：

$$\text{GNP} = \text{GDP} + (\text{本国人在境外的生产}) - (\text{外国人在境内的生产}).$$

对一个大型经济体而言，GDP 与 GNP 之间的差距很小。世界银行用人均 GNP——如今改称为国民总收入（GNI）——来衡量一国国民的富裕程度，把各经济体划分为低收入、中等收入和高收入；中国目前处于上中等收入水平。

2.2　循环流转与基本恒等式

为什么各种度量产出的方法会给出同一个答案？原因在于：一个主体的支出正是另一个主体的收入，经济中的各笔支付在一个闭环里彼此追逐。这个闭环就是*循环流转*（circular flow），如图 2.1 所示。

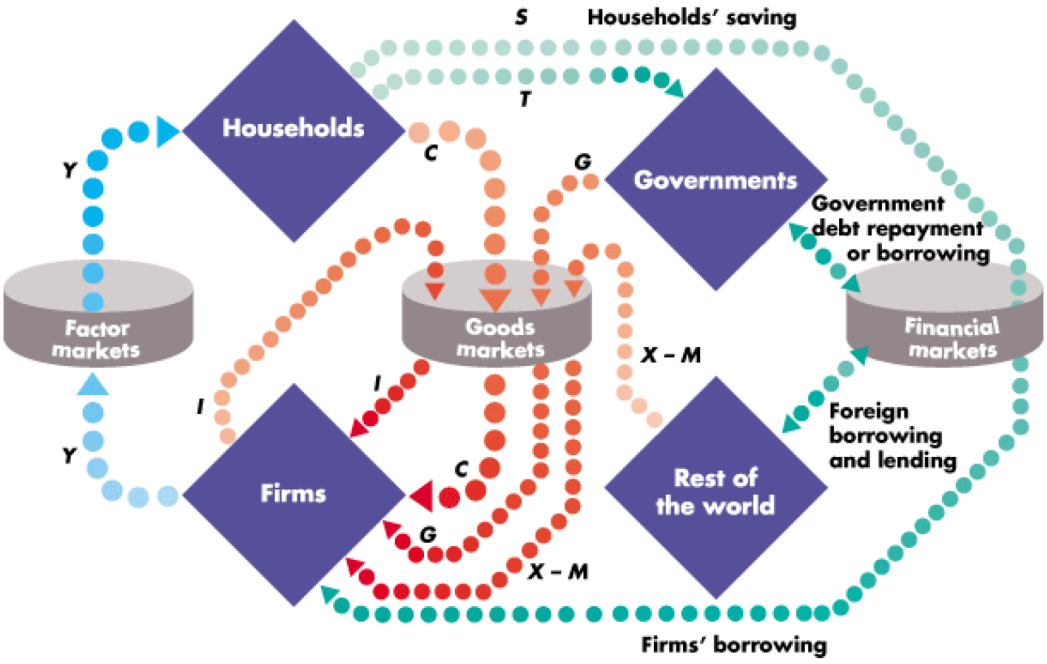

图 2.1: 收入的循环流转：家庭、厂商、政府与世界其他地区跨越产品市场、要素市场与金融市场进行交易，从而使总收入等于总支出。

分析的核心在于*家庭*。家庭是整个系统的枢纽，因为它既负责收入、也负责支出：广义地看，家庭是要素市场的供给方，也是产品市场的需求方。收入流向家庭，随后分散到消费 C、储蓄 S 与税收 T，最终又以 C、I、G 和净出口的形式重新表现为各项需求。

三个市场组织起整幅图景。*产品市场*关联了主要的宏观经济主体——厂商、家庭、政府和世界其他地区。厂商是这个市场的供给方；需求来自家庭、政府和国外。*要素市场*则闭合了这个循环：厂商在产品市场上赚得的收益，最终要回流去报偿生产要素，而这些要素的供给方——并因此获得相应收入——正是家庭。*金融市场*把储蓄与投资联

系起来，让经济得以向外国人“放贷”、使净出口得以发生，同时也与政府的财政紧密相连。

由于收入与支出本是同一笔支付从两侧看到的两面，总收入（AI）等于总支出（AE）：

$$\mathrm{AI} = Y = C + I + G + (X - M) = \mathrm{AE}. \tag{2.1}$$

恒等式 (2.1) 是*事后*（ex post）成立的：按构造，已实现的收入恒等于已实现的支出。而总需求是*事前*（ex ante）的概念——它是由意愿中的 $C + I + G + \mathrm{NX}$ 加总起来的、计划和打算进行的支出；短期波动恰恰起因于计划需求与已实现产出之间的背离。因此，分解式 $C + I + G + \mathrm{NX} = \mathrm{AD}$ 是一个短期对象，不断波动，且各项是按进行支出的*主体*来划分的。

2.2.1　拉动增长的三驾马车

在像中国这样的经济体里，按主体划分支出会遇到麻烦：国家进行了大量投资，所有制又往往边界模糊——许多国有企业以混合甚至民营所有制运营，因此常常无法判断某笔支出究竟属于公共还是私人。补救之法是：不按*谁*支出、而按支出的*性质*来归类政府支出 G，把 G 的每一部分重新归入消费或投资。政府*消费*并入 C，政府*投资*并入 I。（原则上 I 还可再分为私人投资和政府投资，但统计上无法分开。）按性质划分后，(2.1) 便坍缩为

$$\mathrm{GDP} = C + I + \mathrm{NX}. \tag{2.2}$$

这三项就是著名的拉动经济增长的“三驾马车”：最终消费支出 C、资本形成总额 I，以及货物和服务的净出口 $\mathrm{NX} = X - M$。前两项之和 $C + I$ 是*内需*；净出口是*外需*；并能分别算出各自对一年增长的贡献率。

2.3　通向同一总量的三条路径

恒等式 (2.1) 已经暗示，产出可以用不止一种方式来核算。标准的方法有三种，且按构造它们都得到同一个 GDP。

2.3.1　生产法（增加值法）

定义 2.3: 增加值

一个厂商的*增加值*等于其销售额减去用于中间投入（原材料）的支出。按生产法核算的 GDP，就是把全经济所有生产性机构的增加值加总起来。

这种方法最直接地演绎了 GDP 的定义：增加值恰恰是厂商新创造的价值，把它加总起来就避免了对途经其间的中间品的重复计算。销售额、中间投入、增加值与 GDP 都是流量概念。

有一个细节值得强调：所减去的*仅仅*是中间投入成本，而*不包括*财务成本、用工成本或销售成本等（如工资、利息等支出）。生产法看的是*生产*——也即生产性机构所创

造的价值。它实际上把生产性机构当作 GDP 的来源，而不再单独去核算家庭通过提供要素所创造的价值；从效果上看，它是从整体上把握社会的生产过程。与之相对，若把全经济所有的销售额加总起来，得到的是工农业总产值，其中的中间环节销售被重复计算了；中国征收的增值税，正是把这部分重复计算剥离了出去。

2.3.2　收入法

定义 2.4: 收入法

按收入法核算的 GDP，是经济体中所有主体所赚取收入的加总，因为增加值终归会作为收入归于某个主体。

加总成 GDP 的收入包括：工资收入、利息收入（营运资本的回报）、税收收入，以及会计利润——其中会计利润又包含经济利润和租金两部分。租金这一项刻画的是：用于自己生产的机器设备，本可以拿到市场上租出去，自己使用它，实际上相当于自己生产、自己消费了这项租赁服务。存货也在此处计入，可理解为当年创造、来年创收的价值。

2.3.3　支出法

定义 2.5: 支出法

按支出法核算的 GDP，是对最终产品和服务的支出之和：$C+I+G+(X-M)$。

它的焦点是产品市场。先把政府和进出口搁在一旁：产品市场上的每一件最终品，要么被家庭买走，要么被厂商买走。厂商作为资本品买走的机器设备是投资支出 I；其余几乎都是家庭的消费支出 C——只有一个标准的例外，即购房计入投资、而非消费。再把政府和进出口纳入进来，并不改变什么本质的东西；焦点始终落在产品市场上。按照惯例，家庭购买的物品——住房除外——其价值完全计在购买当期。

三种方法，一个数字

生产、收入与支出，是从三个视角去观察同一个循环流转。生产中创造的增加值会作为收入归于某人，而那笔收入又会被支出。因此三种方法度量的是同一个 GDP，它们的相等是一条核算恒等式，而不是一项经验发现。

2.4　金融市场核算与储蓄—投资恒等式

金融市场连接起同样的四个主体——家庭、政府、厂商和外国人。家庭储蓄 S 是资金的主要来源；厂商从中吸纳资金以融通投资 I；政府出现赤字 $(G-T)$ 时，必须向市场举债；外国人购买本国的净出口 $(X-M)$ 时，同样需要借钱。顺着 GDP 核算的思路，可以把储蓄 S 看作供给金融市场的全部“收入”，把 $I+(G-T)+(X-M)$ 看作这些资金的全部“用途”。

我们可以把这一关系干净地推导出来。收入恰好有三种用途——消费、储蓄和税收——这就是金融流转恒等式

$$Y = C + S + T. \tag{2.3}$$

令收入的用途 (2.3) 等于支出恒等式 $Y = C + I + G + (X - M)$，消去 C，便得到 $S + T = I + G + (X - M)$，即

$$S = I + (G - T) + (X - M)\,. \tag{2.4}$$

这与上文的直觉相符：储蓄为私人投资、加上政府赤字、再加上外国人为购买我国出口而进行的净借款提供资金。若改为解出投资，则有

$$I = S + (T - G) + (M - X)\,. \tag{2.5}$$

可见投资的资金来自三处：

- 私人储蓄 S；
- 政府预算盈余 $T - G$；
- 向外国人的净借款 $M - X$。

这里，单独的 S 是*民间储蓄*，而 $S + (T - G)$ 是*国民储蓄*。

注（读懂符号）.

> (2.5) 中有两项相对于 (2.4) 翻转了符号，值得弄清其缘由。税收 T 相当于二次分配——继首次分配之后、遵循公平原则的再一轮配置，如累进所得税。（注意中国的征税本质上是工资税，而非财产税。）$T - G$ 这一项是说，政府以税收作为资金来源、又向产品市场支出；若有盈余，便*流放*进金融市场；若有赤字，便*向*金融市场举债。$M - X$ 这一项是说，外国人购买我国净出口时须以本币支付，而本币要靠向我们借钱获得，于是抽走了国内金融市场的资金；因此 $M - X = -(X - M)$，正是 (2.4) 中那笔外国借款的镜像。

注（宏观意义上的储蓄）.

> 关于储蓄有三点提醒。第一，*国民储蓄*是总收入减去总消费，中国的国民储蓄率远高于大多数国家；储蓄与投资有极强的正相关关系。第二，宏观意义上的储蓄并不等同于日常生活中“把钱攒起来”的概念——它是进入并在金融市场中流转的资金。第三，外商直接投资（FDI）带来的，与其说是资金，不如说更多是先进的技术和管理经验。

2.5　投资与资本形成

在三驾马车中，投资是最容易被误解的一项，因此值得单独处理。在核算中与投资相对应的度量是*资本形成总额*，它的任务是追踪经济体实物资本存量的变动。

定义 2.6: 资本运动方程

资本是一个会折旧、并由投资这一流量来补充的存量。记 K_t 为资本存量，I_t 为总投资，δ 为折旧率，则

$$\Delta K = K_2 - K_1 = I_1 - \delta K_1, \qquad \text{等价地} \qquad K_2 = (1-\delta)\,K_1 + I_1.$$

这才是真正意义上的投资，它包括两部分：固定资本形成总额，以及存货的变动（后者在总量中占比很小）。投资与单纯的资产买卖之间的区别至关重要。除存货变动之外，只有*形成资本*——这里指实物资本——的支出，才是宏观经济意义上的投资。相比之下，买入一份股票只是转移了一项金融资产的所有权；在此过程中资产价格或许会变动，但并没有新的资本品被生产出来。

注（固定资产投资与资本形成）.

> 有两个相关的统计量不应混为一谈。*固定资产投资*是一个会计范畴，它包括土地购置费、旧设备和旧建筑物的购置费——但这些*并不*计入资本形成总额，因为它们并未创造新的资本。规模以下的项目，以及计算机软件等无形资产投资，同样被排除在外。土地购置费（来自土地拍卖）实际上是地方政府财政的主要收入来源。其结果是，报告的固定资产投资与资本形成总额之间差异很大。此外，跨省的固定资产投资可能被重复计算，因此各省数据加总起来甚至可能*超过*全国总量。中国国家统计局公布的固定资产投资是年初至今的累计值，因为单月值波动剧烈；投资数据比消费数据更难采集、计算也更繁杂，而 1、2 月份的数据合并报告，以消化春节因素的时间扰动。固定资产投资有三项最重要的分项指标：房地产、基建和制造业。

注（买房永远算投资）.

> 无论购买何种类型的房子，都算作投资，而绝不算作消费——持有期足够长、金额足够大，使其具有投资属性。这一点对宏观经济举足轻重：中国的新房销售近来下滑超过 20%，而整体经济的疲弱也与房地产的下行密切相关。

不妨在此先约定一组日后会反复用到的要素价格记号。令 r 表示*资本租赁价格*，它*不是*利率这个概念；令 w 表示劳动的租赁价格（即工资）。产出由资本和劳动通过生产函数 $Y = F(K, L)$ 生产出来。原材料与资本之间更深层的对照在于耐久性：原材料是一次性消耗掉的，而资本是耐用品，它的“消耗”恰恰就是折旧，它的补充则通过投资这一过程来完成。人力资本也具有同样的属性——随着时间的流逝，人力资本同样会折旧。

2.6　从 GDP 到可支配收入

GDP 是一个*总*（gross）量：它包含了产出中仅用于替换当年损耗资本的那一部分。同理，总投资 I 也包含了用于替换折旧资本的那一部分。要从这个“总”的生产量出发，逐步逼近人们真正可用于支配的金额，我们需要先剥去折旧，再做一系列进一步的调整。这条链条机械、却值得集中列出。

定义 2.7: 从 GDP 逐级到个人可支配收入

$$\begin{aligned}
\text{NNP} &= \text{GDP} - \text{折旧},\\
\text{国民收入} &= \text{NNP} - \text{间接营业税},\\
\text{个人收入} &= \text{国民收入} - \text{公司利润} - \text{净利息}\\
&\quad + \text{股息} + \text{政府转移支付} + \text{个人利息收入},\\
\text{个人可支配收入} &= \text{个人收入} - \text{个人所得税}.
\end{aligned}$$

净国民生产总值（NNP）从总量中剥去折旧；再减去间接营业税，便得到国民收入，即生产要素所赚得的总额；个人收入则对国民收入做两类调整：一是减去“赚得但未收到”的收入（公司留存收益与净利息），二是加上“收到但并非来自生产”的收入（转移支付、股息、个人利息）；最后扣除个人所得税，剩下的便是个人可支配收入，也就是家庭可自由用于消费或储蓄的金额。

2.7　政府预算与财政目标

政府的账目通过其征收的税款与进行的购买进入循环流转，而二者之差就是预算余额。

定义 2.8: 政府预算余额

$$\text{政府预算余额} = T - G,$$

即税收 T 超过政府支出 G 的部分；为正时是盈余，为负时是赤字。

财政政策可以瞄准两个相当不同的目标。第一是*平衡预算*（balanced budget）策略——量入为出，把 $T - G$ 维持在零。第二则以*经济增长*为财政政策的核心目的，在赤字有助于增长时接受赤字。

注（国家能力与中国的财政保守）.

> 一个国家能追求哪个目标，取决于它的财政能力，而历史颇富启发。英国的光荣革命和《权利法案》大幅提升了国家征税与举债的能力，为工业革命开辟了道路。明清时期的中国，尽管权力更为绝对地集中，征税却很弱、人均税负也很轻，因此其财政能力——以及与之相应的军事与国防力量（即国家能力，state capacity）——也就相应薄弱。今日的中国仍带着这份遗产的影子：它对赤字率保持严格控制、希望负债越低越好，是一种较为保守的财政传统。然而当今的下行压力，使古代那种平衡预算的本能并不契合时势；眼下恰当的策略是增长，而非平衡。中国公布的财政活动占 GDP 的比重并不大，但其*隐性*负债却很大。

2.8　GDP 的产业构成

总产出不仅可以按“谁来支出”归类，也可以按“在何处生产”归类。标准的划分是分成三大产业。

定义 2.9: 三大产业

- 第一产业包括农、林、牧、渔业。
- 第二产业包括工业和建筑业。
- 第三产业即服务业。

注（产业划分的惯例与结构转型）.

几点实务说明。中国把农、林、牧、渔业中的服务计入第一产业，因此第三产业与“服务业”本身略有不同，不过数值上基本相当；农产品是无法用于平衡贸易的。第二产业中的“工业”涵盖三大类：采矿业，制造业，以及电力、热力、燃气及水的生产和供应业。产业占比通常既看占 GDP 的份额、也看就业人数占就业总人数的比例，其变动被用来衡量一国的产业结构转型。中国如今增长最快的是第三产业，约占经济的 60%。

关于把这些占比当作政策目标来读，需要提一句警告。产业结构的变化应被理解为一种结果，而不是一个原因。不能因为某个值得效仿的国家有某种特定的产业占比，就把那套结构照搬照套到自己的经济上、并指望复制其成功。制造业仍是推动增长的强劲引擎；即便我们希望服务业的比重大一些，所需要的也是更好的制造业，而不是更少的制造业——正如环境问题的解药并不是浪漫地退回田园牧歌、一味削减碳排放，而是通过快速增长、让经济在前进中减碳。

注（消费率上升，投资率下降）.

中国如今是消费驱动的经济：它的消费率上升、投资率下降。当短期增长令人失望时，一种诱惑是归咎于消费疲弱、并去刺激消费。但“把结构读作结果而非原因”这条教训，给出的指向恰恰相反。历史上的快速增长，依靠的是快速投资，而不是快速消费。中国的投资率其实远高于美国（约为 40% 对 16%），而“中国投资过度”的说法被夸大了：问题不在于投资太多，而在于投资投错了地方——是投资得不好，而不是不该投资。同样，消费的疲弱也并非消费本身的问题：消费增速放缓反映的是收入与社会保障网的问题，主要原因是可支配收入的快速下降，因而单靠刺激消费是治不好的。

2.9　GDP 的旁证指标

在某种意义上，GDP 是一个理论构造，要带着时滞计算和发布。在两次发布之间，分析者依靠一些更及时的现实数据序列作为侧面指标。每一种都带有有用的信号，也都带有各自特有的噪声。

1. **用电量与货运量**。两者都追踪实物层面的活动，但也都取决于经济的**结构**：完全有可能出现用电量和货运量增速放缓、而整体经济活动反而加快的情形，只要活动转向了能耗与运输强度较低的部门。

2. **新增信贷**。这是一个月度序列，但与央行的货币政策、宏观审慎政策以及金融发展水平紧密相连；它的变动未必反映实体活动。

3. **采购经理人指数（*PMI*）**。一个经过季节性调整的月度指数，以 50% 作为扩张与收缩的分界线。PMI 是每月最早可获知的综合性宏观读数。其中的**新订单**分项是一个先行指标，可预测未来的生产，并与实体经济序列高度相关；央行也在自己的预测中采用 PMI。

4. **社会消费品零售总额**。指实物商品零售加上餐饮服务。它**并不**等同于国民收入核算中的消费 C——出行、教育等服务未被计入——但它更易于采集，主要从企业和商户处采集。

注（社零掩盖了什么，以及耐用品与易耗品）.

> 社零之中最重要的单项是汽车，正如猪肉在消费价格指数中举足轻重。消费可分为**易耗品**（一次性消费）和**耐用品**；持有时间最长、金额最大的购买通常是汽车，其中新能源车是主要的驱动力量——而它带有一部分投资成分。疫情极大地压缩了消费的场景与手段。易耗品和服务的消费无法递延，而耐用品消费可以推迟，有时还会在日后产生补偿性消费——婚恋（金银珠宝）和汽车是最典型的例子。（由此还有“口红效应”：经济下行时转向小额奢侈品的消费降级。）

第三章　价格、实际产出与通胀指数

第 二 章的国民经济核算用一个数字概括了一个经济体一年内生产的全部产品的价值。可这个数字是用货币来衡量的，而货币是一把会滑动的尺子：GDP 上升 10% 既可能意味着经济体多生产了 10% 的产品，也可能意味着产量分毫未变、只是价格涨了 10%，或是介于二者之间的任何组合。要判断一个国家究竟有没有增长，要把它和自己的过去相比，或是和另一个国家相比，我们都得先把支出变化中反映“产出更多”的那一部分，同反映“价格更高”的那一部分区分开来。把这两者拆开，正是本章要做的事。

我们分两步走。第一步，把名义 GDP 拆成一个实际数量和一个价格水平，由此得到 GDP 平减指数（GDP deflator），并把名义增长干净利落地分解为实际增长加上通货膨胀。有了这个工具，我们终于可以对 GDP 做跨期和跨国的比较；其间我们会停下来讨论几个实务上的报告惯例——同比与环比、以及经季节调整的折年率——正是它们使公布的增长数字之间具有可比性。第二步，我们梳理经济学家实际查阅的那一族价格指数：消费者价格指数（CPI）及其“核心”变体、生产者价格指数（PPI），以及个人消费支出（PCE）指数。一个反复出现的主题是：并不存在唯一的、那个“总”价格水平；每个指数回答的是不同的问题，选错了指数，就会悄无声息地扭曲答案。最后我们把 CPI 和 GDP 平减指数直接放在一起比较，因为二者之间的差异——固定权重对变动权重、进口品对国产品——恰恰是会出现在真实政策辩论中的那类测量细节。

关于本章范围的一点说明。本章谈的是*度量*价格与通胀，而非*解释*它们。整体价格水平为什么会变动、通胀给社会带来什么代价、一个靠印钞的政府所征收的“通胀税”，以及把货币增长同通胀联系起来的货币数量论，这些都留到第 九 章再讲。这里我们只打造尺子；至于被度量的东西为什么会动，其背后的理论留待之后。

3.1　名义 GDP 与实际 GDP

名义 GDP 从一年到下一年的变化里，混杂着两件本应分开的事：数量可能变了，价格也可能变了。要把数量单独分离出来，我们就把产出按某个选定的*基期*的价格固定下来重新计算。这正是名义 GDP 与实际 GDP 之分背后唯一的那个想法。

定义 3.1: 名义 GDP 与实际 GDP

名义 GDP（nominal GDP）按每种产品在其生产当年的现行价格来计价，

$$\text{名义 GDP}_t = \sum_i p_{i,t}\, q_{i,t},$$

其中 $p_{i,t}$ 与 $q_{i,t}$ 是产品 i 在第 t 年的价格与数量。*实际 GDP*（real GDP）则把同样的这些数量按某个基期 0 的固定价格来计价，

$$\text{实际 GDP}_t = \sum_i p_{i,0}\, q_{i,t}.$$

因为实际 GDP 把价格固定住了，它的每一处变化反映的都是实物产出的变化，而非价格的变化。

注（中国的数据口径）.

> 中国国家统计局公布这两套序列时，用的名称翻译过来分别是“现价 GDP”，也就是名义 GDP；以及“不变价 GDP”，也就是按某个声明的基期来度量的实际 GDP。通常所说的“GDP 增速”和公布的“GDP 指数”指的都是*实际* GDP 的增长。读到任何这类序列时，要问的第一个问题永远是：以哪一年为基期，以及这是一个现价还是不变价的数字？

3.1.1　GDP 平减指数

如果说实际 GDP 抓住的是数量，那么拿名义 GDP 同实际 GDP 一比，剩下的那部分就必然抓住的是价格。这个余项就是整个经济体涵盖面最广的价格指数。

定义 3.2: GDP 平减指数

*GDP 平减指数*是名义 GDP 与实际 GDP 之比，按惯例乘以 100 来标度，

$$\text{GDP 平减指数}_t = \frac{\text{名义 GDP}_t}{\text{实际 GDP}_t} \times 100 = \frac{\sum_i p_{i,t}\, q_{i,t}}{\sum_i p_{i,0}\, q_{i,t}} \times 100.$$

它是一个涵盖进入 GDP 的*所有*产品的价格指数，并以*当年*的数量为权重。

在基期，名义 GDP 与实际 GDP 重合，平减指数恰好等于 100。只要价格自基期以来平均而言是上升的——在任何存在正通胀的经济体里都属常态——分子就大于分母，平减指数便在 100 之上；而在基期之前，它会落在 100 以下。一个好记的口诀是：平减指数在多数时候都大于 100，原因恰恰在于基期通常落在过去。

3.1.2　名义增长的分解

现在我们可以把“名义增长就是实际增长加上通货膨胀”这句含糊的话讲精确了。记 g^N 为名义 GDP 的增长率、g^R 为实际 GDP 的增长率、π 为 GDP 平减指数的增长率

（即平减指数通胀率）。由于名义 GDP 是实际 GDP 与平减指数的乘积，它的毛增长因子就是另外两者的乘积：

定理 3.3: 增长分解

毛增长因子满足

$$1+g^{N}=\left(1+g^{R}\right)(1+\pi).$$

展开后丢掉二阶交叉项 $g^{R}\pi$——当两个增长率都很小时这一项也很小——就得到便于使用的近似式

$$g^{N}\approx g^{R}+\pi.$$

证明. 按定义有 名义 GDP = 实际 GDP × (平减指数/100)，于是在相邻两年间取比值，

$$1+g^{N}=\frac{\text{名义 GDP}_{t+1}}{\text{名义 GDP}_{t}}=\frac{\text{实际 GDP}_{t+1}}{\text{实际 GDP}_{t}}\cdot\frac{\text{平减指数}_{t+1}}{\text{平减指数}_{t}}=\left(1+g^{R}\right)(1+\pi).$$

乘开后 $1+g^{N}=1+g^{R}+\pi+g^{R}\pi$，故 $g^{N}=g^{R}+\pi+g^{R}\pi$。对于量级为百分之几的增长率，乘积 $g^{R}\pi$ 的量级仅在百分之零点几，故予以舍去。 □

这个分解写成乘积形式时是精确的，写成加法形式时则只是近似；两者的差距就是那个交叉项，它只在增长率或通胀率很大时才要紧。这与支配利息的复利逻辑是同一回事：基期一旦确定，在正常年份里，名义 GDP 会越来越偏离实际 GDP，平减指数也越来越大，因为每一年的价格上涨都在前一年的基础上复利累积。

3.2　GDP 的跨期与跨国比较

我们之所以费这番功夫把价格从 GDP 里剥离出来，是因为我们想要比较：今年比去年、一国比另一国。正是实际 GDP 让这些比较变得有意义。

3.2.1　跨期比较与跨国比较

要把一个经济体同自己的过去相比，就必须把价格因素去掉，而这正是实际 GDP 所做的：实际 GDP 的上升是产品和服务数量上的真实增加，而不是通胀造成的假象。要做跨国比较，我们还需额外把数字换算成同一种货币（并且，做福利比较时，往往还要通过购买力平价来修正价格水平的差异，这一精细化处理在此暂且不论）。于是 GDP 既为追踪长期增长、又为在某一时点上给经济体排序，提供了一个共同的基准。

即便在使用这把尺子时，也值得把它的局限放在眼前。GDP 记录的是生产的市场价值；它对这份产出如何分配、对财富（而非收入）、对环境退化、对腐败，以及对人们是否因这些被它计入的活动而更快乐，统统是缄默的。这些都不会让 GDP 变得无用——跨国家、跨数十年来看，从预期寿命到识字率，GDP 同我们真正在乎的那些东西都稳健地正相关。正确的结论不是 GDP 可以被抛弃、或被某个更好的单一数字取代，而是人类福祉本就是多维的，最好同时参照若干个指标来读，而 GDP 是其中信息含量最高的指标之一。

3.2.2　同比、环比与折年率

高频数据带来了一个年度数字会掩盖的报告难题。某一季度的产出，既可以同*去年同季度*相比（同比增长），也可以同*紧邻的上一*季度相比（环比增长），而这两者讲述的是不同的故事。

- **同比**增长拿相同季度同相同季度相比，因而会自动抵消掉规律性的季节模式——比如说，零售额总会在某个节日前后激增。它的短处是慢：因为它把中间整整一年的变化都平均了进去，所以在揭示经济的转折点上可能反应迟缓。
- **环比**增长则及时，变化一发生就被它记录下来；但也正因如此，它被季节性所污染。要让它可用，就必须做*季节调整*，即在计算增长率之前，先用统计方法把年内规律性的季节模式剔除掉。

注（本应吻合、却未必吻合的两套序列）.

> 这两种口径并非彼此独立——一套自洽的季度水平数据，应当能让同比与环比增长率相互对得上——可实务中，中国公布的同比和环比 GDP 数字有时却无法相互印证，这提醒我们：季节调整这一步本身就是一种估计，对高频宏观数据应当抱着一些容忍度去读。

有若干经济体——美国、日本和欧元区都在其列——会报告一个*经季节调整的折年率*（seasonally adjusted annualized rate，SAAR）。其想法是，取当季经季节调整后的环比增长，然后追问：*假如*经济按这个季度速度连续增长四个季度，会得到怎样的年增长率？写成公式，若 g 为环比增长率，则折年率为

$$\mathrm{SAAR} = (1+g)^{4} - 1.$$

折年率既及时又直观——它把一个季度的读数用人们熟悉的年率单位讲出来——但由于它把单个季度的增长复利了四次，它也把那个季度的噪声放大了，所以 SAAR 比同比增长波动得厉害得多。

注（被熨平的数据与“伤疤效应”）.

> 统计机构可能会刻意“削峰填谷”，让公布出来的序列看上去比真实经济更平滑。熨平*产出*序列是一回事——GDP 还吸收得了——但底层那个砸向*收入*的冲击却没那么容易被糊弄过去：一次真正的衰退会永久性地削减受影响者的收入，而支出下跌得还要更多，一方面是因为收入降了，另一方面是因为家庭会以预防性储蓄来应对冲击，从而持久地改变其消费与储蓄习惯。一次急剧的收缩在收缩本身过去很久之后，仍留在经济活动水平上的那道持久创伤，就叫做*伤疤效应*（scar effect）。

3.3　实务中的价格指数

GDP 平减指数涵盖面广，但它公布得不勤、又有滞后，而且它也不是观察生活成本、或观察对货币政策最为相关的那些压力时被盯得最紧的指数。实务中，经济学家会查阅一小族价格指数，每一个都是为不同的目的而构造的。我们逐个来看。

3.3.1　消费者价格指数

定义 3.4: 消费者价格指数（CPI）

消费者价格指数度量的是一篮固定的、能代表典型家庭消费的产品与服务的成本。每一类消费被赋予一个反映其在代表性家庭预算中占比的权重，指数随时间追踪这一篮固定的、加权的产品的成本。一言以蔽之，它是对生活成本的一种衡量。

CPI 的构造就是为了回答一个问题：本月再去买典型家庭从前买的那同一篮消费品，要多花（或少花）多少？由于这一篮商品及其权重被固定住了——在中国，它们大约每五年才修订一次，而且各细类的权重并不对外公布，不过可以从其他公布的数字间接倒推出来——这个指数就把价格的变化同人们所买东西的变化分离开了。

注（篮子里装了什么——一幅中国快照）.

> 食品在中国 CPI 里占很大的权重；单是猪肉，历史上就一直是权重最高的单一商品，而由于养猪遵循一个政策很难驯服的生产周期，猪肉价格大幅波动，把指数也一并拖着走。范围更广的食品烟酒类大约占指数的四分之一。这里有一条一般性的规律：食品消费占比越低，生活水准就越高——这就是恩格尔定律，即越富裕的家庭，花在食品上的收入比例越小，花在一切能提升生活品质的东西上的比例越大。居住主要通过租金进入 CPI：因为中国自有住房比例高，买房被当作投资而非消费，所以房价的涨跌并不直接进入 CPI，于是房地产价格的迅速攀升或崩塌可能根本不会在指数里显现。出于同样的道理，股票和债券价格——它们是资产，不是消费品——也被排除在外。

一个标准惯例把 CPI 通胀高于 2% 看作真正的通货膨胀，把低于 2% 看作通胀放缓、渐渐过渡到通货紧缩。CPI 在实务上最大的优点是频率：它每月公布，这使它在实时分析上远比按季公布、又有滞后的平减指数有用。

注（CPI 并非全貌）.

> 只靠 CPI 来讨论经济，得到的是一幅不完整的图景。直到 2000 年代中期，美国 CPI 一直波澜不惊，而房价和次级抵押贷款市场却剧烈波动；这个指数根本没看到那场最终在 2008 年酿成危机的金融风险的累积。事后吸取的教训是：必须把消费价格同资产价格——房地产和金融资产——放在一起看，决策者必须留意一组更广的宏观指标、留意系统性风险，这就是如今所称的宏观审慎（macroprudential）政策立场。

3.3.2　成本推动型与需求拉动型通货膨胀

价格为什么上涨，与它涨得有多快同等要紧，因为压力的来源决定了政策能否对它有所作为。把两条渠道分开来看是有益的。

定义 3.5: 成本推动型与需求拉动型通货膨胀

成本推动型通货膨胀（cost-push inflation）源自*供给侧*：投入品成本——能源、食品、原材料——的上升，即便在需求不变时也把价格往上推。*需求拉动型通货膨胀*（demand-pull inflation）源自*需求侧*：更旺盛的支出，把价格拉着，去对抗只能缓慢调整的供给。

这两者在数据里有各自的特征印记。食品和能源价格在很大程度上由供给状况驱动——收成、天气、石油市场——所以能源价格的急剧上升正是成本推动型通胀的教科书案例。相比之下，制造业产出是在相对稳定的供给条件下生产的，所以制成品价格的变动通常被归因于需求的变化，这是需求拉动型通胀的教科书案例。这一区分不只是描述性的：我们将会看到，它告诉我们货币政策有望应对的是哪一种通胀。

3.3.3　核心 CPI 与货币政策的逻辑

定义 3.6: 核心 CPI

核心 CPI（core CPI）是把*食品和能源*成分从篮子里剔除后重新计算的消费者价格指数。通过剔除那些最易波动、由供给驱动的价格，它比整体 CPI 更平滑地追踪价格的底层趋势。

核心 CPI 的动机，恰恰就在成本推动／需求拉动之分。食品和能源是最易受供给冲击影响的价格——即成本推动那一部分——它们也是波动最大的。把它们剔除后，留下的指数波动更小，而且就其构造而言，主要反映的是需求侧的压力，而非转瞬即逝的供给冲击。

这正是一家中央银行所需要看到的，而其缘由是结构性的。

中央银行为何盯住核心通胀

货币政策通过总*需求*起作用：宽松或收紧，改变的是经济体想要花掉多少。它不作用于供给侧——它没法让石油变便宜、也没法让收成变大。因此，中央银行应当对它实际能够影响的那部分通胀——也就是需求拉动型通胀——做出反应。由于核心 CPI 把由供给驱动的、成本推动的那一部分剔除掉了，作为货币政策的向导，它比整体 CPI 更可靠。

这条原则的实际威力，在它被无视时看得最清楚；接下来两则注记记录了几个通胀来源被误诊的事例。

注（中国 2008：朝着供给冲击去收紧）.

2008 年初一个异常寒冷的冬天，让中国南方大量生猪死亡；猪肉价格飙升，单月 CPI 读数一度达到 8% 上下。在那年春天的全国人民代表大会上，领导层提出要收紧基础货币的供给。但这场通胀压倒性地是成本推动型的——是一次砸向猪肉的供给侧冲击，底层需求几乎没变——所以收紧货币是用错了工具，收效甚微。更糟的是，美国金融危机当时正开始波及中国，而出口对中国经济又如此重要，审慎之举本应是

> 准备宽松。2008 年下半年增长明显放缓，下行压力加大，秋天的党代会便顺势提出要放松。短短一年之内立场来了个彻底反转，暴露出当时的政策框架是多么缺乏前瞻性。

注（中国 2018 与猪肉难题）.

> 一句广为流传的俏皮话道出了 2018 年猪肉价格的飙升：“加上猪肉，全是通胀；去掉猪肉，全是通缩。”货币政策碰不到供给侧，可面对一个很高的总体通胀数字——以及它所引发的公众与市场反应——中央银行可能感到不得不收紧。危险在于，朝着供给驱动的通胀去收紧，会制造出滞胀：通胀既是成本推动型的，便治不好，而经济却被进一步往下压。

注（美国的供给侧通胀）.

> 同一时期美国的一段经历讲出了同样的道理。那一时期的通胀是由能源短缺、劳动力供应紧张和物流系统不畅所驱动的——全都是供给侧的问题。比方说，劳动力供应之所以紧张，部分缘于优厚的福利制度压低了劳动参与率，而非需求过热。需求侧的货币政策同一个供给侧的问题严重不匹配，与其说能解决通胀，不如说有加剧通胀之虞。

3.3.4　PCE 指数

定义 3.7: 个人消费支出（PCE）指数

个人消费支出价格指数是美国联邦储备体系偏好的消费通胀度量。它的构造与 CPI 大体相同、走势也与之大体一致，但由美联储单独计算，而美联储在其政策权衡中对通胀给予极高的权重。

美联储宁愿自行维护一套消费价格度量，而不直接采用公布的 CPI，这件事本身就表明：一家中央银行在价格指数的选择上有多么认真。PCE 与 CPI 彼此走势贴得很近，但在覆盖范围和权重上有所不同；二者之间的差距在正常时期很小，却真实存在，而一家盯住通胀的中央银行，会在意自己手里拿的是哪一把尺子。

3.3.5　生产者价格指数

定义 3.8: 生产者价格指数（PPI）

生产者价格指数度量的是出厂环节的价格——生产者就其产出所收到的价格，而非消费者所支付的零售价格。出厂价格大致是成本加利润加税金，它也是商人为了转售而进货时所支付的批发价格。

CPI 和 GDP 平减指数看的都是零售价格，即买家在最终销售环节面对的价格；PPI 则往上游看一步，看的是生产者收取的价格。这使它成为一个不同而又互补的工

具。PPI 不含农产品，与需求紧密相连、又与企业利润直接挂钩，因而对经济周期相当敏感，这让它对企业的投资决策颇具信息含量。

它的关键分界线是零。当 PPI 为正时，产出价格在上涨、生产者的利润空间在扩大；当它转负时，生产者利润被挤压，生产与投资的意愿随之收缩。这甚至会*自我强化*：下跌的生产者价格压低利润和投资，进而削弱经济活动、把价格进一步推低，在短期内把经济拖入一个恶性循环——或者，把符号反过来，拖入一个良性循环。

注（一次通缩式的读法）.

> 在一个具有代表性的月份——2022 年 10 月——中国 CPI 略高于 1%，而 PPI 已经转负。消费价格通胀疲软、生产者价格则干脆下跌，消费需求和投资需求都很弱，经济正滑向通货紧缩。把这两个指数放在一起读，对这场放缓的诊断，比单看其中任何一个都要清晰。

3.4　CPI 与 GDP 平减指数之比较

现在我们手上有了两个涵盖面广的价格指数——CPI 与 GDP 平减指数——值得把它们仔细比较一番，因为二者在三个要紧的方面有所不同，也因为这一比较揭示出 CPI 身上一个系统性的偏误。

固定权重与替代偏误。　CPI 建立在一篮*固定*的商品之上：它问的是，年复一年地买相同的数量，要花多少钱，不管相对价格怎么变。GDP 平减指数则相反，它以*当年*的数量为权重来给商品加权，而这些数量会随相对价格的变化而调整。这个差异是有方向的。当某种商品变得相对更贵时，家庭会从它身上替代开来、转向更便宜的替代品，所以维持给定生活水准的真实成本，其上升幅度小于那一篮旧的、固定的商品的成本上升幅度。由于把数量固定住了，CPI 无视了这种替代，因而*高估*了生活成本的上升——它患有*替代偏误*（substitution bias）。而平减指数通过当年权重把替代嵌了进去，便不存在此问题。出于同样的效果，CPI 对新产品的问世、以及既有产品质量的提升都反应迟缓，而这两者都在不抬高那一篮固定商品成本的情况下提升了福利。

进口品。　两个指数在来源范围上也不同。GDP 平减指数只覆盖*国内生产*的商品——GDP 按定义就把进口排除在外。CPI 覆盖的则是消费者实际购买的东西，其中包含进口品。比方说，进口石油价格的跳涨会直接抬高 CPI，但它进入 GDP 平减指数，仅限于它影响到国产产出价格的那部分。

商品的国内范围。　最后，两个指数覆盖经济中不同的切片。CPI 只追踪消费品。GDP 平减指数追踪 GDP 中的一切——不仅是消费，还有投资、政府购买和净出口。消费价格是平减指数的一个重要构成，但也只是其中一个构成；平减指数的覆盖面要宽广得多。两个指数回答的是不同的问题，在任何给定的年份里，它们都可能彼此分道而行。

两个指数怎么读

CPI 是一个固定篮子的生活成本指数，每月公布，包含进口品，并倾向于通过替代偏误、以及对新产品与改良产品的迟缓纳入而高估通胀。GDP 平减指数是一个以当年为权重、只含国内、涵盖整个 GDP 的价格指数，它把替代纳入了进来，但公布得没那么勤、还有滞后。各自适用于不同的问题：看生活成本、要及时读数时用 CPI，要看实际 GDP 背后那个全经济范围的价格水平时用平减指数。

本章为我们配齐了尺子——名义量对实际量、平减指数，以及消费者价格指数和生产者价格指数——并给了我们一种实用的判断力，知道在何时哪一把才是对的那把。这些工具度量的，是整体价格水平及其变化率，也就是通货膨胀。至于那个水平为什么会动、通胀对借款人和放款人各自做了什么、政府靠印钞攫取的那笔收入，以及把货币同价格联系起来的货币数量论，则是第 九 章的主题；不过我们现在要先转向核算账户只是隐约提到的另一个大总量——劳动力市场与失业。

第四章　劳动力市场与失业

在宏观经济学家关注的所有数字里，失业率大概是离研讨室最远的一个。GDP 增长率多一个百分点是一种抽象，而一个找不到工作的人却是实实在在的。在许多民主国家，失业率被读作对政府的一纸判决，从利率决策到延长失业救济，一长串政策都或明或暗地与它挂钩。本章要搭建的，正是诚实地读懂这个数字所需要的一整套工具：劳动年龄人口如何被分门别类，谁算失业、谁不算；为什么离开劳动参与率单看失业率几乎毫无意义；以及为什么一个健康的经济体应当存在正的失业，而不是把失业降到零。

思路是直截了当的。失业率是一个比率，我们必须先说清楚分子和分母里到底装着什么，所以第一步是对人口做划分。随后我们定义两个核心比率——失业率与劳动参与率——并说明为什么二者必须放在一起读。会计口径理清之后，我们按成因把失业分成三类：摩擦性失业、结构性失业与周期性失业。这一分类引出本章的核心概念——自然失业率（natural rate of unemployment），它进而界定了经济学家所说的“充分就业”以及经济体的潜在产出（potential output）。全程中我们都会留意那些制度与数据质量上的问题，正是它们让这些数字在实践中远比理论上更难解读，在中国的语境下尤其如此。

4.1　人口的划分

要把“失业”变成一个可度量的量，我们必须先确定谁才有资格被算作失业者。官方统计分两步建立这一划分，依次问两个问题：这个人能不能工作？如果能，那他愿不愿意工作？

第一层划分是看**有无劳动能力**。人口中有一部分人之所以在劳动力市场之外，不是出于选择，而是出于处境或规则：儿童、被收容者、全日制学生、现役军人、在押人员。无论他们自己愿不愿意，这些群体在第一层就被排除在外，因为他们并不能自由地去市场上谋一份工作。

在有劳动能力的人当中，第二层划分是看**有无劳动意愿**。一个既有能力又愿意工作的人，被算入*劳动力*；一个有能力工作、却不愿意工作的人——选择不去领薪工作的家庭主妇、靠积蓄提早退休的人——则被算作*非劳动力*。因此，劳动力是两个条件的交集：既有能力，又有意愿。

定义 4.1: 劳动力及其构成

在劳动年龄、未被收容的人口中：

- *劳动力*是所有既*有能力*工作、又*有意愿*工作的人。它进一步分为*就业者*与*失业者*。
- *失业者*是那些有能力、也有意愿工作——正在积极寻找工作——但当下没有工作的人。
- *非劳动力*指有能力工作、但在现行条件下不愿意工作的人。这一群体包括家庭主妇等创造了价值却没有产生可计量市场产出的人，也包括*灰心丧气的劳动者*（discouraged workers）。

这一划分有两个特征值得强调，因为官方口径恰恰会在这两处误导我们。

第一，失业者并不等于“没有工作的人”。退休老人和幼儿同样没有工作，可两者都不算失业：退休老人是不愿意，幼儿是没能力。要被算作失业，一个人必须同时越过两道门槛——有能力且有意愿——并且仍然没有工作。失业是想工作的意愿与得到工作的机会之间的落差。

第二，灰心丧气的劳动者卡在边界上，恰好暴露了这套口径的一处微妙弱点。所谓灰心丧气的劳动者，是指那些原本愿意接受一份工作、却因确信无工可寻而放弃了主动寻找的人。按照调查实际采用的意愿标准——它以“是否在积极找工作”为关键——灰心丧气的劳动者被划归非劳动力，而不是失业者。于是他们同时从失业率的分子和分母里消失了。当经济下行把大量求职者推入灰心丧气的状态时，所度量的失业率反而可能下降，尽管劳动力市场实际上恶化了——原因正是这些人被重新归类到了劳动力之外。这也是为什么失业率不能孤立地读，我们下文还会回到这一点。

注（计量失业的两种口径，以及中国为何切换）.

中国长期使用城镇*登记*失业率，它只统计那些走进当地劳动部门、为领取救济而登记为失业的人。由于登记本身既不完整、又受激励驱动，这一口径对真实的劳动力市场松弛程度反映得很差。中国后来开始按国际通行做法、依据住户调查逐月公布城镇*调查*失业率，从而同时提升了数据的可靠性与国际可比性。一个更宽泛的提醒值得直白说出：中国的劳动力市场数据历来弱于相应的产出和价格数据，官方规划中的就业目标往往定得偏稳健（即偏低），以确保能够可靠地完成。在就业数据内部，两个群体——工人与应届大学毕业生——在政治上最为敏感，也最受密切关注。

注（失业与政治周期）.

不同经济体赋予失业的分量差别很大。在许多西方民主国家，失业率是核心的选举与政策变量，大量宏观经济政策都与它挂钩。中国的政策框架对失业数字本身的反应要间接得多，更多依靠其他大的实体经济指标；但这并不意味着就业在那里不重要，因为劳动力市场状况与其他每一个宏观经济问题都紧密交织。差别在于决策者盯着哪一个刻度盘，而不在于就业重不重要。

4.2　两个核心比率

划分一旦固定，用来概括它的两个比率就只是简单的比值了。第一个度量劳动力内部的松弛程度，第二个度量劳动力相对于原则上能供给劳动的人口有多大。

定义 4.2: 失业率

*失业率*是劳动力中处于失业状态的比例，

$$u = \frac{\text{失业人口}}{\text{劳动力人口}} = \frac{\text{失业人口}}{\text{失业人口} + \text{就业人口}}.$$

它的分母是劳动力，*而非*全部人口：不愿意或没能力工作的人在分子和分母里都被排除在外。

定义 4.3: 劳动参与率

劳动参与率（labor-force participation rate）是有劳动能力的人口中实际进入劳动力的比例，

$$\text{LFPR} = \frac{\text{劳动力人口}}{\text{劳动年龄人口}}.$$

它度量的是，在有能力参与劳动的人当中，有多少比例*愿意*去劳动。

这两个比率回答的是不同的问题，而且各自都对另一个所看见的东西视而不见。失业率只往劳动力内部看，它对“有多少有能力的人选择待在劳动力之外”只字不提；劳动参与率只数有多少人进了劳动力，它对“进来的人过得怎样”只字不提。任何一个单独报出来，都可以被读成截然相反的意思。

看失业率一定要参考劳动参与率

失业率下降只有在劳动参与率稳住的前提下才是好消息。如果人们离开劳动力——提前退休、重返校园，或干脆放弃寻找而成为灰心丧气的劳动者——那么 u 的分子和分母会同时缩小，于是即便劳动力市场正在恶化，失业率也可能下降。*失业率低并不意味着劳动力供给充足。*

最清楚的例证是一个福利和转移支付都很优厚的经济体。当不工作也过得舒服时，许多有能力的劳动者会选择不参与；劳动参与率下降，而由于灰心丧气者和自愿赋闲者都不在劳动力之内，所度量的失业率也随之下降。然而结果不是劳动力市场的繁荣，而是它的萎缩——美国就曾呈现出恰好这样的格局：劳动参与率下行，失业率却很低，同时又抱怨劳工不足。失业率看上去很强劲，劳动参与率却揭示出，可供使用的劳动力池子已经变薄了。

4.2.1　非农就业

基于调查的比率并非唯一会撼动市场的劳动力市场数据。在美国，最受密切关注的单项发布是*非农就业*（nonfarm payrolls，NFP），即农业部门之外有薪工作岗位数量的

月度变化。

定义 4.4: 非农就业（NFP）

非农就业统计的是经济中除农业部门以外的受薪雇员数量。其头条数字是环比月度变化，被解读为净新增（或净流失）的就业岗位。

农业被刻意剔除。农业就业具有强烈的季节性和波动性，随播种与收获的日历起伏，这些起伏对劳动力市场的根本健康状况几乎说明不了什么。把它去掉之后，留下的是对“经济到底创造或损失了多少岗位”更干净的读数。由于非农就业是对岗位创造既及时又直接的度量，一次大幅偏离预期的发布，可能给金融市场带来巨大的冲击。

4.3　为什么劳动力市场对“人”可能无法出清

我们现在从度量转向机制：**为什么**会存在失业？在一个理想化的竞争性市场里，过剩会把价格压低，直到供给量等于需求量，没有任何东西卖不出去。失业的持续存在告诉我们，劳动力市场并不会如此干净地出清。作为一条组织原则：**劳动力市场上出现失业，说明市场上存在摩擦**（*friction*）。三种失业，就是三种不同类型的摩擦。

4.3.1　摩擦性失业

第一种摩擦很简单，就是把一个劳动者匹配到一份工作需要时间。**摩擦性失业**（frictional unemployment）是人们在寻找并被分配到岗位的过程中出现的失业，即便在一个其他方面都健康的经济体里也存在。

定义 4.5: 摩擦性失业

摩擦性失业是因匹配劳动者与工作岗位需要时间而产生的失业。劳动者出于与自身能力无关的原因失去或离开工作，而要找到合适的新匹配——合适的企业、职业或地点——需要时间去寻找和谈判。

有些摩擦性失业是个人层面的：处于两份工作之间的劳动者正经历一段匹配期，不断比较各种录用机会，直到合意的那个出现。有些则是行业层面的：当某个行业重组、技术变迁，或经济活动从一个地区迁往另一个地区时，劳动者必须跨企业、跨职业或跨地点流动，而这种再配置无法瞬间完成。在这个意义上，摩擦性失业是一个不断创造与毁灭岗位的动态经济不可避免的副产品。它不是需求疲软的症状——劳动者失去工作的原因与他们是否胜任毫无关系——而是**寻找与匹配需要时间**这一无法消除的事实的症状。

4.3.2　结构性失业

劳动力市场是一个特殊的市场，因为它交易的商品依附于一个人，而人并不会像一蒲式耳小麦那样被压价到使市场出清的工资水平。**结构性失业**（structural unemployment）是因为工资被维持在高于市场出清水平之上而持续存在的失业：此时劳动供给量超过需求量，一部分愿意工作的人找不到工作。它有三个经典来源。

定义 4.6: 结构性失业

*结构性失业*是因工资被维持在其市场出清水平之上、从而造成劳动长期供过于求而持续存在的失业。它的三个标准来源是工会、最低工资立法与效率工资。

工会与集体谈判。　工会代表其成员进行集体谈判，谈判桌上的头号议题就是涨工资。从一个本来会在均衡工资处出清的市场出发，一个被谈判（或被行政力量强加）抬到该水平之上的工资，提高了那些保住饭碗者的报酬，却减少了企业愿意提供的岗位数量，从而让一部分人陷入失业。实际上，工会保护的是它的*内部人*——那些在更高工资下仍然就业的成员——代价则由那些被高工资挤出工作的外部人承担。

最低工资。　立法规定的最低工资是同一机制的政府版本。如果它被设定在低技能劳动的市场出清工资之上，就会在提高保住饭碗者收入的同时，把生产率最低的劳动者挤出就业——一项出发点良好、却可能办坏事的政策。这正是民粹主义的陷阱：一项听上去纯粹是体恤劳动者的措施，由于忽视了企业会如何反应，反而可能让最脆弱的劳动者处境更糟。它提出了一个真正棘手的问题——多高的工资下限或福利支持才算合理？——对此，市场力量给出一个答案，政治压力给出另一个答案。

效率工资。　第三个来源更为微妙，因为在这里是企业*主动*抬高工资。按照*效率工资*（efficiency wage）理论，一家对工人努力程度掌握信息不完备的企业，可能会发现支付高于市场出清水平的工资反而有利可图，因为这能提高生产率：工资溢价降低了人员流动、吸引了更优秀的应聘者，并让现有工人有了“怕失去”的东西，于是他们更卖力、更少偷懒。企业以更高的工资支出换取每个工人更高的努力。但同样这份买来努力的溢价，也意味着工资位于出清水平之上，因此从总量上看，被雇佣的工人少于竞争性工资下本应雇佣的数量——而这部分过剩，便以失业的形式显现出来。

注（结构性失业关乎“人”，而非商品）.

> 三个来源——工会、最低工资与效率工资——共有一个根源：劳动力市场交易的是活生生的人，而不是匿名的货物。工资向下具有黏性，合同与法律对其构成约束，公平感与士气会影响努力程度，而关于谁在卖力工作的信息又是不完备的。这些“人”的特征，恰恰是一个无摩擦的商品市场所缺少的，也正是劳动力市场不会简单出清的原因。

4.3.3　周期性失业

前两类失业是经济正常运转的固有属性，第三类则是经济周期的属性。*周期性失业*（cyclical unemployment）是随宏观经济总体状况起落而升降的失业：衰退收缩了对劳动的需求、把工人推出岗位，繁荣则反其道而行之。

定义 4.7: 周期性失业

*周期性失业*是由经济周期中总需求波动所引起的那部分失业。它在衰退中为正——此时对劳动的需求下降；在经济过热的繁荣中则可能为负。

在繁荣时期，周期性失业实际上可以转为负值：企业招不到足够的工人，雇员可能被迫加班，工作量超过他们自愿选择的水平。不过这种过热通常被认为是暂时的，而非经济的一种持久状态。

4.4　自然失业率、充分就业与潜在产出

现在我们可以把三类失业组装成本章的核心组织思想了。摩擦性失业与结构性失业是*长期*特征：它们反映了劳动力市场永久性的构造——寻找摩擦、制度、信息——即便在一个完全没有需求冲击的经济体里也会持续存在。相比之下，周期性失业是*短期*成分，是随繁荣与衰退来去的那一部分。自然失业率把长期的那一部分单独剥离出来。

定义 4.8: 自然失业率

*自然失业率*是摩擦性失业与结构性失业之和：

$$u^* = u_{\text{摩擦}} + u_{\text{结构}}.$$

它是周期性失业为零时所通行的失业率。由于它由长期的、结构性的因素所驱动，自然失业率无法被短期宏观经济政策所撼动。

自然失业率就是把经济周期“关掉”之后剩下来的东西。一个有用的看法是：当*实际* GDP 等于*潜在* GDP 时，外生需求冲击基本为零，周期性失业为零，于是所观测到的失业率恰好就是自然失业率——摩擦性失业与结构性失业之和。既然自然失业率由长期力量所决定，短期稳定政策就无法把它降下来；短期政策真正应当瞄准的，只有*周期性*那一部分。

这就使经济学家所说的“充分就业”这个容易被误读的词变得更加清晰。

充分就业指周期性失业为零，而非失业率为零

*充分就业*是周期性失业为零、所度量的失业率等于自然失业率 u^* 的状态。它*并不*意味着失业率为零。即便处于充分就业，也仍然存在摩擦性失业（人们仍在两份工作之间流动）和结构性失业（市场的某些部分工资仍高于出清水平）。一个所度量失业率为零的经济体并不健康，那将是一个劳动者完全没有余地在岗位之间流动的经济体。

充分就业在产出上的对应物是*潜在 GDP*。

定义 4.9: 潜在产出与潜在增长率

潜在 GDP 是经济体在充分就业时所生产的产出水平——也就是当实际产出剥离了短期冲击、失业处于其自然率时的产出。**潜在增长率**是去除短期波动之后经济体的长期增长率，由资本积累和技术进步所驱动。

注（潜在产出与其说是测量，不如说是一种信念）.

潜在产出与潜在增长率都是理论建构：我们从未观测到一个被干净地剥去了冲击的经济体。在实践中，对潜在增长率的估计更接近于一种**信念**，通常是通过总结过去的经验、再向前外推而形成的。当一个政府宣布——比如说——要在某一年之前把经济总量翻一番（折算下来约为每年平均百分之五的增速）时，它表达的是关于潜在增长率的一种看法，而不是在报告一个测量到的事实。自然失业率也具有同样的地位：它真实而重要，但它是被推断出来的，而不是从某个刻度盘上读出来的。

把时间维度归纳一下：摩擦性失业与结构性失业是长期成分，由劳动力市场的结构所固定，短期政策无从触碰；周期性失业是短期成分，是需求管理政策唯一有望撼动的那一部分。充分就业则是那个短期成分已被压到零、只剩自然失业率的那一点。

4.5　人口结构与劳动参与：两段旁论

劳动参与率引出的问题会自然地通向人口学，从中国的经验里有两点观察值得记下来，尽管它们都不能干净利落地纳入上文的失业会计框架。

注（为什么中国女性的劳动参与率如此之高）.

尽管有着女性相夫教子的悠久文化传统，中国的女性劳动参与率却位居世界前列。有两股力量朝同一个方向推动。第一是**家庭收入水平**：在单一劳动者无法养活一家人的地方，夫妻双方都必须工作才能勉强维生，参与与否其实算不上一种选择。第二是**技术变迁**：随着生产从体力劳动转向脑力劳动，女性所能从事的工作种类、以及她们从事这些工作的生产率，都与男性趋于接近，旧有的职业分工因而被侵蚀。当体力劳动越来越不重要、脑力劳动越来越重要时，男性体力上的历史溢价便逐渐消退。

同样这种向脑力劳动的转移也重塑着生育，而这里的经济学值得写下来。考虑一个在消费 C 与孩子数量 N 之间做选择的家庭，其效用为 $U(C,N)$，预算约束为

$$C + NP = w\,(1 - tN),$$

其中 w 是工资，P 是抚养一个孩子的物品成本，t 是抚养每个孩子所需的时间。$w\,(1-tN)$ 一项是家庭在扣除了为抚养 N 个孩子而从工作中抽走的时间之后的收入：每个孩子占用父母工作时间的比例为 t，因此每个孩子除了直接成本 P 之外，还附带 wt 的**机会成本**，即放弃的工资。

注（数量与质量的权衡，以及人口转型）.

随着工资 w 上升来读这条预算约束，便勾勒出了人口转型的轨迹。在工业化的早期阶段，w 迅速增长，而孩子的时间成本仍然不高，于是家庭对收入提高的反应是生**更多**的孩子。当 w 进一步上升，每个孩子的时间成本 wt——即父母放弃的工资——开始占据主导，孩子在“随收入而水涨船高”的那一种资源上变得昂贵起来：父母的时间。于是家庭转向少生孩子、在每个孩子的**人力资本**（human capital）上投入更多；这正是经典的**数量与质量的权衡**（quantity–quality tradeoff）。随之而来的生育率下降便是人口转型，它意味着人口老龄化不是政策的意外，而是一个建立在脑力劳动之上的社会在发展过程中基本无法回避的一个阶段。

第五章　索洛增长模型

为什么今天一个普通美国人的富裕程度，大约是两百年前普通人的五十倍？为什么今天有些国家的富裕程度是另一些国家的二十倍？这些关于**长期**的问题——关于生活水平在数十年乃至数代人之间的缓慢漂移——正是增长理论的研究领域，而它们的意义极为重大：正是人均产出的持续增长，提升了消费、延长了预期寿命，并通过把经济总量这块蛋糕做大，使得即便在不平等程度很高的地方，再分配与社会稳定也成为可能。我们后面要研究的经济周期，也就是短期波动，最好被理解为围绕这条上升趋势线的偏离；用一句有用的口号来说，周期 = 增长 + 冲击。要理解周期，我们首先需要一套关于趋势的理论。

本章发展第一个、也是最简单的这样一套理论：**索洛增长模型**（Solow growth model）。在引入它之前，我们先停下来搭好所有宏观动态模型共享的一般框架：一条时间线、一个概括经济当前所处状态的状态变量、经济所选择的控制变量，以及一条把状态推向未来的运动方程。索洛模型是这个框架最精简的一个实例。它标志性的简化处理——也是把它与第 六 章新古典模型区分开来的那唯一一条假设——在于储蓄率是外生固定的，而非由优化的家庭选择得到的。我们将搭起基准模型、求出它的稳态、追问哪个稳态是最优的（黄金律）、刻画政策改变储蓄率后经济如何调整，最后把模型扩展到容纳人口增长与技术进步。这个模型不能解释一切，但它能解释相当多的东西，并且它把增长理论其余部分所要回答的问题组织了起来。

5.1　宏观经济学的动态框架

微观经济学在很大程度上是静态的：消费者选择一个消费束，厂商选择一个产量，我们再比较均衡。而宏观增长在本质上是**动态**的。其核心对象不是单个选择，而是一串通过时间相连的选择，把它们连起来的装置就是**资本**的积累。因此，每一个动态宏观模型都从铺设一条时间线开始。

解决宏观问题一定要先有一条时间线

要分析任何宏观增长问题，先写下一条以时期 $t = 1, 2, \ldots$ 为标号的时间线。每一期，经济体继承一笔资本存量，做出生产出产出、并决定带入下一期多少资本的种种选择，如此循环往复。

资本存量扮演着特殊的角色。在第 t 期之初，存量 K_t 是固定的——它就是过去的投资所遗留下来的那个数——所以从第 t 期的视角看，它是**给定**的，是今天的经济体无

法改变的一个数。这种变量被称为**状态变量**（state variable）：它概括了经济体为了决定眼下该怎么做、所需要知道的关于自身过去的一切。面对给定的状态，经济体选择如何把当期产出在消费与储蓄之间分配。这些选择就是**控制变量**（control variable）：消费 C_t 与储蓄 S_t 可以自由设定，而一旦设定，下一期的资本存量就被确定了。决定下一期存量的那条规则，就是**资本的运动方程**（law of motion of capital），它刻画了存量如何随时间积累。

定义 5.1: 状态变量与控制变量

在一个动态模型中，**状态变量**是一个在每期之初已被预先决定、并概括了经济体相关历史的量——这里就是资本存量 K_t。**控制变量**是经济体在当期内自由选择的量——这里就是消费 C_t 与储蓄 S_t。状态约束着控制，控制又决定下一期的状态。

为把这些想法落到实处，考虑一个去掉了家庭与厂商、只展示支配任何经济体的会计恒等式的两期例子。产出由资本生产，$Y_i = F(K_i)$；第一期产出在消费与储蓄之间分配，$C_1 + S_1 = Y_1$；储蓄被投资出去，$I_1 = S_1$；资本以 δ 的速率折旧、又由投资补充，故 $K_2 = (1-\delta)K_1 + I_1$；第二期家庭消费其产出再加上储蓄的总回报，$C_2 = Y_2 + (1+r)S_1$。把这些汇总起来，

$$\begin{aligned} C_1 + S_1 &= Y_1, \\ C_2 &= Y_2 + (1+r)S_1, \\ K_2 &= (1-\delta)K_1 + I_1, \qquad I_1 = S_1, \\ Y_i &= F(K_i). \end{aligned}$$

这些就是经济体机械运行所遵循的规律；它们不涉及任何优化。这恰恰是索洛模型工作的层次——它给产出如何分配施加一条固定规则，然后任凭会计恒等式把经济体推向未来。而第 六 章的新古典模型则会加上微观基础，从一个最大化终身效用的家庭、以及在一般均衡中出清的厂商与市场出发，推导出消费-储蓄的分配。具体而言，优化的家庭求解

$$\max \sum_{t=1}^{\infty} \beta^{t-1} U(C_t),$$

约束条件就是同样那几条运动方程。但无论我们是否加上优化，结构都是一样的：每一期 K_t 是状态，C_t 与 S_t 是控制。

这一结构有一个深刻的后果。站在第 t 期、手握状态 K_t，每选定一个 C_t 就钉死了一个 S_t，进而钉死了 K_{t+1}。于是从今天的状态到明天的状态的映射，可以写成单独一个函数，

$$K_{t+1} = g(K_t),$$

称为**政策函数**（policy function）。一期接一期地顺着政策函数走，就描画出整个经济体的最优路径。这个映射有两个值得点名的特征。第一，下一期的状态只依赖于这一期的状态，而不依赖经济体是经由怎样的整段历史走到这里的；这个问题具有**马尔可夫**（Markov）性质。第二，每个日期所面对的决策问题结构完全相同——不同的只是所继承的状态取值——所以这个问题是**递归的**（recursive）。这条乍看令人生畏的无穷期选

择序列，于是坍缩成同一个问题、以不同起点被反复求解。正是这种递归性使得动态宏观可处理，它支撑起本书中的每一个增长模型。

5.2　索洛模型：基本假设

现在我们把这个框架专门化到索洛的经济中。这个模型作出六条贯穿始终的假设，我们把它们集中放在一处。

假设 5.2: 索洛模型

1. **总量生产函数**。产出由资本与劳动生产，$Y = F(K, L)$。劳动力 L 保持不变；唯一会变动的量是资本 K。

2. **规模报酬不变（CRS）**。F 是一次齐次的：把两种投入都乘以任一倍数 λ，产出也乘以 λ。

3. **边际报酬递减与稻田条件**。每种投入都是有生产力的，但报酬递减，

$$F_K > 0, \quad F_L > 0, \qquad F_{KK} < 0, \quad F_{LL} < 0,$$

且当某种投入趋于消失时其边际产出变得无界，

$$\lim_{K \to 0} F_K(K, L) = \infty, \qquad \lim_{L \to 0} F_L(K, L) = \infty.$$

4. **无政府的封闭经济**。净出口与政府支出均为零，$\mathrm{NX} = G = 0$。

5. **储蓄率外生且不变**。产出中一个固定份额 $s \in (0, 1)$ 被储蓄起来，$I = sY$，其中 s 由模型之外给定。

6. **资本积累**。资本以 δ 的速率折旧、又由投资补充，

$$K' = (1 - \delta)K + I,$$

其中撇号表示下一期。

在动用它们之前，其中有两条假设值得评说。

规模报酬不变（假设 2）正是让我们能够在总量经济与一个代表性的“劳均”描述之间自由穿梭的东西。在 CRS 下，一家厂商被拆分或合并都不改变其生产率，所以我们既可以把整个社会看作一家雇用 L 个工人与 K 台机器的巨型厂商，也可以等价地看作单独一个代表性工人操作着 $k := K/L$ 单位资本。在齐次性性质中取 $\lambda = 1/L$，

$$\frac{Y}{L} = F\left(\frac{K}{L}, 1\right) =: f(k),$$

这就定义了**集约型**（劳均）生产函数 $y = f(k)$。稻田条件（假设 3）保证了内点解：因为当资本稀缺时其边际产出极大、当资本充裕时其边际产出极小，经济体总是被从 $k = 0$

与 $k=\infty$ 这两个角点拉开，拉向一个合理的内点稳态。

封闭经济假设（假设 4）让我们能把储蓄与投资划上等号。在 $\mathrm{NX}=G=0$ 下，产出恒等式读作 $Y=C+I$，而收入被分为消费与储蓄，$Y=C+S$；两式相减便得到封闭经济的基本条件

$$S=I.$$

化为劳均形式即 $y=c+i$，其中 $i=sy$、$c=(1-s)y$。

索洛之所以为“索洛”

这个模型的简单之处——也是把它与新古典模型区分开来的那一条假设——在于**储蓄率 s 是外生且不变的**。家庭并不通过权衡当前效用与未来效用来选择储蓄多少；收入中一个固定份额被机械地储蓄起来。第 六 章的新古典模型则用一个最优的、最大化效用的消费-储蓄选择来取代这一点。正是在这个确切的意义上，索洛描述的是一个储蓄规则被硬性写死的计划经济。

5.3　基准模型：无人口增长

我们从最简单的情形入手，即劳动力恒定。由于 L 不变，劳均的运动方程与总量形式具有相同的样子。把 $K'=(1-\delta)K+I$ 除以常数 L，

$$k'=(1-\delta)k+i.$$

于是劳均资本存量从一期到下一期的变化为

$$\Delta k=k'-k=i-\delta k=s\,y-\delta k=s\,f(k)-\delta k. \tag{5.1}$$

式 (5.1) 是索洛模型的心脏。它是劳均资本的运动方程，读懂它就是读懂整个模型。当劳均总投资 $sf(k)$ 超过折旧损耗的资本 δk 时，劳均资本上升；不足时则下降。其动态完全是跨期的：今天在消费与投资之间的分配决定了明天的资本存量，进而决定明天的产出，如此往复。这正是增长与微观经济学静态比较的对照之处——在宏观增长里，戏份全在运动方程上。

5.3.1　稳态

当劳均资本不再变化时，经济体便安顿下来。在式 (5.1) 中令 $\Delta k=0$，得到**稳态**条件

$$s\,f(k^*)=\delta k^*. \tag{5.2}$$

定义 5.3: 稳态

索洛经济的一个**稳态**（steady state）是劳均资本的一个水平 k^*，在该水平上总投资恰好补足折旧掉的资本，$sf(k^*)=\delta k^*$，从而 $\Delta k=0$。在稳态处，每一个劳均变量——k^*、$y^*=f(k^*)$、$c^*=(1-s)f(k^*)$、$i^*=sf(k^*)$——都随时间保持不变。

稳态用图来看最为直观。图 5.1 对着劳均资本 k 画出三条曲线：生产函数 $f(k)$，因报酬递减而递增却上凸；投资曲线 $sf(k)$，与前者同形，只是按 s 缩小了；以及折旧线 δk，一条过原点的射线。因为在原点附近 $sf(k)$ 起步时高于 δk（稻田条件使得投资在那里斜率无穷），而最终又被这条直线反超（报酬递减把 f 压平），二者恰好在 k^* 处相交一次。

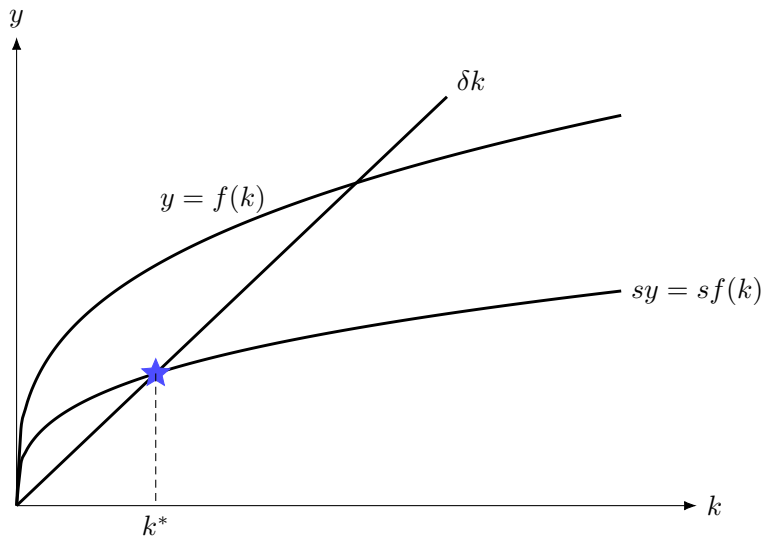

图 5.1: 基准索洛图。经济体收敛到稳态 k^*，在那里投资曲线 $sf(k)$ 与折旧线 δk 相交；在 k^* 左侧投资超过折旧、k 上升，在右侧则下降。

这个稳态是**稳定的**：无论经济体被推离 k^* 多远，它都会回来。在 k^* 左侧，投资 $sf(k)$ 位于折旧 δk 之上，故 $\Delta k>0$，资本朝 k^* 增长；在右侧，折旧占上风，$\Delta k<0$，资本又缩回 k^*。这一收敛背后的经济学，正是把生产函数压弯的那同一个报酬递减：当资本稀缺时，每多一台机器都极有生产力、轻易就赚回了自身的折旧，于是存量不断累积；随着资本积累，边际产出下降，直到一台新机器的产出恰好只够补上整个存量的折旧，增长便就此停止。

单靠资本积累的增长终将停止

在基准索洛模型中，劳均资本的积累最终会停下来。由于资本的边际产出递减、而折旧却线性增长，存在一个有限的 k^* 使二者相抵，经济体便停滞在那里。单靠资本深化，其本身无法维持增长。

这一结论带出一个关于收敛的著名推论。两个技术、储蓄率与折旧率相同的经济

体，共享同一个 k^*，因而无论起点在哪里，都会收敛到相同的劳均产出水平。一个起步时资本很少的后发穷国，最初增长很快——它稀缺的资本赚取高回报——但随着逼近共同的稳态而放慢。早期发展在很大程度上是一个资本积累的故事，但正因为这台发动机会熄火，各国之间任何**持久**的收入差异都必定来自资本以外的某种东西。于是基准模型过于简单，无法解释我们在世界上观察到的那些持续不散的差距；找出它缺了什么，将驱动下文的各项扩展。

5.4　黄金律

更高的储蓄率会带来更高的稳态资本存量 k^*——可是，更富的稳态就是更幸福的稳态吗？这个经济体里的福利是用消费而非资本来衡量的，从这个角度看答案并不显而易见。劳均稳态消费为

$$c^* = (1-s)f(k^*) = f(k^*) - \delta k^*,$$

其中第二个等号用了稳态条件 $sf(k^*) = \delta k^*$，把储蓄换成了投资。这样写出来，c^* 就是生产函数 $f(k^*)$ 与折旧线 δk^* 之间的竖直缺口：所生产的一切之中，没被折旧吃掉的那部分可供消费。

提高 s 对这个缺口有两个相反的影响。一方面，它把经济体推向更大的 k^*，在那里经济体生产得更多、$f(k^*)$ 更高、可分配的东西也更多。另一方面，维持更大的资本存量需要把更多产出仅仅用于补足折旧 δk^*，这会侵蚀消费。在两个极端之间——在什么都不储蓄（$k^* = 0$，没有产出）与储蓄一切（$c^* = 0$，全部产出都重新投资）之间——的某处，消费达到最大。那个最优的稳态就是**黄金律**（Golden Rule）。

定理 5.4: 黄金律条件

在所有稳态中，劳均消费在满足

$$f'(k_g^*) = \delta$$

的资本存量 k_g^* 处达到最大。黄金律就是带来这一稳态的那个储蓄率；把这一最优条件与稳态条件 $sf(k_g^*) = \delta k_g^*$ 联立，

$$s_g = \frac{\delta k_g^*}{f(k_g^*)}.$$

证明. 稳态消费 $c^*(k^*) = f(k^*) - \delta k^*$ 是储蓄率所挑选的那个稳态资本存量的函数。对 k^* 求最大，

$$\frac{\mathrm{d}c^*}{\mathrm{d}k^*} = f'(k^*) - \delta = 0 \quad \Longrightarrow \quad f'(k^*) = \delta.$$

从几何上看这是一个切线问题：消费是 $f(k^*)$ 与 δk^* 之间的缺口，而这个缺口在生产函数斜率等于折旧线斜率 δ 之处最宽。f 的凹性保证了这是一个最大值。把稳态条件 $sf(k_g^*) = \delta k_g^*$ 在 k_g^* 处读出来，就得到 s_g 的表达式。 □

图 5.2 把这一几何关系画得明明白白。稳态消费是 $f(k^*)$ 与 δk^* 之间的竖直距离，而这段距离在 k_g^* 点处最大——在那里生产函数与折旧线平行，也就是 $f'(k_g^*)=\delta$，资本的边际产出等于折旧率。

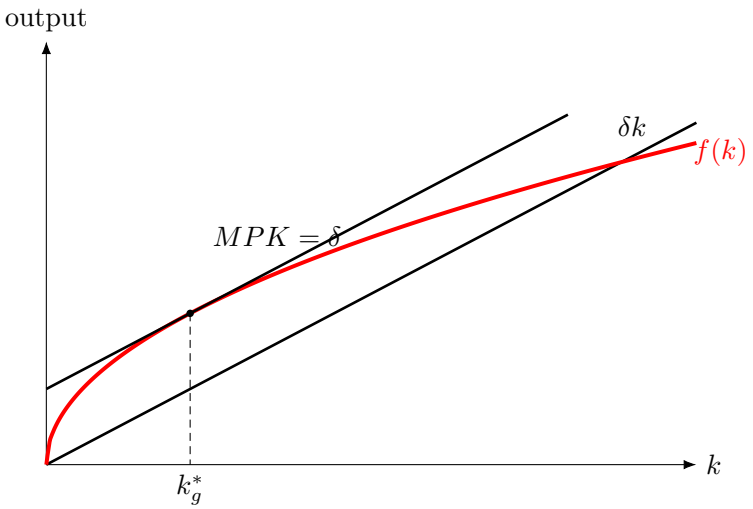

图 5.2: 黄金律。稳态消费是产出 $f(k^*)$ 与折旧 δk^* 之间的缺口；它在 k_g^* 处达到最大，那里资本的边际产出等于折旧率，$f'(k_g^*)=\delta$。

注（[]）.

> “等价”折旧] 黄金律条件 $f'(k_g^*)=\delta$ 中的 δ 应当被读作劳均资本被耗损的**等价**速率。在基准模型里它就是物理折旧 δ。一旦加入人口增长，它就变成 $\delta+n$；再加入劳动增进型技术进步，它就变成 $\delta+n+g$。同一个条件——资本的边际产出等于等价折旧——自始至终成立；改变的只是我们往“等价折旧”里代入什么。

5.4.1　柯布-道格拉斯情形下的黄金律

当生产为柯布-道格拉斯型时，黄金律储蓄率会呈现出一个格外干净的形式。令

$$Y=AK^{\theta}L^{1-\theta},\qquad 0<\theta<1,$$

于是集约型生产函数为 $y=f(k)=Ak^{\theta}$。

例（柯布-道格拉斯下的黄金律储蓄率）.

> 证明在 $Y=AK^{\theta}L^{1-\theta}$ 下，黄金律储蓄率为 $s_g=\theta$。

解.

> 集约型函数为 $f(k)=Ak^{\theta}$，其边际产出为 $f'(k)=\theta Ak^{\theta-1}$。黄金律条件 $f'(k_g^*)=\delta$ 给出
>
> $$\theta A(k_g^*)^{\theta-1}=\delta \quad\Longrightarrow\quad \delta=\theta\,\frac{A(k_g^*)^{\theta}}{k_g^*}=\theta\,\frac{f(k_g^*)}{k_g^*}.$$

代入储蓄率公式，

$$s_g = \frac{\delta k_g^*}{f(k_g^*)} = \frac{\theta\, f(k_g^*)/k_g^* \cdot k_g^*}{f(k_g^*)} = \theta.$$

$s_g = \theta$ 这个结果好记：黄金律储蓄率等于资本在产出中的份额。在竞争性经济里资本获得其边际产出的报酬，所以 θ 恰好是国民收入中归于资本的那个比例。这条规则说的是，把资本所赚取的那份收入精确地储蓄并再投资出去，不多也不少。储蓄超过资本份额就会过度积累——多出来的机器赚不回自身的损耗；储蓄少于资本份额则会让经济体达不到使消费最大化的存量。在这个意义上，黄金律要求在资本上的投资相对于资本对产出的贡献而言恰好划算（cost-effective）。

两项比较静态分析把这幅图补完整。更高的折旧率 δ 会拉低黄金律消费：经济体必须留出更多产出仅仅用来维持其资本，能享用的就更少了。而更高效的技术——每个 k 上都有更高的 $f(k)$——会抬高黄金律消费，因为经济体生产得更高效；不过，与之相伴的黄金律储蓄率究竟是上升还是下降则是含糊的，要视 f 的形状而定。

5.5　储蓄率变化后的动态调整

稳态告诉我们经济体最终落在哪里；它并没有告诉我们经济体如何走到那里。假设一位仁慈的社会规划者判断当前储蓄率过高——经济体已越过黄金律、过度积累了——于是把它调低到 s_g。消费、投资、产出与资本会随时间作何反应？答案完全取决于一个区分：哪些变量是**流量**、哪些是**存量**。

存量缓慢调整，流量瞬间跳跃

流量（按单位时间度量的过程量，如消费 c、投资 i 或产出 y）能在参数一改变的那一刻不连续地跳变。而**存量**（随时间积累起来的量，如资本 k）不能跳变；它只能通过投资这股流量被逐渐建立或耗减。追踪任何宏观调整，归根结底就是把这一区分理清楚，并搞清楚每个参数是怎样起作用的。

在储蓄率降到 s_g 的那一刻，劳均资本 k 还停在它原来那个更高的取值上——存量不能瞬间改变。由于 k 暂时被钉住，产出 $y = f(k)$ 在冲击当下也不变。但那份固定产出的分配方式立刻就变了：投资 $i = s_g y$ 因 s_g 变小而立即**下降**，消费 $c = (1 - s_g)y$ 则相应地立即**跳升**。两者都是流量，所以都在冲击当下移动；又因为产出在那一刻不变，投资的下降量恰好等于消费的上升量。

跳跃过后，缓慢的调整开始了。在更低的储蓄率下，投资如今不足以补足折旧，$s_g f(k) < \delta k$，故 $\Delta k < 0$，劳均资本朝新的、更低的稳态漂落。随着 k 下降，产出 $y = f(k)$ 也随之下降，并紧跟着 k，因为 y 通过生产函数与 k 绑定在一起。随着产出下降，消费与投资——如今都是一份不断缩小的产出的固定比例——也逐渐随之下滑。消费在冲击当下纵身跃起之后，又缓缓回落到它新的稳态水平，但这一水平仍然高于此前（这正是迁向黄金律的初衷）。图 5.3 展示了这四条时间路径：流量 i 与 c 在政策变动当期的陡然不连续跳跃、存量 k 的平滑下滑，以及紧跟着 k 的产出 y。

这条路径还有一个值得点名的特征：向新稳态的收敛是缓慢而减速的。随着 k 接近它新的歇脚点，投资与折旧之间的缺口缩小，故 Δk 缩小，经济体越来越轻地朝稳态

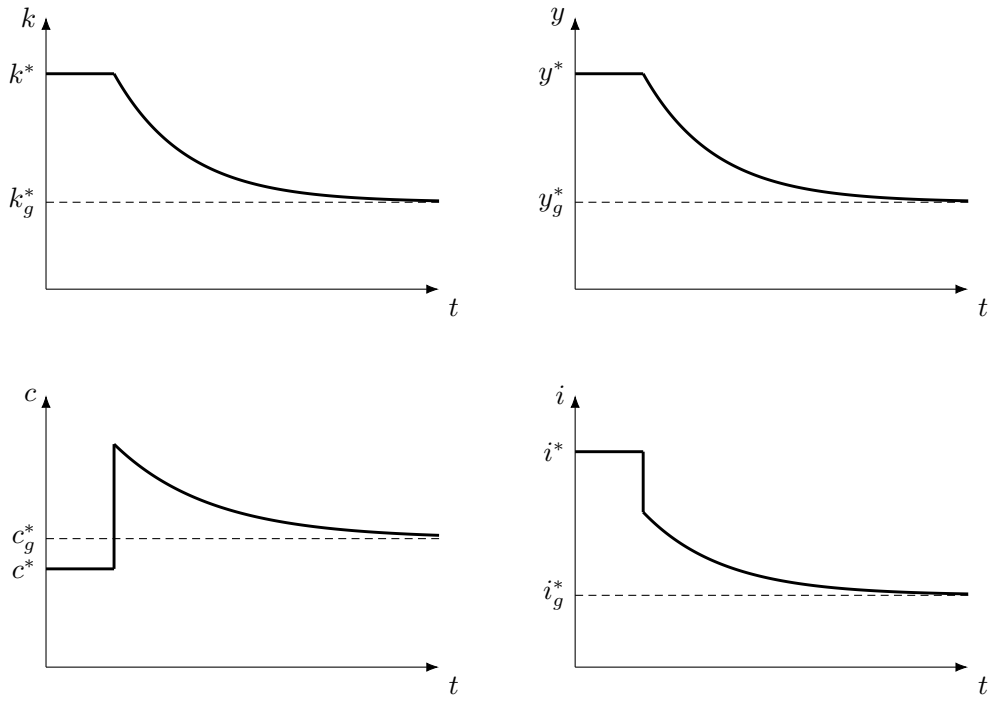

图 5.3: 储蓄率被砍到 s_g 之后，资本 k、产出 y、消费 c 与投资 i 的时间路径。流量 i 与 c 在冲击当下不连续地跳跃（投资下降、消费上升）；随后存量 k 缓慢下降，拖着 y、c、i 一道朝新的稳态走低。

蹭去。抵达新的均衡是许多期才能完成的工作，而非一蹴而就。

5.6　扩展的索洛模型

基准模型让劳动力恒定、并假设没有技术变化。这两点都明显与事实不符，而放松它们正是让模型能够谈论持续增长的关键。指导原则就是动态框架中所强调的那条：新的因素只通过**运动方程**进入；静态关系并不改变。在伸手去拿代数工具之前，先把直觉想清楚。

5.6.1　人口增长

设劳动力以恒定速率 n 增长，

$$\frac{\Delta L}{L} = n.$$

总量的运动方程毫发未动，$K' = (1-\delta)K + I$。改变的是向劳均形式的换算，因为工人数本身如今在上升。记 $k = K/L$，并除以下一期更大的劳动力，

$$k'(1+n) = (1-\delta)k + i.$$

减去 k 并化简，

$$\begin{aligned}\Delta k = k' - k &= i - \delta k - nk' \\ &= i - \delta k - nk + n\left(\frac{K}{L} - \frac{K'}{L'}\right) \\ &\approx i - (\delta + n)k,\end{aligned}$$

其中 $n\,(K/L - K'/L')$ 一项是两个微小变化之积，在一阶近似下可以忽略。于是我们得到人口增长情形下的核心结果，

$$\Delta k \approx sf(k) - (\delta + n)k. \tag{5.3}$$

其解读立竿见影，而且恰是直觉所预言的那样。与基准模型相比，劳均资本被耗损的等价速率从 δ 升到了 $\delta + n$。除了替换掉用坏的机器（δk 一项），经济体如今还必须**为每期加入劳动力的新工人配备资本**（nk 一项），仅仅为了不让劳均资本下降。说到底，人口增长起的作用就像额外的折旧。稳态条件变为 $sf(k^*) = (\delta + n)k^*$，黄金律条件变为 $f'(k_g^*) = \delta + n$。

注（[）.

> 人口增长、马尔萨斯陷阱与 CRS 的局限] 由于资本的回报受边际报酬递减支配，而人口增长又不断稀释着每个工人所拥有的资本，更快的人口增长会拉低稳态劳均资本、从而拉低劳均产出——人人都更穷了。推到极致，这就是**马尔萨斯陷阱**（第 七 章）的逻辑：收入的任何增益都被更庞大的人口耗散掉。这一机制稳稳地立在规模报酬不变的假设之上。现实中人的群集可以产生正的外部性——思想、专业分工、稠密的市场——以至于社会可能呈现**规模报酬递增**（IRS）。况且，生育是一种主动的选择，把人口增长当作一个内生的决策、而非一个外生的参数来处理，才是更令人满意的建模立场，而这一立场正是索洛框架所不采取的。

5.6.2　劳动增进型技术进步

要产生生活水平的持续增长，我们必须加入技术进步。索洛是通过一个刻意特殊化的装置来做这件事的：技术作为劳动的乘数进入生产函数，

$$Y = F(K, E \cdot L),$$

其中 E 度量**劳动效率**（efficiency of labor），并以恒定速率 g 增长。乘积 $E \cdot L$ 就是**有效工人**（effective workers）的供给。我们通常把技术进步想象成不断上升的全要素生产率、改善着所有投入的使用；而把它建模为**劳动增进型**（labor-augmenting）则是一个便于处理的选择，它让经济体能够安顿到一条干净的平衡路径上。（如果技术也增进资本，那就相当于直接改进全要素生产率了。）

运动方程依旧不变，$K' = (1 - \delta)K + I$。现在我们把一切按**有效**工人来度量，定义 $k = K/(EL)$、$i = I/(EL)$。由于有效劳动以 $n + g$ 的速率增长（人口以 n、效率以 g），把运动方程除以下一期的有效劳动，

$$k'(1 + n)(1 + g) = (1 - \delta)k + i.$$

展开并丢掉二阶项 ng，

$$k'(1+n+g) = (1-\delta)k + i,$$

经由与之前相同的步骤整理后得到

$$\Delta k \approx sf(k) - (\delta + n + g)k. \tag{5.4}$$

如今等价折旧率是 $\delta + n + g$：每个有效工人的资本被物理折旧耗损、被人口增长稀释、又被劳动效率的增长再稀释一次。在稳态处令 $\Delta k = 0$ 得到

$$(\delta + n + g)\, k^* = sf(k^*) = i^*.$$

关键的问题在于，这对真实的生活水平意味着什么——而生活水平是按**人均**、而非按有效工人来度量的。在稳态处 $k^* = K/(EL)$ 不变，所以**人均**资本与产出为

$$\frac{K}{L} = k^* E = k^* E_0 (1+g)^t, \qquad \frac{Y}{L} = y^* E = y^* E_0 (1+g)^t.$$

尽管把 g 加进等价折旧会拉低以有效工人为单位度量的稳态 k^*（从而拉低 y^*），但乘性因子 $E_0(1+g)^t$ 无界增长，完全占据主导。于是人均产出与人均资本永远以 g 的速率增长。这是这个模型的核心成就：**劳动增进型技术进步是人均收入持久增长的唯一来源**。与此同时，总产出 Y 等于 EL 乘以它的有效工人取值，故以 $n+g$ 的速率增长。

注（[）.

> 作为公共品的技术] 为什么劳动增进型技术抬高的是所有人的收入，而不是自己索取一份要素报酬？模型中的关键在于，效率 E 是**非竞争性的**（non-rivalrous）：我用一个想法，并不妨碍你也用它。资本与劳动是竞争性要素，在要素市场上交易、赚取回报；而技术在这个意义上并不是一种要素，而是一种公共品，其好处归于全体劳动者。我们可以把 E 想象成一个乘数，让社会用 L 个人就实现本来需要 $E \cdot L$ 个人才能达到的成效；当我们盘点人类福祉时，又把这 E 个有效工人合并回 L 个真实的人，于是增益就以人均收入的不断上升显现出来。现实中专利、保密或其他摩擦使技术部分地变得可排他，结果便会与这个无摩擦的模型有所偏离。

5.6.3　稳态增长率：小结

模型的三个版本之间的差别，仅在于长期究竟什么在增长。集约型变量（按有效工人）在每个稳态里都按构造保持不变；差别在于人均与总量的增长率，汇总于表 5.1。

表 5.1: 索洛模型三个版本在稳态下各变量的长期增长率。

稳态下	基准	人口增长	人口与技术增长
按有效工人	不变	不变	不变
人均	不变	不变	g
总量	不变	n	$n+g$

这张表把上文的讨论一一读出。当劳动与技术都不增长时，一切都是平的。当只有

人口增长时，长期来看人均产出依旧是平的——经济体更大了，但人均上并没有更富——而总产出以 n 增长。只有当劳动增进型技术进步在场时，人均收入才会增长，速率为 g，总量则以 $n+g$ 增长。

5.7　关于索洛模型的更多讨论

索洛框架尽管简单，却支撑得起一组丰富的进一步观察。我们把最有用的几条收集在这里。

5.7.1　经济增长的三个阶段

从经验上看，各国似乎要经历三个大的阶段，每个阶段由一种不同的增长机制支配。

1. **贫困陷阱**（人均 GDP 约低于 \$3,000）。劳均资本极低，所以资本的边际产出——从而投资的回报——极高。处于这一阶段的经济体可以仅靠积累资本就快速增长，这个过程称为**资本深化**（capital deepening）。诸如中国这样的快速追赶式增长，恰恰是基准索洛模型对一个资本稀缺经济体所作的预言，而非什么反常现象。

2. **中等收入转型阶段**。随着资本深化，报酬递减开始显现，靠积累带来的轻松增长逐渐消退。要继续增长，经济体必须从要素积累转向创新驱动的增长。这一转型很难——转型失败就是所谓的**中等收入陷阱**——驾驭它通常需要持续的制度改革。

3. **高收入阶段**。这里增长压倒性地来自技术创新，因为进一步的资本深化已收效甚微。在这个区制里，是 g 这一项、而非储蓄率，在驱动生活水平。

5.7.2　增长核算与索洛余值

为了度量增长究竟来自何处，设产出沿着一条带全要素生产率 z 的柯布-道格拉斯路径走，

$$Y = z\,K^{\alpha}L^{1-\alpha}.$$

取对数再微分，得到**增长核算**（growth accounting）分解，

$$\frac{\mathrm{d}Y}{Y} = \frac{\mathrm{d}z}{z} + \alpha\,\frac{\mathrm{d}K}{K} + (1-\alpha)\,\frac{\mathrm{d}L}{L}. \tag{5.5}$$

产出增长是生产率增长、加上资本增长与劳动增长各自贡献之和，每项都以其收入份额为权重。历史上 $\alpha \approx 1/3$。化为劳均形式，$y = zk^{\alpha}$，分解坍缩为

$$\frac{\mathrm{d}y}{y} = \frac{\mathrm{d}z}{z} + \alpha\,\frac{\mathrm{d}k}{k}.$$

z 一项就是**索洛余值**（Solow residual）：把资本与劳动可度量的贡献扣除之后，产出增长中剩下的那部分。它之所以称为余值，正是因为它无法被直接观测，而是从式 (5.5) 中反推出来的，$\mathrm{d}z/z = \mathrm{d}Y/Y - \alpha\,\mathrm{d}K/K - (1-\alpha)\,\mathrm{d}L/L$。

定义 5.5: 索洛余值

索洛余值 z（常被称为全要素生产率）是产出中不能由资本与劳动的可度量数量所解释的那部分。它捕捉了在给定投入下抬高产出的一切，包括

- 人力资本（凝结在工人身上的教育与技能）；
- 技术进步；
- 外部性（公共品的溢出，与那种可排他、收费的物品相反）；
- 制度质量（厂商的组织与管理知识、专利保护、对腐败的治理）。

请注意它与上一节劳动增进型 E 的对照：E 只改善劳动，而余值 z 乘的是整个生产函数，因此同时抬高资本与劳动的生产率。理解 z 的一个有用方式是，它本身是一种**结果**——是对教育、研究与制度的投资所形成的均衡产物——所以人们可以用分析任何其他均衡对象时所用的同一套供求逻辑，去分析它的水平与增长。促进增长的政策杠杆，直接从余值的内容里读出来：提高储蓄率、投资人力资本、鼓励技术进步、建立正确的制度。

5.7.3　平衡增长路径与收敛

扩展模型的稳态是一条**平衡增长路径**（balanced growth path，BGP）：沿着这条轨迹，产出、资本与消费全都以恒定（且彼此一致）的速率增长。和基准稳态一样，BGP 是一个吸引子——若经济体被撞离它，各变量会随时间收敛回来。

两个生产率度量有助于在路径上给经济体定位。**劳动生产率**与**资本生产率**分别为

$$\frac{Y}{L} = z\left(\frac{K}{L}\right)^{\alpha}, \qquad \frac{Y}{K} = z\left(\frac{K}{L}\right)^{\alpha-1}.$$

注（[）.

> 劳动生产率不是工资] 劳动生产率 Y/L 不应与平均工资相混淆，因为这里的 L 是**在业**工人数，而非整个人口；失业、劳动参与率与人口结构在劳均产出与人均收入之间打入了一道楔子。

索洛逻辑现在交出它最锋利的一课。资本-劳动比 K/L 的增长**不**可持续——报酬递减保证了它会逐渐熄火；而余值 z 的增长**是**可持续的，且它同时抬高劳动生产率与资本生产率。成熟经济体主要靠不断上升的全要素生产率增长；比如今天的美国，其增长大半要归功于余值、而非资本深化。

最后，模型对收敛有一个重要的限定。基准模型那种无条件收敛——人人最终一样富——只在各经济体共享相同的基本面时才成立。由于各国的制度与技术路径不同，经济体收敛到的并不是单独一个共同的稳态，而是收敛到**各自**的平衡增长路径。那些共享相似制度与技术的国家，确实会收敛到一条共同的路径，这一现象被称为**条件收敛**或**俱乐部收敛**（conditional / club convergence）。技术进步的内生性——g 本身取决于政

策、制度与研究这一事实——恰恰是索洛模型留而未解的，而那正是后来的增长理论的起点。

索洛能解释什么、不能解释什么

索洛模型解释了为什么资本贫乏的经济体先快速增长而后放慢、为什么资本积累无法维持增长，以及为什么生活水平的长期增长必须来自技术进步。它不能解释那技术进步从何而来，也不能解释为什么有些国家的技术与制度改善得比另一些更快。它并不能解释一切——但它能解释相当多的东西，并且通过用优化的家庭取代索洛那条机械的储蓄规则，它框定了新古典模型（第 六 章）将要接手的问题。

第六章　增长的微观基础：新古典增长模型

第 五 章的索罗模型带我们走了很远：仅凭一条资本积累方程，它就解释了为什么穷国比富国增长得快、为什么经济会收敛到稳态、以及为什么长期增长归根结底要靠技术进步而非单靠节俭。但这份解释力是有代价的。储蓄率 s 完全是从模型之外硬塞进来的，是一个靠假设固定下来的常数。一个无论如何都把固定比例的收入存起来的经济，本质上就是一个**计划**经济：没有人被问及存这么多是否符合自身利益，模型也对由此得到的资本存量究竟是偏多、偏少还是恰到好处保持沉默。

新古典增长模型补上了索罗所缺的那一块——*微观基础*。我们让经济中住进做最优化决策的主体：家庭选择消费与储蓄以最大化效用，厂商选择投入以最大化利润，并要求各市场出清。于是储蓄率不再是我们外生强加的参数，而是家庭在市场给定的回报下、在今天的消费与明天的消费之间权衡所**内生得出的结果**。值得玩味的是，当我们把这套最优化推演到底，得到的长期预言竟与索罗的几乎一模一样：稳定的稳态、收敛、以及由技术驱动的增长。不同之处在于，如今这些预言扎根于个体选择，因此我们可以提出索罗根本无从提起的福利问题——竞争性结果是否有效率？计划者能否做得更好？——并给出回答。

我们分阶段搭建这套工具。先从一个静态一般均衡经济入手，在最简单的设定里把家庭最优化、厂商最优化与市场出清的逻辑钉死。然后引入时间：一个两期禀赋经济带出欧拉方程与消费平滑；一个世代交叠经济把这套两期逻辑变成一套真正的资本积累理论，配上运动方程与稳态；最后，一个无穷期的吃蛋糕问题及其生产版本——社会计划者问题——补全全局，让我们得以陈述把竞争与效率绑在一起的两条福利定理。

6.1　静态一般均衡模型

我们从完全没有资本积累的情形起步。在单一一期之内资本存量无法改变——投资需要时间才能转化为生产能力——所以我们把 K 固定住，研究在这一期之内代表性家庭与代表性厂商如何在商品与劳动两个市场上相互作用。这就把问题剥到只剩它的一般均衡内核：每一侧都是做最优化的主体，价格不断调整直到供给等于需求。

假设 6.1: 静态新古典环境

- 各主体彼此相同，因此我们可以只研究单一的*代表性*家庭与单一的*代表性*厂商。
- 家庭在给定价格下最大化效用；厂商在给定价格下最大化利润。
- 市场出清：在每个市场上，需求等于供给。
- 资本存量 K 在本期内是外生的（它是一个状态变量，不是选择变量）。

一开始就值得指出，宏观经济学在侧重点上将与读者在微观经济学里见过的有所不同。家庭的两类选择支配着此后的一切。*劳动与闲暇*之间的选择——工作多少——本质上是*静态*的：它在单一一期之内就解决了。*消费与储蓄*之间的选择——为未来留下多少——本质上是*动态*的，而正是这一选择，一期接一期地堆积起资本存量 K。静态模型把第一类选择单独拎出来；本章其余部分则展开第二类。

6.1.1　家庭：消费与闲暇

代表性家庭从消费 C 与闲暇 l 中获得效用，

$$U = U(C, l),$$

我们假定通常的良态条件：两种物品都是值得拥有的，$U_C > 0$ 且 $U_l > 0$；边际替代率递减（无差异曲线是凸的）；并且消费与闲暇都是正常品。

家庭被赋予固定的时间禀赋 H，把它在闲暇 l 与工作 $H - l$ 之间分配。我们把消费品的价格标准化为一，并令 w 表示实际工资，于是 w 同时也是闲暇的价格——每一小时不工作，家庭就要付出 w 单位被放弃的消费作为代价。除工资收入之外，家庭还获得厂商分配的利润 π，并缴纳一笔总额税 T。其预算约束为

$$C = w(H - l) + \pi - T.$$

把闲暇的工资价值移到左边，约束就呈现出它最有启发性的形式，

$$C + w\,l = wH + \pi - T,$$

这读作一个*完全收入*（full income）预算：家庭的全部资源，即右边，等于其全部时间禀赋的价值加上利润减税之后的余额，而它把这些资源花在消费上、以及花在以 w 计价的闲暇上。右边是一个家庭视为既定的数。

在此约束下最大化 $U(C, l)$，标准的相切条件是闲暇与消费之间的边际替代率等于闲暇的相对价格，

$$\text{MRS} = w \quad \Longrightarrow \quad C^*,\ l^*.$$

于是家庭的最优劳动供给为 $N^s = H - l^*$。

6.1.2　厂商：生产与劳动需求

代表性厂商用资本与劳动生产产出，

$$Y = z\,F(K, N^d),$$

其中 z 是全要素生产率，N^d 是雇佣的劳动。我们假定 F 呈现规模报酬不变（CRS），正因如此我们才可以把整个生产侧当作单一的价格接受厂商来处理：在 CRS 下单个厂商的规模无从确定，而总量层面表现得就像一个代表性生产者。技术具有为正但递减的边际产出，

$$F_K > 0, \quad F_N > 0, \qquad F_{KK} < 0, \quad F_{NN} < 0.$$

注（对课堂讲义的一处更正）.

> 讲义把二阶条件写成了 $F_{KK} > 0$ 与 $F_{NN} > 0$。这是笔误：边际产出递减——正是这一特征让模型表现良好、并在下文的增长版本里给出收敛——要求二阶导数为负，即 $F_{KK} < 0$ 与 $F_{NN} < 0$。一阶导数为正；二阶导数为负。

在本期之内，K 是一个固定的状态变量，所以厂商只选择雇佣多少劳动。它的问题是

$$\max_{N^d} \left\{ z\,F(K, N^d) - w\,N^d \right\}.$$

一阶条件令劳动的边际产出等于实际工资，

$$\text{MPN} = z\,F_N(K, N^d) = w \quad \Longrightarrow \quad N^{d*}.$$

注意这两个条件之间的分工：MRS $= w$ 是家庭的最优化条件，MPN $= w$ 是厂商的。课堂上的速记“MPN $=$ MRS $= w$”把两者压成了一行，但它们分属不同的主体；只有在均衡中、当唯一的工资 w 调整到让两侧都满意时，它们才同时成立。

6.1.3　均衡与瓦尔拉斯定律

一个均衡就是一个让两个市场都出清的工资 w。在劳动市场上，家庭的供给等于厂商的需求，

$$H - l^* = N^{d*}.$$

在商品市场上，家庭消费的恰好就是厂商生产的，

$$C^* = Y^*.$$

这两个出清条件并不相互独立。由家庭的预算约束可知，它所需求之物的价值始终等于它所供给之物的价值；把各市场加总，超额需求必然彼此抵消。这就是**瓦尔拉斯定律**（Walras' law）：在一个有两个市场的经济里，只要一个出清，另一个也必然出清。

瓦尔拉斯定律

在一个有 n 个市场的一般均衡中，总超额需求的价值恒等于零。因此，只要 $n-1$ 个市场出清，最后一个市场就自动出清。在我们这个两市场经济里，**一旦劳动市场出清，商品市场也随之出清**——我们只需找到那个使劳动供给等于劳动需求的唯一工资 w 即可。

可以证明，存在一个**唯一**的工资 w 使得 $H-l^*=N^{d*}$。在该工资下，厂商的利润 π 被确定，家庭的完全收入被确定，整个配置 (C^*,l^*,N^{d*},Y^*) 也随之落定。

6.1.4　工资变动的收入效应与替代效应

均衡对工资变动如何反应？由于工资既是家庭每工作一小时的收入、又是闲暇的价格，w 的变动会引发两股相互角力的力量。

设想 w 上升。**收入效应**：家庭如今更富了，又因为消费与闲暇都是正常品，它两者都想要更多——消费上升，闲暇也上升（工作减少）。**替代效应**：闲暇相对于消费变贵了，于是家庭从闲暇转向消费——消费上升，闲暇下降（工作增加）。

两种效应把消费推向同一方向，所以工资上升**毫无疑义地抬高消费**。但它们把闲暇——从而把劳动供给——推向相反方向，所以对工作小时数的净效应一般是不确定的。劳动供给随工资上升还是下降，取决于哪种效应占上风。

有一个干净的特例让这种抵消恰好成立。把时间禀赋标准化为一（$H=1$，于是当 $\pi=T=0$ 时 $C=w(1-l)$），并取对数偏好，

$$\max_{C,l}\ \log C+\beta\log l\qquad\text{s.t.}\qquad C+w\,l=w.$$

例（对数效用：收入效应与替代效应相互抵消）.

求解家庭的问题，并证明劳动供给与工资无关。

解.

把预算约束 $C=w(1-l)$ 代入目标函数，使之成为仅关于 l 的函数：

$$U(l)=\log\big(w(1-l)\big)+\beta\log l=\log w+\log(1-l)+\beta\log l.$$

一阶条件为

$$\frac{\mathrm{d}U}{\mathrm{d}l}=\frac{-1}{1-l}+\frac{\beta}{l}=0\quad\Longrightarrow\quad\beta(1-l)=l\quad\Longrightarrow\quad l^*=\frac{\beta}{1+\beta}.$$

于是劳动供给为

$$1-l^*=\frac{1}{1+\beta},$$

这是一个常数：工资 w 已彻底退出。在对数效用下，闲暇上的收入效应与替代效应恰好相互抵消，所以工作小时数与工资无关。工资上升的全部作用都流向了消费：由

预算约束，$C^* = w(1-l^*) = w/(1+\beta)$ 随 w 一比一地上升。

这一刀刃般的结论并非代数上的巧合；对数效用恰恰是替代弹性与收入弹性调得使两种效应相互抵消的那个特例。这正是对数效用在本章里反复出现的原因：它给出最简单的闭式解，把我们关心的动态机制从恼人的、作用于劳动的财富效应中分离出来。

6.2　两期禀赋经济

现在我们以最简单的形式引入时间。暂且把生产和厂商剥离，考虑一个恰好活两期、且每一期都获得一笔外生收入的家庭——一份天上掉下来的禀赋（windfall），像吗哪一样。这里没有什么可生产，也没有劳动可供给；唯一的决策是何时消费。仅凭这副最朴素的设定，就足以引出动态宏观经济学的核心对象——欧拉方程。

家庭最大化终身效用

$$\max_{C_1,\,C_2}\; U(C_1,C_2) = \log C_1 + \beta \log C_2,$$

其中 $\beta \in (0,1)$ 是**贴现因子**（discount factor）——一个接近 0.95 的数，刻画了人的不耐：明天的一单位效用不如今天的一单位值钱。

定义 6.2: 贴现因子

贴现因子 $\beta \in (0,1)$ 是家庭对下一期效用相对于本期效用所赋予的权重。β 越大表示越有耐心。与之对应的**贴现率** ρ 满足 $\beta = 1/(1+\rho)$。

收入在第 1 期到来为 Y_1，在第 2 期到来为 Y_2，二者皆为外生。家庭只能通过以市场利率 r 储蓄一笔金额 S 来在两期之间转移资源。它逐期的预算约束为

$$\begin{aligned} C_1 + S &= Y_1, \\ C_2 &= Y_2 + (1+r)\,S. \end{aligned}$$

第一期的储蓄在第二期赚得毛回报 $1+r$。和静态模型一样，从家庭的视角看我们把 r、Y_1、Y_2 都当作既定：决定利率的宏观力量超出了这单个家庭的控制范围。

储蓄 S 是联结两期的桥梁。把它消去——从第二个约束解出 S 代入第一个——便把两条单期预算约束合并成单一的**跨期预算约束**，

$$C_1 + \frac{C_2}{1+r} = Y_1 + \frac{Y_2}{1+r}.$$

其含义正是人们所期望的：终身消费的现值等于终身收入的现值。

6.2.1　欧拉方程

在最优处，家庭无法通过把边际一单位消费在两期之间挪动来提高效用。相切条件是跨期边际替代率等于毛利率——即以明天的消费衡量的、今天消费的相对价格：

$$\mathrm{MRS} = -\left.\frac{\mathrm{d}C_2}{\mathrm{d}C_1}\right|_{U} = \frac{\mathrm{MU}_1}{\mathrm{MU}_2} = 1 + r.$$

在 $U = \log C_1 + \beta \log C_2$ 下我们有 $\mathrm{MU}_1 = 1/C_1$ 与 $\mathrm{MU}_2 = \beta/C_2$，于是

$$\frac{1/C_1}{\beta/C_2} = \frac{C_2}{\beta C_1} = 1 + r,$$

整理后即得**欧拉方程**（Euler equation）。

定理 6.3: 消费的欧拉方程（两期，对数效用）

最优消费路径满足

$$\frac{C_2^*}{C_1^*} = \beta(1+r), \qquad \text{等价地} \qquad C_2^* = \beta(1+r)\, C_1^*.$$

证明. 由跨期预算约束 $C_1 + C_2/(1+r) = W$，其中 $W := Y_1 + Y_2/(1+r)$ 是终身财富，拉格朗日函数为

$$\mathcal{L} = \log C_1 + \beta \log C_2 + \lambda\left(W - C_1 - \frac{C_2}{1+r}\right).$$

一阶条件为

$$\frac{1}{C_1} = \lambda, \qquad \frac{\beta}{C_2} = \frac{\lambda}{1+r}.$$

第一式除以第二式消去 λ，得 $\dfrac{C_2}{\beta C_1} = 1 + r$，即 $C_2^* = \beta(1+r)\, C_1^*$。 □

不构造跨期预算约束，也能得到同一个条件：直接把约束代入并对 S 求最大，

$$U(S) = \log\big(Y_1 - S\big) + \beta \log\big(Y_2 + (1+r)S\big), \qquad \frac{-1}{Y_1 - S} + \frac{\beta(1+r)}{Y_2 + (1+r)S} = 0,$$

回代 $C_1 = Y_1 - S$ 与 $C_2 = Y_2 + (1+r)S$ 后，依然得到 $C_2 = \beta(1+r)C_1$。

6.2.2　消费平滑

欧拉方程是理解家庭如何把消费铺展到时间之上的钥匙。假定有那么一刻既无不耐也无利息，即 $\beta = 1$ 且 $r = 0$。那么欧拉方程读作 $C_2^* = C_1^*$：家庭选择在两期消费完全相同的量，无论其收入的时间分布如何。这就是**消费平滑**（consumption smoothing）。

为什么凹性意味着平滑

只要效用是严格凹的（边际效用为正且递减，$U'>0$，$U''<0$），家庭就偏好一条平滑的消费路径，而非一条现值相同却起伏的路径。从高消费的一期取走一单位消费、挪到低消费的一期，会提高效用，因为失去的边际效用小于得到的边际效用。平滑正是边际效用递减在行为上的内涵。

一旦把不耐与利息重新放回，路径会倾斜，但不会断裂。欧拉方程 $C_2^*=\beta(1+r)C_1^*$ 说家庭是围绕因子 $\beta(1+r)$ 来平滑的。直觉正如人们所料：把一单位储蓄留到下一期会赚得利息，资源因而增长 $(1+r)$ 倍，这把消费向未来倾斜；但下一期的效用要被 β 折现，这又把它拉回当下。当 $\beta(1+r)=1$ 时这两股力量平衡，消费持平；当 $\beta(1+r)>1$ 时回报盖过不耐，消费随时间上升；当 $\beta(1+r)<1$ 时不耐取胜，消费随时间下降。

6.2.3　值函数与影子价格

一旦参数 (Y_1,Y_2,r) 被固定下来，这场最优化就钉死了唯一的解 (C_1^*,C_2^*,S^*)，从而也钉死了唯一的最大化效用。把这个最大化效用看作那些参数的函数，它就是**值函数**（value function），

$$U^*=V(Y_1,Y_2,r).$$

值函数记录了在给定处境下家庭所能达到的最好结果。它的导数携带着经济含义。对第一期收入求导，$\partial V/\partial Y_1$，衡量第 1 期多一单位天降收入会把终身福利抬高多少；这恰恰是第 1 期资源的**影子价格**（shadow price）——即最优化里的拉格朗日乘子 λ。在最优处，乘子等于每种物品的边际效用与其价格之比，

$$\frac{\partial U/\partial x_i}{p_i}=\lambda \qquad \text{对每个选择} x_i,$$

这正是那个熟悉的论断：在均衡中，家庭把每一元钱的边际效用在所有边际上拉平。下文的无穷期问题里我们将再次遇见值函数及其影子价格，到那时这种递归的思考方式将变得不可或缺。

6.3　世代交叠模型

两期经济教会了我们单个家庭如何把消费配置到时间之上，但它没有生产、没有资本，所以还称不上一套**增长**理论。世代交叠（OLG）模型把这一点补上：它把两期的人生一个叠一个地摞起来，使得在任何时点上，年轻一代与年老一代都并存。年轻人工作并储蓄；他们的储蓄变成下一期所继承的资本。如此一来，两期模型里那个动态的储蓄决策就成了资本积累的引擎。

这一结构如图 6.1 所示。每个个体活两期。年轻时她供给一单位劳动、赚取工资，消费其中一部分、把其余储蓄起来；年老时她不再工作，靠储蓄度日（此时储蓄已化为她租给厂商的资本）。尽管任何一个人都只活两期，世代交叠的链条却无尽地延伸下去，所以这个模型实际上是一个**无穷期**经济——只是我们一次只截取一代来看罢了。

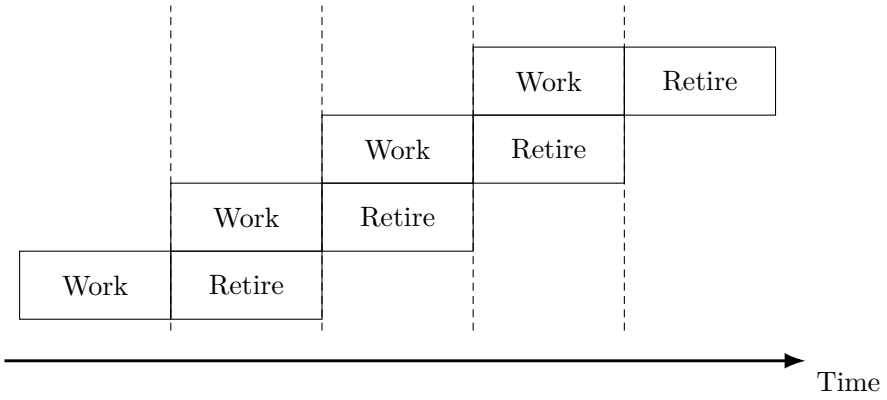

图 6.1: 世代交叠的时间线：每一代年轻时工作并储蓄，年老时靠积累下来的储蓄度日，于是在每个时点上，一代正在工作的年轻人都与一代已退休的年老人交叠在一起。

6.3.1　家庭

一个在 t 期出生的个体，年轻时供给一单位劳动，故其工资收入为 w_t。她的预算约束为

$$C_t + S_t = w_t,$$
$$C_{t+1} = (1 + r_{t+1})\, S_t,$$

其中 S_t 是储蓄，r_{t+1} 是这笔储蓄所赚得的净回报。关键的会计联结是：一代人的储蓄以资本的形式持有，所以年轻人的储蓄变成下一期的资本存量，

$$S_t = K_{t+1}.$$

正是这一个方程，把一连串两期问题变成了一套积累理论：明天社会动用的资本，恰恰就是今天年轻人选择留出来的那部分。

在对数偏好下，

$$U(C_t, C_{t+1}) = \ln C_t + \beta \ln C_{t+1},$$

储蓄决策就是我们刚刚解过的那个问题。欧拉方程 $C_{t+1}/C_t = \beta(1 + r_{t+1})$ 连同预算约束，给出一条明确的储蓄法则。

例（OLG 模型中的储蓄）.

在对数效用与上述预算约束下，证明年轻人储蓄的是工资的一个固定比例，从而

$$K_{t+1} = \frac{\beta}{1+\beta}\, w_t.$$

解.

把约束代入目标函数。在 $C_t = w_t - S_t$ 与 $C_{t+1} = (1 + r_{t+1})S_t$ 下，

$$U(S_t) = \ln(w_t - S_t) + \beta \ln\big((1 + r_{t+1})S_t\big).$$

关于 S_t 的一阶条件为

$$\frac{-1}{w_t - S_t} + \frac{\beta}{S_t} = 0 \quad \Longrightarrow \quad \beta(w_t - S_t) = S_t \quad \Longrightarrow \quad S_t = \frac{\beta}{1+\beta}\, w_t.$$

毛回报 $1 + r_{t+1}$ 被消掉了——这又是对数效用的一个后果——所以储蓄是工资的固定比例 $\beta/(1+\beta)$，与利率无关。由于 $S_t = K_{t+1}$，

$$K_{t+1} = \frac{\beta}{1+\beta}\, w_t.$$

我们保留 K 而不把它代换掉，因为它是联结市场两侧的那个变量——家庭通过储蓄供给它，厂商为生产而需求它——它同时也是联结相邻两期的那个变量。

6.3.2　厂商

厂商每期租用资本、雇佣劳动以最大化利润，

$$\max_{K,\,L}\; F(K, L) - w\, L - r\, K,$$

采用柯布–道格拉斯的、劳动增进型技术

$$F(K, L) = K^{\alpha}\,(zL)^{1-\alpha}, \qquad 0 < \alpha < 1.$$

一阶条件令每种要素的价格等于其边际产出，

$$w_t = (1-\alpha)\, z_t^{1-\alpha}\, K_t^{\alpha}\, L_t^{-\alpha},$$
$$r_t = \alpha\, z_t^{1-\alpha}\, K_t^{\alpha-1}\, L_t^{1-\alpha}.$$

6.3.3　市场均衡与运动方程

在均衡中，每种要素的价格在供给侧与需求侧是相同的，所以我们可以把厂商的工资代入家庭的储蓄法则。结果是一个关于资本的差分方程，

$$K_{t+1} = \frac{\beta}{1+\beta}\, w_t = \frac{\beta}{1+\beta}\,(1-\alpha)\, z_t^{1-\alpha}\, L_t^{-\alpha}\, K_t^{\alpha}.$$

由于问题里处处都没有出现人口变量、且我们已把每个年轻人的劳动供给标准化为一单位，我们令 $L_t = 1$，并把每个量都读作人均量。暂且把生产率固定在常数水平 $z_t = z$，运动方程就变成

$$K_{t+1} = \frac{\beta(1-\alpha)}{1+\beta}\, z^{\,1-\alpha}\, K_t^{\alpha}.$$

由于指数 α 小于一，右边是 K_t 的一个凹的、递增的函数，并且与 45° 线恰好相交一次（图 6.2）。这个交点就是稳态，在那里资本自我复制，$K_{t+1} = K_t = K_{\text{ss}}$。

例（稳态资本）.

从运动方程解出 K_{ss}。

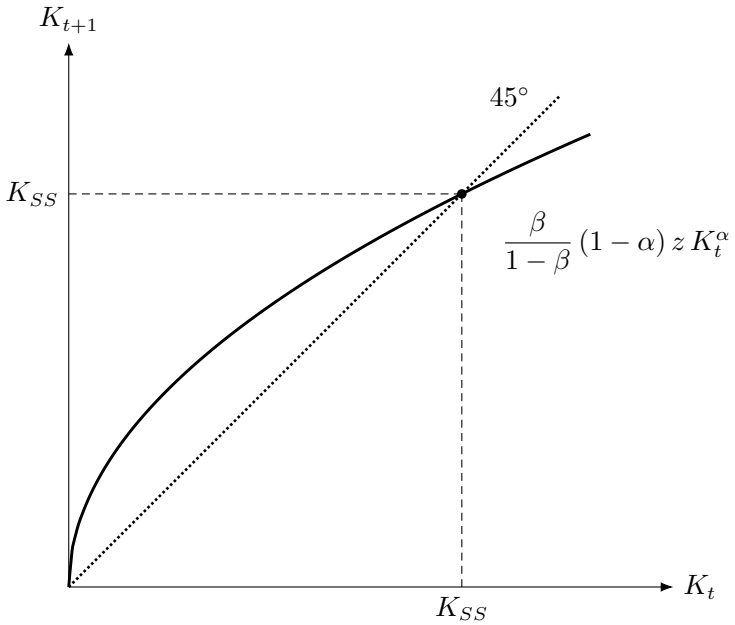

图 6.2: 资本的运动方程 $K_{t+1} = f(K_t)$。由于 $\alpha < 1$，曲线是凹的，且与 45° 线恰好相交一次于稳态 K^*；从任何为正的初始点出发，经济都收敛到该点。

解.

在稳态处 $K_{\text{ss}} = \dfrac{\beta(1-\alpha)}{1+\beta}\, z^{1-\alpha}\, K_{\text{ss}}^{\alpha}$。两边同除以 K_{ss}^{α}，

$$K_{\text{ss}}^{1-\alpha} = \frac{\beta(1-\alpha)}{1+\beta}\, z^{1-\alpha} \quad \Longrightarrow \quad K_{\text{ss}} = z\left(\frac{\beta(1-\alpha)}{1+\beta}\right)^{\frac{1}{1-\alpha}}.$$

通过一期接一期地施用运动方程，我们实际上已把两期的 OLG 结构延展到了无穷多期。又由于没有任何人口项出现，每个变量都带有人均的含义。运动方程还重现了索罗的那些教训：一个起步资本很少的国家位于图 6.2 中靠左的位置，那里曲线很陡，所以它的资本——以及产出——增长得很快；相似的经济收敛到相同的稳态。其深层原因，再一次，是 $\alpha < 1$ 所编码的资本边际产出递减。

6.3.4　稳定稳态与不稳定稳态

值得就差分方程的一个一般要点停一停，因为并非每个不动点都是归宿。考虑任意一阶差分方程

$$x_{t+1} = f(x_t), \qquad t = 1, 2, \ldots$$

这就是那个典范的*运动方程*，或曰马尔可夫链结构：明天的取值只取决于今天的取值，而不取决于全部历史。稳态就是一个不动点，

$$x_{\text{ss}} = f(x_{\text{ss}}),$$

即 f 的图像与 45° 线相交之处。但解出 $x = f(x)$ 只告诉我们稳态*存在*；它并不告诉我们经济是否真的会朝它移动。

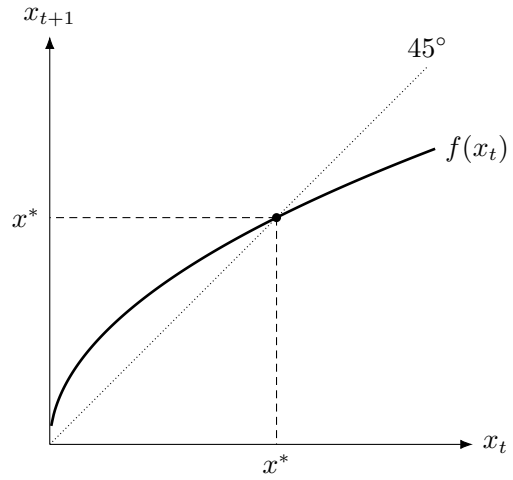

图 6.3: 一个*稳定*的稳态：运动方程穿过 45° 线时斜率的绝对值小于一，所以在附近起步的轨迹会被拉回到交点。

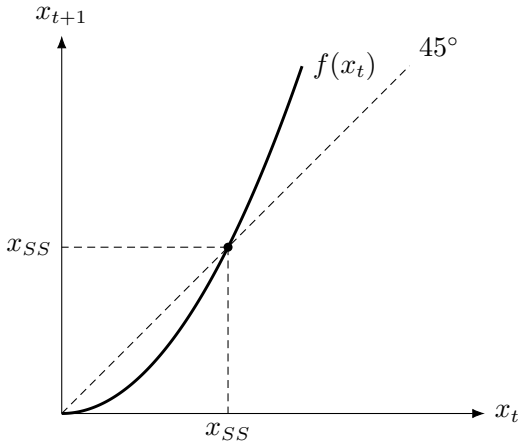

图 6.4: 一个*不稳定*的稳态：运动方程穿过 45° 线时斜率大于一，所以交点排斥而非吸引附近的轨迹。

稳定性由 f 在交点处的斜率决定。若曲线从上方穿过 45° 线、斜率绝对值小于一（图 6.3），轨迹会被拉回交点，稳态便是*稳定*的——它是一个真正的吸引子。若曲线穿过时斜率大于一（图 6.4），交点会*排斥*附近的路径，稳态便是*不稳定*的：它解出了 $x = f(x)$，但经济永远不会在那里安顿下来。我们这个凹的资本运动方程在交点处的斜率为 $\alpha\, K_{\text{ss}}^{\alpha-1} \cdot (\text{常数}) < 1$，所以 K_{ss} 是稳定的——属于图 6.3 那个好的情形。

6.3.5　生产率增长

至此我们把生产率冻结在了 $z_t = z$。但 z 衡量的是劳动的生产效率——技术、组织、知识技能——它并不停滞不前。改设生产率以恒定速率 g 增长，

$$z_{t+1} = (1+g)\, z_t.$$

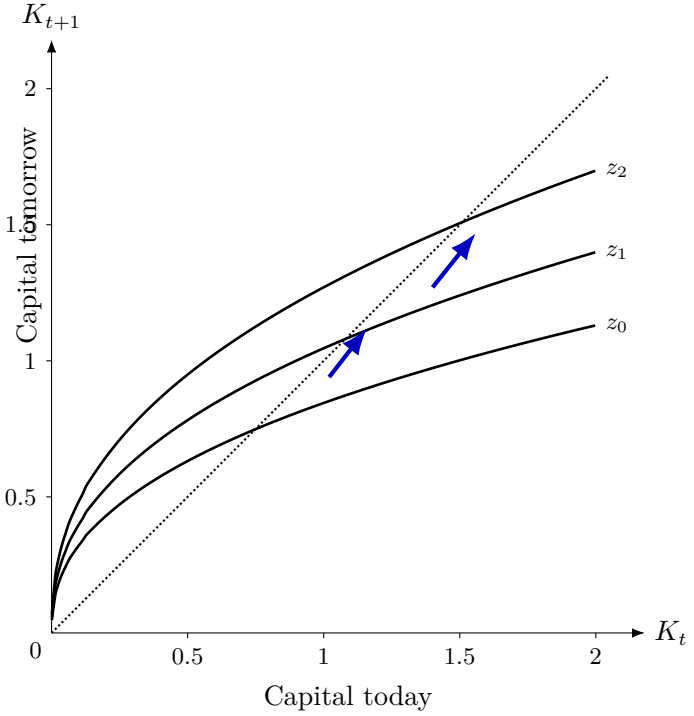

图 6.5: 生产率增长使运动方程逐期上移：随着 z_t 上升，曲线 $K_{t+1} = f(K_t)$ 向外移动，拖着那个移动的目标 $K_{\text{ss}}(z_t)$ 稳步右移，并把资本存量沿一条增长路径拉着走。

运动方程仍读作

$$K_{t+1} = \frac{\beta}{1+\beta}\, w_t = \frac{\beta(1-\alpha)}{1+\beta}\, z_t^{1-\alpha}\, L_t^{-\alpha}\, K_t^{\alpha},$$

若令 $L_t = 1$ 并暂时把 z_t 当作固定，与此前同样的代数给出一个目标水平

$$K_{\text{ss}} = z_t \left(\frac{\beta(1-\alpha)}{1+\beta} \right)^{\frac{1}{1-\alpha}}.$$

但如今 z_t 自身在增长，所以严格地说 $K_{t+1} \neq f(K_t)$ 并不对应一个不随时间变化的 f，上面那个“K_{ss}”也就不是一个真正的稳态。它毋宁是一个*移动的目标*：一条随 z_t 上升而向上漂移的增长路径（图 6.5）。若资本存量偏离了这条路径，边际报酬递减那股力量仍会把它拉回路径之上。于是在长期，资本以与生产率相同的速率增长，

$$K_{t+1} = (1+g)\, K_t,$$

因为 K_{ss} 与 z_t 成比例。

把这种平衡增长代入生产函数，取 $L_t = 1$，

$$Y_{t+1} = K_{t+1}^{\alpha}\,(z_{t+1}L_{t+1})^{1-\alpha} = \big[(1+g)K_t\big]^{\alpha}\big[(1+g)z_tL_{t+1}\big]^{1-\alpha} = (1+g)\,Y_t,$$

所以产出也以生产率的增速 g 增长。长期增长由技术驱动，与索罗一模一样。

Y 与 K 同速增长，对比值与回报率有着鲜明的含义。比值 Y/K 趋于一个常数，因为两者都以 g 增长。资本的回报率于是也是恒定的，

$$r_t = \alpha\, z_t^{1-\alpha}\, K_t^{\alpha-1}\, L_t^{1-\alpha} = \frac{\alpha\, Y_t}{K_t},$$

是那个（恒定的）产出–资本比的固定倍数。相形之下，工资随生产率增长，

$$w_t = \frac{(1-\alpha)\,Y_t}{L_t},$$

随人均产出的上升以速率 g 上升。

战后增长的卡尔多典型化事实

新古典模型重现了二战以来发达经济体的若干宽泛规律：

1. 产出–资本比 Y/K 大致恒定；
2. 资本的实际回报率 r 大致恒定；
3. 资本的收入份额 α 大致恒定；
4. 人均产出 Y/L 与人均资本 K/L 以共同的速率 g 增长。

索罗模型与新古典模型都与这些事实相符；不同的是，新古典模型是从最优化中赢得它们，而非靠假设给定它们。

6.4　吃蛋糕问题

现在我们从两期人生转向一个具有无穷视野的单一决策者。**吃蛋糕问题**（cake-eating problem）是最干净的无穷期最优化：一个消费者拥有一笔资源的存量——一块蛋糕——必须决定以多快的速度把它吃掉。它没有生产、没有价格，只有一笔随消费而缩小的存量，它教给我们的，是本章所有动态模型背后那套递归结构。

消费者最大化经过贴现的终身效用

$$\max_{\{C_t\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t\, U(C_t)$$

受制于资源的转移方程

$$C_t + H_{t+1} = (1-\delta)\,H_t, \qquad \forall t,$$

其中 H_t 是第 t 期初的蛋糕存量，δ 是折旧率——蛋糕每期会变质一点。这里的术语极为基本，并在动态宏观经济学中处处复现。

定义 6.4: 状态变量与控制变量

状态（state）变量是决策者在一期之初所继承、且在本期之内无法改变的东西；这里它是存量 H_t。控制（control）变量是她在本期之内所选择的东西；这里它是 C_t（等价地是 H_{t+1}，因为二者被资源约束联结起来）。每个变量都先后扮演这两种角色：今天的控制 H_{t+1} 成为明天的状态。

6.4.1　求解该问题

给每一期的约束附上一个乘子 λ_t，构造拉格朗日函数

$$\mathcal{L} = \sum_{t=0}^{\infty} \beta^t U(C_t) + \sum_{t=0}^{\infty} \lambda_t \big((1-\delta)H_t - C_t - H_{t+1}\big)\,.$$

关于 C_t、C_{t+1} 以及那个起桥梁作用的存量 H_{t+1} 的一阶条件为

$$\begin{aligned} C_t: &\quad \beta^t U'(C_t) = \lambda_t, \\ C_{t+1}: &\quad \beta^{t+1} U'(C_{t+1}) = \lambda_{t+1}, \\ H_{t+1}: &\quad \lambda_{t+1}(1-\delta) = \lambda_t. \end{aligned}$$

存量 H_{t+1} 既出现在第 t 期的约束里（带负号），又出现在第 $(t+1)$ 期的约束里（乘以 $1-\delta$），这正是为什么它的一阶条件把两个乘子拴在一起。消去乘子便得到欧拉方程，

$$\frac{U'(C_t)}{\beta\, U'(C_{t+1})} = 1 - \delta.$$

特化到对数效用 $U(C_t) = \log C_t$，于是 $U'(C) = 1/C$，

$$\frac{C_{t+1}}{C_t} = \beta(1-\delta).$$

定义 6.5: 政策函数

政策函数（policy function）把当前状态映射为最优控制——它勾勒出经过最优化的决策实际所沿行的那条路径。这里它把消费表示为所继承存量的函数。

欧拉方程 $C_{t+1} = \beta(1-\delta)\,C_t$ 是消费的运动方程；与资源约束合起来便给出政策函数。

例（吃蛋糕的政策函数）.

> 在对数效用下，把消费求成存量的函数，并验证它确实解了该问题。设初始总存量为 $H_1 = H$。

解.

猜测消费与存量成比例，$C_t = \gamma\, H_t$，其中 γ 是一个待定常数。资源约束随之给出存量的转移，

$$H_{t+1} = (1-\delta)H_t - C_t = (1-\delta-\gamma)\, H_t,$$

从而 $C_{t+1} = \gamma\, H_{t+1} = \gamma(1-\delta-\gamma)\, H_t$。施加欧拉方程 $C_{t+1}/C_t = \beta(1-\delta)$，

$$\frac{\gamma(1-\delta-\gamma)H_t}{\gamma H_t} = 1-\delta-\gamma = \beta(1-\delta) \quad\Longrightarrow\quad \gamma = (1-\delta)-\beta(1-\delta) = (1-\beta)(1-\delta).$$

于是政策函数与由此得到的路径为

$$C_t = (1-\beta)(1-\delta)\, H_t, \qquad H_{t+1} = \beta(1-\delta)\, H_t.$$

从 $H_1 = H$ 出发，存量以几何方式衰减，$H_t = \big(\beta(1-\delta)\big)^{t-1} H$，所以消费为

$$C_1 = (1-\beta)(1-\delta)\, H, \qquad C_t = (1-\beta)(1-\delta)\, \big(\beta(1-\delta)\big)^{t-1}\, H.$$

若 $\delta = 0$，这就坍缩成经典的吃蛋糕答案 $C_t = (1-\beta)\,\beta^{\,t-1} H$：每期把剩余的固定比例 $1-\beta$ 吃掉。

注（对课堂讲义的一处更正）.

讲义把消费比例写成了 $1-\beta(1-\delta)$，给出 $C_1 = \big(1-\beta(1-\delta)\big)H$。正确的常数是 $(1-\beta)(1-\delta)$。两者仅在特例 $\delta = 0$ 时一致，此时都化为 $1-\beta$；当折旧为正时，上面经过验证的答案才是对的——只需直接拿它对照欧拉方程便可核实。

6.4.2　吃蛋糕问题教会了我们什么

这道习题的价值是概念性的，而非计算性的。有两条教训会延续下去。

第一，辨明谁是状态、谁是控制。这里的存量 H 所扮演的角色，恰恰就是此前各模型里*储蓄*所扮演的角色：它是从过去带到现在的那笔资源。认出这一对应关系，正是让我们把所有这些模型看作同一主题之变奏的关键。

第二，留意约束代表着什么。在前面几节那些受预算约束的问题里，相对价格登场、并支配着各种权衡。在吃蛋糕问题里则根本没有价格——约束是一个纯粹的*资源约束*，只陈述这种物品在物理上有多少可用。这恰恰是当一个单独的决策者（而非市场）来配置资源时、约束所采取的形式。它是通往计划者问题的桥梁，我们现在就转向计划者问题。

概念重于代数

动态宏观里反复出现的纪律是：要分清什么是*状态*、什么是*控制*，以及一个约束反映的是*价格*（市场配置）还是纯粹的*资源*（计划者配置）。数学形式是次要的；对状态、控制、约束的经济解读，才是让这些模型可以相互比较的东西。

6.5　社会计划者问题与福利定理

在竞争性一般均衡中，家庭最大化效用、厂商最大化利润，整个配置由供给与需求通过价格的相互作用所驱动。现在改设想一个仁慈的**社会计划者**，他掌握经济中的全部信息，直接配置资源，根本不用价格。两种配置何时会相同？这个问题的答案是经济学中最深刻的结论之一。

计划者求解

$$\max_{\{C_t\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t U(C_t)$$

受制于经济的资源约束

$$c_t + i_t = y_t,$$

它说产出被划分为消费与投资——并且关键地请注意，这里没有出现任何价格。产出由资本与劳动、用厂商所用的同一技术生产出来，

$$y_t = F(k_t, n_t),$$

而资本通过

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

积累起来，k_0 给定。吃蛋糕问题是 $F \equiv 0$ 的特例，在其中唯一的“资源”就是上一期剩下的存量；计划者问题把它推广开来，让剩下的存量变得**有生产力**——仿佛没吃掉的蛋糕被借了出去，赚来一笔额外的产出 $F(k_t)$。

令 $n_t = 1$，把资源约束与积累方程合起来消去 i_t，

$$c_t + \big(k_{t+1} - (1-\delta)k_t\big) = F(k_t),$$

于是计划者问题为

$$\max_{\{c_t\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t U(c_t) \qquad \text{s.t.} \qquad c_t + k_{t+1} = F(k_t) + (1-\delta)k_t, \quad k_0 \text{ 给定.}$$

6.5.1　求解计划者问题

为求清晰，取完全折旧的情形 $\delta = 1$，于是约束读作 $c_t + k_{t+1} = F(k_t)$。拉格朗日函数为

$$\mathcal{L} = \sum_{t=0}^{\infty} \beta^t U(c_t) + \sum_{t=0}^{\infty} \lambda_t \big(F(k_t) - k_{t+1} - c_t\big),$$

一阶条件为

$$\begin{aligned} c_t: &\quad \beta^t U'(c_t) = \lambda_t, \\ c_{t+1}: &\quad \beta^{t+1} U'(c_{t+1}) = \lambda_{t+1}, \\ k_{t+1}: &\quad \lambda_{t+1} F'(k_{t+1}) = \lambda_t. \end{aligned}$$

消去乘子得到计划者的欧拉方程，

$$\frac{U'(c_t)}{\beta\, U'(c_{t+1})} = F'(k_{t+1}).$$

这正是吃蛋糕的欧拉方程，只不过把 $1-\delta$ 换成了资本的边际产出 $F'(k_{t+1})$：资源“增长”的速率如今是技术的回报，而不再是一个固定的留存率。

例（柯布–道格拉斯的社会计划者）.

在 $F(k_t) = A\,k_t^{\theta}$ 与对数效用 $U(c_t) = \log c_t$ 下，解出资本的运动方程与稳态。

解.

在对数效用下欧拉方程变成

$$\frac{c_{t+1}}{\beta\, c_t} = F'(k_{t+1}) = A\,\theta\, k_{t+1}^{\theta-1}.$$

用约束给出的 $c_t = Ak_t^{\theta} - k_{t+1}$ 代入，得到一个只含资本的方程，

$$\frac{Ak_{t+1}^{\theta} - k_{t+2}}{\beta\,(Ak_t^{\theta} - k_{t+1})} = A\,\theta\, k_{t+1}^{\theta-1}.$$

这个方程的形式提示我们猜测 $k_{t+1} = B\,k_t^{\theta}$，其中 B 为常数。那么 $c_t = Ak_t^{\theta} - Bk_t^{\theta} = (A-B)k_t^{\theta}$ 且 $c_{t+1} = (A-B)k_{t+1}^{\theta}$。代入欧拉方程，

$$\frac{(A-B)k_{t+1}^{\theta}}{\beta(A-B)k_t^{\theta}} = \frac{B^{\theta}}{\beta}\, k_t^{\theta^2-\theta} \;\overset{!}{=}\; A\theta\,(Bk_t^{\theta})^{\theta-1} = A\theta\, B^{\theta-1}\, k_t^{\theta^2-\theta}.$$

k_t 的幂次相符，证实了猜测；令系数相等，

$$\frac{B^{\theta}}{\beta} = A\theta\, B^{\theta-1} \quad\Longrightarrow\quad B = A\beta\theta,$$

所以运动方程为

$$k_{t+1} = A\beta\theta\, k_t^{\theta}.$$

令 $k_{t+1} = k_t = k_{\mathrm{ss}}$ 并除以 k_{ss}^{θ}，

$$k_{\mathrm{ss}}^{1-\theta} = A\beta\theta \quad\Longrightarrow\quad k_{\mathrm{ss}} = (A\beta\theta)^{\frac{1}{1-\theta}}.$$

和 OLG 模型一样，生产率水平 A 的增长会把这个目标向上推移，于是稳态沿一条增长路径移动，而非停在原地。

这个计划者配置与索罗稳态的结构完全相同，但它如今是从跨期最优化中导出的，而非从一个假设的储蓄率得来。这正是新古典纲领的全部要旨所在：结论存活了下来，但它们扎根于选择。又因为模型是从明确的偏好与技术搭起来的，它的参数可以被重新诠释以涵盖更多东西——比如，可以通过让生产率随机化来引入经济周期，$y_t = z\,A\,k_t^{\theta}$，其中 z 是随机的（从某个分布中抽取，或在高、低两种状态之间切换）。单一代表性主

体的计划者解随后便能推广到整个经济，因为每个主体都彼此相同。

6.5.2　两条福利定理

现在我们可以陈述计划者配置与竞争性均衡之间的关系。在合适的条件下两者重合，而联结它们的桥梁，恰恰是计划者问题里不包含的那组价格。计划者的资源约束携带着影子价格——即乘子 λ_t——而这些影子价格，正是能够把计划者配置去中心化（decentralize）的那些竞争性价格。

定理 6.6: 两条福利定理

1. **第一福利定理**。一个竞争性均衡配置是帕累托最优的：在合适的价格下，那个去中心化的结果——每个家庭最大化效用、每个厂商最大化利润——已经无法在不损害某个主体的前提下改善另一个主体了。
2. **第二福利定理**。任何一个帕累托最优配置——尤其是一个仁慈的社会计划者所会选择的那个——都可以被支撑为一个竞争性均衡，只要价格设置得当（并且在必要时，用一次性总额转移来重新分配初始财富）。

其经济内涵是：市场与计划，尽管组织原则截然相反，在这个无摩擦的世界里却抵达同一个归宿。计划者按资源约束来配置、不理会价格；市场按价格来配置、不理会任何中央指令。然而计划者最优化中隐含的影子价格，恰恰就是那组使家庭与厂商——仅仅出于各自的私利行事——自发选出计划者那套配置的市场价格。正是这一等价关系，许可我们在此后整门课程里去解一个计划者问题（往往要容易得多），并从中读出竞争性均衡。

注（这是要往哪里去）.

> 新古典模型给了储蓄率一个微观经济学的基础，并借此给了我们一种方法去追问市场结果是否有效率。福利定理在这个基准经济里回答“是”。宏观经济学其余的大部分内容，正是研究当这些定理**失效**时会发生什么——当货币、黏性价格、外部性或缺失的市场在竞争性结果与计划者理想之间打入一道楔子，为政策改善市场腾出空间时。第 七 章的马尔萨斯经济是第一个这样的背离；货币摩擦与凯恩斯摩擦紧随其后。

第七章　人口、土地与马尔萨斯经济

第 五 章的索罗模型，以及第 六 章为它补上微观基础的新古典版本，都暗含两个未曾言明的承诺：它们只用资本与劳动来构造产出；它们把人口规模当成从外部交给经济的东西——一个由建模者随手写下、以速率 n 增长的数字。对于这两个模型本来要解释的世界——一个紧约束在于可累积的资本、人们越富裕反而生得越少的现代经济——这两个承诺都站得住脚。可一旦面对它之前的那个世界，二者就都失灵了。在前工业的农业社会里，稀缺要素并不是资本，而是**土地**：土地是固定的，无法被累积。人口也不是外生的：一个社会里究竟有多少人，本身就是一项经济决策的结果——是各个家庭关于养育多少孩子的决策。

本章研究的，正是把这两件事重新放回中心的模型。它建立在一个一旦看清便难以再忽视的典型事实之上：纵观漫长的农业时代，技术进步并没有让普通人过得更好。更好的种子、更好的犁、更好的灌溉，最终没有体现为更高的生活水平，而是体现为**更多的人**以和从前一样的水平活着。其中的机制正是托马斯 · 马尔萨斯（Thomas Malthus）在 1798 年描述的那一套：当某项技术改良提高了土地的产量，家庭会短暂地变富，于是多养孩子；人口随之增长；固定的那块土地被更多张嘴瓜分；人均产量又被压回到原来的位置。土地是紧约束，而人口正是这条约束所要按住的那个变量。整个模型的用意，就是把这个陷阱说精确——然后再追问：人类最终究竟靠什么逃出了它。

我们分三步走。先把土地重新放回生产函数，说明为什么一个固定要素会把日益增加的投入变成不断递减的回报——这就是农业的**过密化**（*involution*）现象。接着写出并求解家庭的最优化问题，导出人口的运动方程，定位出马尔萨斯陷阱真正咬住经济的那个稳态：生活水平被钉死在维生水平，任何技术改进都被全数花在了多出来的人口上。最后，我们把养育孩子的成本修改为还包含一项**时间**成本，它会把收入与生育之间的关系压平、从而拆掉这个陷阱；再叠加一个数量—质量的取舍，把这条关系彻底翻转过来——这便是人口转变。

7.1　土地作为紧约束要素

我们目前见过的增长模型，其生产函数里只有资本 K 和劳动 L 两个要素。要刻画一个农业经济，我们再加入第三个要素——土地 T，并写下一个柯布—道格拉斯技术

$$F(K, L, T) = A\, K^{\alpha} L^{\beta} T^{\,1-\alpha-\beta}, \tag{7.1}$$

其中各指数均为正、且相加等于一，因此该技术对三种要素整体而言是规模报酬不变的。A 是全要素生产率，代表技术的水平。

在现代经济里，土地几乎无足轻重。相对于生产所需，土地实在太多，土地约束从来咬不住，产出由资本和劳动主宰，于是土地份额 $1-\alpha-\beta$ 可以略去而几乎没有损失——这恰恰是第 五 章和第 六 章把它丢掉的原因。可在农业社会里情形正好相反。那里可耕地的供给实际上是固定的，一旦不断增长的人口把它全都开垦殆尽，土地就变成了**硬约束**。此时往一块固定的地上不断追加劳动力，就会狠狠撞上边际报酬递减规律：每多一双手为总产出带来的增量越来越小，到最后劳动的边际产出远远跌到了平均产出之下。

定义 7.1: 农业过密化

过密化（亦即“高水平均衡陷阱”）指的是这样一种局面：一个固定要素——此处即土地——逼着人们投入越来越多的劳动，却只换来越来越小的边际产出增量。人们把心力与巧思一股脑倾注进去，只为从同一块地里多榨出一点点；总产出甚至可能还在上升，但追加劳动的边际产出已坍缩，**人均**产出陷入停滞。那种极尽精细、劳力饱和的耕作方式，正是一个社会被死死压在土地约束上的可见症状。

注（一个熟悉模式的名字）.

“过密化”（involution，中文译作**内卷**）这个词近来已经跑出了农业史的研讨课，被用来形容上班族为同一批固定的奖赏越拼越狠。其经济骨架是完全一样的：一个固定要素、不断加码的努力，以及一项已被压到几乎为零的努力边际回报。马尔萨斯模型正是这一模式最早、最农业的那个版本。

至此我们可以陈述这个模型存在的理由——它要解释的那个核心经验规律。

马尔萨斯典型事实

在农业社会里，人们很难看到任何持续的人均生活水平的改进。技术进步带来的，反而是一个在生活水平不变的前提下**规模更大的人口**。一个具有这种性质的经济——更好的技术所带来的收益被更多的人、而非更富的人吸收掉——就被称为**马尔萨斯（*Malthus*）经济**。

本章余下的任务，就是把这个事实从最优化行为中推导出来，让我们不只是知道它会发生，而是确切地理解它为什么发生。

7.2　家庭的最优化问题

考虑农业经济中的一个代表性家庭。它在乎自己的消费 C_t，也在乎自己养育的孩子数量 n_t，并按速率 $\beta\in(0,1)$ 对未来进行贴现。其终生效用为

$$\max_{\{C_t,\,n_t\}_{t=0}^{\infty}}\ \sum_{t=0}^{\infty}\beta^t\left(\log C_t+\log n_t\right). \tag{7.2}$$

对数设定承载了两层值得停下来体会的经济含义。孩子直接进入效用，刻画的是农业现实中的一个事实：孩子既是天伦之乐，也是一笔资产——田里多一双手、老来多一份保

障。而效用取对数，意味着消费与孩子是弹性为一的不完全替代品，于是家庭总会两样都要一些，并按固定比例在二者间分配自己的资源——这一点我们马上就会看到。

每一期家庭都面对一个预算约束。养孩子是有代价的：每养一个要花掉 b 单位的消费品。记 c_t 为每个家庭的消费、y_t 为每个家庭的收入，则花在消费和孩子上的资源不能超过收入，

$$c_t + b\,n_t = y_t, \tag{7.3}$$

其中 $b > 0$ 是养育一个孩子的物质成本。

把这个家庭问题和上一节的土地约束系在一起的，是收入 y_t 的生成方式。设 P_t 为人口规模——它将是我们的*状态变量*，即那个缓慢移动、概括了经济全部历史的存量。土地总量是固定的；把它标准化为一。那么人均土地就只是人口的倒数，

$$L_t = \frac{1}{P_t}. \tag{7.4}$$

人均产出取决于每个人能耕作多少土地。我们现在把生产函数 (7.1) 特化到以劳动与土地为起作用的要素的农业情形。为了让记号清楚，把劳动投入写成工人数 N_t（即先前称作 L 的那个符号），把固定的土地存量写成 $T = 1$，于是人均土地就是 $L_t = T/P_t = 1/P_t$，与 (7.4) 一致——从这里起，字母 L 表示人均土地，而不再表示劳动。把两要素产出 $F(N_t, T)$ 除以工人数，就得到一个关于人均土地递增且凹的人均产量，

$$y_t = \frac{F(N_t, T)}{N_t} = f(L_t) = z_t\, L_t^{\theta} = \frac{z_t}{P_t^{\theta}}, \qquad 0 < \theta < 1, \tag{7.5}$$

其中 z_t 是农业技术的水平（扮演上文 A 的角色），θ 是产量对人均土地的弹性。最后一个等号用到了 (7.4) 中的 $L_t = 1/P_t$。方程 (7.5) 是整个故事的发动机：**人口越多，人均土地越少，因而人均收入越低**。这正是那个固定要素伸进每一个家庭预算里的通道。

假设 7.2: 马尔萨斯经济

本模型依赖以下一组维持性假设：

- 一个代表性家庭，其偏好 (7.2) 定义在消费 C_t 与孩子数量 n_t 之上，按 $\beta \in (0,1)$ 贴现。
- 每养一个孩子花费 b 单位的消费品，得到预算约束 $c_t + b\,n_t = y_t$。
- 土地总量固定、标准化为一，故人均土地为 $L_t = 1/P_t$。
- 一个凹的人均产量 $y_t = z_t L_t^{\theta} = z_t / P_t^{\theta}$，其中 $0 < \theta < 1$，z_t 为技术水平。
- 人口通过生育演化：下一期人口等于本期人口乘以每个家庭的孩子数。

7.2.1　静态选择：如何分配收入

家庭的完整问题是动态的，因为今天养育的孩子会扩大明天的人口，从而通过 (7.5) 压低明天的收入。但*如何把给定的收入 y_t* 在消费和孩子之间分配，却是一个纯粹的期内

静态问题：固定住 y_t，期内收益 $\log c_t + \log n_t$ 只取决于第 t 期的选择。因此我们可以逐期求解这个分配。

例（期内分配）.

给定收入 y_t，选择 c_t 与 n_t 以最大化 $\log c_t + \log n_t$，约束为预算 $c_t + b\,n_t = y_t$。

解.

构造拉格朗日函数

$$\mathcal{L} = \log c_t + \log n_t + \lambda\,(y_t - c_t - b\,n_t)\,.$$

一阶条件为

$$\frac{\partial \mathcal{L}}{\partial c_t}: \quad \frac{1}{c_t} = \lambda, \qquad \frac{\partial \mathcal{L}}{\partial n_t}: \quad \frac{1}{n_t} = \lambda\, b.$$

第一式除以第二式消去乘子，得到 $n_t b = c_t$：家庭花在养孩子上的钱，恰好和花在自己消费上的钱一样多。这正是对数—对数偏好的特征——每种用途都拿到一份固定的预算份额，此处即各占一半。把 $c_t = b\,n_t$ 代回预算 $c_t + b\,n_t = y_t$，得 $2b\,n_t = y_t$，于是

$$n_t^* = \frac{y_t}{2b}, \qquad c_t^* = \frac{y_t}{2}.$$

这两条政策规则既干净又直观。收入的一半被消费掉；另一半花在养孩子上，按单位成本 b 折算，可以买到 $y_t/(2b)$ 个孩子。关键在于，**生育随收入而上升**：越富的家庭养越多的孩子。仅凭这一条行为事实，再加上土地约束，就足以生成马尔萨斯陷阱。

7.3　人口动态与稳态

现在让静态规则去喂动态。每个家庭养 n_t 个孩子，故下一期的人口为

$$P_{t+1} = n_t\, P_t. \tag{7.6}$$

（把 n_t 读作每个家庭的孩子数；在每个家庭一名成年人的设定下，它也就是整个人口的总繁殖系数。）把最优生育 $n_t^* = y_t/(2b)$、再把产量关系 $y_t = z_t/P_t^{\theta}$ 代入，就得到一个关于状态 P_t 的单一差分方程，

$$P_{t+1} = \frac{y_t}{2b}\,P_t = \frac{z_t}{2b}\,P_t^{1-\theta}. \tag{7.7}$$

指数 $1-\theta$ 严格介于零与一之间，故右端关于 P_t 既凹又增：这个映射有唯一且稳定的正不动点。人口太少时会增长（土地充裕、收入高、家庭多生孩子）；人口太多时会收缩（土地稀缺、收入低、家庭少生孩子）。无论从哪一边出发，经济都被驱向同一个歇脚点。

7.3.1　求解稳态

在稳态上，人口不再变化，$P_{t+1} = P_t = P_{\text{ss}}$，技术固定在 $z_t = z$。把这一条件加到 (7.7) 上，

$$P_{\text{ss}} = \frac{z}{2b} P_{\text{ss}}^{1-\theta} \quad \Longrightarrow \quad P_{\text{ss}}^{\theta} = \frac{z}{2b} \quad \Longrightarrow \quad P_{\text{ss}} = \left(\frac{z}{2b}\right)^{1/\theta}.$$

其余一切都靠代入得到。由 (7.5)，稳态人均收入为

$$y_{\text{ss}} = \frac{z}{P_{\text{ss}}^{\theta}} = \frac{z}{z/(2b)} = 2b,$$

再用最优规则便得到稳态的生育率与消费，

$$n_{\text{ss}} = \frac{y_{\text{ss}}}{2b} = 1, \qquad c_{\text{ss}} = \frac{y_{\text{ss}}}{2} = b.$$

定理 7.3: 马尔萨斯稳态

在土地存量固定、技术为 z、且面对上述家庭问题的情形下，经济收敛到唯一的稳态，其人口为

$$P_{\text{ss}} = \left(\frac{z}{2b}\right)^{1/\theta},$$

而在该稳态上，人均收入、生育率与人均消费分别为

$$y_{\text{ss}} = 2b, \qquad n_{\text{ss}} = 1, \qquad c_{\text{ss}} = b.$$

这个稳态有两个特点值得着重强调。第一，$n_{\text{ss}} = 1$：每个家庭恰好养育一个存活的孩子，于是人口刚好自我替代、停止增长。这正是对均衡最自然的读法——稳定的人口并非被假设进来，而是被经济学*生产*出来的。当土地被瓜分到收入跌至 $y_{\text{ss}} = 2b$ 时，各家恰好只养得起一个孩子，人口于是稳定下来。第二，也更引人注目：稳态的生活水平 $c_{\text{ss}} = b$ 仅仅被钉在养育孩子的成本上。农业经济里的生活水平徘徊在一个由生物学与繁衍成本决定的位置，而不取决于这个社会技术有多高明。

7.3.2　为何技术买来的是人口，而非繁荣

这个模型的回报，正在于它对技术进步的论断。设技术水平 z 永久上升——更好的种子、更重的犁、新的轮作制。看看那些稳态表达式，问问什么在动。

稳态人口上升：

$$P_{\text{ss}} = \left(\frac{z}{2b}\right)^{1/\theta} \uparrow \qquad \text{当} z \uparrow .$$

但人均收入、生育率与人均消费却是

$$y_{\text{ss}} = 2b, \qquad n_{\text{ss}} = 1, \qquad c_{\text{ss}} = b,$$

其中没有一个含 z。技术从每一个福利度量里都消失了。**更好的技术抬高了这块土地所**

能供养的人数，却抬不高其中任何一个人所享的生活水准。这恰恰就是我们一开始要解释的那个典型事实，如今它是被推导出来的、而非被断言的：在马尔萨斯世界里，进步是以人口的形式支付出去的。

把这套机制顺着模型走一遍，可以干净地陈述成一条链。更高的 z 在人口给定时抬高收入 y_t；更高的收入抬高生育 $n_t^* = y_t/(2b)$；更多的孩子扩大人口 $P_{t+1} = n_t P_t$；更大的人口意味着更少的人均土地 $L_t = 1/P_t$；而更少的人均土地又把收入 $y_t = z/P_t^{\theta}$ 压回去。经济就沿着这条链滑动，直到收入回到 $2b$、生育回到一——只不过此时人口更大了。更好的技术带来的那点暂时繁荣，到头来被完全转化成了多出来的嘴。

注（控制变量会跳，状态变量只爬）.

> 这里值得留意不同变量的调整速度。生育与消费是控制变量：家庭每一期都瞬间重置它们，所以当 z 跳升时，收入和生育可以随之跳升。人口则是状态变量，是一个只能通过出生与死亡缓慢变化的存量。于是一次技术改进会带来即刻的繁荣井喷与一波生育潮，随后是人口向新的、更高的稳态长久而缓慢的攀升——在这段攀升中，早先的繁荣被一点点侵蚀殆尽。好光景是真实的，却是过渡性的；多出来的人口则是永久的。

7.4　逃出陷阱：孩子的时间成本

刚才搭起来的这个模型冷酷无情：每一步进步都被人口吞没，生活水平永远逃不出维生线。然而它们分明逃出去了——从工业化开始。这套机制里一定有什么东西断掉了。模型最重要的那个不足，恰恰指向了答案。

我们曾假设孩子的唯一成本就是物质成本 b。但养孩子还要花时间——在一个劳动能挣到工资的经济里，那些时间本可以用来工作。因此一个孩子的真实成本，是花在他身上的物质加上照料他时放弃的工资。设 w_t 为工资，每个孩子占用一单位时间中的份额 m，则每个孩子的放弃收入为 $m\,w_t$。把预算按家庭的时间禀赋以工资计价重写，便成为

$$c_t + (b + m\,w_t)\; n_t = w_t. \tag{7.8}$$

右端是家庭把全部时间禀赋都用来工作时的价值；$(b + m w_t)$ 这一项则是一个孩子的完全成本——物质加机会成本。这里决定性的新特征在于，这个完全成本随工资上升：在更富的经济里，一个孩子所耗用的时间更值钱，于是孩子变得更贵。

像之前一样求解期内问题——对数—对数偏好再次把支出在消费与孩子之间对半分——得到

$$c_t^* = \frac{w_t}{2}, \qquad n_t^* = \frac{w_t}{2\,(b + m\,w_t)}.$$

现在追踪一下工资上升时会发生什么。消费无上限地上升，但生育不会。当 $w_t \to \infty$ 时，物质成本 b 相对于时间成本变得微不足道，于是

$$n_t^* = \frac{w_t}{2\,(b + m w_t)} \longrightarrow \frac{1}{2m}.$$

生育不再随收入失控暴涨；不断上升的孩子机会成本给它封了顶。那条曾经驱动马尔萨斯陷阱的失控反馈——越富意味着越多孩子、越多孩子意味着越穷——已在源头被拆

除。随着工资继续上升，家庭愿意投在孩子单纯数量上的资源份额不再增长，转而开始把收入替换到别的用途，包括对每个孩子投入更多。生育在收入上升时这样被压弯、封顶，便是人口转变。

人口转变

一旦孩子既有物质成本又有时间成本，孩子的完全价格 $(b + mw)$ 便随工资上升。于是当经济变富时，生育不再无界地上升——它转而被压平，并渐近趋于 $1/(2m)$，而不是随收入暴涨。那条失控的马尔萨斯反馈被打断，人口会收敛，而不再永远跟着技术一路向上。富裕国家里观察到的生育率彻底下降，则来自另一个更进一步的机制，即下文注释中展开的数量—质量取舍。这二者合在一起，就是人口转变（*demographic transition*）。

注（为何时间成本只压弯生育，以及由谁来收尾）.

朴素的对数—对数模型加上时间成本，给出的生育是 $n^* = w/[2(b + mw)]$，它其实关于工资是递增的，只是上有 $1/(2m)$ 的界。所以时间成本本身只是移除了那条失控的马尔萨斯反馈、给生育封了顶；它并不能单独造出我们在富裕国家观察到的生育下降。真正带来彻底下降的那一块，是数量—质量取舍。一旦家庭还能选择对每个孩子投入多少——教育、健康、人力资本——更高的工资就会在边际上让质量更具吸引力，家庭于是从“多生、少投”替换向“少生、多投”。因此人口转变最好被理解为孩子的数量与质量之间的取舍：人力资本进入了生产函数，孩子的质量变得值得购买，被选择的生育率随之下降。

同一套逻辑还解释了现代经济内部一个有充分记录的模式：越富的家庭、尤其是收入越高的女性，往往孩子越少。当脑力劳动取代体力劳动成为生产的关键投入、当女性工资上升时，孩子所需时间的机会成本随之攀升——而一旦数量—质量取舍让家庭得以替换向“投资于每个孩子”，被选择的孩子数量便下降。上升的时间成本压平了收入—生育关系；数量—质量取舍则把它倒转过来。那个孩子廉价、繁荣只会催生更多孩子的马尔萨斯世界，已被彻头彻尾地翻了过来。

注（关于人口政策的一点说明）.

人们很容易把马尔萨斯模型读成压制人口增长的依据，二十世纪的人口控制项目有一部分理由正是这样讲的。但能为这种主张背书的，是那个农业模型——其中产出被一个固定要素封顶，多一个人就是同一块地上多一张要喂的嘴。而一旦经济离开了农业——一旦它像索罗模型和新古典模型那样靠资本、想法和人力资本运转——这套逻辑就不再适用。人并不只是一块固定大饼的消费者；他们产生巨大的正外部性：发明、分工，并扩张着马尔萨斯模型当作常数按住的那个技术 z 本身。把人口当成纯粹的负担，等于把马尔萨斯式的直觉硬塞进一个早已超越它的世界。这个模型是对农业过去的忠实描述，而不是对工业现在的处方。

至此，土地与内生人口都已在手，我们也就走完了本课程长期、实物的那一面：索罗模型、它的微观基础版本，以及先于现代增长的那个马尔萨斯世界。接下来的几章将从实物经济转向名义经济——货币、银行与价格水平——从第 八 章开始。

第八章　货币、银行与货币政策

到目前为止，宏观经济学里我们研究过的几乎一切——产出、资本、劳动、增长——原则上都可以在完全不提“货币”的情况下讲清楚。第 五 章和第 六 章的索洛模型与新古典模型，整套都是用实际产品来书写的。然而，没有任何一个现代经济体能离开货币运转；而政府在短期内为管理宏观经济所*真正做*的事，绝大部分都是通过中央银行来完成的。本章讲的就是这套“管道系统”：货币是什么，银行体系如何制造出其中的大部分，以及中央银行用哪些工具来扩张或收缩信贷供给。

这个主题不可避免地比之前的模型更偏制度性。凡是有干净结构的地方——资产的流动性排序、货币乘数、央行资产负债表——我们都会把它精确地讲出来。但货币政策的很大一部分，是“谁、通过哪种法律工具、对谁、做了什么”的问题，而这些安排在各国之间差异极大。本章贯穿始终的两个参照体系是：美国，其央行主要通过国债市场运作；中国，其央行主要直接通过商业银行运作。我们会比较细致地处理中国特有的这套机制，一来因为课程如此，二来因为它更清晰地展示了一家央行如何借助银行渠道来驾驭一个经济体。除非另有说明，下文中的“央行”指的都是中国人民银行（PBoC）。

有两个看似该放在本章的主题，被安排到了别处。消费者价格指数和生产者价格指数（CPI 与 PPI），即我们对价格水平的度量，已在第 三 章建立。货币数量论、通胀税与铸币税——这些是*印钞*所带来的后果，而非*管理*货币的机制——则是第 九 章的内容。这里我们只谈制度。

8.1　货币与流动性

理解货币最好的方式，是把它看作一种*资产*——一件可以持有的、有价值的东西——它与其他资产的区别，首要在于一项性质：流动性。一项资产的流动性，指的是它能在多大程度上被又快、又便宜、又以可预期的价值兑换成商品。手中的现金是这条谱系的极端情形：它处处被接受、即时生效、按面值兑付。一张政府债券、一份共同基金份额或一只上市股票同样是价值储藏，但要把它变成可花的购买力，需要时间，而且可能被迫在不利价位上卖出。货币的定义性特征，就是它处在这条流动性谱系的高流动性一端。

让一项资产被接受为货币的，归根结底是*信用*（credibility）：人们之所以肯收下它来交换，只是因为他们确信别人也会再从自己手里收下它。这种接受面越广，资产的流动性就越高。图 8.1把主要的金融资产沿这一维度排了序。

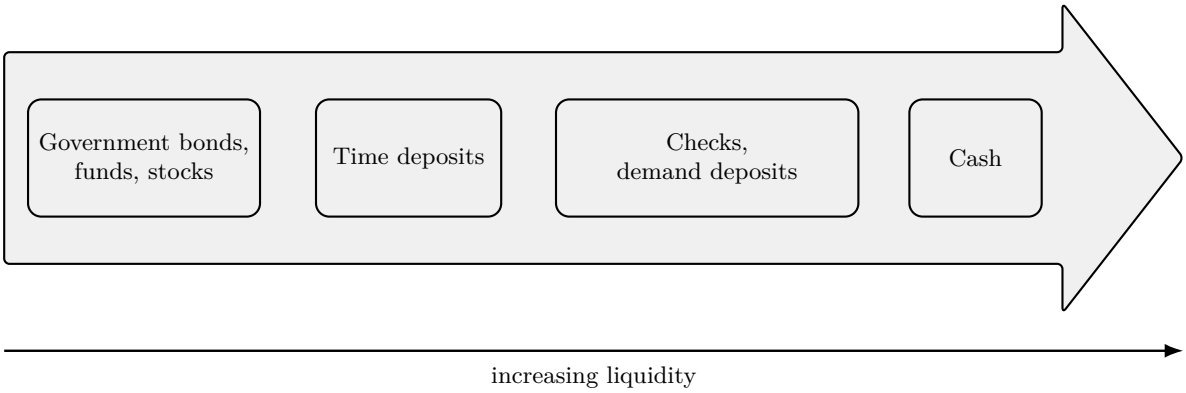

图 8.1: 流动性谱系：从最左端流动性最低的资产（债券、基金份额、股票），到最右端流动性最高的资产——现金。

定义 8.1: 流动性

一项资产的*流动性*（liquidity），是指它能在不损失价值的前提下被转换为普遍接受的支付手段的难易程度。货币按定义就是流动性最高的一类资产；而现金是货币中流动性最高的形态。

一个有用的推论是：流动性是被定价的。一项流动性较低的资产，必须提供一个*溢价*（premium）——更高的预期收益，或一个价格折让——才能与流动性更高的资产并存被人持有；否则没人愿意忍受持有它的不便。“低流动性资产相对高流动性资产折价交易”这一条想法，本身就组织起了银行体系与央行所做的大部分事情。

注（[）.

> 银行业即一场流动性转换] 银行贷款是有抵押的：借款人质押一项低流动性资产（一套房子、一只债券、一座工厂），换回高流动性的资金。借款人那笔不易变现的财富，实际上被换成了流动的购买力，而即便没有任何新的实物资产被创造出来，实体经济可动用的*总*流动性也上升了。央行在更上一层做的是完全相同的事：它以政府债券等抵押品向商业银行放款，把银行那些流动性较低的持有物，换成流动性最高的那种资产。调控全社会的流动性，正是央行日常的本职工作。比如，当财政部征税时，它是在从私人部门抽走流动性；央行通常会对应投放基础货币来对冲，使这一轮征税不会造成一次意料之外的货币紧缩。

8.2　货币层次

由于“货币”涵盖了流动性各异的资产，统计部门并不会用单独一个数字来报告它。它们报告的是一族嵌套的*货币层次*（货币总量，monetary aggregates），每一层都在里一层之上，再添上一层流动性更低、更像存款的工具。图 8.2展示了标准的嵌套关系，$M_0 \subset M_1 \subset M_2$。

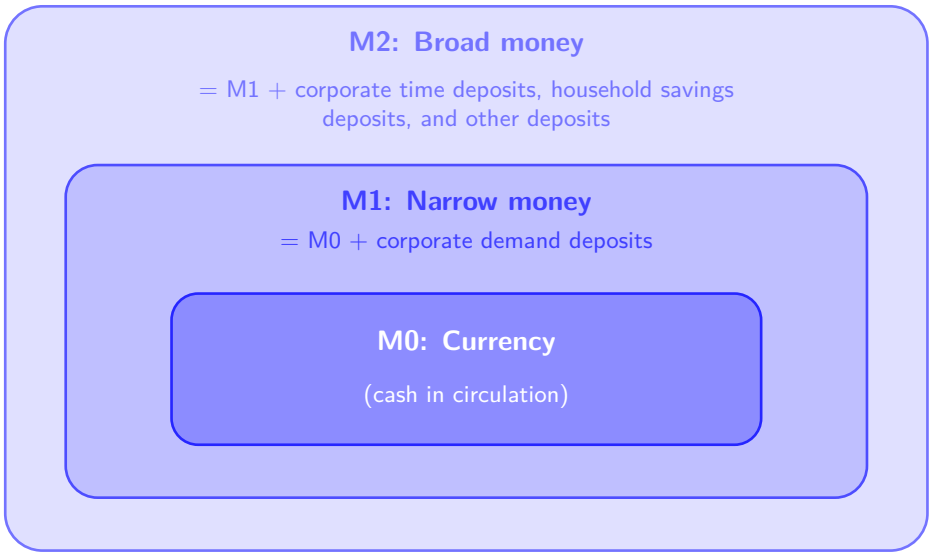

图 8.2: 嵌套的货币层次：通货 M_0 套在狭义货币 M_1 之内，而 M_1 又套在广义货币 M_2 之内。

定义 8.2: 货币层次

货币层次按流动性把货币排序：

- M_0 ——流通中的通货：处在银行体系之外的实物现金，是流动性最高的部分。
- M_1 ——“狭义货币”：M_0 加上活期（可开支票的）存款，这类存款能在短时间内动用。
- M_2 ——“广义货币”：M_1 加上定期存款与储蓄存款，这类存款流动性略低。

每一层都严格包含它里面的各层。

通货部分 M_0 规模很小，主要受短期、季节性因素（节假日、发薪周期）驱动。对政策真正重要的层次是 M_2。现金只占 M_2 中很小的一部分；现代经济中绝大部分“货币”根本不是纸钞，而是*存款余额*——银行账面上的记账。一句话说：货币就是账面上的钱。

注（[）.

> 持有的货币与社会上的货币] M_0、M_1、M_2 度量的都是公众*实际持有*的货币。这既不等同于整个体系所能创造出的货币总量，也不等同于“财富”。这些层次量度的是某一时点上流通中的流动性债权存量，仅此而已；不要把它们读作经济中全部购买力的一次普查。

8.2.1　为何 M_2 是政策目标

M_2 是观察货币增长的主视角。M_2 的**增速**是货币政策的头条度量；在中国，每年“两会”常把这一目标定在大约百分之十。作为经验法则，M_2 的增速被期望与名义 GDP 的增速大体匹配——新增的货币恰好够为新增的交易融资，不多不少。但严格说来，这条规则不可能机械地成立，因为货币政策的全部要义恰在于**逆周期**：央行会刻意让货币在衰退期增长得快于产出、在繁荣期增长得慢于产出。一个每季度都简单地跟住名义 GDP 的货币供给，根本就没在做任何稳定经济的工作。

注（[）.

> M_2/GDP 之比很容易被误读] 中国的 M_2 与 GDP 之比超过 200%，而美国的不到 100%。人们很容易由此断定中国流动性泛滥、而美国没有。但这一比较并不支持这个结论，理由有好几条。
>
> 第一，两国央行的行事风格大不相同：中国人民银行在扩张货币供给上一向很收敛，而美联储则是大开大合（危机时大幅降息，事后再反向操作）。第二，也更根本的，是 M_2 被创造出来的**渠道**不同。在美国、英国这样的普通法经济体中，**直接金融**占主导：企业主要靠发行债券和股票来筹钱，这些都不会穿过、也不会推高银行存款类的货币总量。中国则依赖**间接金融**：企业向商业银行借款，而每一笔这样的贷款都会派生出一笔与之对应的存款，于是一个信贷驱动的经济体，机械地就会背负一个相对 GDP 更大的 M_2。商业银行是中国金融体系的中枢。第三，M_2 是**存量**，GDP 是**流量**；二者之比混用了量纲，只有它们的**增速**才可能被有意义地比较。最后，改革开放以来的快速市场化，把许多原本非市场化的活动逐步货币化了，这也出于与“过剩”无关的原因把 M_2 推了上去。

8.2.2　贷款与社会融资

M_2 与银行信贷同涨同落，因为在间接金融体系中，新增贷款是新增广义货币的主要来源。除了 M_2 的两大教科书驱动因素——基础货币的供给与货币乘数（在下一节展开）——之外，M_2 增长还可能因**需求**侧的原因而停滞：实体经济可能不想借（信贷需求疲弱），或者银行可能不愿放（“信贷紧缩”，银行囤积流动性）。因此，把两个相关的概念分开来谈是有用的。

宽货币与宽信用

宽货币（“货币宽松”立场）指的是央行增加基础货币的供给。**宽信用**（“信用宽松”立场）指的是实体经济确实更容易获得融资了——信贷的可得性变宽了。二者并不是一回事：如果宽货币没能转化为宽信用，那就说明**货币传导机制**受阻，多投放的基础货币会淤积在银行体系里，到不了企业和居民手中。

有几个公开发布的数据序列，能让分析者直接观察信贷渠道。**人民币新增贷款**是观测间接金融的关键指标，分为住户贷款与企业贷款两块。住户贷款以消费贷为主；近年来消费贷与信用卡余额的违约率急剧上升。一个更宽口径的度量是**社会融资规模**。

定义 8.3: 社会融资规模

社会融资规模度量的是金融机构向实体经济——向企业与个人——提供的全部融资，并不包括金融机构彼此之间的借贷。它既涵盖通过银行体系的间接融资（贷款、信托贷款、未贴现银行承兑汇票——后两者属于“表外”），也涵盖通过资本市场的直接融资（债券与股票）。普通的银行贷款记为“表内”；信托贷款与委托贷款记为“表外”。

市场尤其紧盯三个头条数字——M_2、人民币新增贷款、社会融资规模——通常在每月的 10 日至 15 日之间公布。

8.3　货币的类型

货币说到底，是一个社会共同同意去讲的一个故事。是什么赋予某个特定物件以货币价值，这在历史上一直在变，而这些差异本身很有启发性。

商品货币。　商品货币有内在价值：物件本身就值点东西（黄金、白银、铜），所以它的货币价值锚定在它的材料价值上。因此，商品货币的供给量由生产技术决定——比如能开采出多少白银——无法随意扩张。

注（[）.

> 纸币比你想的要古老] 2022 年是交子问世一千周年，交子是中国早期的一种纸币。中国长期受困于贵金属稀缺，古代大多使用铜钱，白银很少；宋朝甚至试用过铁钱，其一大弊病是会生锈。纸币的早早出现本身就是一个信号：它恰恰出现在商业活动已经活跃到现有的贵金属存量不足以结清所有交易的时候。纸币正是对流动性短缺的一种回应。

法定货币。　法定货币没有内在价值——一张钞票不过是印了字的纸——它被接受的原因有二。其一是发行者的信用：发行机构可信到足以让它的票据流通（历史上，那些声誉极高的私人银行的票据）。其二，也是现代货币的根基，是国家的信用：主权力量强制其被接受，票据由央行发行。央行甚至可以把发行权委托给指定的、可靠的商业银行——香港没有央行，其货币就由发钞银行发行。

流通中的现金是央行的负债，是持有者的资产——这一点我们待会儿读央行资产负债表时会用到。布雷顿森林体系瓦解以来，货币就成了纯粹的法定（信用）货币，不再可以兑换成黄金。

信用货币可以随意创造

商品货币的供给由生产能力（白银的开采速度）固定。法定（信用）货币的供给则是人的决定：它基本可以无限制地扩张，而一个信用货币体系只要国家的信用还在，就能维系。一旦那份信用崩塌——公众不再信任本币、转而投奔外币——结果就是恶性通货膨胀。所以这种对货币“透支”的余地虽是真实的，却被信心所

限。

数字货币与加密货币。　一种加密货币是被私下认可的货币，其供给由算法固定。由于供给固定、而物件本身又一文不值，它的价格波动极大。它的定义性特征是*供给固定*与*去中心化*——没有央行——而不在于它仅仅是“数字的”（电子支付与线上转账同样是数字的，但那些是对普通银行存款的债权）。在这一点上，比特币很像黄金：黄金本身不带来任何收益，唯一的回报就是其价格的变动，这正是为什么二者都是投机的对象，而不像股票或债券那样是会产生收入的投资。这类资产把价值储藏功能与交易功能结合在一起，因而带有一定的价值；而在地缘政治风险加剧的世界里，这种价值还可能增长。

8.4　货币传导机制

一项央行行动是如何抵达一个普通借款人的？答案——*传导机制*——在不同国家要经过不同的中间市场。

	传导链条
中国	央行 → **商业银行** → 社会
美国	央行 → **国债市场** → 社会

这一对比，恰恰就是前面那个直接金融与间接金融的区别。在美国，央行作用于国债市场——一条直接金融渠道——其余的经济部门再根据国债收益率重新定价。在中国，央行作用于商业银行——一条间接金融渠道——银行再把这一脉冲传递给企业和居民。

美国的这条渠道立足于一条资产负债表恒等式。*国债是央行的资产，现金是央行的负债。*当美联储*买入*债券时，它的资产负债表*扩张*，并用新创造的货币来支付债券价款，从而抬高货币供给。当它*卖出*债券时，资产负债表*收缩*，货币被从流通中抽走。所以买债即放松、卖债即收紧。美国国债市场的主要交易对手本身也是商业银行；其独特之处在于美联储是作为又一个普通交易方平等地参与买卖，而非直接向银行放款。由于现金只占 M_2 中很小的份额，央行里最重要的部门并不是负责现金发行的，而是执行这些市场操作的货币政策司。

8.5　准备金制度与货币乘数

现在我们触及现代银行业的那个核心机械事实：*创造出大部分货币的，是商业银行，而不是央行。*它们靠准备金制度来做到这一点。一家银行收到一笔存款，只被要求把其中一小部分作为*准备金*留下，余下的可以放贷出去；这笔贷款会在别处变成一笔存款，又被大体放贷出去，如此往复。一单位基础货币撑起的是其若干倍的存款。乘数量化的就是这个“若干倍”。

定义 8.4: 货币乘数

记 C 为公众持有的通货，D 为存款，R 为银行准备金。广义货币为 $M = C + D$，而基础货币（也称高能货币）为 $B = C + R$。货币乘数就是下面这个比值

$$\text{货币乘数} \;=\; \frac{M}{B} \;=\; \frac{C+D}{C+R} \;=\; \frac{c+1}{c+r},$$

其中 $c = C/D$ 是通货存款比，$r = R/D$ 是准备金率。第二个等号是把分子分母同除以 D 得到的。

这个推导值得做一遍，因为代数本身就是全部内容所在。从 $M = C + D$ 与 $B = C + R$ 出发，把 M/B 的分子分母同除以存款 D：

$$\frac{M}{B} \;=\; \frac{C/D + D/D}{C/D + R/D} \;=\; \frac{c+1}{c+r}.$$

这个表达式的两个特征承载了全部经济学含义。第一，通货 C 是从银行体系里漏出去的货币；它躺在钱包里，无法被再次放贷，所以通货存款比 c 越高，乘数就越小。实践中 c 很小，因为公众持有的现金相对存款而言很少。第二，也是那个政策杠杆，就是准备金率。

定理 8.5: 准备金率越低，乘数越高

保持通货存款比 c 不变，货币乘数 $\dfrac{c+1}{c+r}$ 关于准备金率 r 严格递减。r 越低，银行就能把每笔存款中更大的比例放贷出去，从而支撑起更多轮的存款派生，于是从同样的基础货币里生成更大的 M_2。

证明. 将 $m(r) = (c+1)/(c+r)$ 关于 r 求导：

$$\frac{\mathrm{d}m}{\mathrm{d}r} = -\frac{c+1}{(c+r)^2} < 0,$$

因为 $c + 1 > 0$。故 m 随 r 上升而下降、随 r 下降而上升。 □

这正是为什么法定准备金率是一项货币政策工具：在基础货币不变的情况下下调它，会机械地抬高 M_2。而基础货币本身由央行供给。

定义 8.6: 基础货币

基础货币（高能货币）是央行直接创造出来的货币——“最初的钱”。在间接金融体系里，它在央行向商业银行放款时进入经济。央行向银行释放的一切资金都是基础货币。由于它随后会被货币乘数放大，基础货币会生成一个远大于其自身的 M_2。因此，M_2 的增长主要来自两个来源：基础货币的供给与准备金率。

8.5.1　银行为何会把准备金压得过低：道德风险与准备金红线

准备金是在安全与利润之间做权衡。持有更多准备金的银行更安全、但赚得更少，因为闲置的准备金几乎不生息；持有更少准备金的银行赚得更多、但更容易遭遇挤兑。每家银行只在意自己的利润，而利润是私有的、排他的。然而，安全却带有**外部性**：一家银行的倒闭可能在整个体系中层层传染。由于知道金融体系会动员起来救助某个陷入困境的成员，每家银行都有动机搭便车——把自己的准备金率压得更低、攫取那份利润，而让系统性的安全由别人来提供。这是一个经典的**道德风险**（moral hazard），它还会孳生逆向选择：在没有管制的情况下，活下来的银行恰恰是风险最高的那些，因为它们赚得比稳健的同行多。

注（[）.

> 准备金红线即一条互助规则] 制度上的应对，是一条协调一致的准备金**红线**。各家银行实际上结成了一个带有红线的互助会：如果某个成员陷入困境时，其准备金水平**高于**红线，互助会就有义务去救它；若其准备金**低于**红线，则不予救助。这条红线恢复了持有足额准备金的激励。历史上，央行所扮演的正是这个最后贷款人、协调救助的角色，它由各成员银行所有，就像一个俱乐部或行业协会。比如美联储就是有股东的——那些成员银行——它把大部分利润上缴财政部、并分一部分给那些股东；然而无论是股东还是总统都无权指挥它的操作。一家现代央行立足的是国家的信用，而不是它的所有权。

8.6　中央银行

央行本身也是一家银行，但它不以盈利为目的——尽管它在现实中盈利能力极强。它只与商业银行（以及少数指定机构）打交道，其利润来自**利差**，主要是它就准备金支付的利率与它从放贷便利中所赚的更高利率之间的差（在中国，即准备金利率与 MLF 利率之间的利差）。

注（[）.

> 谁可以在央行开户] 只有银行和少数非银机构可以在央行开设账户。这里的“银行”包括商业银行与**政策性银行**，后者中最重要的是国家开发银行（国开行）。政策性银行为基础设施这类规模大、利润薄、战略地位重要的项目提供资金；国开行可以自行发债（国开债），其资金最终由央行提供。中央财政部也可以开设账户——这相当于国库。央行还会指定一份**一级交易商**名单，它们通常是在央行总部开户的大型商业银行；央行主要与这些一级交易商交易，再由它们向银行体系的其余部分转融通资金。

由于央行不可能持续审查每家商业银行的现金，银行把准备金存放在**设于**央行的准备金账户里，而准备金水平低于法定要求的银行会受到处罚。由于一家银行的存款时刻在波动，稳健的银行会在法定最低限之上多持有一些**超额准备金**作为缓冲。央行会对超额准备金支付利息，但——为了防止银行仅靠把资金停泊在央行就获利——超额准备金利率不可能超过法定准备金利率，而且超额准备金事实上还设有上限。

超额准备金利率是利率走廊的下限

超额准备金的利率确定了政策利率走廊的*下限*：没有银行会以低于它在央行所能无风险赚到的利率去拆出隔夜资金。当超额准备金利率下降时，市场利率会随之下降。整个体系中较低的超额准备金率，意味着流动性吃紧——银行手头的闲钱很少。

注（[）.

央行代表谁的利益] 央行代表的是*货币资产的持有者*——所有持有货币或类货币债权的人——而*断然不*代表政府本身。当它放贷、把低流动性资产换成高流动性资产时，它服务的是这些债权人。它对独立性的坚持，恰恰是一种拒绝直接服务于政府融资需要的态度：一家被财政当局俘获的央行，只会去为政府赤字货币化。在中国，央行不可以直接*购买*国债，但可以把国债作为*抵押品*；而它发行的货币也不必与抵押品一一对应——它基本可以相机地创造信用货币。一些国家偏向于财政货币化，这一想法扎根于*现代货币理论*（MMT，Modern Monetary Theory），该理论认为一个信用充足的主权国家，其货币发行的上限比通常想象的要高得多，而央行应当配合财政政策。中国人民银行强力捍卫自己的独立性，坚决反对财政货币化。

8.6.1　法定职责与组织结构

简而言之，货币政策的运作，是通过设定三个工具——货币供应量、政策利率、准备金率——来驾驭三个目标：M_2、社会融资规模、人民币新增贷款。一家央行的法定职责框定了它要实现什么，而目标的排序很关键。

中国人民银行的职责：币值稳定优先

中国人民银行是政府部门，在国务院的指导下制定和实施货币政策。它的目标是*维持币值稳定，并以此促进经济增长*。币值稳定——也就是控制通货膨胀——是*首要*目标；经济增长在其后。一条实务性的推论是：凡是收紧的，往往是央行自己发起的；而凡是宽松的，往往是政府向它施压促成的。

央行并不对经济增长直接负责；在中国，主管经济增长的牵头机构是国家发展和改革委员会，其次是财政部。央行对利率的调整也很谨慎：如果无风险利率定得太高，有风险的创业投资就会被挤得没有发展空间。

美国联邦储备体系尽管有私人股东，却行使着国家机构的职能。它的理事会总部设在华盛顿，由七名理事组成，而美联储主席的任期被刻意与总统任期错开。主席不是内阁成员，总统对美联储没有直接的行政命令权。

定义 8.7: FOMC 与联邦基金利率

联邦公开市场委员会（FOMC）制定美国的货币政策。它有 12 名投票委员：7 名理事，加上纽约联邦储备银行行长，再加上其余 11 位地区储备银行行长中轮值的 4 位（$7+1+4=12$）。它大约每六周开一次会。当美联储“加息”或“降息”时，它作用的是联邦基金利率——美国各银行之间的隔夜同业拆借利率。

在市场化的体制下，美联储无法真的指定联邦基金利率；它宣布的是一个目标区间。为了把实际利率维持在区间之内，它使用公开市场操作来调节流动性，而这进一步会移动诸如十年期国债之类较长期工具的价格。实际上美联储很少直接向商业银行放款，也无法靠下令来固定大多数利率——它只能去引导一个区间。相比之下，中国的央行则与银行频繁交易，可以直接确定利率。

注（[）.

> 应对危机与对货币政策的过度依赖] 美联储应对危机的标准套路，是大幅降息、向体系灌入流动性，并接受其后的再通胀：2001 年互联网泡沫破裂与“9 · 11”之后、2008 年金融危机、以及 2020 年的疫情，都是这个模式。课程反复出现的一个主题是：西方经济越来越过度依赖货币政策，因为财政政策很难施展——像 2008 年之后的危机复苏，主要就是靠央行续命。一个治理上的对比把这一点讲得更尖锐：美国的货币政策由一个绝缘的技术官僚精英集团掌控，它快、但不民主；而对面是民主、却迟缓的国会。草根民主与精英专制之间的平衡在许多场合反复出现；危机当头时，美国只能通过这条快速的、技术官僚式的货币渠道来行动。

8.6.2　基于规则的政策与相机抉择的政策

各国央行在可预测程度上各不相同。

定义 8.8: 基于规则的政策与相机抉择的政策

在基于规则（rule-based）的政策下，央行遵循一条透明的、事先承诺好的规则，因而市场可以预判它的动作，央行也不会让预期落空。最典型的例子是泰勒规则（Taylor rule），它把目标利率设为通胀与失业（或产出）缺口的函数。在相机抉择（discretionary）的政策下，目标和工具被刻意保持模糊，因而市场无法对下一步动作形成精确预期。

美联储收紧的意图是公开摆明的——它大体遵循一条以通胀和失业为锚、经过优化的泰勒规则——这使得美国的政策是基于规则的。中国的政策目标则相对不透明，通过利率、准备金率、LPR 等一套工具来推进，这使得它是相机抉择的。

注（[）.

> 近年美联储周期的一条粗略时间线] 互联网泡沫破裂与“9 · 11”导致美国在 2001 年大幅降息，2005 年之后才恢复。2007 年的次贷危机迫使又一轮深度降息，并把利率长期维持在 0.25% 附近。2017 年之后，经济充分恢复，美联储进入加息周期；2019 年，正值繁荣，特朗普总统却屡次强势要求降息。2022 年的加息比 2008 年之后那

> 一轮更陡峭，并且是由市场两侧共同驱动的：需求侧的经济复苏，以及供给侧的压力——劳动参与率下降，加上俄乌战争触发的能源价格飙升。如果通胀缓解，美联储可能会提前退出加息路径。

8.7　央行的资产负债表

央行靠改变自身资产负债表的规模与构成来驾驭经济。会计账目是对它所作所为最干净的概括。

如何读一张央行资产负债表

在央行的资产负债表上，流通中的现金是负债，而债券（及其他买入的资产）是资产。因此：扩表——买入资产，并创造准备金和现金来支付——会向体系增添流动性，把利率往下压；缩表则抽走流动性，把利率往上推。

下面是两张参照性的资产负债表，以 T 形账户呈现。

美联储

资产	负债
债券（以国债与 MBS 为主）	流通中现金（主要） 银行准备金

注（[）.

> MBS 是怎么进入美联储资产负债表的] 2008 年危机之后——其直接诱因是被打包成住房抵押贷款支持证券（MBS）的次级债——美联储于 2009 年买入了大量 MBS 来救市。这正是为什么 MBS 会与国债并列，成为美联储资产负债表上的一项主要资产。

中国人民银行

资产	负债
外汇占款 债券	流通中现金（主要） 银行准备金

中国人民银行资产负债表上最醒目的特征，是资产侧那一大笔外汇占款，主要是为了维持金融体系的稳定而持有的。下面讲中国的政策工具箱时，我们会看到这些外汇占款的累积是如何在十多年里驱动了央行的资产负债表。

8.8　常规与非常规工具

除了设定准备金要求、开展市场操作之外，央行还备有一套梯度分明的工具，从例行的一直延伸到应急的。

贴现窗口。　除公开市场操作之外，美联储也会直接向金融机构放款；这就是*贴现*，其贷款利率即*贴现率*。贴现率通过投标确定（尽管美联储对它有很大影响力），银行再据此决定借多少。

量化宽松。　当经济转入下行——尤其是当政策利率已经接近零时——美联储就转向*量化宽松*（QE）：大规模直接买入国债与住房抵押贷款支持证券（MBS）。

定理 8.9: 量化宽松如何运作

通过大批量买入长期证券，央行抬高了它们的*价格*，从而压低了它们的*收益率*。较低的长期收益率会拉低长期贷款的利率，使长周期的投资更具吸引力。于是，当短端利率已经触底时，QE 通过作用于收益率曲线的长端来放松金融条件。

流动性陷阱。　宽货币并不总是带来增长。

定义 8.10: 流动性陷阱

当过旺的流动性无法支撑实体经济增长、反而推高资产价格时，就出现了*流动性陷阱*（liquidity trap）。在流动性陷阱中，银行拿到的基础货币流不到实体经济——传导机制受阻——于是新增的流动性是去抬高金融资产与房地产价格，而不是去为新增产出融资。

注（[）.

> 2008 年的一条教训：要盯资产价格，而不只是 CPI 和 PPI] 2008 年危机的一条核心教训是：判断通胀不能只看 CPI 和 PPI，还必须盯住*资产价格*。危机前奉行的泰勒规则并没有把它们纳入进来，于是错失了在房地产中累积起来的资产价格通胀。

负利率。　欧洲和日本的结构性问题很严峻：它们缺乏国内的增长引擎，其央行只要条件允许就必须宽松，却即便在复苏期也难以收紧。它们的非常规应对，是*负利率*。这可以采取两种形式：央行与商业银行之间为负的隔夜拆借利率（比如 -0.1%），或者为负的超额准备金利率（比如 -0.1%，这是一项旨在把流动性从央行挤出、推向实体经济的收费）。债券出现负收益率也是可能的，不过那是一个结果，而非一项政策工具。

如果一个经济体在货币收紧时仍能保持增长，那就说明它有强劲的、自我维持的增长动力——这正是欧洲和日本所缺乏的那种健康情形。

8.9　中国的政策工具箱

中国的货币政策几乎完全通过银行运作，配有一套丰富而具体的工具。在介绍工具本身之前，课程先强调了两个贯穿性的区分。

第一个区分——*价格型政策*与*数量型政策*——结果并不像听上去那么泾渭分明，因为任何一笔数量操作都附带着一个价格，二者是绑定在一起的。更有用的区分，是*总量政策*与*结构政策*之分。

总量政策与结构政策

货币政策按定义就是*总量型*的：它没有指定的方向，作用于整个经济体（降息、买国债、向商业银行放款——并不指定资金最终去向何处）。相比之下，*结构*型政策把支持定向投放到某个特定的行业或用途。中国以往的总量宽松往往流向房地产，于是要让政策见效就难免“大水漫灌”整个经济；后来转向结构型工具，则是为了对目标行业实现“精准滴灌”。

8.9.1　法定准备金率

经典的数量型工具是*法定准备金率*（RRR）。在基础货币不变的情况下，下调 RRR 会抬高 M_2（通过抬高乘数，如上所证），上调 RRR 则会压低 M_2。

注（[）.

2015 年以来的差别化准备金率] 准备金率是第一个被赋予结构型形态的工具。从 2015 年起，中国采用了*差别化*的法定准备金率——一种强调宏观审慎管理、关注系统性风险的结构型政策。大型银行面对更高的准备金要求；小型银行面对更低的要求；城市与农村商业银行则享有优惠待遇。

注（[）.

为何 2012 年前后会那样使用 RRR] 中国的货币政策在 2012 年前后有过明显的转折。2012 年*之前*，国际收支表现为大量盈余，由贸易和外商直接投资（FDI）流入所驱动。国际收支盈余会造成*输入型通货膨胀*：中国不得不被动吸纳流入的外来资金，央行被迫扩张其资产负债表（即资产侧的外汇占款），从而冒着经济过热与通胀的风险。因此央行的主要任务是*回收*流动性，具体做法是逐步*上调* RRR，并*发行央票*——央行的借款凭证，在香港市场上交易。（顺带一提：人民币交易是一个封闭体系。离岸人民币 CNH 主要在香港交易；在岸人民币 CNY 则在境内管理；CNH 通常波动得更快、更大，尽管两者都可以被干预。央行通过在离岸发行央票，可以买进大量人民币、收回输入的流动性——但若持续这样做，离岸人民币池子会趋于枯竭。）

2012 年*之后*，国际收支转为平稳、甚至出现了适度流出：金融账户出现大额流出（FDI 流入减少、中国对外投资上升，以及来自中国股市和债市的证券投资流出）。人民币回流央行，流动性收缩，加之当时经济下行，央行的任务于是反转过来——现在要去*投放*流动性，主要靠*下调* RRR。更宏观的一条道理是：出口对中国的经济增长极为重要，而贸易逆差会造成严重挫伤；中国一直保持着贸易顺差，只在贸易战的几个月里偶有逆差。

8.9.2　公开市场操作：回购

中国的公开市场操作主要通过*回购协议*（回购）来进行，有两种形式，而且其命名与美国的用法正好相反。

正回购与逆回购（中国命名）

在中国，当*央行向商业银行放款*时，这一操作是*逆回购*；当*央行向商业银行借款*时，则是*正回购*。（这与美国的惯例正好相反。）逆回购需要抵押品，通常是国债——这又是一次以低流动性资产换取高流动性资产的过程，而*不是*直接的买入或卖出。标准期限是 7 天（最常见）、14 天和 28 天；没有隔夜回购。

回购是短期资金融通。每天逆回购的规模通常在百亿量级，于上午 9 时许公布，每天都有新发放的和到期的；在重要节假日和季末、月末，央行可能做 14 天逆回购以满足资金结算需求。如果当局希望对某一类债券表示支持，它们可能会*扩大合格抵押品的范围*：比如当市场担心地方政府融资平台（城投）债券违约时，央行接受更多此类债券作抵押，就发出了一个积极的信号。

公开市场操作利率即逆回购利率。它名义上由投标决定，实则事先就由央行商定并固定下来，央行直接“降”这个利率——这与只设定一个目标区间的美联储形成对比。

8.9.3　MLF 与 LPR

定义 8.11: 中期借贷便利（MLF）

中期借贷便利（MLF，俗称“麻辣粉”）是一项一年期的放款便利。央行通常在每月的 15 日开展 MLF，同时宣布规模与利率。

MLF 利率与逆回购利率同向变动——通常央行会先下调 MLF 利率——这两者是最重要的政策利率：当我们说“降息”或“加息”时，指的就是 MLF 与逆回购利率。两者之中 MLF 更有力。在紧急情况下，央行会打破常规、先下调逆回购利率，甚至不在通常的次日早晨那个时点：2020 年 2 月 8 日，逆回购利率在上午 9:15 公布，赶在股市 9:30 开盘之前，于是这一操作当天就能影响市场。央行也可以加做一次 MLF 来稳定市场——比如 2020 年 11 月末，在河南一家煤炭集团的债券违约可能引发动荡之后。

定义 8.12: 贷款市场报价利率（LPR）

贷款市场报价利率（LPR）是商业银行向实体经济放贷的利率——具体说，是向其*最优质*（prime，即最好的）客户给出的利率；其他客户在 LPR 的基础上加点。LPR 基于报价，并由 *MLF 利率加上一个点差*确定：

$$\text{LPR} = \text{MLF 利率} + \text{点差}.$$

其运作按月度循环展开：每月 15 日开展 MLF 并确定其利率，每月 20 日确定 LPR 报价。央行向 18 家商业银行询价，剔除极端报价后，取点差的算术平均。LPR 改革之后，政策向实体经济的传导路径为

$$\text{MLF 利率} \;\to\; \text{LPR} \;\to\; \text{贷款利率}.$$

注（[]）.

LPR 取代了旧的基准利率] 旧的贷款基准利率只是靠行政命令“空降”下来的；LPR 则立足于实际的市场报价，中国如今已不再使用基准利率。在 LPR 当中，MLF 利率扮演着昔日基准利率的角色。这项改革提高了市场化程度，但央行对 LPR 仍有很强的影响力：银行本身没什么动机主动把 LPR 报低，所以每到 20 日前后，央行会频繁地与它们沟通，以传递自己的政策意图。一年期 LPR 反映短期贷款；五年期 LPR 主要反映房贷。两者通常同向变动，但也可能被非对称下调——单独下调五年期 LPR，发出的信号是政策正转向支持房地产。

8.9.4　SLF、PSL 与再贷款

定义 8.13: 常备借贷便利（SLF）

常备借贷便利（SLF，俗称“酸辣粉”）是一项一个月以内的短期便利，央行借此向金融机构放款。它向所有金融机构开放，提供应急流动性，但利率非常高。7 天 SLF 利率被视为利率走廊的上限。实践中它用得很少。

所有金融机构都可以在央行开户，但级别有别：小银行可能只能在央行的分支机构开户，而一级交易商则在总部开户。以超额准备金利率为下限、SLF 利率为上限，二者共同界定了短期利率理应在其中交易的利率走廊。

定义 8.14: 抵押补充贷款（PSL）

抵押补充贷款（PSL）是一项定向（结构型）便利，发放给政策性银行——主要是国家开发银行。它通过政策性银行这条渠道，为有针对性的政府优先事项提供资金。

注（[）.

PSL 与房地产周期] PSL 在房地产“去库存”行动中居于核心——通过让城中村居民迁入来填补未售房产的库存。这一行动规模如此之大，以致拉动了房价的普遍上涨，尤其在二三线城市。2018 年中央经济工作会议以“房住不炒”的口号收紧房地产立场之后，PSL 规模缩减。2022 年它再度上升，主要是为国开行发动的基建投资融资；2022 年 7 月之后，面对大规模的房地产债务违约及其带来的社会维稳风险，央行向国开行发放 PSL，为“保交楼”工作提供资金。

既然国开债的收益率与国债收益率相近，为何不干脆通过普通的政府债券来为这些优先事项提供资金？因为财政的规模是事先定好的：2022 年初预算定得很小，而更宽松的财政立场无法在年中临时出台，于是这股财政脉冲只能借道国开行来释放。国开行因而在一定程度上替代了财政当局，绕开了全国人大的程序——这本质上就是央行把资金引向财政用途。其底层逻辑仍然是 MMT：只要国家的信用站得住，政府债务就被当作基本不受约束，由无限的货币发行权所支撑。

定义 8.15: 再贷款

再贷款是一项结构型便利，央行借此通过商业银行把资金引向特定的企业：商业银行向某个指定的企业放款，央行随后再以相同的数额向该商业银行放款。再贷款的规模超过五万亿元；例子包括支农、支小微企业、支抗疫的再贷款。

8.9.5　银行间隔夜拆借

资金一旦到达商业银行，银行就在银行间市场上管理它的流动性，在那里银行通过各种产品（包括回购交易）彼此借入和借出资金。有一点很重要，要把两件事分清楚。

银行间拆借与央行回购

银行间隔夜拆借的存在，是为了让银行彼此融通资金；央行回购的存在，则是为了让央行注入或回收流动性。央行回购没有隔夜期限——其最短期限是 7 天——而银行间拆借天然就是当天的隔夜业务。

定义 8.16: DR001 与 DR007

DR001 是隔夜银行间拆借利率——相当于美国联邦基金利率的中国对应物。严格说来，它是银行之间以利率债为质押的隔夜回购的加权平均利率。*DR007* 则是对应的 7 天加权平均回购利率。两者都是实际成交的利率。（相比之下，Shibor 只是一种报价，并不代表已实现的成交价格。）

这些基准利率把整条链条闭合起来。央行作用于基础货币和政策利率（逆回购、MLF）；银行通过银行间市场（DR001、DR007）传导这一脉冲，再借助 LPR 以及建立在其上的各笔贷款，把它传向实体经济。这条从央行、经由银行、抵达企业与居民的链条，正是本章开篇所讲的货币传导机制，如今已被一件工具一件工具地填充完整。

第九章　货币数量论与通货膨胀

第 八 章讲述了货币是如何被创造出来的——中央银行与商业银行如何携手扩张在经济中流通的通货与存款。本章顺势追问下一个问题：当这一存量增长时，价格水平会发生什么？简短的回答，也是古典货币经济学的核心思想，是：价格水平归根结底是一种货币现象。当货币数量的增长快于商品数量时，每一单位货币所追逐的商品比从前更多，价格便随之攀升。价格水平持续而广泛的上涨，就是我们所说的**通货膨胀**（inflation），而在长期内，它由货币的增长率所驱动。

在建立把这一点说清楚的模型之前，我们先停下来讨论一个学宏观经济学的人几乎总是搞反的前置问题：**通货膨胀究竟为什么重要？**流行的抱怨是，通货膨胀侵蚀我们的购买力，让我们变穷。我们将看到，这种直觉若照字面理解，在很大程度上是一种混淆——在一个无摩擦的世界里，收入会随着价格一同上涨，通货膨胀真正的实际成本另有所在。弄清这些成本究竟在哪里，以及哪里甚至还存在某些抵消性的好处，能让我们更清楚地理解货币政策能做什么、不能做什么。随后本章发展出**货币数量论**（quantity theory of money），它由方程 $MV = PY$ 概括；由此推导出通货膨胀率的一个简洁表达式；再把同一方程改写成一种货币需求理论；最后以**古典二分法**（classical dichotomy）与**货币中性**（neutrality of money）收尾——后者断言，货币在长期内不影响实际量。这一断言只有在价格已充分调整时才成立；中性在短期（价格具有粘性时）的失效，恰恰是通向后续各章凯恩斯模型的入口。

9.1　作为一种税的通货膨胀

观察通货膨胀的一个有用的初步视角是财政视角。当政府无法或不愿通过普通的征税或借债来为其支出筹资时，它可以干脆印钞来付账。它每发行一单位新货币，就稀释了已经在流通中的每一单位货币的购买力。货币的持有者由此在实际意义上变穷了，而这些货币所支配的资源被转移给了政府。其效果与对货币余额征收的一种税无异。

定义 9.1: 铸币税（通胀税）

铸币税（seigniorage）是政府通过发行货币获得的实际收入。等价地，它就是**通胀税**（inflation tax）：当货币存量增长、价格水平上升时，公众持有的货币的实际价值下降，而这部分失去的购买力归于货币的发行者。这里的“税基”是公众选择持有的实际货币余额 M/P，“税率”则是侵蚀它们的通货膨胀率 π。

这就是认为通胀相当于在征收通胀税的传统观点。它值得记住，因为它解释了为什

么财政承压的政府如此频繁地诉诸印钞机，也解释了为什么恶性通货膨胀（hyperinflation）的发作几乎总是财政崩溃的发作：通胀税是最后的收入来源。它还把货币数量论的核心问题用财政语言框定下来——如果政府不断扩张 M 来为支出筹资，这对 P 会有什么影响?

9.2　通货膨胀的成本

如果通货膨胀转移了资源、扭曲了决策，那么这些成本究竟由谁、在何处承担? 把经济理论所识别出的真正一阶成本，与公众最为恐惧、但实际上更多是表象而非实质的那种成本区分开来，是有益的。

9.2.1　真实的成本

1. **鞋底成本（交易成本上升）**。当预期价格会持续上涨时，持有货币的代价变高，于是人们设法节约自己的货币余额——更频繁地往返银行、把财富存放于带息资产或实物资产、缩短合同的期限。长期合同变得更难订立，因为双方都必须不断把变动的价格水平考虑进来。这个名称让人联想到额外跑银行磨损的鞋底；更广义地说，它涵盖了在价格稳定时本无必要、却因管理货币持有而耗费的全部实际资源。

2. **菜单成本（价格调整的成本）**。企业必须实实在在地更改其标价——重印菜单和目录、重贴货架标签、重新设定系统——而每一次这样的改动都消耗实际资源。通货膨胀率越高，价格就必须越频繁地修订，这些成本也就越大。

3. **资源配置被扭曲**。价格是引导资源流向其最有价值用途的信号。当总体价格水平在上涨、且不同商品以不同速度上涨时，就难以分辨某种商品*相对*价格的变化（这本应重新引导资源）与总体价格水平的变化（这本不该引导资源）。*价格的信号作用失效*，资源因此被错误配置。

4. **债务人与债权人之间的再分配**。当通货膨胀高于签订贷款时所预期的水平时，实际支付的*实际*利率被压低。债务人用价值低于预期的货币偿还，而债权人收到的实际价值少于合同所承诺的。因此通货膨胀把财富从债权人（贷方）再分配给债务人（借方）：*借方得益、贷方受损*。

9.2.2　为什么“通货膨胀缩水了我的收入”不是一阶成本

请注意上面这份清单里*缺*了什么：大多数人恐惧通货膨胀的理由，也就是价格上涨缩水了他们收入的购买力、从而让他们变穷。这一顾虑在理论上站不住脚。如果经济中所有价格按比例同步上涨，那么工资、租金和利润——它们本身也是价格——就理应同比例上涨。一个工资和食品账单都翻了一倍的工人，处境并没有变差。实际收入，即名义收入与价格水平之比，保持不变。统一且被充分预期到的通货膨胀，本身并不会减少任何人的实际收入。

注（流行的担忧在何时变成真问题）.

上面这个理论论证假定所有价格和收入都同步、即时地一起变动。现实在两方面偏离了这一假定。第一，通货膨胀往往是*结构性*的：不同价格以不同速度上涨，因此某个特定家庭所购买的篮子的涨速，可能快于它那一份特定的收入来源。第二，工资和许多价格具有*粘性*——它们不会立即调整。当商品价格上涨、而某工人的名义工资被合同锁定一年时，这位工人的实际收入在这段期间里确实下降了。正是这些摩擦，让流行的担忧在短期里有了着力点。它们也是凯恩斯主张的基础，即财政政策和货币政策在短期内能够推动实际产出：如果价格能即时调整，就没有可供利用的短期实际效应了。我们将在本章末尾讨论货币中性时回到这一点，并在第 十 章至第 十二 章中将其充分展开。

9.3　通货膨胀的收益

通货膨胀并非全是成本。适度的通货膨胀率带来两项值得明确指出的好处，二者都经由上面那些成本所通过的同一渠道发挥作用。

1. **它刺激了信贷需求**。通过压低存量债务的实际利率，通货膨胀使借贷在实际意义上变得更便宜，从而鼓励企业和家庭举债与投资。更高的信贷需求能够支撑支出与经济活动。

2. **它增加了劳动力市场的弹性**。名义工资是出了名地难以下调：工人抵制直接的减薪，合同与惯例使其*向下刚性*。然而劳动力市场有时需要实际工资下降——例如当一家企业面临需求疲软、否则就不得不裁员时。通货膨胀提供了一条出路。如果价格水平上涨而名义工资保持不变，那么*实际*工资就在没有任何人被正式减薪的情况下下降了。更低的实际工资使企业雇佣工人的成本下降，从而提高劳动需求、支撑就业。如此一来，适度的通货膨胀就让劳动力市场运转得更顺畅。

9.4　货币数量论

我们现在从“通货膨胀为什么重要”转向“是什么决定了通货膨胀”。古典的回答是*货币数量论*，它从一个把货币数量与它所完成的交易价值联系起来的会计恒等式出发。

定义 9.2: 数量方程

数量方程（quantity equation，又称交换方程）指出，货币数量乘以它的流通速度，等于交易的货币价值：

$$M \times V = P \times T,$$

其中 M 是名义货币存量，V 是*货币的流通速度*（velocity of money，即每单位货币在一个时期内平均转手的次数），P 是价格水平，T 是交易量。

交易量 T 很难直接衡量，所以在实践中我们用实际产出（实际 GDP）来替代它，记为 Y。经此替换，右端的 $P \times Y$ 恰好就是*名义* GDP，于是数量方程就成为我们全书将要使用的形式：

$$M \times V = P \times Y. \tag{9.1}$$

把这个恒等式转化为一种理论的，是一组关于哪些项可以自由变动、哪些项在短期内固定的陈述。

假设 9.3: 货币数量论的假设

产出由供给侧决定。 实际产出 Y 由经济的潜力——它的资本、劳动和技术——所决定，因此在短期内被视为给定，与货币存量无关。

流通速度由习惯决定。 流通速度 V 由经济的交易技术与支付习惯（人们如何被支付、多久支付一次、以及如何持有货币）所支配。这些习惯只会缓慢改变，因此 V 在短期内大致不变。

注（流通速度与利率）.

> 流通速度并非字面意义上的常数。当持有货币的机会成本上升时它会上升：当利率攀升——也就是流动性的价格上涨——时，人们相对于其支出会持有更少的货币，于是每单位货币转手得更频繁，V 随之增加。我们的假设仅仅是，在短期内，V 相较于我们所关心的 M 与 P 的波动而言变动甚微。V 对利率的这种依赖，正是第 9.5 节的流动性偏好改写所要明确刻画的。

把 V 与 Y 固定为背景，方程 (9.1) 就把价格水平直接同货币存量绑在了一起。要干净利落地看清这一点，对两边取对数，

$$\log M + \log V = \log P + \log Y,$$

再对时间求导。由于 $\log X$ 的时间导数就是增长率 $\Delta X/X$，这便得到

$$\frac{\Delta M}{M} + \frac{\Delta V}{V} = \frac{\Delta P}{P} + \frac{\Delta Y}{Y}, \tag{9.2}$$

或用百分比变化的语言来写，

$$\%\Delta M + \%\Delta V = \%\Delta P + \%\Delta Y.$$

方程 (9.2) 把数量方程表达为一种增长率之间的关系：货币增长率与流通速度增长率之和，等于通货膨胀率与实际产出增长率之和。由此可得两条结论。

货币数量论的两条结论

1. **通货膨胀由货币增长所驱动。**在流通速度与实际产出的增长率固定不变时，货币增长率 $\Delta M/M$ 的任何上升都会一对一地传递到通货膨胀率 $\Delta P/P$。持续的通货膨胀，归根结底是持续的货币创造的后果。

2. **货币增长追随名义 GDP 增长。**由于 $\Delta V/V \approx 0$，货币增长 $\Delta M/M$ 与名义 GDP 的增长 $\%\Delta P + \%\Delta Y$ 大致同步变动。在现实中，中央银行扩张货币供给的速度，大体上与名义产出的增长相匹配。

在 (9.2) 中令流通速度增长 $\Delta V/V$ 为零，并回忆通货膨胀率 π 就是价格水平的增长率，$\pi = \Delta P/P$，我们就能直接解出通货膨胀：

$$\pi = \frac{\Delta M}{M} - \frac{\Delta Y}{Y}. \tag{9.3}$$

定义 9.4: 由货币增长决定的通货膨胀率

在流通速度不变的条件下，通货膨胀率等于货币增长与实际产出增长之差，

$$\pi = \frac{\Delta M}{M} - \frac{\Delta Y}{Y}.$$

方程 (9.3) 给出了对通货膨胀一个粗略却富有启发的估计。可以这样理解它：理论上，流通中的货币数量“理应”与经济的实际产出同步增长，使得货币恰好够在价格不变的情况下完成那些新增的实际交易。货币增长中*超出*实际增长的那一部分，没有额外的商品可供追逐，于是溢出为更高的价格。这一超额的部分，就是通货膨胀率。

9.5　货币需求与流动性偏好

到目前为止，我们一直把数量方程读作一种价格水平理论。同一方程经过重排，也是一种*货币需求*理论——关于人们愿意持有多少货币。把 $MV = PY$ 重排以分离出实际余额，

$$\frac{M}{P} = \frac{Y}{V}. \tag{9.4}$$

左端的 M/P 是*实际货币供给*——货币存量的购买力——在均衡时它等于公众所需求的实际余额数量。因此数量方程刻画了*流动性偏好*：维持经济体运转所需要的货币量。由此可得两点观察。第一，这一方程从根本上讲的是货币*需求*。第二，它刻画的是一种*均衡*：当货币市场出清时，方程中的 M 同时代表货币的供给与对货币的需求。

凯恩斯把 (9.4) 的右端推广为一个明确的货币需求函数。他不再以固定的流通速度把实际余额钉在 Y/V 上，而是让对实际余额的需求通过一个*流动性偏好*（liquidity preference）函数 L 依赖于收入与利率：

$$\left(\frac{M}{P}\right)^{d} = L(i, Y) = L\left(r + \pi^{e},\, Y\right). \tag{9.5}$$

这里 i 是名义利率，由费雪关系（Fisher relation）它等于实际利率 r 与预期通货膨胀 π^{e} 之和，即 $i = r + \pi^{e}$。两个自变量的符号至关重要：

1. **货币需求随收入 Y 递增**。更高的实际收入水平意味着更多的支出、更多需要融资的交易，因此需要更多的货币来完成它们。于是 L 随 Y 上升。
2. **货币需求随名义利率 i 递减**。利率是持有货币的机会成本：以现金或无息账户形式持有的货币，放弃了债券及其他资产所能提供的回报 i。当 i 上升时，持有货币的代价变高，于是人们持有得更少——等价地说，流通速度上升，每单位货币承担更多的交易。于是 L 随 i 下降。

注（为什么货币需求用的是预期通货膨胀）.

> 对于一项今天就要做出的、关于在接下来一个时期内持有多少货币的决策而言，真正重要的机会成本是名义利率，而相关的名义利率，是那个为储蓄者预期会发生的通货膨胀作补偿的利率，而非事后才实现的通货膨胀。货币需求是一个事前（ex ante，即前瞻性）的概念，所以它依赖于预期通货膨胀 π^e，这正是为什么 L 的第二个自变量写作 $r+\pi^e$ 而非 $r+\pi$。

在市场的另一侧，名义货币供给 M 由中央银行设定。我们把它视为一个外生的、独立的政策决策：货币当局通过行政手段选择 M，而这一选择被模型的其余部分当作给定。货币市场的均衡要求实际货币供给等于对它的实际需求，

$$\frac{M}{P} = L(i, Y) = L\left(r + \pi^e, Y\right). \tag{9.6}$$

在名义货币供给 M 由政策固定的情况下，利率 i 与价格水平 P 会进行调整，以使 (9.6) 成立。这一均衡条件就是我们将在第 十一 章构建的 LM 关系的引擎。

9.6　古典二分法与货币中性

我们现在可以陈述货币理论的核心长期命题，以及它在短期内可能失效的确切含义——通往后续一切内容的桥梁。

定义 9.5: 古典二分法

古典二分法是把经济变量分离为实际量（产出、就业、相对价格、实际利率）与名义量（货币存量、价格水平、名义工资）。这一二分法认为，名义变量不会影响实际变量：货币数量的变化推动价格水平，却不触动实际配置。

如果古典二分法成立，那么价格水平就不能改变实际经济行为。货币存量的变化以及产生它的政策，对产出、就业或消费都没有影响——它们只是把所有名义量一起按比例放大或缩小。这时我们就说货币是**中性**的。

定理 9.6: 货币中性

如果名义货币存量的变化使所有名义变量按同一比例变化、而让所有实际变量保持不变，那么货币就是**中性**的。当经济中每一个价格都在同一时间、朝同一方向、按同一比例进行调整，从而没有任何相对价格被扰动时，中性成立。

中性成立的条件是苛刻的：它要求所有价格——商品价格、工资、资产价格——都即时且按比例地一起变动。当这一条件被满足时，货币存量翻倍只是简单地使价格水平以及每一项名义工资和合同翻倍，而实际工资、实际产出和实际利率都原封不动。这与第 八 章简单货币模型中遇到的中性结果是同一个，如今则在整个经济的层面上陈述出来。

然而现实充满摩擦。价格与工资具有粘性：它们被提前设定、被不频繁地修订，且并非全都同时变动。当货币存量变化、而价格尚未跟上时，相对价格被暂时扰动，于是

货币的变化确实推动了实际量。因此在短期内，货币不是中性的，货币政策能够影响产出与就业的水平。

中性是长期性质，而非短期性质

- **短期**：价格具有粘性，古典二分法失效，货币不中性。货币政策与财政政策能够推动实际产出与就业。
- **长期**：只要给予足够的时间，每一个有粘性的价格都会调整，直到经济达到一个不再残留任何价格刚性的新均衡。古典二分法得以恢复，货币是中性的。政策只推动价格水平，而非实际量。

这一区分是本课程余下部分的组织原则。经济存在一个长期均衡，它锚定于潜在产出、由供给侧支配，在那里货币是中性的、古典二分法成立。它也存在短期波动——经济周期——在此期间，有粘性的价格让需求侧的扰动（以及更难以控制的供给侧冲击，例如全要素生产率的变动）把经济推离潜在水平。随着时间推移，价格、利率及相关变量的调整，会把一个偏离了潜在产出的短期均衡重新拉回潜在产出。

注（凯恩斯论短期与长期）.

> 凯恩斯的理论正是建立在短期与长期这一区分之上的，二者以价格是否具有粘性来划分。在短期，供给是可变的、价格是刚性的，于是货币政策与财政政策可以移动总需求、从而改变产出水平；货币不是中性的。在长期，价格完全可变，总供给在潜在产出处保持刚性，货币是中性的。凯恩斯框架的承诺在于：政策能够熨平经济周期的短期波动，尽管它无法改变经济的长期潜力。后续各章的模型——凯恩斯交叉（第 十 章）、IS–LM（第 十一 章）以及总需求与总供给（第 十二 章）——就是把这一承诺付诸运转的机器。

第十章　凯恩斯交叉与乘数

前几章的增长模型告诉我们经济长远会走向何处。在索洛模型与新古典框架中（第 五 章与第 六 章），以及每当我们写下

$$Y = F(K, L, \text{TFP})$$

时，产出都是由供给一侧决定的：取决于资本存量、劳动力数量，以及把二者转化为产品的技术水平。对于**长期**而言，这是正确的图景——此时价格与工资已有足够时间调整，经济运行在其生产潜力之上。但它对另一些问题闭口不谈：产出为什么会逐年起落，衰退为什么发生，政府又为什么会认为一次减税就能把经济拉出低谷。要回答这些问题，我们需要一个**短期**模型——在其中，决定产出的不是经济能生产多少，而是买家愿意花多少钱。

本章发展的正是这样一个最简单的模型，它出自凯恩斯之手。其组织性恒等式，是我们在国民收入核算中早已熟悉的产出需求侧分解（第 二 章）：

$$Y = C + I + G + \text{NX},$$

只是如今要把它读作关于**计划**支出的陈述，而非已实现的会计记录。我们将逐项说明消费、投资、政府购买与净出口这四个分量如何依赖于收入，把它们加总为一条计划支出曲线，再施加“计划支出等于产出”这一要求。由此得到的便是**凯恩斯交叉**（Keynesian cross）——一张其交点钉住均衡产出的图。它最具分量的教益是**乘数**（multiplier）：由于任何一笔额外支出都会变成某人的收入并被再次部分花出，对需求的一点小小推动会让产出移动得比这点推动本身更多。

关于这个模型究竟是什么，要先说一句提醒的话。凯恩斯的框架在本质上是**静态**的，它没有清晰的“未来”概念：消费只对**当期**收入作出反应，投资被处理成与当期产出无关，而利率、终生财富之类的跨期因素，要么被压下不谈，要么被统统折叠进一个行为参数里。这是一个实实在在的局限，后面的章节（第 十一 章的 IS–LM 模型与第 十四 章的动态分析）会把这里缺失的东西补回去一部分。然而模型的粗糙本身也正是它历久弥新的源头：它抓住了一种力量——由当期收入产生的支出——而短期数据恰恰最强烈地呈现出这种力量，并且它把这种力量装进了一个足够简单、可以用来推理的形式之中。

10.1　产出的两种时间视野

值得把贯穿下文的那一组对照明白地说出来。长期之下，产出是供给侧的量：$Y = F(K, L, \text{TFP})$ 由要素存量与技术状态固定，需求至多只能决定价格。短期之下，价格黏滞、要素可能闲置，因果方向便反了过来。需求的水平——家庭、企业、政府与外国想要购买多少——决定了企业实际生产多少，从而决定产出与就业。

会计恒等式 $Y = C + I + G + \text{NX}$ 在每一期都按定义成立，因为国民收入核算记录的是已实现的购买。凯恩斯的做法，是把右端重新诠释为**计划**支出或**意愿**支出，再追问：什么时候这些计划与它们所预设的产出彼此自洽？由于支出对应着需求，我们就把计划支出当作经济的总需求，

$$\text{AD} \;=\; C + I + G + \text{NX},$$

并不再多加辩白地把二者视为等同。这一等同并不绝对严谨——在更完整的模型里，总需求还会对价格水平和利率作出反应——但在短期中，需求的波动如此之大、又远比潜在产出来得迅速，以至于就本章的目的而言，我们大可把产出的短期变动当作需求的变动，无须再区分一条短期曲线与一条长期曲线。

10.2　计划支出的各个分量

现在我们为四个分量各写下一个行为方程，逐项搭建起计划支出曲线。全章中，Y 记收入（也即产出），T 记扣除转移支付后的净税收，因而 $Y - T$ 就是**可支配收入**——家庭手中可供花费或储蓄的那部分收入。

10.2.1　消费

一个家庭消费多少由什么决定？原则上，决定因素很多：它的当期收入、它预期的未来收入（极端而言是它的整个**终生**收入）、它积累的财富（一个存量，与消费正相关），以及利率、对未来的贴现率这类跨期项。凯恩斯的模型有意忽略了这其中的绝大部分。由于它没有真正的动态结构——没有可以享用所储蓄资源的未来期——它无法把利率或终生收入表示为独立的力量。消费被设定为只依赖于**当期**可支配收入：

定义 10.1: 凯恩斯消费函数

家庭消费是当期可支配收入的线性函数，

$$C \;=\; C_0 + \text{MPC}\,(Y - T)\,,$$

其中 $C_0 > 0$ 是**自发型消费**（autonomous consumption）——即可支配收入为零时仍会发生的消费量；$\text{MPC} \in (0, 1)$ 是**边际消费倾向**（marginal propensity to consume），即每多一元可支配收入中被花掉的比例。$\text{MPC}\,(Y - T)$ 一项是**引致消费**（induced consumption），即由收入驱动的那部分。

关键的建模约定是：*所有变量都是当期的*——方程里没有预期未来收入这一项。未来、财富或利率在塑造行为时所起的任何作用，都被整体吸收进 MPC 这一个参数里。一个有耐心、富有或对未来乐观的家庭，无非是拥有一个不同的 MPC；至于这个数字从何而来，模型并不交代。

注（一个粗糙却合乎数据的模型）.

> 这条消费函数显然是不完整的——它略去了财富、预期和利率，而这些都是更丰富的理论会纳入的因素。凯恩斯本人并不否认这一点。他的辩护是经验性的：这条函数是从观察、而非从第一原理中搭起来的，而可支配收入事实上正是能解释总消费变动最多的那个单一因素。我们应当把 $C = C_0 + \mathrm{MPC}\,(Y - T)$ 读作*基础*模型，再随应用所需向其追加更多项（一个利率效应、一个财富效应）。

注（减税与中国的边际消费倾向）.

> MPC 的大小有着直接的政策含义，而且这些含义因制度而异。中国的税收减免历来主要通过削减企业税来实现——税收的大头也正是由企业贡献的——而非通过对个人的转移支付；向家庭直接发放现金补贴的做法并不常见。再叠加上相对偏低的家庭 MPC，这意味着一笔给定的减税往往只能微弱地传导到消费上。相比之下，以投资为基础的刺激则能在两个边际上同时发力：短期内它拉动就业与需求，长期内它又能通过增添生产能力而改变供给一侧。

储蓄不过是可支配收入中未被消费的那一部分，因此消费函数立刻就蕴含了一条*储蓄函数*：

$$S \;=\; (Y - T) - C \;=\; -C_0 + (1 - \mathrm{MPC})\,(Y - T)\,.$$

斜率 $1 - \mathrm{MPC}$ 是*边际储蓄倾向*：每一元可支配收入非花即存，所以这两个倾向之和恒为一。

10.2.2　投资

一家企业只在某个投资项目有利可图时才会去做它。自然的判据是净现值：把项目未来的收益流与成本流贴现回当下，当前者超过后者时就投资，

$$\mathrm{NPV}(I) \;=\; \sum_{t=1}^{\infty} \frac{\mathrm{Income}_t}{(1+r)^t} \;-\; \sum_{t=1}^{\infty} \frac{\mathrm{Cost}_t}{(1+r)^t}.$$

麻烦出在收益一侧。未来收益高度不确定，且大多难以量化刻画；在实践中，这一决策靠的更接近“直觉”而非计算——也就是凯恩斯所说的著名的*动物精神*（animal spirits），他以动物追逐猎物的本能冲动作类比。成本一侧则更易于把握。投资必须融资，而融资成本随利率 r 上升，于是利率越高，能跨过门槛的项目就越少。因此我们把投资建模为利率的减函数，

$$I = I(r), \qquad I'(r) < 0,$$

而不去认定某个具体的函数形式——凯恩斯本人也没有提出任何具体形状。

投资不依赖于当期收入

在凯恩斯模型中，投资仅是利率的函数：$I = I(r)$，关于 r 递减。它不依赖于当期产出、家庭收入或财富。在这个静态模型里，这是一个有意为之的简化，目的是让 I 与 Y 的决定因素彼此分开。

注（现实中的保留意见）.

"投资独立于当期收入"这一假设是一个干净的理论立场，而非对现实世界的字面描述。现实当中，投资与产出强相关：它会对未来需求的预期、对政府政策作出反应，也与决策者的财富和信心、与内部资金的可得性相关。把 I 当作外生于 Y，最好读作一种"一次只孤立一条渠道"的办法；一旦我们让投资对利率作出反应，而这个利率又由第 十一 章的 IS–LM 模型与产出联合决定，这一假设便被放松了。

10.2.3　政府购买

政府对商品与服务的购买 G，被视为某种政治与行政过程的产物，而非对当期经济变量的反应。它由决策定下，不由模型定下，因此我们把 G 当作外生的：一个由政策制定者选定的常数。税收 T 同样处理，当作一个由模型之外设定的政策工具。

10.2.4　净出口

净出口是出口减去进口，

$$\mathrm{NX} = X - M.$$

出口 X 是外国人对本国产品的购买，因而依赖于外国收入，并与之正相关；从本国模型的视角看，它是外生的。进口 M 是本国对外国产品的购买，随本国可支配收入上升，

$$M = \mathrm{MPM}\,(Y - T),$$

其中 MPM 是边际进口倾向（marginal propensity to import），即每多一元可支配收入中花在外国产品上的比例。在大部分分析中，把这种收入依赖性压下、同样把 NX 当作一个外生常数，往往更方便，也更常见；我们在推导乘数时就会这样做，并在一则注记中说明对收入敏感的情形会如何修正结果。

10.3　计划支出与凯恩斯交叉

把四个分量加起来，便得到作为收入函数的计划支出曲线。把依赖收入的各项明确写出来，

$$\mathrm{AD} \;=\; C + I + G + \mathrm{NX} \;=\; \underbrace{C_0 + I(r) + G + X}_{\text{自发部分}} \;+\; \underbrace{(\mathrm{MPC} - \mathrm{MPM})}_{\text{斜率}}\,(Y - T)\,.$$

这条曲线随收入递增，但斜率严格小于一：收入每多一元，计划支出只上升 MPC − MPM <1，因为这多出的一元里有一部分被储蓄、一部分外漏到了国外。在一张纵轴为计划支出、横轴为收入 Y 的图中，这是一条向上倾斜、但比 45° 线更平缓的直线。（若把进口折进外生常数里，斜率就只是 MPC <1；真正要紧的，只是它为正且小于一。）

均衡要求计划彼此自洽：企业生产的恰好是买家想买的，从而计划支出等于产出，不存在非意愿的存货堆积或消耗。

定义 10.2: 短期均衡

当计划总支出等于产出时，经济处于短期均衡，

$$\mathrm{AD} \;=\; Y,$$

也即计划支出曲线与“支出等于收入”的那条 45° 线相交之处。

45° 线是所有“需求量等于产出量”的点的轨迹；它正是界定均衡的约束条件。由于计划支出曲线的斜率小于一，而 45° 线的斜率恰为一，两条线必定恰好相交一次。它们的交点就是均衡产出水平。这张图便是凯恩斯交叉，如图 10.1 所示。

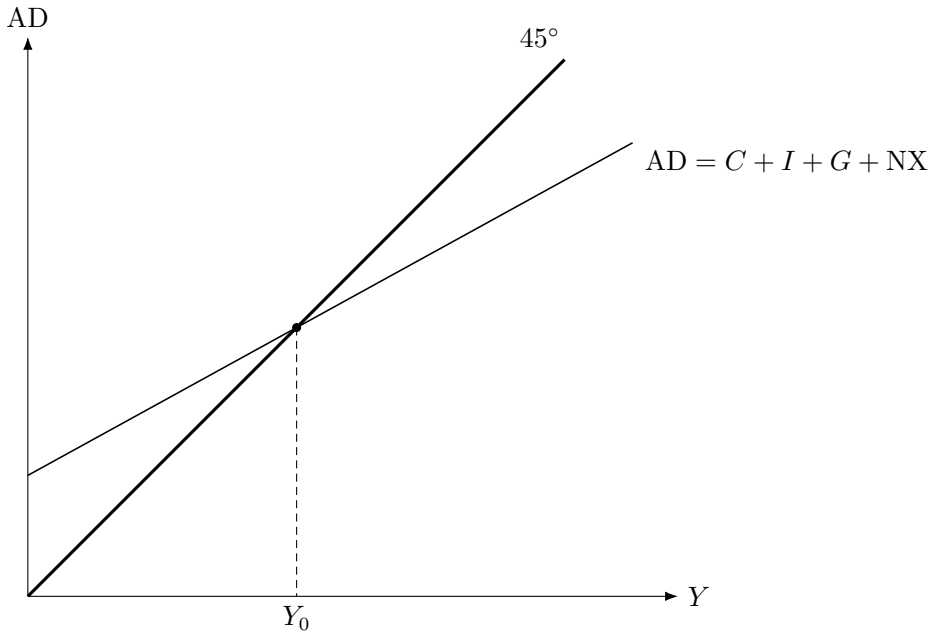

图 10.1: 凯恩斯交叉：均衡产出出现在计划支出曲线（斜率小于一）与“支出等于收入”的 45° 线相交之处。

10.3.1　均衡的稳定性

这个交点不仅仅是方程恰好相抵的一点；它还是一个稳定的休止点，经济从两侧都会被推向它。其调整机制经由存货与企业的生产决策运作。设产出一时高于均衡水平，$Y_1 > Y_0$。在 Y_1 处，计划支出线位于 45° 线之下，于是计划支出不及产出：买家想要

的比企业已生产的少。商品作为未售存货堆积起来，企业便缩减生产，产出向 Y_0 漂回。反过来，若产出过低，$Y_2 < Y_0$，计划支出超过产出，存货被消耗，企业扩大生产，产出又升回 Y_0。从两侧看，经济都收敛到那个交点，如图 10.2 所示。

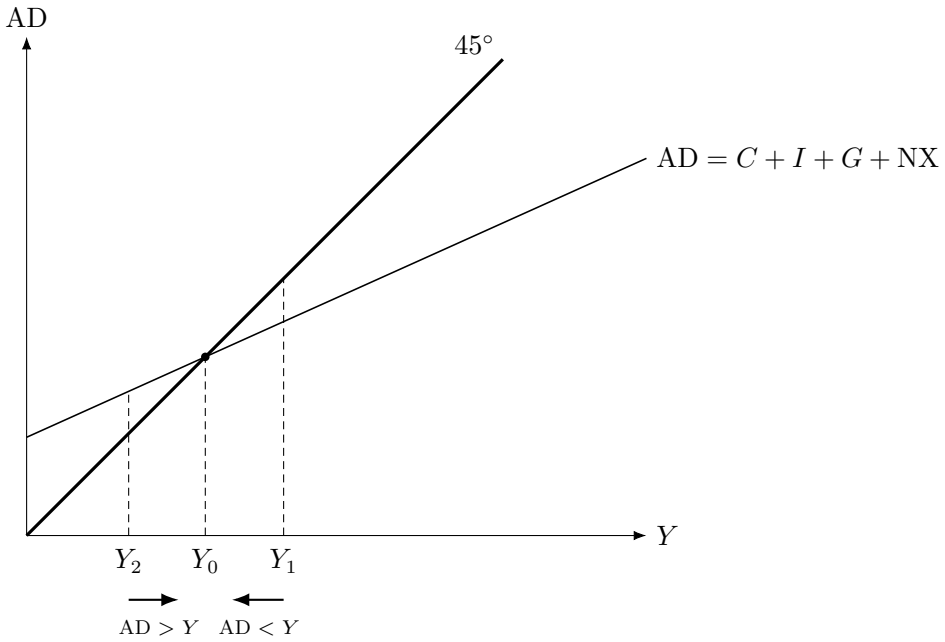

图 10.2: 凯恩斯交叉均衡的稳定性：在 Y_0 之上，计划支出不及产出，企业缩减；在 Y_0 之下，支出超过产出，企业扩张；产出从两侧都收敛到 Y_0。

注（均衡产出未必等于潜在产出）.

> 长期来看，凯恩斯交叉的均衡本应与潜在产出重合——也就是供给所决定的那个水平 $Y = F(K, L, \text{TFP})$。短期之下却没有这样的保证。经济可以稳定地停在潜在产出之下，伴随闲置产能与失业劳动；也可以停在其上，经济过热。模型给出的均衡与长期模型给出的潜在产出之间的这道缺口，正是短期需求管理之所以有理由存在的依据。

10.4　乘数

两条线的斜率相对关系所确定的，不止是它们会相交这一点。它还支配着当计划支出曲线移动时，交点会*移动多远*。设某项自发支出上升，使整条曲线向上平移了 ΔD。由于这条曲线比 45° 线更平缓，交点在横向上走过的距离 ΔY 会超过引起它的纵向平移 ΔD：自发支出的小幅增加，换来了产出更大幅度的增加。这种放大就是*乘数效应*，而平移被放大的倍数就是*乘数*。其几何关系如图 10.3 所示。

要在代数上求出乘数，就把 NX 当作外生常数，并对计划支出曲线

$$\text{AD} \;=\; C_0 + \text{MPC}\,(Y - T) + I + G + \text{NX}$$

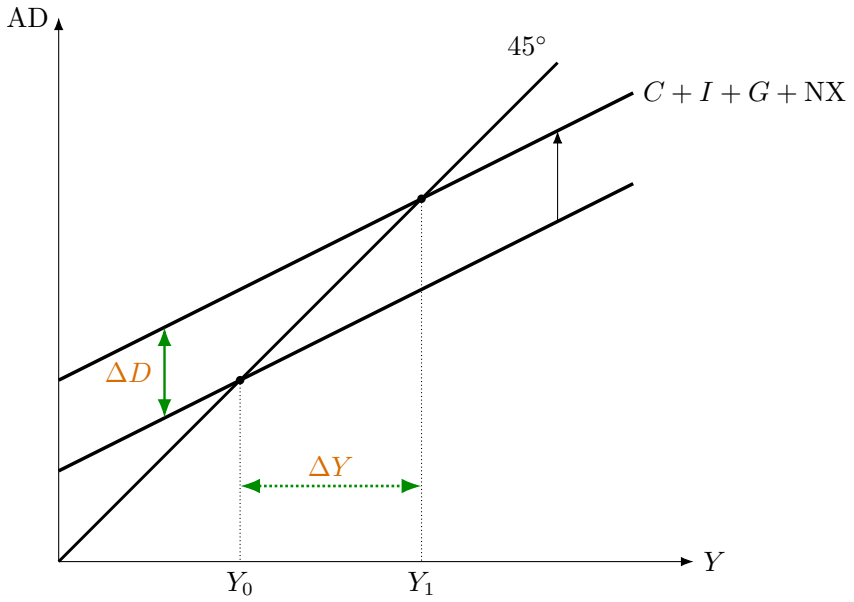

图 10.3: 乘数：计划支出曲线向上平移 ΔD 时，均衡产出上升 $\Delta Y > \Delta D$，原因是这条曲线比 45° 线更平缓。

施加均衡条件 $\mathrm{AD} = Y$。令 $\mathrm{AD} = Y$ 并解出 Y：

$$\begin{aligned} Y &= C_0 + \mathrm{MPC}\,(Y - T) + I + G + \mathrm{NX} \\ Y - \mathrm{MPC}\,Y &= C_0 - \mathrm{MPC}\,T + I + G + \mathrm{NX} \\ (1 - \mathrm{MPC})\,Y &= C_0 - \mathrm{MPC}\,T + I + G + \mathrm{NX}, \end{aligned}$$

即得均衡产出的简化式解，

$$Y = \frac{1}{1 - \mathrm{MPC}}\,(C_0 - \mathrm{MPC}\,T + I + G + \mathrm{NX})\,. \tag{10.1}$$

现在，每个乘数都可以读作 Y 对相应外生变量的偏导数。

定理 10.3: 凯恩斯乘数

在均衡式 (10.1) 中，产出对各外生分量的反应为：

$$\text{支出乘数：}\quad \frac{\partial Y}{\partial G}=\frac{\partial Y}{\partial I}=\frac{\partial Y}{\partial C_0}=\frac{1}{1-\mathrm{MPC}}>1,$$

$$\text{税收乘数：}\quad \frac{\partial Y}{\partial T}=\frac{-\mathrm{MPC}}{1-\mathrm{MPC}}<0,$$

$$\text{平衡预算乘数：}\quad \left.\frac{\mathrm{d}Y}{\mathrm{d}G}\right|_{\mathrm{d}T=\mathrm{d}G}=1.$$

对任意 $\mathrm{MPC}\in(0,1)$，支出乘数都大于一；税收乘数为负，其绝对值 $\dfrac{\mathrm{MPC}}{1-\mathrm{MPC}}$ 大于一当且仅当 $\mathrm{MPC}>\frac{1}{2}$。

证明. 支出乘数由对 (10.1) 求导得到：由于 C_0、I、G、NX 都以系数 $1/(1-\mathrm{MPC})$ 进入，它们各自的乘数都是 $1/(1-\mathrm{MPC})$，而这个值因 $0<\mathrm{MPC}<1$ 而大于一。其经济内涵就是一轮接一轮的“漏出”故事：多出的一元自发支出变成某人的一元收入，其中比例 MPC 被再次花出，再花出的部分又有 MPC 被花出，如此往复。引致出来的产出是一个等比级数

$$1+\mathrm{MPC}+\mathrm{MPC}^2+\cdots=\frac{1}{1-\mathrm{MPC}}.$$

对 (10.1) 关于 T 求导，得税收乘数 $-\mathrm{MPC}/(1-\mathrm{MPC})$，它为负，因为增税削减可支配收入、从而削减支出。它的绝对值恰好比支出乘数小一个因子 MPC：税收的变化要先穿过消费函数才能作用于需求，所以它的第一轮效应是 $\mathrm{MPC}\,\Delta T$，而不是整个 ΔG。解 $\mathrm{MPC}/(1-\mathrm{MPC})>1$ 得 $\mathrm{MPC}>1-\mathrm{MPC}$，即 $\mathrm{MPC}>\frac{1}{2}$。

至于平衡预算乘数，令支出与税收同时增加同一数额，$\Delta G=\Delta T\equiv\Delta$。由刚求得的两个乘数，

$$\Delta Y=\frac{1}{1-\mathrm{MPC}}\,\Delta G-\frac{\mathrm{MPC}}{1-\mathrm{MPC}}\,\Delta T=\frac{1-\mathrm{MPC}}{1-\mathrm{MPC}}\,\Delta=\Delta.$$

一笔完全由增税来融资的政府购买增加，使产出恰好上升等于这笔增加的数额：平衡预算乘数为一。 □

支出乘数与税收乘数之差背后的直觉值得多花些笔墨，因为它支撑着整场财政政策辩论。G 或 I 的增加会把计划支出曲线整额地向上推，因而被整个因子 $1/(1-\mathrm{MPC})$ 放大。减税则相反，它只能间接地抬高支出：家庭拿到了多出的可支配收入，却只花掉其中的比例 MPC，把其余存了起来。于是曲线只向上平移 $\mathrm{MPC}\,|\Delta T|$，所以就同样一元钱而言，税收是更弱的工具。平衡预算的结果随之道出一件惊人的事：哪怕是一笔完全靠增税来支付的刺激，也并非中性——它仍然会一比一地抬高产出，因为支出注入足额起作用，而税收的抽离却被较低的储蓄部分吸收掉了。

注（对收入敏感的净出口）.

若保留进口对收入的敏感性、不把它折进常数，即 $\mathrm{NX} = X - \mathrm{MPM}(Y - T)$，则均衡条件变为 $Y = C_0 + \mathrm{MPC}(Y - T) + I + G + X - \mathrm{MPM}(Y - T)$，解之得

$$Y = \frac{1}{1 - \mathrm{MPC} + \mathrm{MPM}} \left(C_0 - (\mathrm{MPC} - \mathrm{MPM})T + I + G + X \right).$$

现在支出乘数是 $1/(1 - \mathrm{MPC} + \mathrm{MPM})$，比之前更小：进口是一笔额外的漏出，所以每一轮再支出失去的不只是被储蓄的那部分，还有花到国外的那部分。定性的教益——一个大于一的正支出乘数、一个绝对值更小的负税收乘数——则保持不变。

乘数为何对政策重要

乘数告诉我们：支出、税收或投资变动所产生的效果，并不局限于其自身的规模——自发需求的一点小变化撬动了产出的一个更大变化。在评估其中任何一项时，都必须算上这种被乘数放大的效果，而不只是直接效果。两个乘数之间的不对称还带有一层政策意味——在同样的预算成本下，政府购买的增加比同等规模的减税更能拉动产出。这正是为什么当经济面临下行压力时，典型的应对是一种偏重支出的财政扩张。我们将在第 十三 章详谈财政政策。

然而凯恩斯交叉只讲了短期故事的一半。在计算投资时，我们一直把利率 r 当作固定的，连带把整个自发部分也当作固定的。但利率本身是一个均衡价格，由货币市场决定。让 r 变动，就会描出一整族均衡产出水平——每一个利率对应一个凯恩斯交叉均衡——而把这一族曲线与货币市场结合起来，正是第 十一 章 IS–LM 模型的任务。

第十一章　IS–LM 模型

第 十 章的凯恩斯交叉，是在**给定**利率的前提下确定产出的：它告诉我们计划支出与 45° 线在何处相交，却把利率 r 当作外界递进来的一个数字。然而 r 本身就是一个均衡价格——它是流动性的价格——并且会随产出的变化而变化。IS–LM 模型把这个回路闭合起来。它仍然假定价格水平固定（我们依旧处在短期），转而去寻找使两个市场**同时**出清的 (Y, r) 组合：一个是商品市场，由 *IS* 曲线刻画；另一个是货币市场，由 *LM* 曲线刻画。我们将会看到，财政政策作用于前一条曲线，货币政策作用于后一条曲线，而两者的相互作用，正是使上一章那个简单乘数被高估的原因。

11.1　IS 曲线

IS 曲线是使商品市场处于均衡——即计划支出等于产出——的全部 (Y, r) 组合的集合。曲线上的每一点，都对应着某个特定利率下的凯恩斯交叉均衡点。要看清它的形状，回忆一下：利率上升时投资下降，即 $I = I(r)$ 且 $I'(r) < 0$。把投资写成 r 的函数，再去解凯恩斯交叉，可得

$$Y^{*} = \frac{1}{1 - \mathrm{MPC}} \big(C_0 - \mathrm{MPC} \cdot T + I(r) + G + \mathrm{NX} \big),$$

于是利率下降会抬高投资，使计划支出线向上平移，并经由乘数放大，抬高均衡产出。因此，沿着商品市场的均衡轨迹，产出与利率朝相反方向移动。

定义 11.1: IS 曲线

IS **曲线**是使计划支出等于产出（商品市场出清）的 (Y, r) 组合的轨迹。由于利率越低则投资越高、从而均衡产出越高，IS 曲线**向下**倾斜。

图 11.1 把生成它的两幅图叠在一起。除利率以外，凡是会移动凯恩斯交叉均衡的因素——G 的变动、税收的变动、自发消费或自发投资的变动、净出口的变动——都会使**整条** IS 曲线移动；而利率本身的变动，则是沿着曲线**移动**。

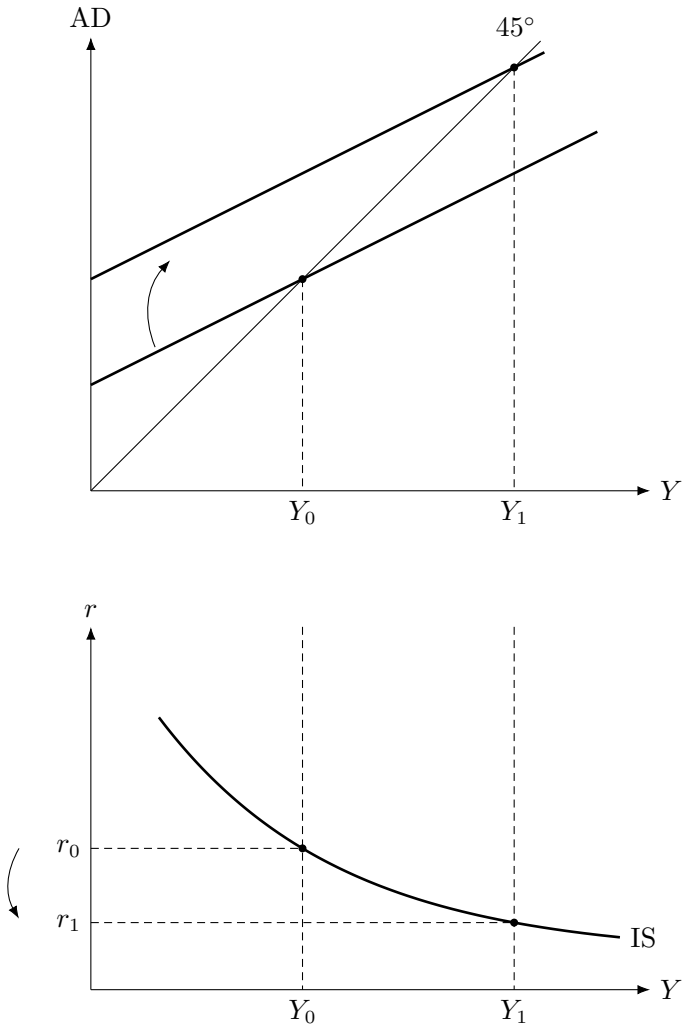

图 11.1: IS 曲线的推导。上图中，利率从 r_0 降到 r_1 使投资增加，把计划支出线向上推，凯恩斯交叉均衡从 Y_0 移到 Y_1。把每一个利率与它所对应的产出画在一起（下图），就描出了向下倾斜的 IS 曲线。

11.2　LM 曲线

LM 曲线来自货币市场。货币数量论 $MV = PY$ 可以读作一个关于货币需求的命题：为了在流通速度 V 之下完成名义交易量 PY，公众希望持有

$$M^d = \frac{PY}{V}.$$

这里流通速度并非常数；它随利率上升而上升，即 $V = V(r)$、$V'(r) > 0$，因为利率越高，持有不生息货币的机会成本越大，于是每一单位货币的周转次数增加，所需持有的货币就越少。与此相对，货币供给由央行设定，与利率无关：在图中它是一条垂直线。

图 11.2 中的均衡，钉住了在给定产出与价格下使货币市场出清的那个利率。现在让产出上升。Y 越高意味着交易越多，于是在每一个利率水平上货币需求都更大，需求曲线向右移动。供给既然固定，利率就必须上升，把货币需求压回到既有的货币存量

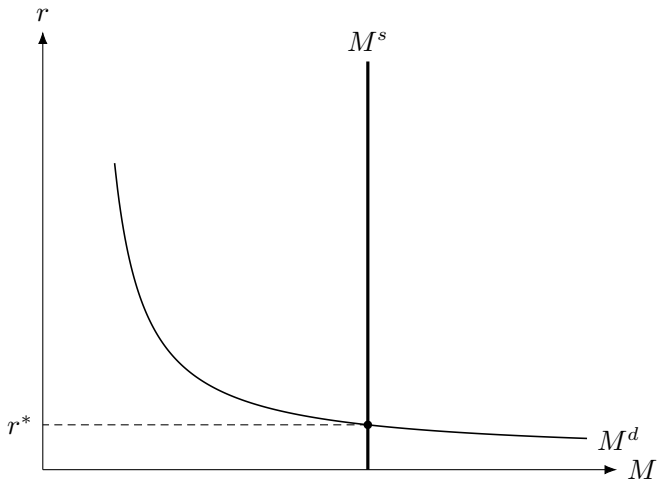

图 11.2: 货币市场。货币供给 M^s 是垂直的（由央行设定）；货币需求 M^d 随利率向下倾斜。两者的交点确定了均衡利率。

上。

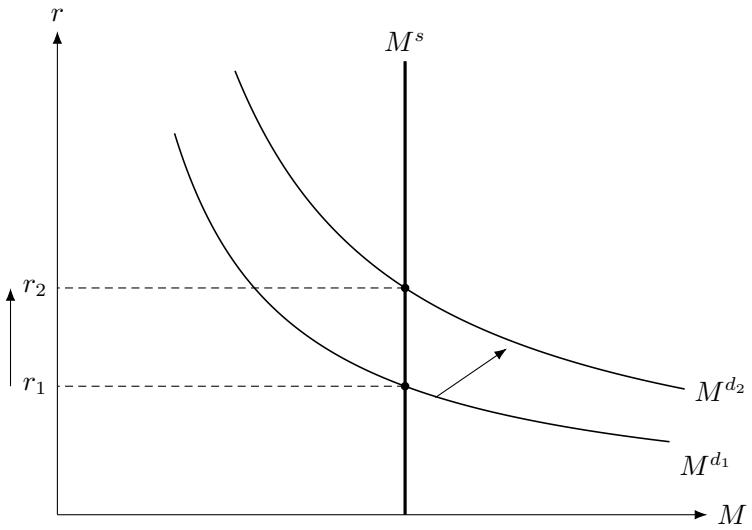

图 11.3: 产出上升使货币需求曲线向右移动；货币供给固定，均衡利率随之从 r_0 上升到 r_1。

把这些 (Y, r) 组合收集起来——产出越高，越需要更高的利率来维持货币市场的平衡——就描出了 LM 曲线。

定义 11.2: LM 曲线

LM **曲线**（流动性偏好–货币供给，liquidity preference–money supply）是在给定实际货币供给下使货币市场出清的 (Y, r) 组合的轨迹。由于产出越高则货币需求越大、从而均衡利率越高，LM 曲线**向上**倾斜。

什么会使整条 LM 曲线移动？任何改变实际货币供给 M/P 的因素。名义货币供

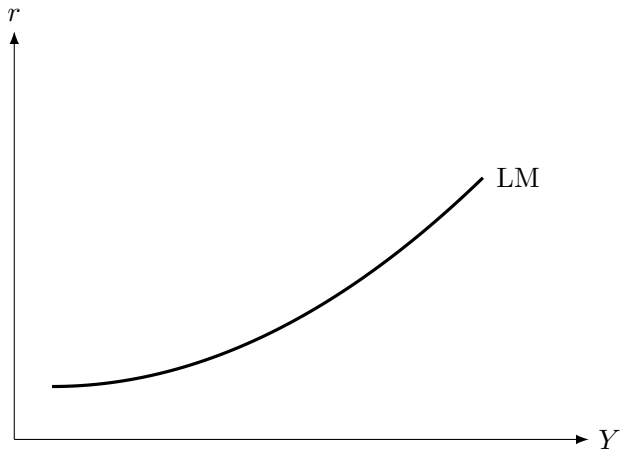

图 11.4: LM 曲线：使货币市场出清的产出与利率组合的轨迹。它向上倾斜，因为产出越高则货币需求越大，从而使市场出清的利率越高。

给增加，意味着在任意产出水平上，市场都在更低的利率上出清，于是 LM 曲线向下（等价地，向右）移动。

价格水平上升的作用方向恰好相反。这里真正相关的量是**实际**货币供给 M/P：P 越高，M/P 就越小，其效果与削减名义货币供给一模一样，使货币需求相对供给上升，把 LM 曲线向上推。

两条曲线怎么读

IS 曲线刻画商品市场，由**财政**变量（G、T、自发支出）使其移动。LM 曲线刻画货币市场，由**货币**变量——名义货币供给与价格水平——使其移动，二者都是通过实际货币供给 M/P 起作用的。

11.3　均衡

把两条曲线画在同一张图上，横轴为产出，纵轴为利率，唯一的交点给出使**两个**市场都出清的组合 (Y_0, r_0)。因果关系是双向传递的：利率通过投资帮助决定产出，产出又通过货币需求帮助决定利率。

这个均衡是稳定的。假设某种力量把产出推到 Y_0 之上——比如把政策利率定得过低。在那个更高的产出上，货币需求会超过固定的供给，利率被抬升，从而抑制投资，直到产出回落到均衡。利率或产出的变动会让经济沿着曲线**移动**；任何其他决定因素的变动则会使整条曲线整体平移。

11.4　IS–LM 中的财政政策

考虑政府购买增加。在商品市场上，这使 IS 曲线向右移动。如果利率仍停留在 r_0，产出就会一路升到 Y_1——这是完整的凯恩斯交叉乘数作用在更大的 G 之上的结果。但是，沿着新的 IS 曲线移向它与未变动的 LM 曲线的交点，讲述的是一个更冷静的故事：

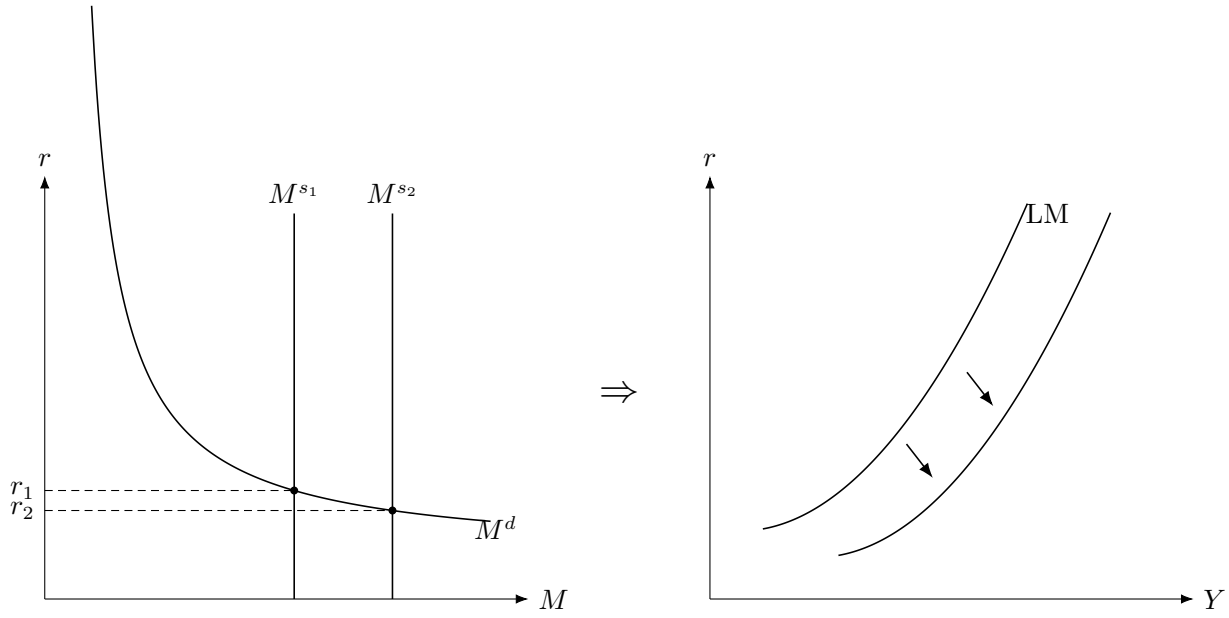

图 11.5: 货币供给增加（从 M^{s_1} 到 M^{s_2}）使每一产出水平下的市场出清利率下降，LM 曲线向下、向右移动。

随着产出上升，货币需求上升，利率被推高到 r_2，而更高的利率又挤出私人投资，把产出从 Y_1 拉回到 Y_2，其中 $Y_0 < Y_2 < Y_1$。

定义 11.3：（第一层）挤出效应

挤出效应（crowding-out effect）是指财政扩张通过货币市场引致利率上升，由此造成的私人投资减少。它使实际实现的产出增量（$Y_2 - Y_0$）小于朴素乘数的预测（$Y_1 - Y_0$）。

减税也以同样的方式（只是经由税收乘数而非支出乘数）使 IS 曲线向外移动，同样的挤出逻辑依旧成立。可以统一概括为：**财政政策作用于 *IS* 曲线**。

11.5　IS–LM 中的货币政策

现在让央行增加货币供给。IS 曲线不受影响；LM 曲线向下移动，因为对任意产出而言，更多的货币在更低的利率上就能使市场出清。在初始产出处利率本会下降；更低的利率刺激投资，产出沿 IS 曲线上升到 Y_1，并伴随一个新的、更低的均衡利率 r_1。

价格水平上升则让同一套机制反向运转。由于它压低了实际货币供给 M/P，它使 LM 曲线向上移动，抬高利率、压低产出。

最后这个实验值得停下来细想，因为它是 IS–LM 与下一章总需求曲线之间的关键枢纽：**对货币市场而言，更高的价格水平等价于更小的货币供给**。于是，固定名义货币存量、只变动 P，就描出了价格水平与均衡产出之间的一种关系——而这恰恰就是总需求。

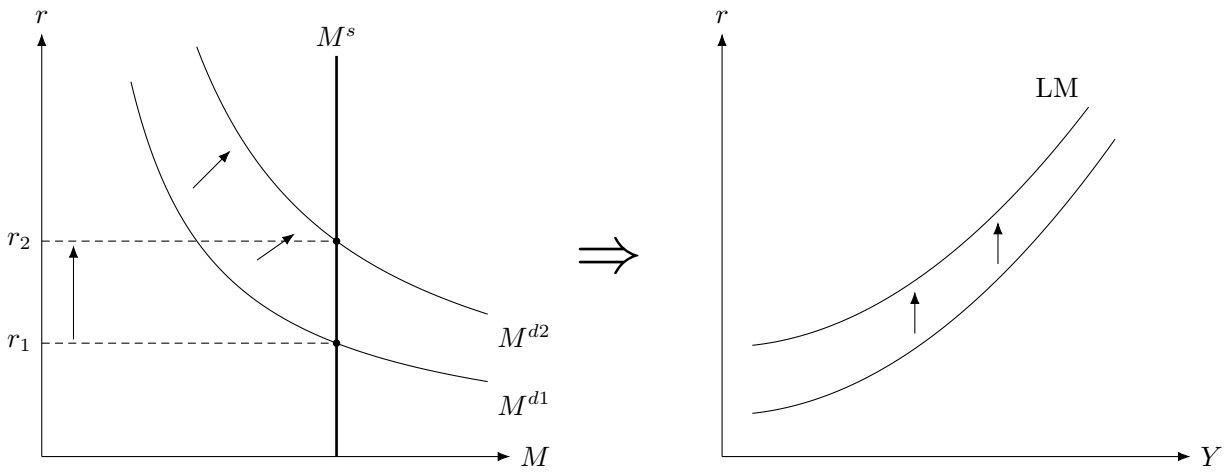

图 11.6: 价格水平上升抬高名义货币需求（压低实际货币供给 M/P）；每一产出水平下的市场出清利率上升，LM 曲线向上移动。

小结

财政政策使 IS 曲线移动；货币政策（以及价格水平）使 LM 曲线移动。财政扩张抬高产出，但也抬高利率，从而挤出投资；货币扩张则通过压低利率来抬高产出。在每一种情形下，利率都会调整，使两个市场同时出清。

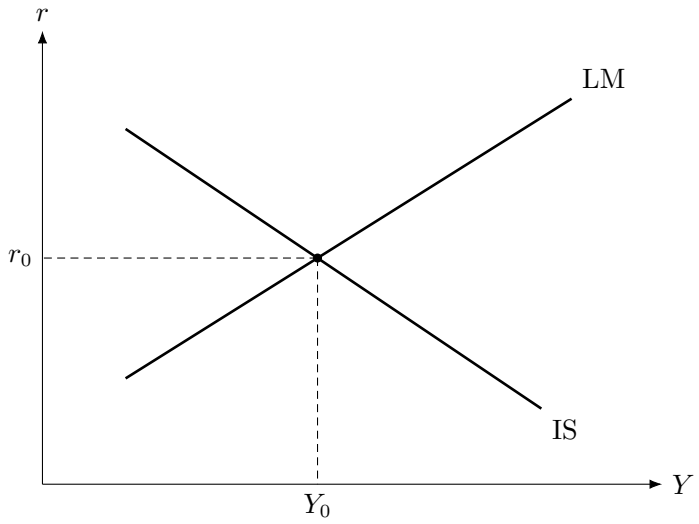

图 11.7: IS–LM 均衡：向下倾斜的 IS 曲线与向上倾斜的 LM 曲线相交于唯一组合 (Y_0, r_0)，在该点商品市场与货币市场同时出清。

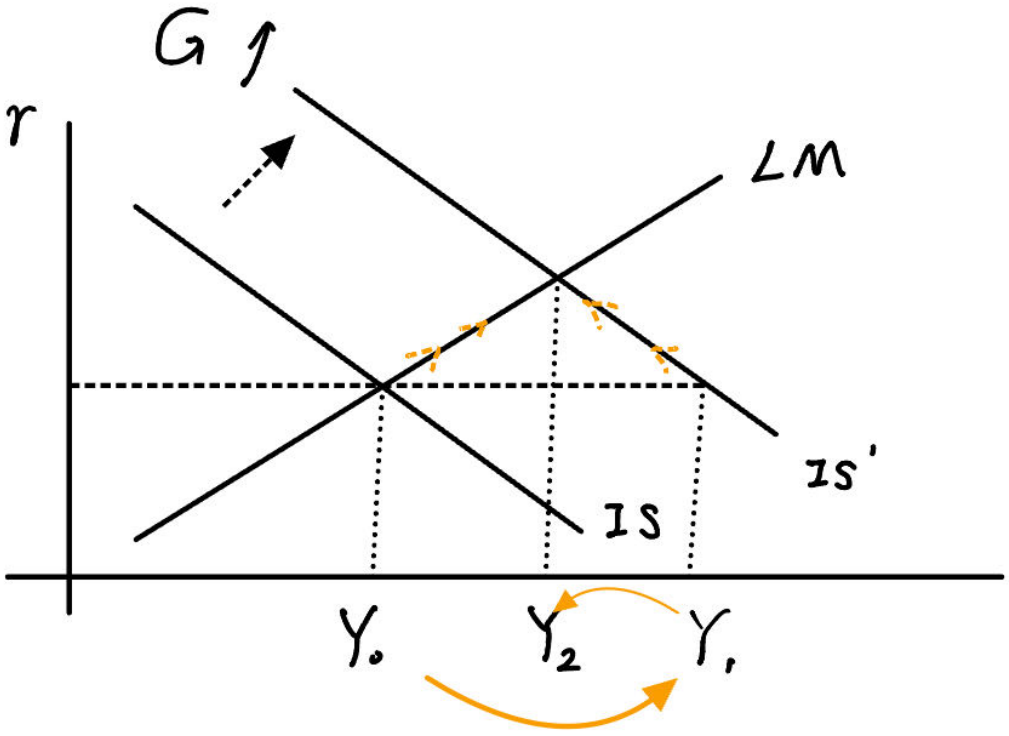

图 11.8: 政府购买增加使 IS 曲线向右移动。若利率固定，产出会升到 Y_1；但产出上升引致利率被推高到 r_2，更高的利率挤出了投资，于是新均衡落在 Y_2，其中 $Y_0 < Y_2 < Y_1$。$Y_1 - Y_2$ 这段缺口就是第一层挤出效应。

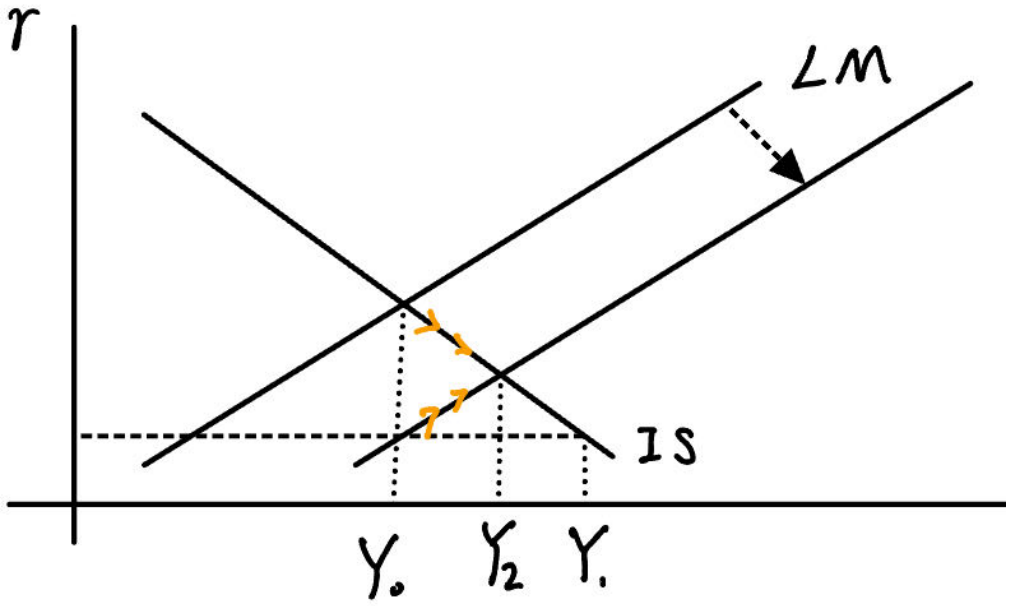

图 11.9: 货币供给增加使 LM 曲线向下移动。产出从 Y_0 升到 Y_1，利率从 r_0 降到 r_1。

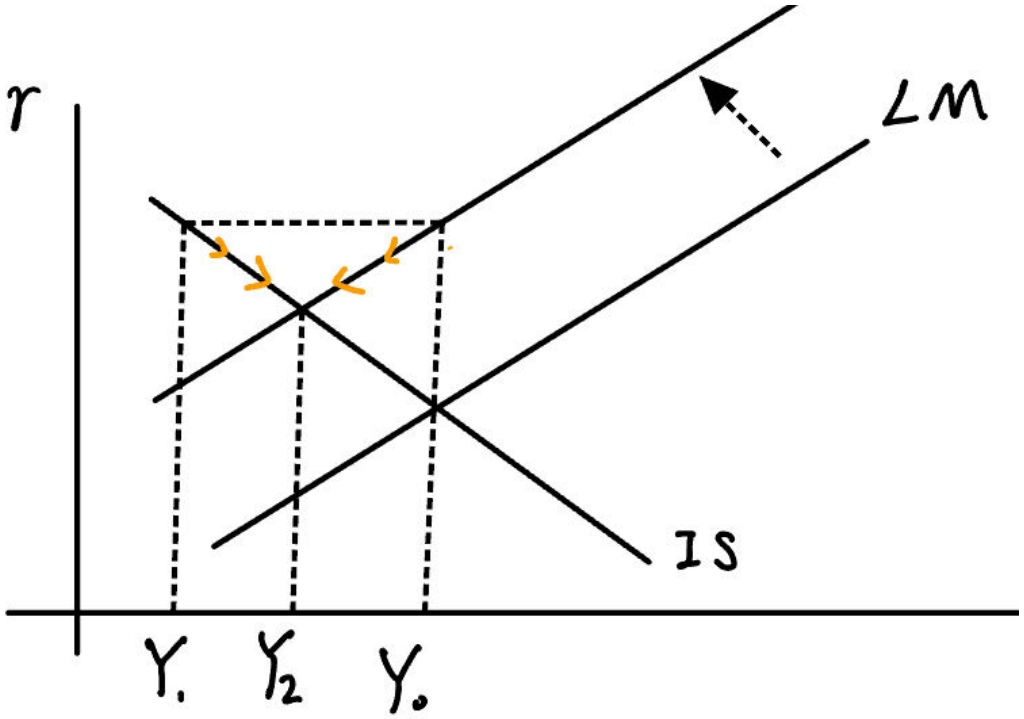

图 11.10: 价格水平上升（实际货币供给 M/P 下降）使 LM 曲线向上移动：利率升到 r_1，产出降到 Y_1。

第十二章　总需求、总供给与供给侧

IS–LM 模型把价格水平固定了下来。要补全短期的全貌，我们就得让价格动起来，问一问产出与价格究竟是如何共同决定的。答案是一条由 IS–LM 继承而来的总需求曲线，再加上一条刻画厂商如何对价格作出反应的总供给曲线。二者的交点决定了短期均衡；而价格已经充分调整后的长期均衡，则落在潜在产出处那条垂直的长期供给曲线上。有了这套工具，我们终于可以把一次财政扩张完整地追踪到底——先是乘数效应，再是两次挤出效应——并看清为什么在长期里，需求刺激改变的只是价格水平。

12.1　总需求曲线

把 IS 方程与 LM 方程联立起来，消去其中的利率，剩下的便是价格水平与均衡产出之间的一个单一关系。它背后的机制在上一章末尾已经见过：价格水平上升会使实际货币供给 M/P 缩小，把 LM 曲线向上推，抬高利率，从而压低投资和产出。因此均衡产出是价格水平的一个单调减函数，$Y^d = Y(P)$。

注（总需求曲线为什么向下倾斜）.

> 总需求曲线并不是微观经济学里那个故事——买家看到价格更高就少买。在这里，货币可以是完全中性的、公众可以是完全理性的，曲线却依然向下倾斜：价格水平上升降低了实际货币余额，抬高了利率，从而挤出了投资。这条向下的斜率讲的是货币市场和利率，而不是相对价格。

定义 12.1: 总需求曲线

总需求（*aggregate demand*，AD）曲线是所有与 IS–LM 均衡相容的 (Y, P) 点的轨迹。它向下倾斜，是因为价格水平上升会减少实际货币余额，从而抬高利率、压低投资和产出。货币政策与财政政策都会使总需求曲线平移；而价格水平本身的变化，则是沿着曲线移动。

12.2　总供给曲线

在长期，产出由资本、劳动和技术决定，与价格水平无关，因此长期总供给（LRAS）曲线在潜在产出 $\overline{Y}$ 处垂直。在短期，价格是粘性的，供给会对它作出反应。当价格水平上升快于成本（工资的调整存在滞后）时，利润空间被拉大，厂商便把生产扩张到潜

在水平之上，于是短期总供给（SRAS）曲线向上倾斜。它的斜率衡量的是供给对价格有多敏感：经济中闲置的资源越多，SRAS 就越平缓；当资源被充分利用时，SRAS 就几乎垂直。

三线同点的均衡

在完整的长期均衡中，长期供给曲线、短期供给曲线和总需求曲线交于同一个点，产出恰好处在潜在水平 $\overline{Y}$。

12.3　财政扩张的完整过程

现在我们可以把政府购买的增加，同时放进三张图里追踪——凯恩斯交叉、IS–LM 与 AD–AS——看着产出依次经过四个水平。

经济最初处于潜在产出 Y_0 的长期均衡。这时政府购买增加。

1. **乘数效应**（$Y_0 \to Y_1$）。计划支出上升，经由乘数作用，产出本会达到 Y_1——前提是利率和价格水平都保持固定。（在 IS 曲线确定其水平平移量时，我们固定了利率；在 AD 曲线确定其平移量时，我们固定了价格水平。）

2. *IS–LM*，**第一次挤出**（$Y_1 \to Y_2$）。让利率对上升的货币需求作出反应，r 的上升挤出了私人投资，产出回落到 Y_2。

3. *AD–AS*，**第二次挤出**（$Y_2 \to Y_3$）。在 Y_2 处，总需求仍然大于短期供给，于是价格水平上升。更高的价格水平缩小了实际余额，进一步抬高利率，挤出更多投资；产出最终稳定在 $Y_3 < Y_2$。这便是**第二次挤出效应**。

4. **从短期到长期**（$Y_3 \to Y_0$）。随着时间推移，成本（尤其是工资）向更高的价格水平靠拢，利润空间被压缩，短期供给曲线向左平移。每一次左移都重新造成需求超过供给的局面，又把价格往上推，直到供给与需求在潜在产出处重新相遇。

两次挤出效应

第一次挤出：需求扩张抬高货币需求，抬高利率，压低投资。第二次挤出：同一次需求扩张抬高了**价格水平**，价格水平降低实际余额，再次抬高利率，进一步压低投资。乘数因此被削减了两次，对产出的净影响可能很小。

有几点告诫值得明明白白地说出来。价格**只有**在供给与需求失衡时才会变动。IS–LM 均衡是完整均衡的必要条件，却不是充分条件——价格水平还必须让商品市场上的供给与需求出清。而长期的教训是冷峻的：当供给垂直时，单纯的需求刺激只会抬高价格水平，而不会增加产出。

注（“长期来看，我们都死了”）.

> 凯恩斯这句俏皮话是一种辩护，而非让步。凯恩斯主义政策瞄准的是**短期**：它是逆周期的，意在把经济迅速拉出衰退的泥潭，而不是在繁荣之上再人为制造一场繁荣。两次挤出效应究竟会不会留下乘数的大部分，取决于各条曲线相对斜率的大小——

> 平缓的 LM 曲线或 AD 曲线，又或者大量闲置的资源，都会让刺激政策的效果基本保留下来。当 LM 曲线平缓到即便产出大幅变动也几乎推不动利率时，经济就处在**流动性陷阱**（*liquidity trap*）之中，货币政策失灵，而财政政策恰恰最为有效。

12.4　供给侧

上面的实验全都在移动需求。供给的移动则表现得很不一样。沿着一条给定的短期菲利普斯曲线（Phillips curve），通货膨胀与失业反向变动，因此需求驱动的繁荣是用更高的价格换取更低的失业。而一次负向的**供给冲击**，会打破这种取舍关系。

当短期供给收缩时，产出下降（失业率上升）与价格水平上升（通货膨胀上升）同时发生——这就是被称为**滞胀**（*stagflation*，即 *stagnation* 与 *inflation* 的合成）的“双输”局面。需求政策对它无能为力：刺激会加剧通胀，收缩则会加剧失业。解药只能来自供给侧——提高厂商生产的意愿和能力。这正是上世纪八十年代供给学派（supply-side school）的逻辑：削减边际税率，恢复人们工作和投资的激励，并利用第 十三 章给出的那个事实——越过某一点之后，更低的税率未必意味着更少的财政收入。

长期供给的增加——生产能力的真正增长——是上面的镜像：产出上升，价格水平下降。这正是经济增长的深层含义：用更低的价格水平获得更多的产出，是这个模型里唯一的“免费午餐”。

分析一次短期供给冲击时，我们把长期供给固定不变；短期的图仍然可以把长期的变化作为两张静态截面的比较显示出来。

12.5　短期供给曲线的微观基础

短期供给为什么会向上倾斜？三个模型给出三种理由，但它们都导向同一个简化式，

$$Y = \overline{Y} + \alpha\,(P - P^{e}), \qquad \alpha > 0,$$

其中产出只有在价格水平超出预期时，才会高于潜在水平。

12.5.1　粘性工资模型

工人和厂商在价格水平揭晓之前就签下名义工资合同，把名义工资固定为 $W = \omega\,P^{e}$，其中 ω 是目标实际工资。于是实现的实际工资为

$$\frac{W}{P} = \omega\,\frac{P^{e}}{P}.$$

当 $P > P^{e}$ 时，实际工资低于目标，厂商多雇人，产出升到潜在水平之上；只有当 $P = P^{e}$ 时，就业和产出才处在它们的潜在水平。该模型预言实际工资是**逆周期**的——在繁荣中下降——这一点在长期里与数据相违背，不过在较短的时间跨度上，这套机制仍有几分作用。

12.5.2　不完全信息模型

这里所有工资和价格都是完全弹性的，但每个生产者只观察到自己那种商品的名义价格，看不到总体价格水平（要滞后一段时间才能知晓）。供给取决于**相对**价格；由于缺少价格水平的信息，厂商便用自己的预期 P^e 来替代。如果总体价格水平上升而厂商还看不到这一点，它就会把自己商品名义价格的上升误当成一个有利的**相对**价格信号，于是多生产，结果总产出超过了潜在水平。价格水平只能事后才知道，这是一个相当符合现实的假设。

12.5.3　粘性价格模型

产出与价格同向变动，但因果既可以从 P 流向 Y，也可以同样轻易地从 Y 流向 P。每家厂商都想根据价格水平和需求状况来定价，

$$p = P + \alpha\,(Y - \overline{Y}), \qquad \alpha > 0,$$

即产出高于潜在水平时就多要价。可以随意改价的厂商就照这条规则办。而必须提前承诺价格的厂商（因为菜单成本、长期合同或客户关系），则转而依据预期来定价，

$$p = P^e + \alpha\,(Y^e - \overline{Y}),$$

为简便起见，假设它们预期产出处于潜在水平。设有比例为 s 的厂商价格粘性，其余 $1-s$ 完全弹性。均衡价格水平是两类厂商价格的加权平均，

$$P = s\,P^e + (1-s)\big[P + \alpha(Y - \overline{Y})\big],$$

整理后得到

$$P = P^e + \left[\frac{(1-s)\,\alpha}{s}\right](Y - \overline{Y}), \qquad \text{即} \qquad Y = \overline{Y} + \frac{s}{\alpha(1-s)}\,(P - P^e).$$

如果每家厂商都是弹性的（$s = 0$），名义收入的增加就会全部传导到价格上：货币是中性的，供给垂直。正是粘性定价者的存在，才给了供给曲线一个正的斜率。与粘性工资模型不同，粘性价格模型意味着实际工资是**顺周期**的：一次负向需求冲击会让弹性厂商降价、让粘性厂商减产并减少劳动需求，而劳动需求曲线的左移压低了实际工资——这一点与数据吻合得更好。

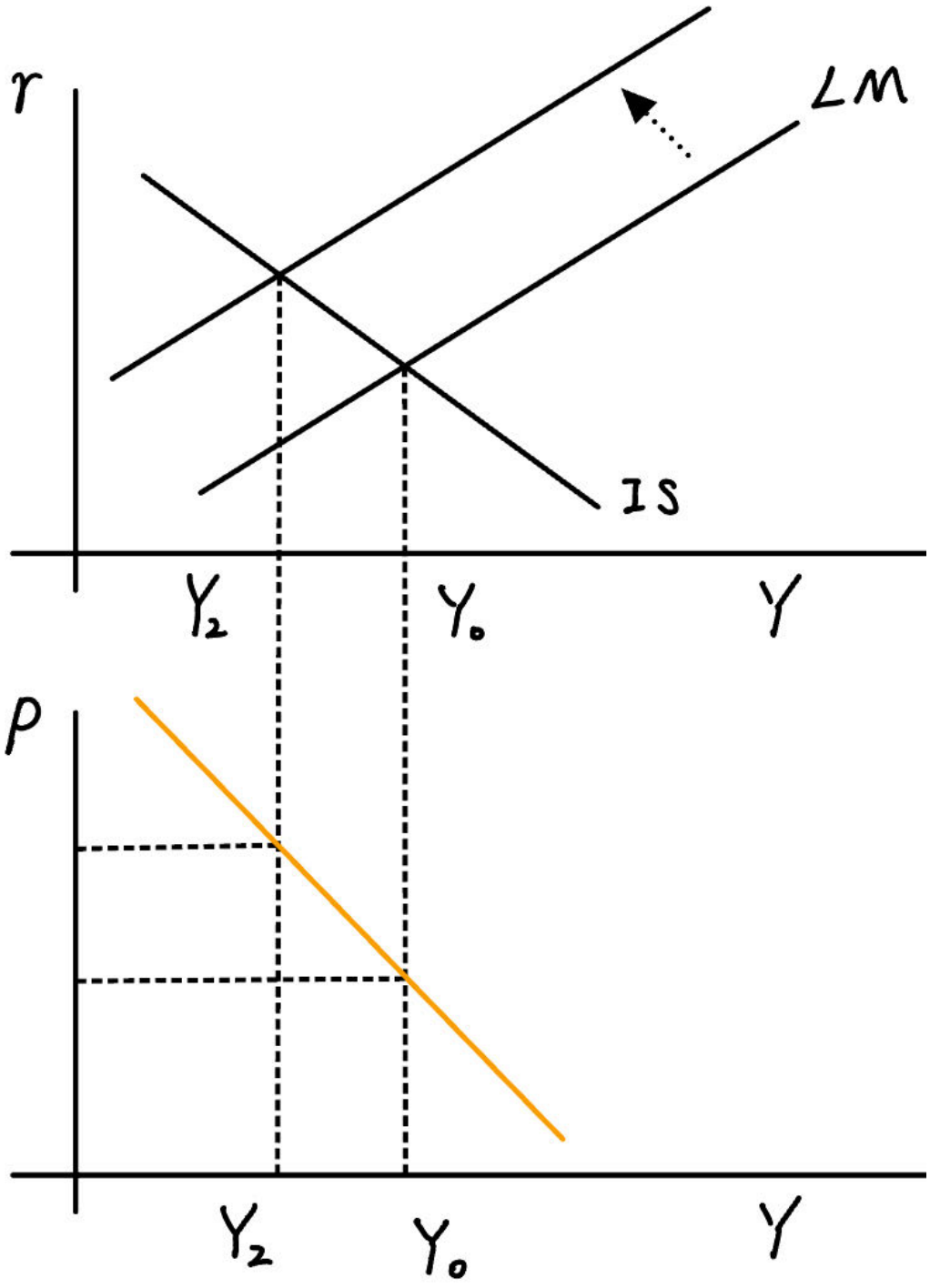

图 12.1: 总需求曲线的推导。在上图中，价格水平上升降低了实际货币供给，使 LM 曲线上移，产出从 Y_0 减少到 Y_2。把每个价格水平与其对应的产出画在一起（下图），便描出了向下倾斜的总需求（AD）曲线。

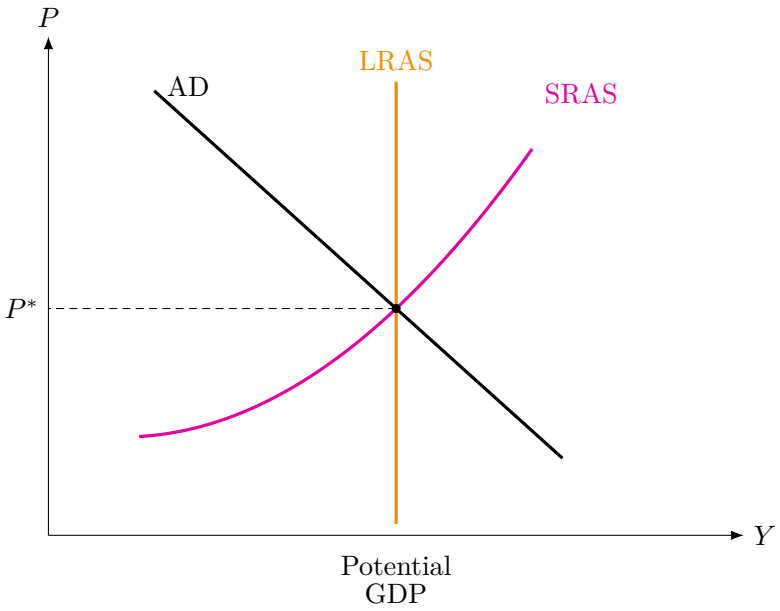

图 12.2: 总供给与总需求。长期总供给（LRAS）在潜在产出处垂直；短期总供给（SRAS）向上倾斜；总需求（AD）向下倾斜。在完整的均衡中，三条曲线交于同一点。

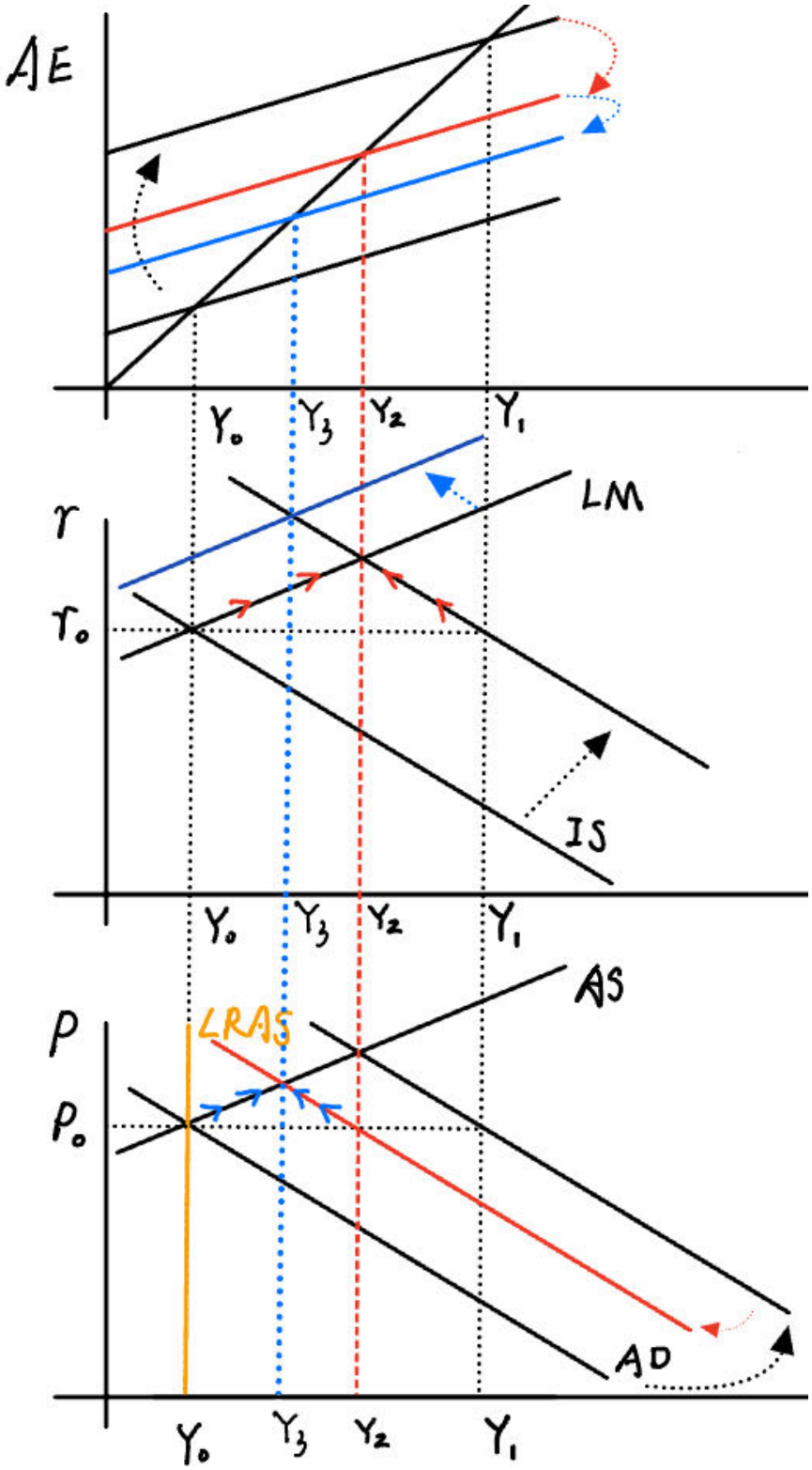

图 12.3: 财政扩张的完整调整过程。黑色：在利率和价格水平都固定的情况下，乘数效应把产出抬到 Y_1。红色：考虑进流动性偏好后，利率上升，第一次挤出效应把产出拉回 Y_2。蓝色：再让价格水平上升，第二次挤出效应把产出拉回 Y_3。在长期，成本上升又把产出送回潜在水平 Y_0。整个过程中始终有 $Y_0 < Y_3 < Y_2 < Y_1$。

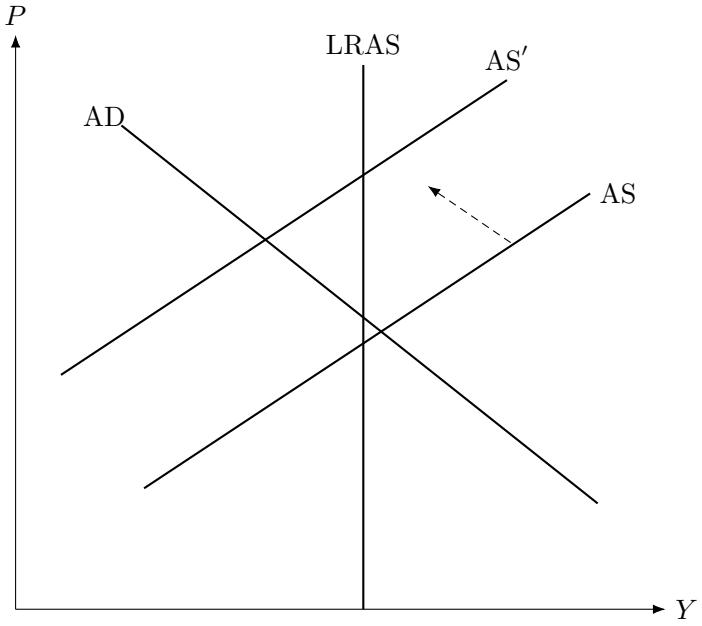

图 12.4: 一次负向供给冲击。短期供给曲线左移：产出下降，价格水平却同时上升——这就是滞胀。

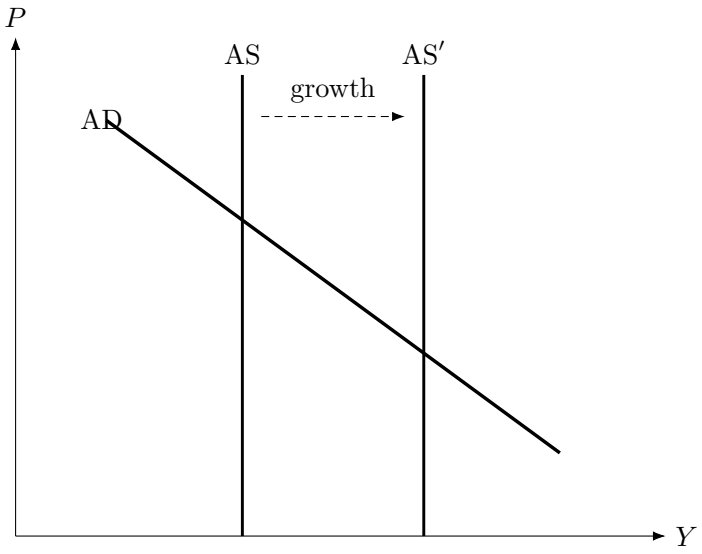

图 12.5: 长期供给增长。供给曲线右移：产出上升，价格水平下降——这就是真正经济增长的“双赢”。

第十三章　财政政策：李嘉图等价、拉弗曲线与政府债务

凯恩斯的分析框架把一次财政扩张看作 IS 曲线的一次移动，至于这笔钱是靠税收还是靠借债来筹的，则交由某种未加说明的组合去解决。古典传统对此提出异议：筹资方式根本不是细枝末节。一个有前瞻性的家庭明白，今天的赤字就是明天的税收，并据此安排自己的计划——以至于在足够强的假设之下，征税与发债之间的选择对任何实际变量都不产生影响。本章在一个干净的两期模型里发展这一论断，它被称为**李嘉图等价**（Ricardian equivalence）；进而辨明它所依赖的那些假设，由此看清它在现实中失效的种种理由；随后转向被这一等价结论重新框定、而非彻底了结的两个问题：税率究竟能走多高才会变得自拆台脚（拉弗曲线），以及政府债务是不是一种负担。

13.1　李嘉图等价

古典经济学家——斯密、李嘉图——认为市场能解决大多数问题，政府在很大程度上是不必要的。然而他们无法解释经济大萧条，凯恩斯正是在这里登场，提出用政府支出来撑起需求的主张。古典学派的回应——后来由新古典学派给出了严谨的微观基础——是：政府支出无论靠税收还是靠债务来筹措，都会挤出私人消费，因为理性的消费者看穿了赤字：今天卖出的一张债券，明天必须靠一笔税收来赎回。

考虑一个由单一代表性消费者和一个政府构成的经济体，存续两期。政府只在第一期支出，$G_1 = G$，$G_2 = 0$，并用第一期税收 t_1 与债券 b 的某种组合来为之筹资，即 $G = t_1 + b$。在第二期，它征收一笔税 $t_2 = (1+r)b$ 来连本带息偿还债券。政府的预算可以用现值写成

$$G = t_1 + \frac{t_2}{1+r},$$

即税收的现值等于支出的现值。

定义 13.1: 李嘉图等价

李嘉图等价是这样一个命题：对于给定的政府支出路径，在税收筹资与债券筹资之间的选择，不改变消费者的预算集——从而不改变消费与福利。重要的只是支出的现值，而非它如何筹措。

消费者的偏好为 $U(C_1, C_2) = u(C_1) + \beta\, u(C_2)$，面临各期的预算约束

$$C_1 = y_1 - t_1 - s, \qquad C_2 = y_2 - t_2 + (1+r)s,$$

其中 s 是私人储蓄。由于唯一的资产就是政府债券，在均衡中私人储蓄等于发行的债券，$s = b$。把两期的预算约束合并以消去 s，再代入政府预算约束，便得到消费者的一生预算约束

$$C_1 + \frac{C_2}{1+r} = y_1 + \frac{y_2}{1+r} - \left(t_1 + \frac{t_2}{1+r}\right) = y_1 + \frac{y_2}{1+r} - G.$$

消费的现值等于收入的现值减去支出的现值。G 在 t_1 与 b 之间的拆分方式已经完全消失了。

关键所在

消费者的一生预算只取决于 G 本身，而与 G 是靠税收还是靠债券筹得无关。两种 G 相同的筹资方案给出相同的 (C_1, C_2)。这就是李嘉图等价。

例（对数效用）.

取 $u(C) = \log C$，于是消费者求解

$$\max_{s}\ \log(y_1 - t_1 - s) + \beta \log\big(y_2 - t_2 + (1+r)s\big).$$

之所以对 $s(=b)$ 求最优，是因为收入是外生给定的、税收由政府决定，消费者能够根据自己的偏好来确定储蓄量，而储蓄即投资于债券。一阶条件为

$$\frac{1}{y_1 - t_1 - s} = \frac{\beta(1+r)}{y_2 - t_2 + (1+r)s},$$

其解为

$$b = s^* = \frac{\beta(1+r)(y_1 - t_1) - (y_2 - t_2)}{(1+\beta)(1+r)}.$$

比较两种极端的筹资方案。在*纯税收筹资*下，$G = t_1$，$b = 0$，$t_2 = 0$，最优要求 $G = y_1 - \dfrac{y_2}{\beta(1+r)}$。在*纯债券筹资*下，$b = G$，$t_1 = 0$，$t_2 = (1+r)G$，代入后恰好给出同一个条件 $G = y_1 - \dfrac{y_2}{\beta(1+r)}$。两种情形下政府支出相同，因此 C_1 与 C_2 完全相同：等价成立；由同样的论证，它对介于两者之间的任何税收—债券组合也都成立。

13.1.1　那些假设，以及它们为何失效

等价是一个定理，而像任何定理一样，它有自己的前提。

假设 13.2: 李嘉图等价的前提

1. **完全预见／理性**：消费者预见到 $t_2=(1+r)b$，并对两期一并进行优化。
2. **消费者存活两期**：从赤字中获益的，正是那个承担未来税收的同一主体。
3. **纳税人与债券持有人是同一个消费者**。
4. **税是总量税**。

每一条假设一旦失效，都会以一种富有启发的方式破坏等价。

- *有限寿命*。如果消费者活不到缴未来那笔税的时候，那么债券筹资的赤字会让当代人多消费，并迫使下一代少消费；消费不再在两代之间被平滑。（利他的遗赠通过把两代人联系起来，可以部分地恢复等价。）
- *纳税人 $\neq$ 债券持有人*。这会产生两个效应。一是*分配*效应：现在少买债券的人多消费，而将来不得不多买债券的人少消费。二是*对储蓄与资本*的效应：当赤字未被额外的私人储蓄完全对冲时，政府借债吸走了本会用于投资的资金，使资本存量减少。
- *扭曲性税收*。如果税不是总量税，而是收入相关的税或从价税（例如累进税），它就会改变相对价格，把资源配置扭曲到偏离有效配置之外。现在征税制造的是当下的扭曲；先借债、将来再征税制造的是未来的扭曲。由于扭曲在何时发生现在变得要紧，筹资选择不再中性——这正是*税收平滑*（tax smoothing）的依据：把扭曲性的税薄薄地摊到各期，并倚重债务来做到这一点。

13.2　拉弗曲线

一旦税收是扭曲性的，那么不仅它的时点，连它的*水平*都变得要紧，提高税率也未必能提高收入。一个静态的劳动供给模型把这一点说清楚了。一位消费者在消费 C 与闲暇 l 之间权衡，效用为 $U(C,l)=C+\log l$，面临对劳动收入征收的比例税 τ；设工资为 w、总时间单位化为 1，则预算约束为

$$C=(1-\tau)(1-l)\,w.$$

对 l 最大化 $C+\log l$，一阶条件为 $-(1-\tau)w+1/l=0$，于是最优的闲暇与劳动供给为

$$l^{*}=\frac{1}{(1-\tau)w},\qquad 1-l^{*}=1-\frac{1}{(1-\tau)w}.$$

把预算约束改写为 $C+(1-\tau)w\,l=(1-\tau)w$ 便可看出：更高的 τ *降低*了闲暇的有效价格 $(1-\tau)w$，于是由替代效应闲暇上升、劳动供给下降。税收收入是税率乘以（被征税的）劳动收入，

$$R=\tau w\,(1-l^{*})=\tau w-\frac{\tau}{1-\tau}.$$

这两项有一个清晰的读法：τw 是*潜在收入效应*，即消费者一直工作时所能取得的收入；而 $-\dfrac{\tau}{1-\tau}$ 是*扭曲效应*，其中 τ 是税率、$1-\tau$ 是它所诱发的扭曲。

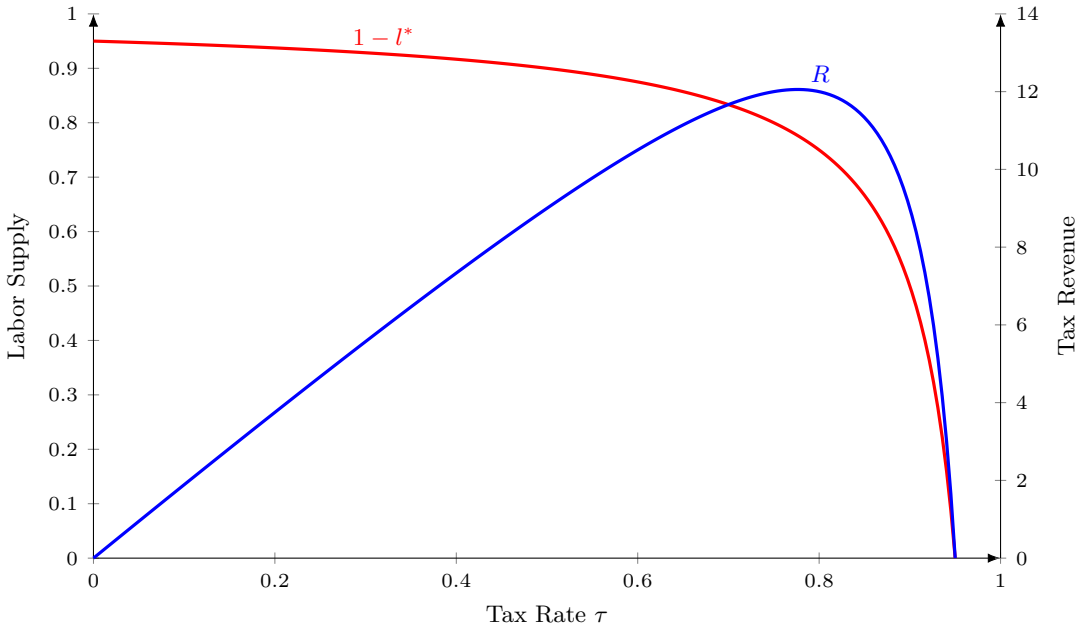

图 13.1: 拉弗曲线。随着税率上升，劳动供给（左轴）下降，先缓后陡；税收收入（右轴）画出一条倒 U 形，在税率效应占上风时上升，一旦不断收缩的税基占上风便转而下降。

求导可得 $\mathrm{d}R/\mathrm{d}\tau = w - 1/(1-\tau)^2$，它在使收入最大化的税率处为零；越过该点，收入便下降。图 13.1 展示了这条倒 U 形：越过峰值之后，税基收缩得比税率上升得更快，于是收入下降。所谓*禁止性*（prohibitive）税率，是指越过峰值的税率，此时劳动供给已经塌缩到提高税率反而降低总收入的地步；在那里，对最优配置的扭曲极为严重。

现实启示

高税率有很强的扭曲性，因此政府应当把扭曲*平滑*到各期，并更多地倚重债务。在战争时期，当支出必须骤增时，教科书式的处方是只适度提高税率，而用债券为其中的大部分融资——这恰恰是税收平滑的实践。

13.3　政府债务

关于公共债务，有三个反复出现的问题。

*政府债务可持续吗？*一个主权国家的债务不同于一个家庭的债务。只要国家存在，它的债务原则上就可以无限地滚动下去——发新债以赎旧债——而永远不必全额偿还。把债务与 GDP 相比是常见的做法，但并不完美：GDP 是流量而债务是存量；政府在债务背后持有大量资产；并且只要政权及其信用维系，债务就有担保。大体上说，债务是可控的，尽管即便财政尚有余地，评级机构的一次降级也可能袭来。

*债务是经济的负担吗？*当政府借债来支出时，有些人今天少消费，而负担通过未来的税收在各代人之间分摊。通常的场合是经济萧条，政府支出以支撑需求。对于公众而言，持有这些债券本身就是一项正收益的资产，只要债务不断滚动，它就是安全的。

*是否存在最优的债税比？*在李嘉图等价之下，税收与债券是等价的，因此并不存在最优的比例。但等价依赖于现实世界所违背的那些假设——主要是总量税——所以一旦税收是扭曲性的，税收平滑的论证便重新登场，于是确实存在一个真实的权衡。从历史上看，美国在第二次世界大战前后主要靠税收为自己筹资；自 1970 年代以来，天平已偏向债务，税收收入平稳，但开支扩张、赤字持续。

13.4　社会保障

社会保障支出是政府预算的一大部分，而它的设计回响着同一个筹资问题。

定义 13.3: 两种社会保障体制

确定给付制（defined-benefit，即现收现付制）用今天劳动者的缴费来为今天的退休者发放——年轻人供养老年人。*确定缴费制*（defined-contribution，即完全积累制）把每位劳动者的缴费拿去投资，并在其退休时连同投资收益返还给他——现在的自己供养未来的自己。

关于现收现付制有一个关键事实：“账户”里的钱并不真的留作储备，它早已花在了当下的退休者身上。这样一种制度对投资风险是稳健的，但对人口结构的变化极为脆弱。随着人口老龄化——更少的劳动者供养更多的退休者，老年抚养比不断上升——现收现付制趋于赤字。补救之道无一令人愉快：提高缴费率会加重劳动者与企业的负担；削减给付在政治上不可能；推迟退休年龄或许在所难免。完全积累制避开了人口结构的挤压，却转而承担通胀与投资风险。

第十四章　菲利普斯曲线、预期与政策

总供给曲线告诉我们：只有当价格意外地高于人们的预期时，产出才会越过潜在水平。把这个意外翻译成通胀与失业的语言，它就成了菲利普斯曲线（Phillips curve）——通胀与失业之间那条短期的取舍菜单，而政策似乎正可以加以利用。但它究竟能不能被利用，完全取决于预期；而预期这个问题——公众如何形成预期、政策又能否塑造预期——正是现代货币与财政政策辩论的主线：政策该主动还是被动，该依规则还是凭相机抉择。本书的最后一章，便顺着这条线索从菲利普斯曲线一路走到泰勒规则。

14.1　菲利普斯曲线

从经验上看，在一个完整的经济周期里，通胀与失业总是朝相反的方向变动。现代的、引入了预期的菲利普斯曲线把这一关系写成

$$\pi - \pi^e = -\beta\,(u - u^n) + v, \qquad \beta > 0,$$

其中 $u - u^n$ 是周期性失业（失业率对其自然率的偏离），v 是供给冲击，π^e 是预期通胀。当失业率低于自然率时，实际通胀就高于预期。

定义 14.1: 菲利普斯曲线

引入预期的菲利普斯曲线（expectations-augmented Phillips curve）把意外通胀同周期性失业联系起来：$\pi - \pi^e = -\beta(u - u^n) + v$。它所刻画的取舍是通胀与周期性失业之间的取舍，并且是相对于预期通胀而言的。

14.1.1　推导

菲利普斯曲线其实就是改头换面的供给曲线。从短期总供给出发，$Y = \overline{Y} + \alpha(P - P^e)$，解出价格水平，

$$P = P^e + \frac{1}{\alpha}\,(Y - \overline{Y}).$$

由于这条关系是从数据中读出来的，我们把供给冲击 v 作为残差项附在后面，再对上一期的价格水平 P_{-1} 取差分，

$$P - P_{-1} = (P^e - P_{-1}) + \frac{1}{\alpha}\,(Y - \overline{Y}) + v,$$

用通胀率写出来，便是 $\pi-\pi^e=\dfrac{1}{k\alpha}\,(Y-\overline{Y})+v$，其中 k 是一个比例常数。最后引用**奥肯定律**（Okun's law）——这条经验规律说产出与失业总是同向变动：粗略地说，失业率每多上升一个百分点，大约对应损失两个百分点的 GDP。奥肯定律让我们可以用周期性失业替换产出缺口，

$$\frac{1}{k\alpha}\,(Y-\overline{Y})=-\beta\,(u-u^n),$$

代回去就得到了菲利普斯曲线。这一关系是相关而非因果，但它把一条政策反复试图加以利用的联系给量化了出来。

14.2　预期

一切都系于 π^e。在信息不完美或价格黏性的世界里，人们是依照自己的预期行事的，于是整条菲利普斯曲线的位置都会随这一预期一起移动。关于预期如何形成，有两套理论，它们给出的政策结论截然不同。

14.2.1　适应性预期

适应性（adaptive）预期是把过去外推得到的。最简单的一种设定取 $\pi^e=\pi_{-1}$，这就把菲利普斯曲线变成了通胀的一条运动方程，

$$\pi=\pi_{-1}-\beta\,(u-u^n)+v.$$

通胀因此具有惯性：在没有冲击的情况下，过去的通胀会传导到当前乃至未来的通胀。这本质上就是对历史数据做回归所得到的结果，而它的软肋也正在于此——它预设过去那些关系会一直延续下去。

注（卢卡斯批判）.

> 用历史数据估计出来的关系去预测政策效果是不可靠的，恰恰因为政策本身会改变预期。用卢卡斯（Lucas）的话说："Forecasting the effects of policy changes has often been done using models estimated with historical data. Such predictions would not be valid if the policy change alters expectations in a way that changes the fundamental relationships between variables." 当政策体制发生变化时，历史相关关系会出现**结构断裂**（structural break），适应性预期的预测便随之失效。

14.2.2　理性预期

理性（rational）预期动用的是模型和一切可得的信息，而不只是过去——预期实际上是在预测自己以及所有其他人的决策。因此分析就不能只靠回归，它需要一个一般均衡模型；这个模型虽然未必能厘清单个的因果，却把每一条机制都纳入其中，给出一套自洽的冲击—反应（shock–response）。在理性预期下，原则上可以实现**无痛去通胀**（painless disinflation）：如果中央银行能用一个可信的信号直接改变预期，供给曲线就会移动，通胀随之下降而不必经历衰退——牺牲率（sacrifice ratio）未必成为约束。对

美国而言，理性预期约等于“相信美联储”。而当央行释放的信号过于晦涩、不足以支撑起量化的预期——政策的逻辑难以读懂——理性预期就退化成了小道消息，市场也随之频频震荡。

14.3　自然率假说与延滞性

这一切的背后是**自然率假说**（natural-rate hypothesis）：存在一个潜在产出水平（以及一个自然失业率），经济在那里处于均衡，长期中必将到达、并且终究总会回归，而短期不过是对它的偏离。这是宏观经济学的组织性假设之一。

与之竞争的观点是**延滞性**（hysteresis）：暂时性冲击可以推动长期水平本身，于是潜在产出虽然存在，却并非一个固定不变的数字——它取决于经济自身的历史。疫情就是一个现成的例子。一段长时间的失业会侵蚀工人的技能，等到需求复苏时他们已找不到与之匹配的工作，议价能力随之丧失；旷日持久的衰退会让消费习惯转向预防性（precautionary）储蓄；供应链外移，资本损耗，投资趋于保守。每一条渠道都让一个本应转瞬即逝的冲击留下了永久的印记。

14.4　政策的实施

14.4.1　主动与被动

主动（active）的立场认为，在经济萧条时政府有责任采取行动——减轻民众的痛苦、刺激经济。但有两个问题让主动政策变得复杂。

第一，政策的作用存在**时滞**（lags）。**内部时滞**（inside lag）是从认识到需要行动、到真正出台对策所花的时间——要确认经济确实在收缩，得靠好几项指标共同印证，而执行（尤其是财政政策的执行）又需要协调。**外部时滞**（outside lag）是政策影响到经济所花的时间，而这一影响还可能矫枉过正。第二，经济本身含有**自动稳定器**（automatic stabilizers）——这些制度无需刻意行动、也几乎没有时滞，便能熨平周期。累进所得税在繁荣期会随额外收入抽取越来越高的份额，从而抑制乘数效应；失业保险在衰退期支撑收入、进而支撑消费；在浮动汇率下，强劲的出口繁荣会推升本币、从而给需求降温。

要把主动政策用好，政府必须尽早判断出经济周期所处的阶段——观察那些先于周期变动的领先指标（综合指数、投资、PMI、房地产资金到位量），并在环境变化到历史关系出现断裂时，改用结构模型而非简单外推。它还必须**领先于**公众的预期：如果一个理性的公众早已料到某项政策，这项政策的效果就会被削弱。假设过去的数据显示货币增长会降低失业率，公众于是预期到了这一点；那么当中央银行扩大货币供应时，预期通胀会同步上升，失业率并不会下降——被预料到的政策是中性的，此时需要一项不同的、或者出人意料的政策。

14.4.2　规则与相机抉择

更深一层的选择，是政策究竟应当遵循一条公开的**规则**（rule），还是凭**相机抉择**（discretion）来制定。

定义 14.2: 规则与相机抉择

依规则行事的政策，是依照一条明确、公开的公式来设定和调整的，因而可以被预期。凭相机抉择行事的政策，则视具体情形逐案应对，无法被预期。

最典型的规则是关于联邦基金利率的泰勒规则（Taylor rule），它根据通胀和产出缺口来设定目标利率，

$$r_{\text{ff}} = 2 + 0.5\,(\pi - 2) - 0.5 \cdot \text{GDP gap}, \qquad \text{GDP gap} = 100 \cdot \frac{\overline{Y} - Y}{\overline{Y}},$$

其中 2% 被取作自然通胀率。注意这里的符号约定：缺口被定义为产出低于潜在水平的亏空，因此减去它就意味着——当产出高于潜在水平时（繁荣，$Y > \overline{Y}$，缺口为负），规则会抬高利率；当产出低于潜在水平时则压低利率——这正是稳定经济的反应，等价于加上通常意义上的产出缺口 $(Y - \overline{Y})/\overline{Y}$。这是一条透明而可信的规则：利率由公式确定，具有独立性和可预测性，而不必去追问某次通胀的来源是什么。相比之下，中国追求让 M_2 增速“匹配”名义 GDP 增速——这个目标同货币数量论赋予货币的逆周期角色配合得相当别扭，而它的“匹配”程度、以及它所依赖的对名义 GDP 增速的预测，本身又都是相机抉择的。

支持规则的理由

有若干论据支持规则胜过相机抉择。决策者可能并不具备所需的专业知识；他们的判断可能建立在模糊或多变的原则之上，而这些原则又可能与公众的并不一致；他们还可能存在时间不一致，从而损害可信度（credibility）——而预期、进而政策自身的有效性，恰恰仰赖于这一可信度。在相机抉择的决策者其能力与一致性都存疑的地方，一条设计良好的规则可以做得更好。