ECON 2070 Course Notes

Fernando Duarte

Fall 2025

Contents

Chapter 1

Growth

1.1 Kaldor's Stylized Facts

In 1961, Nicholas Kaldor (1961) proposed six stylized facts of economic growth, broad empirical patterns observed in advanced economies up to the mid-20th century over long periods of time. These original Kaldor facts were:

1. Constant factor income shares: The shares of national income accruing to labor and capital are roughly constant.
2. Steady growth of capital per worker: the stock of capital per worker grows at a roughly constant rate.
3. Steady growth of output per worker: output per worker grows at a roughly constant rate.
4. Stable capital-output ratio: the ratio of capital to output is roughly constant.
5. Stable rate of return on capital: the rate of return on capital remains roughly constant.
6. Cross-country growth variation: different countries exhibit appreciable variations in growth rates of output and output per worker.

Kaldor's intent was not to claim these quantities never change at business-cycle frequency, but that over long periods and across leading industrial nations, these ratios and growth

rates appeared surprisingly stable. His facts became a foundation for models of long-run growth.

Today, however, we have decades of new data and research for both advanced and emerging economies that shed light on whether these stylized facts still hold. Globalization, technological advances, and institutional changes since 1960 have altered growth patterns in many countries. We will examine each of Kaldor's six facts in turn, evaluating their validity in light of more recent empirical evidence.

1.2 Factor Income Shares

In mid-century evidence assembled by Kaldor for the United States and the United Kingdom, labor's share hovered near two thirds and displayed no persistent trend. This stability is no longer universal. In many advanced economies, labor's share trended down from the 1980s—1990s onward; in the United States the decline began in the 1990s and reached historically low levels by the 2010s. Research attributes a portion of the decline to technological change that automates routine tasks, globalization, weaker labor institutions, and rising product-market concentration that lead to higher markups. Measurement and composition issues qualify the universality of the decline: adjusting for self-employment, housing, and intangible capital attenuates or reverses the downward trend in the labor share in some countries.

In emerging and developing economies, patterns are more heterogeneous: a majority have experienced declining labor income shares since the 1990s, but the research has focused on greater global integration and capital deepening rather than within-country automation as dominant drivers. As countries industrialize and join global value chains, rapid investment raises capital intensity and shifts income toward capital, as illustrated by China's marked labor-share decline during the 2000s and similar episodes in other fast-growing Asian economies and commodity exporters. By contrast, economies with slower structural change have tended to display more stable factor shares.

1.3 Growth of Capital per Worker

Kaldor's second fact holds that, over long horizons, the stock of physical capital per worker grows at a roughly constant rate.

More recent evidence indicates that the spirit of this fact remains broadly valid while the literal constancy of the growth rate does not. In mature advanced economies, capital per worker has continued to increase, but its growth rate varies across subperiods. Postwar data for the United States and the United Kingdom show robust capital deepening from the 1950s to the early 1970s, followed by moderation thereafter. In the United States, measures of the nonfarm business capital-labor ratio and capital services per hour slowed after the early-1970s, accelerated during the late-1990s, and weakened again after the early 2000s. Since then, there has not been a sustained return to the pre-1973 pace. [1]

Patterns in emerging and developing economies are more heterogeneous. During catch-up phases, capital per worker often grows at very high rates, whereas in other periods it can be low or volatile. China's investment rate averaged on the order of 40 percent of GDP through the 2000s, and standard national accounts and Penn World Table series record double-digit growth of real capital per worker during that decade. South Korea and Taiwan likewise experienced exceptionally rapid increases in capital per worker during their industrialization in the 1960s—1980s. By contrast, several developing countries have recorded episodes of stagnant or declining capital per worker, and, on average, emerging economies display more volatile and dispersed capital-deepening trajectories than advanced economies. [2]

Several mechanisms help account for the post-1970s downshift in capital deepening among advanced economies. A primary factor is the slowdown in productivity and output growth that began in the 1970s, which reduced the incentive and need for rapid accumulation. In addition, a long-run decline in the relative price of investment goods—interpreted as

1. See, e.g., (Gordon 2016; Fernald 2015; Jorgenson and Stiroh 2000; Oliner and Sichel 2000; Jorgenson, Ho, and Stiroh 2008; Byrne, Oliner, and Sichel 2017; Herrendorf, Rogerson, and Valentinyi 2014; 2019; Jones 2016).

2. See, among others, (Young 1995; Collins and Bosworth 1996; Bank 1993; Jones 2016; Fund 2017; Barro and Sala-i-Martin 2003; Acemoglu 2009; Bank, World Development Indicators; Feenstra, Inklaar, and Timmer 2015).

investment-specific technical change—means that a given flow of real capital can be acquired with smaller expenditures. Since the 2000s, several advanced economies have also exhibited weak business investment despite historically low borrowing costs, a pattern linked in the literature to rising market power, the growing importance of intangible capital, and corporate-governance considerations. [3]

Emerging economies display widely varying accumulation paths during development with the observed growth rate of capital per worker attributed to technology, demographics, institutions, and policy, among other factors. [4]

1.4 Growth of Output per Worker

Kaldor emphasized the remarkable stability of long-run labor productivity growth in industrialized economies, interpreting roughly constant growth of output per worker as evidence for sustained technological progress.

Postwar experience refines the picture. For the United States and the United Kingdom, the drivers of growth in output per worker mirror those of capital per worker already discussed. Decelerations across other advanced economies following the Global Financial Crisis have been marked by unusually weak productivity growth, prompting debates over secular stagnation, measurement, and the pace of innovation. (OECD 2015; Gordon 2016; Byrne, Fernald, and Reinsdorf 2016; Syverson 2017).

Over sufficiently long horizons, however, frontier economies display striking regularities. Many developed countries trace near-linear paths for log GDP per capita, implying similar average growth rates on the order of 2 percent per year and suggesting convergence to a common steady growth path among the leaders. The near-parallelism of these long-run trajectories is notable given large sectoral and technological shifts over the twentieth and twenty-first centuries. [5]

3. See (Greenwood, Hercowitz, and Krusell 1997; Byrne, Oliner, and Sichel 2017; Gutiérrez and Philippon 2017; Crouzet and Eberly 2019; Loecker, Eeckhout, and Unger 2020; Corrado, Hulten, and Sichel 2005; Karabarbounis and Neiman 2014; Eggertsson, Robbins, and Wold 2018; Gordon 2016; Fernald 2015; OECD 2015).

4. See (Solow 1956; Barro and Sala-i-Martin 2003; Acemoglu 2009; Jones 2016).

5. See (Vollrath 2020; Jones 2016).

Cross-country heterogeneity is much larger away from the frontier. The dispersion of growth rates is higher for economies farther behind, and outcomes range from prolonged stagnation to multi-decade booms. East Asian catch-up experiences illustrate the upper tail: South Korea, Taiwan, and later China and Hong Kong sustained multi-decade growth in output per capita far above frontier rates before decelerating as they approached advanced-economy income levels. China, in particular, combined exceptionally high investment with rapid productivity gains after 1980, yielding average growth in GDP per worker well above 5 percent for several decades before its recent slowdown. [6]

1.5 Capital-Output Ratio

Kaldor's fourth fact posits that the capital-output ratio remains roughly constant over long periods.

The broad order of magnitude of this ratio remains similar in advanced economies, but there have been some drifts and important nuances in how we measure capital. Using national accounts data, many developed countries show a capital-output ratio that has been relatively stable, typically in the range of about 2.5 to 3.0 in net terms, or a bit higher in gross terms. Analyses of Penn World Table data confirm no dramatic trend for major economies. The U.S., Canada, U.K., and others had capital-output ratios around the low-to-mid 2's in 1950, edging up only modestly by the 2010s. Slight upward drift might partly be a data artifact related to how international price comparisons are done. When measured using each country's own national accounts, avoiding cross-country purchasing power parity issues, the U.S. capital-output ratio appears even flatter over time. The largest components of the capital stock in the U.S. show no clear long-term trend in their. Only newer categories like intellectual property have trended upward, though that may reflect better accounting of intangibles in recent years.

Some other advanced economies did see changes. The U.K. 's capital-output ratio rose after the 1970s, suggesting a departure from strict constancy. Looking beyond core capital,

6. See (Barro and Sala-i-Martin 2003; Jones 2016; Jones and Romer 2010; Pritchett 1997; Young 1995; Collins and Bosworth 1996; Bank 1993; Feenstra, Inklaar, and Timmer 2015; Bank, World Development Indicators).

Piketty drew attention to a rising wealth-to-income ratio in many wealthy countries. If one includes not just machines and buildings but also housing, land, and other assets, the value of capital is close to the value of wealth. Wealth relative to national income has increased substantially since the 1970s in countries like France, Germany, and the U.K. For example, the wealth-income ratio in Europe grew from about 2.5—3 in 1950 up to 5—6 by 2010, a trend driven largely by booming real estate values and accumulation of private wealth. This suggests that broader measures of capital do not follow Kaldor's fact as neatly, as they have shown an upward trend. However, it is important to distinguish between replacement-cost measures of productive capital, which are used in growth models and correspond more to Kaldor's concept, and market-value measures of total wealth. The former have been much more stable than the latter. [7]

In emerging economies, the capital-output ratio often starts lower and tends to rise during industrialization. When a country undergoes heavy investment-led growth, the capital-output ratio can increase in the earlier industrialization stages. Countries like Japan and South Korea saw their capital-output ratios increase as they poured resources into building factories and infrastructure. In China, some estimates suggest the capital-output ratio rose significantly from the 1970s to recent years amid enormous investment, though high output growth kept it from rising even more. Eventually, as growth moderates and investment rates stabilize, the capital-output ratio in successful emerging economies should level off and perhaps even decline slightly if very high investment rates fall to levels closer to those of developed economies.

1.6 Rate of Return on Capital

Kaldor emphasized the near constancy of the aggregate return to capital over long horizons in industrial economies, an observation that can be expressed transparently with national accounting objects. Let Y^K denote nominal payments to "reproducible private capital net of depreciation", Y nominal value added, K the replacement-cost capital stock, and $\alpha = Y^K/Y$ the net capital share. Define the average net return on reproducible capital as $r = Y^K/K$.

7. Herrendorf, Rogerson, and Valentinyi (2014) and Piketty (2014).

Then, by definition,

$$r = \frac{Y^K}{K} = \frac{\alpha Y}{K} = \frac{\alpha}{K/Y}.$$

Stability of the net capital share α together with stability of the replacement-cost ratio K/Y mechanically implies stability of the average net return on capital (Kaldor 1957; 1961; Solow 1956; Jones 2016).

Long-run evidence is broadly consistent with a roughly stable average return on broad private wealth in advanced economies.

Data for emerging and developing economies are thinner and returns are more volatile, reflecting greater macroeconomic and institutional risk. Nonetheless, the qualitative pattern is consistent with no clear evidence of a systematic long-run trend. (Jones 2016; Reinhart and Rogoff 2009).

Measurement matters for aligning evidence with the theory. The relevant average return is computed as nominal payments to reproducible productive capital divided by the replacement-cost value of that capital, an object that is invariant to the choice of numeraire and avoids mixing market-valuation effects from asset price swings with the quantity of productive capital (Herrendorf, Rogerson, and Valentinyi 2019). This distinction is essential when comparing replacement-cost returns used in growth accounting with market-value returns.

1.7 Cross-Country Growth Variation

Kaldor underscored that countries exhibit persistently different growth experiences, an observation that anticipated modern debates over drivers of long-term growth. The cross-country variation is perhaps the most central empirical regularity in growth economics that still attracts a lot of new research and not a lot of consensus(Kaldor 1957; 1961; Jones 2016). Contemporary cross-country data confirm that long-run growth is far from uniform: a subset of economies has sustained rapid catch-up while many others have grown slowly or stagnated, producing large and persistent income gaps (Pritchett 1997; Feenstra, Inklaar, and Timmer 2015; Bank, World Development Indicators). Dispersion is systematically related to distance

from the technological frontier, with richer countries displaying relatively similar moderate growth rates and poorer countries exhibiting much greater variance (Jones and Romer 2010; Jones 2016; OECD 2015). For example, advanced economies have clustered around modest productivity growth in recent decades, whereas developing economies span outcomes from rapid convergence to prolonged decline (Jones 2016; OECD 2015).

Taking a longer historical view, the global economy first experienced pronounced divergence as industrial leaders pulled away in the nineteenth and early twentieth centuries, followed by partial convergence among late industrializers in recent decades, especially in Asia (Bolt et al. 2018; Jones 2016). The Maddison Project data record rising cross-country dispersion of per capita incomes up to roughly the late twentieth century and stabilization or modest narrowing thereafter, reflecting the acceleration of several large emerging economies (Bolt et al. 2018). Despite this recent narrowing at the top of the distribution, very large income and total factor productivity differences remain, underscoring that growth is not automatic and depends on country-specific fundamentals (Hall and Jones 1999; Caselli 2005; Jones 2016).

A broad body of evidence highlights the roles of institutions, human capital, openness, and technology diffusion in shaping cross-country growth differences. Cross-country accounting points to the centrality of social infrastructure and human capital for explaining productivity and income gaps (Hall and Jones 1999; Caselli 2005; Barro 1991). International technology diffusion is incomplete and heterogeneous, with faster integration and adoption associated with stronger growth (Keller 2004; Comin and Hobijn 2010; Comin and Mestieri 2018). Demographic transitions can further tilt growth trajectories by altering labor force growth and savings behavior, as illustrated by episodes of rapid expansion in parts of East and Southeast Asia (Bloom and Williamson 1998; Jones 2016).

Table 1.1: Evolution of Kaldor's Facts: Contemporary Evidence

Kaldor's Original Fact	**Advanced Economies Update**	**Emerging & Developing Economies Update**
1. Constant labor and capital income shares	• Labor share declined since 1980s • Capital share rose correspondingly • Drivers: automation, globalization, weaker unions • More stable when excluding pure profits	• Many economies saw declining labor shares • Rapid capital deepening shifted income to capital • Global value chains altered distribution • Limited structural change: more stability
2. Constant growth of capital per worker	• Capital per worker continues rising • Growth rate slowed post-1970s • Slower than mid-20th century pace • Reflects productivity and investment slowdown	• Very fast growth during catch-up phases • East Asia: exceptionally rapid K/L increases • Other countries: low or volatile growth • Successful convergers stabilize near advanced rates
3. Constant growth of output per worker	• 2–3% growth mid-20th century • Slowed to 1–2% after 1970s • Further slowdowns in recent decades • Near 2% average over long horizons	• Range: near zero to over 7% • East Asian tigers: 5–7% for decades • Many countries: below 1% growth • Greater variance farther from frontier
4. Constant capital-output ratio	• Ratio typically 2.5–3.5 • Mild upward drift in some countries • Wealth-income ratio risen with housing/assets • Investment rates near 20% of GDP	• Initially low K/Y in poor countries • Rises during industrialization • Heavy investment phases push ratio up • Stabilizes as growth matures

Continued on next page

Kaldor's Original Fact	Advanced Economies Update	Emerging & Developing Economies Update
5. Constant rate of return on capital	• Long-run returns around 4–5% real • Corporate profits and equity returns robust • Risk-free rates declined significantly • Risk premiums and markups increased	• Returns high but volatile • Higher marginal returns than advanced economies • Risks and crises affect realized returns • Convergence toward global average with integration
6. Persistent cross-country growth differences	• Growth rates cluster around 1–2% per capita • Differences persist across countries • Variation over 10–20 year periods • Relatively narrow dispersion	• Wide range of growth outcomes • Some sustain over 5% annual growth • Others below 1% or negative • Greater variation farther from frontier

1.8 Conclusion

Kaldor's six stylized facts provided a stable backdrop for twentieth-century growth theory by highlighting near-constancies in key ratios and growth rates within the industrial core (Kaldor 1957; 1961; Jones 2016). Table 1.1 summarizes how each of Kaldor's facts holds up according to recent evidence in advanced vs. emerging economies. Several elements remain broadly intact in long-run data, notably the approximate stability of the capital-output ratio and of average aggregate returns, as well as pronounced and persistent cross-country growth differences (Herrendorf, Rogerson, and Valentinyi 2014; 2019; Jordà et al. 2019; Pritchett 1997; Bolt et al. 2018). Other elements have been revised by post-1970 evidence, including the decline of labor's income share in many advanced economies and medium-run shifts in the trend growth rates of output per worker and capital per worker (Karabarbounis and Neiman 2014; Elsby, Hobijn, and Sahin 2013; Fund 2017; Gordon 2016; Fernald 2015; Herrendorf, Rogerson, and Valentinyi 2014; Byrne, Oliner, and Sichel 2017). Global integration and technological change have reshaped income distribution and supported rapid catch-up in parts of Asia, while leaving frontier balanced-growth mechanisms as a useful organizing benchmark for mature economies(Baldwin 2016; Young 1995; Collins and Bosworth 1996; Feenstra, Inklaar, and Timmer 2015; Solow 1956; King and Rebelo 1993; Jones 2016).

Academic research has responded by extending the growth canon and proposing updated stylized facts. New emphases on ideas, institutions, population, and human capital account for features that the original list left implicit, including very large income differences and the lack of diminishing returns to human capital at the aggregate level (Jones and Romer 2010). The breakdown of constant factor shares has motivated models with factor-biased technical change and with market power and rents (Acemoglu 2002; 2003; Loecker, Eeckhout, and Unger 2020; Barkai 2020). Multi-sector frameworks that incorporate structural change and relative-price dynamics can reproduce the empirically measured post-1970 slowdowns(Herrendorf, Rogerson, and Valentinyi 2014; 2019; Duernecker, Herrendorf, and Valentinyi 2021). In parallel, finance-and-development work links the easing of financial frictions to innovation-led growth and convergence, supplying microfoundations for

cross-country heterogeneity (Aghion, Howitt, and Levine 2018).

1.9 References

Acemoglu, Daron. 2002. "Directed Technical Change." The Review of Economic Studies 69, no. 4 (October): 781–809. https://doi.org/10.1111/1467-937X.00226. https://academic.oup.com/restud/article-abstract/69/4/781/1556072.

———. 2003. "Labor- and Capital-Augmenting Technical Change." Journal of the European Economic Association 1, no. 1 (March): 1–37. https://doi.org/10.1162/154247603322256756. https://academic.oup.com/jeea/article-abstract/1/1/1/2294416.

———. 2009. Introduction to Modern Economic Growth. Princeton, NJ: Princeton University Press, January. ISBN: 9780691132921. https://press.princeton.edu/books/hardcover/9780691132921/introduction-to-modern-economic-growth.

Acemoglu, Daron, and Pascual Restrepo. 2018. "The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment." American Economic Review 108, no. 6 (June): 1488–1542. https://doi.org/10.1257/aer.20160696. https://www.aeaweb.org/articles?id=10.1257/aer.20160696.

Aghion, Philippe, Peter Howitt, and Ross Levine. 2018. "Financial Development and Innovation-Led Growth." In Handbook of Finance and Development, edited by Thorsten Beck and Ross Levine, 3–30. Cheltenham, UK: Edward Elgar. ISBN: 978-1-78536-050-3. https://doi.org/10.4337/9781785360510.00007.

Autor, David, David Dorn, Lawrence F. Katz, Christina Patterson, and John Van Reenen. 2020. "The Fall of the Labor Share and the Rise of Superstar Firms." The Quarterly Journal of Economics 135, no. 2 (May): 645–709. https://doi.org/10.1093/qje/qjaa004. https://academic.oup.com/qje/article/135/2/645/5721414.

Autor, David H., David Dorn, and Gordon H. Hanson. 2013. "The China Syndrome: Local Labor Market Effects of Import Competition in the United States." American Economic Review 103, no. 6 (October): 2121–2168. https://doi.org/10.1257/aer.103.6.2121. https://www.aeaweb.org/articles?id=10.1257/aer.103.6.2121.

Baldwin, Richard. 2016. The Great Convergence: Information Technology and the New Globalization. Cambridge, MA: Harvard University Press, November. ISBN: 9780674660489. https://www.hup.harvard.edu/books/9780674660489.

Bank, World. 1993. The East Asian Miracle: Economic Growth and Public Policy. New York, NY: Oxford University Press / World Bank, September. ISBN: 9780195209938. https://documents.worldbank.org/en/publication/documents-reports/documentdetail/975081468244550798.

———. 2025. (World Development Indicators). https://databank.worldbank.org/source/world-development-indicators.

Barkai, Simcha. 2020. "Declining Labor and Capital Shares." The Journal of Finance 75, no. 5 (October): 2421–2463. https://doi.org/10.1111/jofi.12842. https://onlinelibrary.wiley.com/doi/10.1111/jofi.12842.

Barro, Robert J. 1991. "Economic Growth in a Cross Section of Countries." The Quarterly Journal of Economics 106, no. 2 (May): 407–443. https://doi.org/10.2307/2937943. https://academic.oup.com/qje/article-abstract/106/2/407/1905452.

Barro, Robert J., and Xavier Sala-i-Martin. 2003. Economic Growth. 2nd ed. Cambridge, MA: MIT Press, October. ISBN: 9780262025539.

Bloom, David E., and Jeffrey G. Williamson. 1998. "Demographic Transitions and Economic Miracles in Emerging Asia." The World Bank Economic Review 12, no. 3 (September): 419–455. https://doi.org/10.1093/wber/12.3.419. https://academic.oup.com/wber/article-abstract/12/3/419/1632238.

Bolt, Jutta, Robert Inklaar, Herman de Jong, and Jan Luiten van Zanden. 2018. Rebasing 'Maddison': New Income Comparisons and the Shape of Long-Run Economic Development. GGDC Research Memorandum GD-174. Groningen Growth and Development Centre, January. https://www.rug.nl/ggdc/historicaldevelopment/maddison/releases/maddison-project-database-2018.

Byrne, David M., John G. Fernald, and Marshall B. Reinsdorf. 2016. "Does the United States Have a Productivity Slowdown or a Measurement Problem?" Brookings Papers on Economic Activity 2016, no. 1 (March): 109–182. https://doi.org/10.1353/eca.2016.0014. https://www.brookings.edu/articles/does-the-united-states-have-a-productivity-slowdown-or-a-measurement-problem/.

Byrne, David M., Stephen D. Oliner, and Daniel E. Sichel. 2017. "How Fast Are Semiconductor Prices Falling?" Brookings Papers on Economic Activity 2017, no. 1 (March): 247–300. https://www.brookings.edu/articles/how-fast-are-semiconductor-prices-falling/.

Caselli, Francesco. 2005. "Accounting for Cross-Country Income Differences." In Handbook of Economic Growth, Volume 1A, edited by Philippe Aghion and Steven N. Durlauf, 679–741. Amsterdam: Elsevier. https://doi.org/10.1016/S1574-0684(05)01009-9. https://www.sciencedirect.com/science/article/pii/S1574068405010099.

Collins, Susan M., and Barry P. Bosworth. 1996. "Economic Growth in East Asia: Accumulation versus Assimilation?" Brookings Papers on Economic Activity 1996, no. 2 (September): 135–204. https://doi.org/10.2307/2534621. https://www.brookings.edu/wp-content/uploads/1997/06/1996b_bpea_collins_bosworth_rodrik.pdf.

Comin, Diego, and Bart Hobijn. 2010. "An Exploration of Technology Diffusion." American Economic Review 100, no. 5 (December): 2031–2059. https://doi.org/10.1257/aer.100.5.2031. https://www.aeaweb.org/articles?id=10.1257/aer.100.5.2031.

Comin, Diego A., and Martí Mestieri. 2018. "If Technology Has Arrived Everywhere, Why Has Income Diverged?" American Economic Journal: Macroeconomics 10, no. 3 (July): 137–178. https://doi.org/10.1257/mac.20150175. https://www.aeaweb.org/articles?id=10.1257/mac.20150175.

Corrado, Carol, Charles Hulten, and Daniel Sichel. 2005. "Measuring Capital and Technology: An Expanded Framework." In Measuring Capital in the New Economy, edited by Carol Corrado, John Haltiwanger, and Daniel Sichel, 11–46. Chicago: University of Chicago Press. ISBN: 978-0-226-11604-4. https://doi.org/10.7208/chicago/9780226116174.003.0002. https://www.nber.org/books-and-chapters/measuring-capital-new-economy/measuring-capital-and-technology-expanded-framework.

Crouzet, Nicolas, and Janice C. Eberly. 2019. Understanding Weak Capital Investment: The Role of Market Concentration and Intangibles. NBER Working Paper 25869. National Bureau of Economic Research, May. https://www.nber.org/papers/w25869.

Dao, Mai Chi, Mitali Das, Zsoka Koczan, and Weicheng Lian. 2017. Why Is Labor Receiving a Smaller Share of Global Income? Theory and Empirical Evidence. IMF Working Paper WP/17/169. International Monetary Fund, July. https://www.elibrary.imf.org/downloadpdf/view/journals/001/2017/169/001.2017.issue-169-en.pdf.

Duernecker, Georg, Berthold Herrendorf, and Ákos Valentinyi. 2021. "The Productivity Growth Slowdown and Kaldor's Growth Facts." Journal of Economic Dynamics and Control 130:104200. ISSN: 0165-1889. https://doi.org/10.1016/j.jedc.2021.104200.

Eggertsson, Gauti B., Jacob A. Robbins, and Ella Getz Wold. 2018. Kaldor and Piketty's Facts: The Rise of Monopoly Power in the United States. NBER Working Paper 24287. National Bureau of Economic Research, February. https://doi.org/10.3386/w24287. https://www.nber.org/papers/w24287.

Elsby, Michael W. L., Bart Hobijn, and Aysegul Sahin. 2013. "The Decline of the U.S. Labor Share." Brookings Papers on Economic Activity 2013, no. 1 (March): 1–63. https://www.brookings.edu/articles/the-decline-of-the-u-s-labor-share/.

Feenstra, Robert C., Robert Inklaar, and Marcel P. Timmer. 2015. "The Next Generation of the Penn World Table." American Economic Review 105, no. 10 (October): 3150–3182. https://doi.org/10.1257/aer.20130954. https://www.aeaweb.org/articles?id=10.1257/aer.20130954.

Fernald, John G. 2015. Productivity and Potential Output Before, During, and After the Great Recession. Working Paper 2014-15. Federal Reserve Bank of San Francisco, June. https://www.frbsf.org/research-and-insights/publications/working-papers/2014/15/.

Fund, International Monetary. 2017. World Economic Outlook, April 2017: Chapter 3. Understanding the Downward Trend in Labor Income Shares. April 2017. International Monetary Fund, April. https://www.imf.org/en/Publications/WEO/Issues/2017/04/04/world-economic-outlook-april-2017.

Gollin, Douglas. 2002. "Getting Income Shares Right." Journal of Political Economy 110, no. 2 (April): 458–474. https://doi.org/10.1086/338747. https://www.journals.uchicago.edu/doi/10.1086/338747.

Gordon, Robert J. 2016. The Rise and Fall of American Growth: The U.S. Standard of Living since the Civil War. Princeton, NJ: Princeton University Press, January. ISBN: 9780691147727. https://press.princeton.edu/books/hardcover/9780691147727/the-rise-and-fall-of-american-growth.

Greenwood, Jeremy, Zvi Hercowitz, and Per Krusell. 1997. "Long-Run Implications of Investment-Specific Technological Change." American Economic Review 87, no. 3 (June): 342–362. https://www.jstor.org/stable/2951349.

Gutiérrez, Germán, and Thomas Philippon. 2017. "Investment-less Growth: An Empirical Investigation." Brookings Papers on Economic Activity 2017, no. 2 (September): 67–140. https://www.brookings.edu/bpea-articles/investment-less-growth-an-empirical-investigation/.

Hall, Robert E., and Charles I. Jones. 1999. "Why Do Some Countries Produce So Much More Output Per Worker Than Others?" The Quarterly Journal of Economics 114, no. 1 (February): 83–116. https://doi.org/10.1162/003355399555954. https://academic.oup.com/qje/article-abstract/114/1/83/1921741.

Herrendorf, Berthold, Richard Rogerson, and Akos Valentinyi. 2014. "Growth and Structural Transformation." In Handbook of Economic Growth, Volume 2B, edited by Philippe Aghion and Steven N. Durlauf, 855–941. Amsterdam: Elsevier. https://doi.org/10.1016/B978-0-444-53540-5.00006-9. https://www.sciencedirect.com/science/article/pii/B9780444535405000069.

———. 2019. "Growth and the Kaldor Facts." Federal Reserve Bank of St. Louis Review 101, no. 2 (February): 85–99. https://research.stlouisfed.org/publications/review/2019/02/13/growth-and-the-kaldor-facts.

Inada, Ken-Ichi. 1963. "On a Two-Sector Model of Economic Growth: Comments and a Generalization." Review of Economic Studies 30 (2): 119–127.

Jones, Charles I. 2016. "The Facts of Economic Growth." In Handbook of Macroeconomics, Volume 2A, edited by John B. Taylor and Harald Uhlig, 3–69. Amsterdam: Elsevier. https://doi.org/10.1016/bs.hesmac.2016.03.002. https://www.sciencedirect.com/science/article/pii/S1574004816300024.

Jones, Charles I., and Paul M. Romer. 2010. "The New Kaldor Facts: Ideas, Institutions, Population, and Human Capital." American Economic Journal: Macroeconomics 2, no. 1 (January): 224–245. https://doi.org/10.1257/mac.2.1.224. https://www.aeaweb.org/articles?id=10.1257/mac.2.1.224.

Jordà, Òscar, Katharina Knoll, Dmitry Kuvshinov, Moritz Schularick, and Alan M. Taylor. 2019. "The Rate of Return on Everything, 1870–2015." The Quarterly Journal of Economics 134, no. 3 (August): 1225–1298. https://doi.org/10.1093/qje/qjz012. https://academic.oup.com/qje/article-abstract/134/3/1225/5435538.

Jorgenson, Dale W., Mun S. Ho, and Kevin J. Stiroh. 2008. "A Retrospective Look at the U.S. Productivity Growth Resurgence." Journal of Economic Perspectives 22, no. 1 (January): 3–24. https://doi.org/10.1257/jep.22.1.3. https://www.aeaweb.org/articles?id=10.1257/jep.22.1.3.

Jorgenson, Dale W., and Kevin J. Stiroh. 2000. "Raising the Speed Limit: U.S. Economic Growth in the Information Age." Brookings Papers on Economic Activity 2000, no. 1 (June): 125–235. https://doi.org/10.1353/eca.2000.0018. https://www.brookings.edu/bpea-articles/raising-the-speed-limit-u-s-economic-growth-in-the-information-age/.

Kaldor, Nicholas. 1957. "A Model of Economic Growth." The Economic Journal 67, no. 268 (December): 591–624. https://doi.org/10.2307/2227704. https://academic.oup.com/ej/article-abstract/67/268/591/5248725.

———. 1961. "Capital Accumulation and Economic Growth." In The Theory of Capital, edited by F. A. Lutz and D. C. Hague, 177–222. London: Palgrave Macmillan. ISBN: 978-1-349-08452-4. https://doi.org/10.1007/978-1-349-08452-4_10. https://ideas.repec.org/h/pal/intecp/978-1-349-08452-4_10.html.

Karabarbounis, Loukas, and Brent Neiman. 2014. "The Global Decline of the Labor Share." The Quarterly Journal of Economics 129, no. 1 (February): 61–103. https://doi.org/10.1093/qje/qjt032. https://academic.oup.com/qje/article/129/1/61/1890096.

Keller, Wolfgang. 2004. "International Technology Diffusion." Journal of Economic Literature 42, no. 3 (September): 752–782. https://doi.org/10.1257/0022051042177685. https://www.aeaweb.org/articles?id=10.1257/0022051042177685.

King, Robert G., and Sergio T. Rebelo. 1993. "Transitional Dynamics and Economic Growth in the Neoclassical Model." The American Economic Review 83, no. 4 (September): 908–931.

Koh, Dongya, Raul Santaeulalia-Llopis, and Yu Zheng. 2024. "Labor Share Decline and the Capitalization of Intellectual Property Products." Review of Economic Dynamics 54 (January): 181–207. https://doi.org/10.1016/j.red.2023.11.003. https://www.sciencedirect.com/science/article/pii/S1094202523000730.

Loecker, Jan De, Jan Eeckhout, and Gabriel Unger. 2020. "The Rise of Market Power and the Macroeconomic Implications." The Quarterly Journal of Economics 135, no. 2 (May): 561–644. https://doi.org/10.1093/qje/qjz041. https://academic.oup.com/qje/article/135/2/561/5687353.

OECD. 2012. OECD Employment Outlook 2012, Chapter 3: Labour Losing to Capital: What Explains the Declining Labour Share? OECD Publishing, July. https://www.oecd.org/en/publications/reports/2012/07/oecd-employment-outlook-2012_g1g1dcdb.html.

———. 2015. The Future of Productivity. Paris: OECD Publishing, July. ISBN: 978-92-64-24853-3. https://www.oecd.org/en/publications/reports/2015/12/the-future-of-productivity_9789264248533-en.html.

Oliner, Stephen D., and Daniel E. Sichel. 2000. "The Resurgence of Growth in the Late 1990s: Is Information Technology the Story?" Journal of Economic Perspectives 14, no. 4 (December): 3–22. https://doi.org/10.1257/jep.14.4.3. https://www.aeaweb.org/articles?id=10.1257/jep.14.4.3.

Piketty, Thomas. 2014. Capital in the Twenty-First Century. Cambridge, MA: Harvard University Press, March. ISBN: 9780674430006. https://www.hup.harvard.edu/books/9780674430006.

Pritchett, Lant. 1997. "Divergence, Big Time." Journal of Economic Perspectives 11, no. 3 (June): 3–17. https://doi.org/10.1257/jep.11.3.3. https://www.aeaweb.org/articles?id=10.1257/jep.11.3.3.

Reinhart, Carmen M., and Kenneth S. Rogoff. 2009. This Time Is Different: Eight Centuries of Financial Folly. Princeton, NJ: Princeton University Press, October. ISBN: 9780691142166. https://press.princeton.edu/books/hardcover/9780691142166/this-time-is-different.

Rognlie, Matthew. 2015. "Deciphering the Fall and Rise of the Net Capital Share." Brookings Papers on Economic Activity 2015, no. 1 (March): 1–69. https://www.brookings.edu/wp-content/uploads/2016/06/2015a_rognlie.pdf.

Solow, Robert M. 1956. "A Contribution to the Theory of Economic Growth." The Quarterly Journal of Economics 70, no. 1 (February): 65–94. https://doi.org/10.2307/1884513. https://academic.oup.com/qje/article-abstract/70/1/65/1885041.

Syverson, Chad. 2017. "Challenges to Mismeasurement Explanations for the US Productivity Slowdown." Journal of Economic Perspectives 31, no. 2 (May): 165–186. https://doi.org/10.1257/jep.31.2.165. https://www.aeaweb.org/articles?id=10.1257/jep.31.2.165.

Vollrath, Dietrich. 2020. Fully Grown: Why a Stagnant Economy Is a Sign of Success. Chicago: University of Chicago Press, January. ISBN: 9780226666808. https://press.uchicago.edu/ucp/books/book/chicago/F/bo44800441.html.

Young, Alwyn. 1995. "The Tyranny of Numbers: Confronting the Statistical Realities of the East Asian Growth Experience." The Quarterly Journal of Economics 110, no. 3 (August): 641–680. https://doi.org/10.2307/2946695. https://academic.oup.com/qje/article-abstract/110/3/641/1859236.

1.10 The Solow-Swan Model

The first growth model compatible with the stylized Kaldor facts was developed independently by Solow and Swan. The central element of their theory is the notion of an aggregate production function:

$$Y(t) = F(K(t), L(t), t) \tag{1.10.1}$$

where Y is aggregate output, K is the aggregate capital stock, L is aggregate employment, and t is the time index that appears separately in the production function to indicate that other factors such as technology may not be constant over time. We assume that the production function has constant returns to scale

$$F(\lambda K(t), \lambda L(t), t) = \lambda F(K(t), L(t), t), \quad \text{for } \lambda > 0. \tag{1.10.2}$$

Aggregate saving in the economy is assumed to be a constant fraction of output

$$S(t) = sY(t), \quad 0 < s < 1, \tag{1.10.3}$$

where s is the constant propensity to save out of income (recall output and income are the same in aggregate). We assume a closed economy with no government purchases of goods and services, and therefore

$$Y(t) = C(t) + I(t), \tag{1.10.4}$$

where C is consumption of the representative household and I is aggregate investment. Equations (1.10.3) and (1.10.4) imply that consumption is also a constant fraction of output

$$C(t) = (1 - s)Y(t). \tag{1.10.5}$$

The change in capital over time is given by investment net of depreciation:

$$\dot{K}(t) = I(t) - \delta K(t), \tag{1.10.6}$$

where $\delta > 0$ is the constant depreciation rate and $\dot{K}(t) := dK(t)/dt$.

The labor force grows at a constant rate n:

$$\dot{L}(t) = nL(t), \tag{1.10.7}$$

where $\dot{L}(t) := dL(t)/dt$, and $L(t_0)$ is the initial level. We assume that population grows at the same rate as the labor force, so "per worker" and "per capita" are the same in this model.

1.10.1 No technological progress

We first look at the case in which the production function does not directly depend on time:

$$Y(t) = F(K(t), L(t)). \tag{1.10.8}$$

Since $Y(t)$ can only depend on t through $K(t)$ or $L(t)$, the production function (1.10.1) does not allow for technological progress.

In addition to constant returns to scale, we assume that the production function has positive

but diminishing marginal products to both factors:

$$F_K, F_L > 0, \quad F_{KK}, F_{LL} < 0, \quad F_{KL} > 0. \tag{1.10.9}$$

These properties and constant returns to scale imply that $F_{KL} > 0$.

A more controversial assumption, but one we will make nonetheless, is that $F(\cdot)$ obeys the so-called Inada conditions (after (Inada 1963))

$$\lim_{K\to 0} F_K = \lim_{L\to 0} F_L = +\infty, \quad \lim_{K\to\infty} F_K = \lim_{L\to\infty} F_L = 0. \tag{1.10.10}$$

These conditions are not innocuous and preclude a number of interesting non–standard cases.

The Solow-Swan model consists of equations (1.10.2)-(1.10.10).

It is convenient to write the model in per–worker terms, also called **intensive form**. Define

$$\begin{aligned} y(t) &:= \frac{Y(t)}{L(t)}, \\ k(t) &:= \frac{K(t)}{L(t)}, \\ c(t) &:= \frac{C(t)}{L(t)}. \end{aligned}$$

Dividing both sides of (1.10.8) by $L(t)$ and using (1.10.2) with $\lambda = 1/L$, we get

$$\frac{Y(t)}{L(t)} = \frac{1}{L(t)} F(K(t), L(t)) = F\left(\frac{K(t)}{L(t)}, 1\right). \tag{1.10.11}$$

We define the intensive form of the production function

$$f(k(t)) := F\left(\frac{K(t)}{L(t)}, 1\right). \tag{1.10.12}$$

and write (1.10.11) as

$$y(t) = f(k(t)). \tag{1.10.13}$$

The properties of F imply the corresponding properties for f:

$$f(0) = 0; \quad f'(k) > 0; \quad f''(k) < 0. \tag{1.10.14}$$

Using the definition of $k(t)$, the chain rule, and (1.10.7), we get

$$\dot{k}(t) = \frac{1}{L(t)} \cdot \dot{K}(t) - \frac{K(t)}{L(t)^2} \cdot \dot{L}(t) \tag{1.10.15}$$

$$= \frac{\dot{K}(t)}{L(t)} - nk(t). \tag{1.10.16}$$

Equations (1.10.3)-(1.10.5) imply that $S = I$. Combining $S = I$, (1.10.3), and (1.10.6) gives

$$\dot{K}(t) = sY(t) - \delta K(t) \tag{1.10.17}$$

Dividing by $L(t)$ and using (1.10.17) and (1.10.13), we arrive at

$$\dot{k}(t) = sf(k(t)) - (\delta + n)\, k(t). \tag{1.10.18}$$

Equation (1.10.18) is the central equation of the Solow-Swan model. If we understand and solve this central equation, we can understand and solve the rest of the model fairly straightforwardly.

1.10.2 Equilibrium

Figure 1.1 shows a diagram of the $\dot{k}(t)$–equation. The vertical axis shows levels whereas the horizontal axis shows capital per unit of effective labor, $k(t)$.

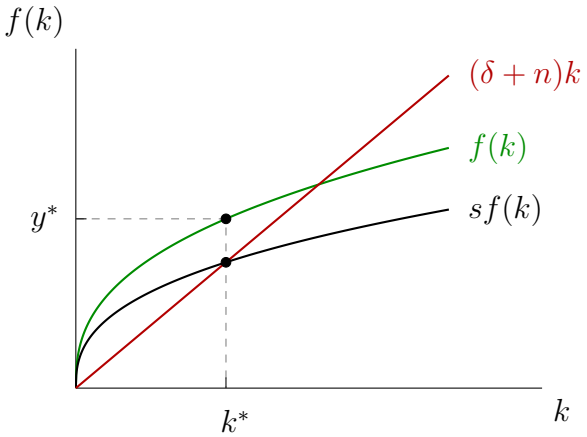

Figure 1.1: The Solow-Swan growth diagram

The two most important lines here are:

The curve $sf(k(t))$ represents the actual investment; it is concave since $f''(k(t)) < 0$. The line $(\delta + n)k(t)$ is referred to as the break-even level of investment. Beyond these curves, however, there is also the $f(k(t))$ curve. Note that the vertical distance between the $sf(k(t))$ and $f(k(t))$ curves equals $c(t) = C(t)/L(t)$, i.e. consumption per unit of labor.

At low levels of $k(t)$, $\dot{k}(t) > 0$, whereas at high levels of $k(t)$, $\dot{k}(t) < 0$. The only steady–state for $k(t)$ in Figure 1.1 occurs at k^*, which is defined by the point where the curve $sf(k(t))$ and the line $(\delta + n)k(t)$ intersect. At k^*, actual investment and breakeven investment are equal so capital per capita remains constant, $\dot{k}(t) = 0$. When $\dot{k}(t) = 0$, $k(t)$ is constant, which implies that $K(t)$ and $L(t)$ grow at a "balanced" rate.

1.10.3 Implications

1.10.4 Cobb-Douglas functional form

If we assume a Cobb-Douglas functional form for the production function $F(K(t), L(t)) = K(t)^{\alpha}L(t)^{1-\alpha}$, we will have the following $\dot{k}(t)$-equation:

$$\dot{k}(t) = sk(t)^{\alpha} - (\delta + n)k(t) \tag{1.10.19}$$

From this expression, we can also derive the growth rate per worker (by using the chain rule):

$$\begin{aligned} \frac{\dot{y}(t)}{y(t)} &= \frac{\frac{\partial(k(t)^{\alpha})}{\partial t}}{k(t)^{\alpha}} = \frac{\alpha k(t)^{\alpha-1}\dot{k}(t)}{k(t)^{\alpha}} = \frac{\alpha\dot{k}(t)}{k(t)} \\ &= s\alpha k(t)^{\alpha-1} - \alpha(\delta + n) = \frac{s\alpha}{k(t)^{1-\alpha}} - \alpha(\delta + n) \end{aligned} \tag{1.10.20}$$

In the short run, the growth rate depends on the initial capital per worker $k(t)$. Economies that begin with low $k(t)$ have $\dot{k}(t) > 0$ and grow faster; as $k(t)$ rises, the economy converges toward the steady state k^* where $\dot{k}(t) = 0$. Starting from k^*, a permanent increase in the saving rate from s to s_{new} changes the steady state. The new steady state k^{**} is higher than k^*. Just after the change in s, the economy is now below its steady-state. This implies that

$\dot{k}(t) > 0$, temporarily lifting the growth rate. Once $k(t)$ reaches the new, higher steady state k^{**}, growth returns to zero for capital per worker (see Figure 1.2). Analogously, a discrete increase in population growth induces a period with $\dot{k}(t) < 0$ until the economy settles at a lower steady-state level of k and, accordingly, a lower level of output per worker. In sum, the model predicts only transitory effects on the growth rate of capital per worker.

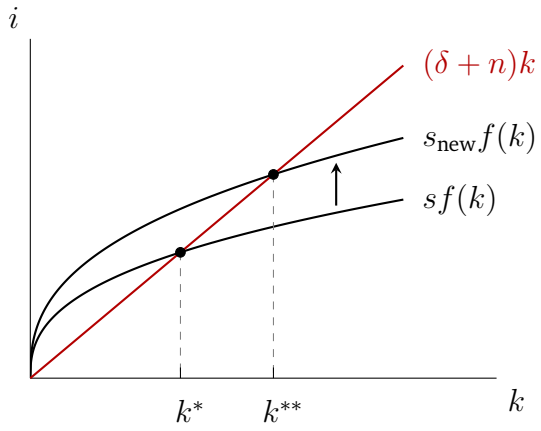

Figure 1.2: Impact on growth rate per worker of an increase in the savings rate

With a Cobb-Douglas production function, we can solve explicitly for the steady-state level of capital:

$$\begin{aligned} \dot{k}(t) = 0 \Longrightarrow s\,(k^*)^{\alpha} &= (\delta + n)k^* \\ \Longrightarrow (k^*)^{\alpha-1} = \frac{(\delta+n)}{s} &\Longrightarrow k^* = \left(\frac{\delta+n}{s}\right)^{\frac{1}{\alpha-1}} = \left(\frac{s}{\delta+n}\right)^{\frac{1}{1-\alpha}} \end{aligned} \tag{1.10.21}$$

where the last step rearranges in order to have the exponent positive $\left(\frac{1}{\alpha-1} < 0\right)$, which helps interpret the result. This expression for the steady-state level shows that k^* will increase with the savings rate s, and decrease with the capital depreciation rate δ and with the population growth rate n. Identical results are found by moving the curves in Figure 1.1.

Using 1.10.21, the steady-state level of output per worker is

$$y^* = (k^*)^{\alpha} = \left(\frac{s}{\delta+n}\right)^{\frac{\alpha}{1-\alpha}} \tag{1.10.22}$$

1.10.5 Golden rule of capital accumulation

What is the impact of a permanent increase in the saving rate s on the steady-state level of consumption per worker c^*? We start from

$$c^* = (1-s)f\left(k^*\right) = f\left(k^*\right) - (n+\delta)k^*. \tag{1.10.23}$$

We know from (1.10.21) that k^* increases with s, but in the expression (1.10.23) there is both a positive and a negative effect of k^* on c^*. When we take the partial derivative, we get

$$\frac{\partial c^*}{\partial s} = \left[f'\left(k^*\right) - (n+\delta)\right]\frac{\partial k^*}{\partial s} \tag{1.10.24}$$

The sign of this expression will be determined by the sign of the term in square brackets (know that $\frac{\partial k^*}{\partial s} > 0$ always holds). Since $f'\left(k^*\right)$ is very large at small levels of k^* (see Figure 1.1), we can infer that $\frac{\partial c^*}{\partial s} > 0$ when k^* is small, whereas at greater levels of k^* it will be the case that $\frac{\partial c^*}{\partial s} < 0$. Hence, there is a level of k^*, referred to as $k^{*,\text{gold}}$, when $\frac{\partial c^*}{\partial s} = 0$:

$$\frac{\partial c^*}{\partial s} = 0 \quad \text{when } f'\left(k^{*,\text{gold}}\right) = n+\delta \tag{1.10.25}$$

Beyond this level of k^*, a further capital accumulation will decrease intensive consumption c^*. The savings rate that achieves $k^{*,\text{gold}}$ is referred to as s^{gold}.

1.10.6 Convergence

The Solow-Swan model has strong predictions about convergence, i.e., that countries that start off with a lower level of $k(t)$ should experience a higher growth rate of output per worker. One way to illustrate this convergence property is as follows.

We keep using the Cobb-Douglas production function $f(k(t)) = k(t)^\alpha$ and write the growth rate of output per worker as:

$$\frac{\dot{y}(t)}{y(t)} = \alpha\left(sk(t)^{\alpha-1} - \delta - n\right) \tag{1.10.26}$$

Since $\frac{Y(t)}{L(t)} = k(t)^\alpha = y(t)$, we can express $k(t)$ as $k(t) = y(t)^{\frac{1}{\alpha}}$. Inserting into the growth

equation yields

$$\frac{\dot{y}(t)}{y(t)} = \alpha\left(sy(t)^{\frac{\alpha-1}{\alpha}} - \delta - n\right) = \alpha\left(\frac{s}{y(t)^{\frac{1-\alpha}{\alpha}}} - \delta - n\right) \tag{1.10.27}$$

Variants of this expression form the basis of the cross-country empirical studies on the determinants of economic growth. The key prediction concerning convergence is that the growth rate of output per worker should decrease with its initial level, $y(t)$. In other words, holding all other factors constant, poorer countries should grow faster than richer ones. A country's growth rate during the convergence process should further increase with its savings rate s, and decrease with the population growth rate n and the capital depreciation rate δ. The growth rate of rich countries that have reached their steady state will depend on the exogenous technological progress parameter.

The convergence result suggests that the poorest countries in the world should experience the highest growth rates. Although we know that many previously poor countries have experienced very fast growth rates in recent decades—think China, India, Botswana—other countries have experienced stagnant growth or even growth collapses. Some countries, such as the Democratic Republic of Congo and Zambia, have even seen their levels of income per capita fall by around half from their peak.

1.10.7 Extensions

The simple version of the Solow growth function presented in this document provides the basic intuition behind the important convergence property and the central role played by physical capital accumulation. It abstracts, however, from numerous factors that are believed to be central for economic growth to occur even in the medium run. Two of the most important of these factors are technological progress and human capital accumulation.

1.10.8 Technological progress

Technological progress can be readily included in the Solow-Swan model. Denote the level of technological knowledge at time t as $A(t)$. Most growth models further assume that technology is primarily labor augmenting, i.e. that $A(t)$ increases workers' level of

productivity. We will refer to the composite production factor $A(t)L(t)$ as effective labor. The aggregate production function then becomes

$$Y(t) = F(K(t), A(t)L(t)) = K(t)^{\alpha}(A(t)L(t))^{1-\alpha} \tag{1.10.28}$$

Let us further assume that the growth rate of technology is exogenously given by

$$\frac{\dot{A}(t)}{A(t)} = g > 0 \tag{1.10.29}$$

The intensive form is now written as $\kappa(t) = K(t)/(A(t)L(t))$ and is referred to as capital per unit of effective labor. The time derivative of $\kappa(t)$ is

$$\begin{aligned} \dot{\kappa}(t) &= \frac{\dot{K}(t)}{A(t)L(t)} - \frac{K(t)L(t)}{(A(t)L(t))^2}\dot{A}(t) - \frac{K(t)A(t)}{(A(t)L(t))^2}\dot{L}(t) \\ &= \frac{sY(t) - \delta K(t)}{A(t)L(t)} - \kappa(t)\frac{\dot{A}(t)}{A(t)} - \kappa(t)\frac{\dot{L}(t)}{L(t)} = s\hat{f}(\kappa(t)) - \kappa(t)(\delta + g + n) \end{aligned} \tag{1.10.30}$$

where $\hat{f}(\kappa(t)) = F(K(t), A(t)L(t))/(A(t)L(t))$. Similar to our earlier analysis, the economy will be in a steady-state equilibrium when $\dot{\kappa}(t) = 0$, which happens at a level $\kappa^* = \left(\frac{s}{\delta+g+n}\right)^{\frac{1}{1-\alpha}}$. The equilibrium level of capital per unit of effective labor thus decreases with the technological growth rate.

Since output per capita can now be expressed as $Y(t)/L(t) = K(t)^{\alpha}(A(t)L(t))^{1-\alpha}/L(t) = A(t)\kappa(t)^{\alpha}$, its growth rate is given by

$$\frac{\dot{y}(t)}{y(t)} = \frac{\dot{A}(t)}{A(t)} + \alpha\frac{\dot{\kappa}(t)}{\kappa(t)} = g + \alpha\left(\frac{s\hat{f}(\kappa(t))}{\kappa(t)} - \delta - g - n\right) \tag{1.10.31}$$

Note that when $\kappa(t) = \kappa^*$, the term inside the parentheses will be equal to zero. The equilibrium growth rate of output per capita is then simply $g > 0$. Hence, rich countries at their equilibrium growth rates will only grow through technological progress. In the Solow model, this growth rate is exogenously given and we cannot say anything very interesting about it. Endogenous growth theory extends this framework to derive this growth rate as the result of intentional human investments in research and development (R&D).

1.11 The Neoclassical Growth Model

The infinite-horizon neoclassical growth model is one of the fundamental workhorse models of economic growth theory. This model, also known as the Ramsey or Cass–Koopmans model, provides a framework for understanding how economies grow over time when savings decisions are made optimally by forward-looking consumers.

The standard neoclassical growth model differs from simpler growth models in one crucial respect: it explicitly models the consumer side and endogenizes savings through consumer optimization. Since the model explicitly models utility functions of consumers, it is also possible to study social welfare. Beyond its use as a basic growth model, this model contains the kernel of what we will later develop into the real business cycle model.

1.11.1 Preferences, Technology and Demographics

Consider an infinite–horizon economy in continuous time. We assume that the economy admits a representative household with instantaneous utility function

$$u(c(t)), \tag{1.11.1}$$

and we make the following standard assumptions on this utility function:

Assumption 1.11.1. $u(c)$ is strictly increasing, concave, twice continuously differentiable with derivatives u' and u'', and satisfies the following Inada–type conditions:

$$\lim_{c\to 0} u'(c) = \infty \text{ and } \lim_{c\to\infty} u'(c) = 0.$$

The representative household represents a set of identical households (with measure normalized to 1). Each household has an instantaneous utility function given by (1.11.1). The number of household members within each household grows at the rate n, starting with $L(0) = 1$, so that the total population is

$$L(t) = \exp(nt).$$

All members of the household supply their labor inelastically. Our baseline assumption is that the household is fully altruistic towards all of its future members, and always makes

the allocations of consumption (among household members) cooperatively. This implies that the objective function of each household at time $t = 0$, $U(0)$, can be written as

$$U(0) := \int_0^\infty \exp(-(\rho - n)t)u(c(t))dt, \tag{1.11.2}$$

where $c(t)$ is consumption per capita at time t, ρ is the subjective discount rate, and the effective discount rate is $\rho - n$, since it is assumed that the household also derives utility from the per-capita consumption of each of its future members.

We also assume throughout that

Assumption 1.11.2. $\rho > n$.

This assumption ensures that there is in fact discounting of future utility streams. Otherwise, (1.11.2) would have infinite value. Assumption 1.11.2 makes sure that in the model without growth, discounted utility is finite. When there is growth, we will strengthen this assumption.

We start with an economy without any technological progress. Factor and product markets are competitive, and the production possibilities set of the economy is represented by the aggregate production function

$$Y(t) = F(K(t), L(t)),$$

where $K(t)$ is physical capital. We impose the following standard assumptions on the production function, familiar from the Solow-Swan model:

Assumption 1.11.3. The production function $F : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ exhibits constant returns to scale: for all $\lambda > 0$, $F(\lambda K, \lambda L) = \lambda F(K, L)$.

Assumption 1.11.4. The production function is twice continuously differentiable, increasing in both arguments, and exhibits diminishing marginal products. The per capita production function $f(k) = F(k, 1)$ is strictly concave and satisfies the Inada conditions:

$$\lim_{k\to 0} f'(k) = \infty \text{ and } \lim_{k\to\infty} f'(k) = 0.$$

The constant returns to scale feature enables us to work with the per capita production

function $f(\cdot)$ such that, output per capita is given by

$$\begin{aligned} y(t) &:= \frac{Y(t)}{L(t)} \\ &= F\left(\frac{K(t)}{L(t)}, 1\right) \\ &:= f(k(t)) \end{aligned}$$

where, as before,

$$k(t) := \frac{K(t)}{L(t)}.$$

We assume that factor markets (for capital and labor) are perfectly competitive.

1.11.2 Firms

The representative competitive firm takes factor prices (R, w) as given and solves

$$\max_{K \geq 0, L \geq 0} \Pi(K, L) := F(K, L) - RK - wL,$$

where F satisfies Assumptions 1.11.3 and 1.11.4. Under free entry and perfectly competitive markets, equilibrium profits are zero at the optimum, so for some optimal input vector (K^*, L^*),

$$\Pi(K^*, L^*) = 0, \qquad \Pi(K, L) \leq 0 \text{ for all } (K, L) \geq 0. \tag{1.11.3}$$

Using $F(K, L) = Lf(K/L)$, write profits as

$$\Pi(K, L) = [f(k) - Rk - w]L, \qquad k := K/L \text{ for } L > 0, \tag{1.11.4}$$

and define the per-worker profit function

$$\phi(k) := f(k) - Rk - w.$$

Because f is strictly concave by Assumption 1.11.4, ϕ is strictly concave, so it has a unique maximizer k^*. If $\max_k \phi(k) > 0$, then by (1.11.4) the firm can choose arbitrarily large L and earn unbounded profits, contradicting (1.11.3). If $\max_k \phi(k) < 0$, the firm sets $L = 0$ and earns zero, again contradicting that (K^*, L^*) is an optimum with positive scale. Therefore

$$\max_k \phi(k) = \phi(k^*) = 0, \tag{1.11.5}$$

and any (K^*, L^*) with $K^*/L^* = k^*$ is profit maximizing with zero profits.

Since ϕ is strictly concave, the unique maximizer k^* satisfies the first-order condition

$$\phi'(k^*) = f'(k^*) - R = 0,$$

so

$$R = f'(k^*). \tag{1.11.6}$$

Evaluating zero profit (1.11.5) at k^* yields

$$w = f(k^*) - Rk^* = f(k^*) - k^* f'(k^*), \tag{1.11.7}$$

where the second equality uses (1.11.6).

To translate (1.11.6)–(1.11.7) into marginal products of F, differentiate $F(K, L) = Lf(K/L)$ using the chain rule. For $L > 0$ and $k = K/L$,

$$F_K(K, L) = \frac{\partial}{\partial K}\big[Lf(K/L)\big] = f'(k), \tag{1.11.8}$$

and

$$F_L(K, L) = \frac{\partial}{\partial L}\big[Lf(K/L)\big] = f(k) - kf'(k). \tag{1.11.9}$$

Combining (1.11.6) with (1.11.8) and (1.11.7) with (1.11.9), evaluated at any optimal pair (K^*, L^*) with $K^*/L^* = k^*$, gives

$$R = F_K(K^*, L^*), \tag{1.11.10}$$

$$w = F_L(K^*, L^*). \tag{1.11.11}$$

1.11.3 Households

The household optimization consists of each household solving a continuous time optimization problem to decide how to use their assets and allocate consumption over time. To prepare for this, let us denote the asset holdings of the representative household at time t by $h(t)$. Then we have the following law of motion for the total assets of the household

$$\dot{h}(t) = r(t)h(t) + w(t)L(t) - c(t)L(t),$$

where $c(t)$ is consumption per capita of the household, $r(t)$ is the risk–free market flow rate of return on assets, and $w(t)L(t)$ is the flow of labor income of the household. Defining per capita assets as

$$a(t) := \frac{h(t)}{L(t)}$$

we obtain:

$$\dot{a}(t) = (r(t) - n)a(t) + w(t) - c(t). \tag{1.11.12}$$

Household assets in this economy consist of capital stock that they rent to firms, $K(t)$, and risk–free bonds, $B(t)$, which are in zero net supply. Since bonds are in zero net supply and there is a representative agent, the equilibrium holdings of bonds are zero in equilibrium. Consequently, assets per capita will be equal to the capital stock per capita (or the capital–labor ratio in the economy), that is,

$$a(t) = k(t).$$

Moreover, with a depreciation rate of δ, the market rate of return on assets will be given by the equilibrium rental rate of capital net of depreciation:

$$r(t) = R(t) - \delta. \tag{1.11.13}$$

Equation (1.11.12) is only a flow constraint. It is not sufficient as a proper budget constraint on the individual. To see this, solve (1.11.12) between times 0 and $T > 0$ to get:

$$\begin{aligned} &\int_0^T c(t)L(t)\exp\left(-\int_0^t r(s)ds\right)dt + \exp\left(-\int_0^T r(s)ds\right)a(T) \\ &= \int_0^T w(t)L(t)\exp\left(-\int_0^t r(s)ds\right)dt + a(0). \end{aligned} \tag{1.11.14}$$

Differentiating this expression with respect to T and rearranging gives (1.11.12). The intertemporal budget constraint (1.11.14) states that the household's asset position at time T is given by the present value of total income plus initial assets minus expenditures, where present values are computed by discounting at the rate $r(t)$.

Now imagine that (1.11.14) applies to a finite-horizon economy ending at date T. In this case, it becomes clear that the flow budget constraint (1.11.12) by itself does not guarantee

that $h(T) \geq 0$. In a finite-horizon economy, we would simply impose $h(T) \geq 0$ as a boundary condition.

In the infinite-horizon case, we need a similar boundary condition. This is generally referred to as the transversality condition. One type of transversality condition is the no–Ponzi–game condition, which takes the form

$$\lim_{t\to\infty} a(t) \exp\left(-\int_0^t (r(s)-n)ds\right) \geq 0. \tag{1.11.15}$$

This condition is stated as an inequality, to ensure that the individual household does not asymptotically tend to a negative wealth. This no–Ponzi–game condition is necessary to ensure proper budget constraints. Furthermore, since utility is increasing, the individual household would never want to have positive wealth asymptotically, so the no–Ponzi–game condition can be alternatively stated as:

$$\lim_{t\to\infty} a(t) \exp\left(-\int_0^t (r(s)-n)ds\right) = 0. \tag{1.11.16}$$

To see where the transversality condition comes from, take the limit of (1.11.14) as $T \to \infty$. The transversality condition implies that

$$\lim_{T\to\infty} \exp\left(-\int_0^T r(s)ds\right) a(T) = 0,$$

so we obtain

$$\int_0^\infty c(t) \exp\left(-\int_0^t (r(s)-n)ds\right) dt = a(0) + \int_0^\infty w(t) \exp\left(-\int_0^t (r(s)-n)ds\right) dt.$$

This equation says that the discounted sum of expenditures must be equal to initial income plus the discounted sum of labor income. Therefore, this equation is a direct extension of (1.11.14) to infinite horizon. This derivation makes it clear that the no–Ponzi–game condition (1.11.16) essentially ensures that the individual's lifetime or intertemporal budget constraint holds in infinite horizon.

Let us start with the problem of the representative household. From the definition of equilibrium, we know that this is to maximize (1.11.2) subject to (1.11.12) and (1.11.16).

Let us first ignore (1.11.16) and set up the current-value Hamiltonian:

$$\hat{H}(a, c, \mu) = u(c(t)) + \mu(t)[w(t) + (r(t) - n)a(t) - c(t)]$$

with state variable a, control variable c and current-value costate variable μ. Applying the maximum principle, we obtain the following necessary conditions:

$$\hat{H}_c(a, c, \mu) = u'(c(t)) - \mu(t) = 0 \tag{1.11.17}$$

$$\hat{H}_a(a, c, \mu) = \mu(t)(r(t) - n) = -\dot{\mu}(t) + (\rho - n)\mu(t) \tag{1.11.18}$$

$$\lim_{t\to\infty}[\exp(-(\rho - n)t)\mu(t)a(t)] = 0$$

and the transition equation (1.11.12).

Moreover, for any $\mu(t) > 0$, $\hat{H}(a, c, \mu)$ is a concave function of (a, c), so the necessary conditions are also sufficient.

Rearranging (1.11.18) yields

$$\frac{\dot{\mu}(t)}{\mu(t)} = -(r(t) - \rho), \tag{1.11.19}$$

which states that the costate variable changes depending on whether the rate of return on assets is currently greater than or less than the discount rate of the household.

Next, (1.11.17) implies that

$$u'(c(t)) = \mu(t).$$

To make more progress, let us differentiate this with respect to time and divide by $\mu(t)$, which yields

$$\frac{u''(c(t))c(t)}{u'(c(t))}\frac{\dot{c}(t)}{c(t)} = \frac{\dot{\mu}(t)}{\mu(t)}.$$

Substituting this into (1.11.19), we obtain the **Euler equation** of this model

$$\frac{\dot{c}(t)}{c(t)} = \frac{1}{\varepsilon_u(c(t))}(r(t) - \rho), \tag{1.11.20}$$

where

$$\varepsilon_u(c(t)) := -\frac{u''(c(t))c(t)}{u'(c(t))}$$

is the inverse of the elasticity of intertemporal substitution. The elasticity of intertemporal substitution summarizes the willingness of individuals to substitute consumption (or labor, or any other attribute that yields utility) across time. The elasticity of intertemporal substitution between dates t and $s > t$ is defined as

$$\sigma_u(t,s) := -\frac{d\log(c(s)/c(t))}{d\log\left(u'(c(s))/u'(c(t))\right)}.$$

As $s \downarrow t$, we have

$$\sigma_u(t,s) \to \sigma_u(t) = -\frac{u'(c(t))}{u''(c(t))c(t)} = \frac{1}{\varepsilon_u(c(t))}.$$

This is not surprising, since the concavity of the utility function $u(\cdot)$—or equivalently, the elasticity of marginal utility—determines how willing individuals are to substitute consumption over time.

Next, integrating (1.11.19), we have

$$\begin{aligned}\mu(t) &= \mu(0)\exp\left(-\int_0^t (r(s)-\rho)ds\right)\\ &= u'(c(0))\exp\left(-\int_0^t (r(s)-\rho)ds\right),\end{aligned}$$

where the second line uses the first optimality condition of the current–value Hamiltonian at time $t = 0$. Now substituting into the transversality condition, we have

$$\lim_{t\to\infty}\left[\exp(-(\rho-n)t)a(t)u'(c(0))\exp\left(-\int_0^t (r(s)-\rho)ds\right)\right] = 0,$$

$$\lim_{t\to\infty}\left[a(t)\exp\left(-\int_0^t (r(s)-n)ds\right)\right] = 0,$$

which implies that the strict no–Ponzi condition, (1.11.16) has to hold. Also, for future reference, note that, since $a(t) = k(t)$, the transversality condition is also equivalent to

$$\lim_{t\to\infty}\left[\exp\left(-\int_0^t (r(s)-n)ds\right)k(t)\right] = 0,$$

which requires that the discounted market value of the capital stock in the very far future is equal to 0. This "market value" version of the transversality condition is sometimes more convenient to work with.

We can derive further results on the consumption behavior of households. In particular,

notice that the term $\exp\left(-\int_0^t r(s)ds\right)$ is a present-value factor that converts a unit of income at time t to a unit of income at time 0. In the special case where $r(s) = r$, this factor would be exactly equal to $\exp(-rt)$. But more generally, we can define an average interest rate between dates 0 and t as

$$\bar{r}(t) = \frac{1}{t}\int_0^t r(s)ds.$$

In that case, we can express the conversion factor between dates 0 and t as

$$\exp(-\bar{r}(t)t),$$

and the transversality condition can be written as

$$\lim_{t\to\infty}[\exp(-(\bar{r}(t) - n)t)a(t)] = 0.$$

Now recalling that the solution to the differential equation

$$\dot{y}(t) = b(t)y(t)$$

is

$$y(t) = y(0)\exp\left(\int_0^t b(s)ds\right),$$

we can integrate (1.11.20), to obtain

$$c(t) = c(0)\exp\left(\int_0^t \frac{r(s) - \rho}{\varepsilon_u(c(s))}ds\right)$$

as the consumption function. Once we determine $c(0)$, the initial level of consumption, the path of consumption can be exactly solved out. In the special case where $\varepsilon_u(c(s))$ is constant, for example, $\varepsilon_u(c(s)) = \theta$, this equation simplifies to

$$c(t) = c(0)\exp\left(\left(\frac{\bar{r}(t) - \rho}{\theta}\right)t\right),$$

and, moreover, the lifetime budget constraint simplifies to

$$\int_0^\infty c(t)\exp(-(\bar{r}(t) - n)t)dt = a(0) + \int_0^\infty w(t)\exp(-(\bar{r}(t) - n)t)dt.$$

Substituting for $c(t)$ into this lifetime budget constraint in the iso-elastic case, we obtain

$$c(0) = \int_0^\infty \exp\left(\left(\frac{(1-\theta)\bar{r}(t)}{\theta} - \frac{\rho}{\theta} + n\right)t\right)dt\left[a(0) + \int_0^\infty w(t)\exp(-(\bar{r}(t) - n)t)dt\right]$$

as the initial value of consumption.

1.11.4 Equilibrium

We can now define an equilibrium in this dynamic economy. We will provide two definitions, the first is somewhat more formal, while the second definition will be more useful in characterizing the equilibrium.

Definition 1.11.1. A competitive equilibrium of the Ramsey economy consists of paths of consumption, capital stock, wage rates and rental rates of capital, $[C(t), K(t), w(t), R(t)]_{t=0}^{\infty}$, such that the representative household maximizes its utility given initial capital stock $K(0)$ and the time path of prices $[w(t), R(t)]_{t=0}^{\infty}$, and all markets clear.

Notice that in equilibrium we need to determine the entire time path of real quantities and the associated prices. This is an important point to bear in mind. In dynamic models whenever we talk of "equilibrium", this refers to the entire path of quantities and prices. In some models, we will focus on the steady–state equilibrium, but equilibrium always refers to the entire path.

Since everything can be equivalently defined in terms of per capita variables, we can state an alternative and more convenient definition of equilibrium:

Definition 1.11.2. A competitive equilibrium of the Ramsey economy consists of paths of per capita consumption, capital–labor ratio, wage rates and rental rates of capital, $[c(t), k(t), w(t), R(t)]_{t=0}^{\infty}$, such that the representative household maximizes (1.11.2) subject to (1.11.12) and (1.11.15) given initial capital–labor ratio $k(0)$, factor prices $[w(t), R(t)]_{t=0}^{\infty}$ as in (1.11.10) and (1.11.11), and the rate of return on assets $r(t)$ given by (1.11.13).

1.11.5 Equilibrium Prices

Equilibrium prices are straightforward and are given by (1.11.10) and (1.11.11). This implies that the market rate of return for consumers, $r(t)$, is given by (1.11.13), i.e.,

$$r(t) = f'(k(t)) - \delta.$$

Substituting this into the consumer's problem, we have

$$\frac{\dot{c}(t)}{c(t)} = \frac{1}{\varepsilon_u(c(t))}\left(f'(k(t)) - \delta - \rho\right) \tag{1.11.21}$$

as the equilibrium version of the consumption growth equation, (1.11.20).

1.11.6 Optimal Growth

Before characterizing the equilibrium further, it is useful to look at the optimal growth problem, defined as the capital and consumption path chosen by a benevolent social planner trying to achieve a Pareto optimal outcome. In an economy that admits a representative household, the optimal growth problem simply involves the maximization of the utility of the representative household subject to technology and feasibility constraints. That is,

$$\max_{[k(t),c(t)]_{t=0}^{\infty}} \int_0^{\infty} \exp(-(\rho - n)t)u(c(t))dt,$$

subject to

$$\dot{k}(t) = f(k(t)) - (n + \delta)k(t) - c(t). \tag{1.11.22}$$

The First and Second Welfare Theorems for economies with a continuum of commodities would imply that the solution to this problem should be the same as the equilibrium growth problem. We can show this equivalence directly.

To do this, let us once again set up the current-value Hamiltonian, which in this case takes the form

$$\hat{H}(k, c, \mu) = u(c(t)) + \mu(t)[f(k(t)) - (n + \delta)k(t) - c(t)],$$

with state variable k, control variable c and current-value costate variable μ. In the relevant range for the capital stock, this problem satisfies all the assumptions for the application of the maximum principle. Consequently, the necessary conditions for an optimal path are:

$$\hat{H}_c(k, c, \mu) = 0 = u'(c(t)) - \mu(t),$$

$$\hat{H}_k(k, c, \mu) = -\dot{\mu}(t) + (\rho - n)\mu(t) = \mu(t)\left(f'(k(t)) - \delta - n\right),$$

$$\lim_{t\to\infty}[\exp(-(\rho - n)t)\mu(t)k(t)] = 0.$$

Repeating the same steps as before, it is straightforward to see that these optimality condi-

tions imply

$$\frac{\dot{c}(t)}{c(t)} = \frac{1}{\varepsilon_u(c(t))} \left(f'(k(t)) - \delta - \rho \right),$$

which is identical to (1.11.21), and the transversality condition

$$\lim_{t\to\infty} \left[k(t) \exp\left(- \int_0^t \left(f'(k(s)) - \delta - n \right) ds \right) \right] = 0,$$

which is, in turn, identical to (1.11.16). This establishes that the competitive equilibrium is a Pareto optimum and that the Pareto allocation can be decentralized as a competitive equilibrium. This result is stated in the next proposition:

Proposition 1.11.1. *In the neoclassical growth model described here, with Assumptions 1.11.3, 1.11.4, 1.11.1 and 1.11.2, the equilibrium is Pareto optimal and coincides with the optimal growth path maximizing the utility of the representative household.*

1.11.7 Steady-State Equilibrium

Now we characterize the steady–state. A steady state equilibrium is defined as an equilibrium path in which capital–labor ratio, consumption and output are constant. Therefore,

$$\dot{c}(t) = 0.$$

From (1.11.21), this implies that as long as $f(k^*) > 0$, irrespective of the exact utility function, we must have a capital–labor ratio k^* such that

$$f'(k^*) = \rho + \delta. \tag{1.11.23}$$

This equation pins down the steady–state capital–labor ratio only as a function of the production function, the discount rate and the depreciation rate. [8] This corresponds to the modified golden rule. The modified golden rule involves a level of the capital stock that does not maximize steady–state consumption, because earlier consumption is preferred to later consumption. This is because of discounting, which means that the objective is not to maximize steady–state consumption, but involves giving a higher weight to earlier

8. In addition, if $f(0) = 0$, there exists another, economically uninteresting steady state at $k = 0$. We ignore this steady state throughout. Moreover, starting with any $k(0) > 0$, the economy will always tend to the steady–state capital–labor ratio k^* given by (1.11.23).

consumption.

Given k^*, the steady–state consumption level is straightforward to determine as:

$$c^* = f(k^*) - (n+\delta)k^*, \tag{1.11.24}$$

which is similar to the consumption level in the basic Solow model. Moreover, given Assumption 1.11.2, a steady state where the capital–labor ratio and thus output are constant necessarily satisfies the transversality condition.

This analysis therefore establishes:

Proposition 1.11.2. *In the neoclassical growth model with Assumptions 1.11.3, 1.11.4, 1.11.1 and 1.11.2, the steady–state equilibrium capital–labor ratio, k^*, is uniquely determined by* (1.11.23) *and is independent of the utility function. The steady–state consumption per capita, c^*, is given by* (1.11.24).

There are also a number of straightforward comparative static results that show how the steady–state values of capital–labor ratio and consumption per capita change with the underlying parameters. For this reason, let us parameterize the production function as follows

$$f(k) = a\tilde{f}(k),$$

where $a > 0$, so that a is a shift parameter, with greater values corresponding to greater productivity of factors. Since $f(k)$ satisfies the regularity conditions imposed, so does $\tilde{f}(k)$.

Proposition 1.11.3. *Consider the neoclassical growth model with Assumptions 1.11.3, 1.11.4, 1.11.1 and 1.11.2, and suppose that $f(k) = a\tilde{f}(k)$. Denote the steady–state level of the capital–labor ratio by $k^*(a,\rho,n,\delta)$ and the steady–state level of consumption per capita by $c^*(a,\rho,n,\delta)$ when the underlying parameters are a,ρ,n and δ. Then we have*

$$\frac{\partial k^*(a,\rho,n,\delta)}{\partial a} > 0, \frac{\partial k^*(a,\rho,n,\delta)}{\partial \rho} < 0, \frac{\partial k^*(a,\rho,n,\delta)}{\partial n} = 0 \text{ and } \frac{\partial k^*(a,\rho,n,\delta)}{\partial \delta} < 0,$$
$$\frac{\partial c^*(a,\rho,n,\delta)}{\partial a} > 0, \frac{\partial c^*(a,\rho,n,\delta)}{\partial \rho} < 0, \frac{\partial c^*(a,\rho,n,\delta)}{\partial n} < 0 \text{ and } \frac{\partial c^*(a,\rho,n,\delta)}{\partial \delta} < 0.$$

The new results here relative to the basic Solow model concern the comparative statics

with respect to the discount rate. In particular, instead of the saving rate, it is now the discount factor that affects the rate of capital accumulation. There is a close link between the discount rate in the neoclassical growth model and the saving rate in the Solow model. Loosely speaking, a lower discount rate implies greater patience and thus greater saving. In the model without technological progress, the steady–state saving rate is

$$s^* = \frac{\delta k^*}{f\left(k^*\right)}.$$

Another interesting result is that the rate of population growth has no impact on the steady state capital–labor ratio, which contrasts with the basic Solow model. This result depends on the way in which intertemporal discounting takes place. Another important result, which is more general, is that k^* and thus c^* do not depend on the instantaneous utility function $u(\cdot)$. The form of the utility function only affects the transitional dynamics (which we will study next), but has no impact on steady states. This is because the steady state is determined by the modified golden rule. This result will not be true when there is technological change, however.

1.11.8 Transitional Dynamics

Next, we can determine the transitional dynamics of this model. Unlike simpler growth models with a single differential equation, the equilibrium here is determined by two differential equations:

$$\dot{k}(t) = f(k(t)) - (n+\delta)k(t) - c(t),$$

and

$$\frac{\dot{c}(t)}{c(t)} = \frac{1}{\varepsilon_u(c(t))}\left(f'(k(t)) - \delta - \rho\right).$$

Moreover, we have an initial condition $k(0) > 0$, also a boundary condition at infinity, of the form

$$\lim_{t\to\infty}\left[k(t)\exp\left(-\int_0^t \left(f'(k(s)) - \delta - n\right)ds\right)\right] = 0.$$

We now study the system diagrammatically using Figure 1.3.

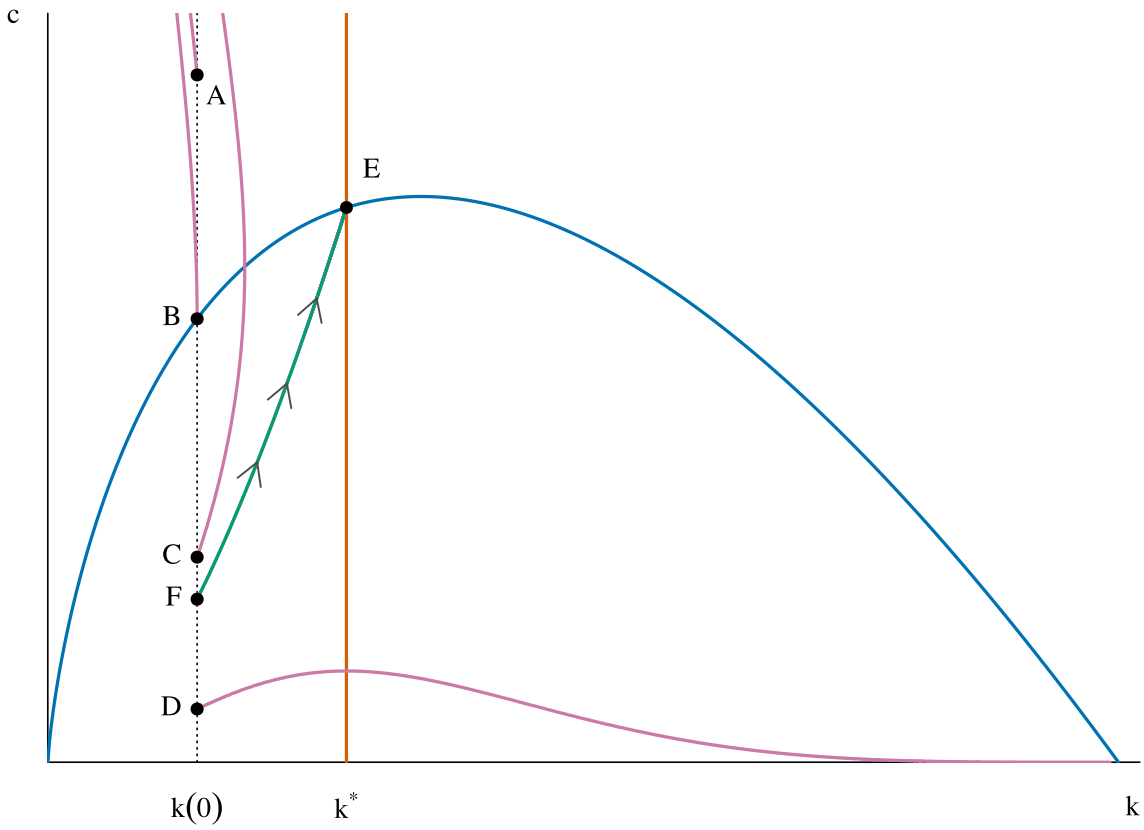

Figure 1.3: Phase diagram for the neoclassical growth model.

The vertical line is the locus of points where $\dot{c} = 0$. The reason why the $\dot{c} = 0$ locus is just a vertical line is that in view of the consumer Euler equation (1.11.21), only the unique level of k^* given by (1.11.23) can keep per capita consumption constant. The inverse U-shaped curve is the locus of points where $\dot{k} = 0$ in (1.11.22). The intersection of these two loci defines the steady state. If the capital stock is too low, steady–state consumption is low, and if the capital stock is too high, then the steady–state consumption is again low. There exists a unique level, k_{gold} that maximizes the steady–state consumption per capita. The $\dot{c} = 0$ locus intersects the $\dot{k} = 0$ locus always to the left of k_{gold}. Once these two loci are drawn, the rest of the diagram can be completed by looking at the direction of motion according to the differential equations. Given this direction of movements, it is clear that there exists a unique saddle path, the one-dimensional manifold tending to the steady state.

All points away from this saddle path diverge, and eventually reach zero consumption or zero capital stock as shown in the figure. To see this, note that if initial consumption, $c(0)$, started above this saddle path, say at $c'(0)$, the capital stock would reach 0 in finite time, while consumption would remain positive. But this would violate feasibility. Therefore, initial values of consumption above this saddle path cannot be part of the equilibrium (or the optimal growth solution). If the initial level of consumption were below it, for example, at $c''(0)$, consumption would reach zero, thus capital would accumulate continuously until the maximum level of capital (reached with zero consumption) $\bar{k} > k_{\text{gold}}$. Continuous capital accumulation towards $\bar{k}$ with no consumption would violate the transversality condition. This establishes that the transitional dynamics in the neoclassical growth model will take the following simple form: $c(0)$ will "jump" to the saddle path, and then (k, c) will monotonically travel along this arm towards the steady state.

An alternative way of establishing the same result is by linearizing the set of differential equations, and looking at their eigenvalues. Recall the two differential equations determining the equilibrium path:

$$\dot{k}(t) = f(k(t)) - (n+\delta)k(t) - c(t),$$

and

$$\frac{\dot{c}(t)}{c(t)} = \frac{1}{\varepsilon_u(c(t))}\left(f'(k(t)) - \delta - \rho\right).$$

Linearizing these equations around the steady state (k^*, c^*), we have (suppressing time dependence)

$$\begin{aligned}
\dot{k} &= \text{ constant } + \left(f'\left(k^*\right) - n - \delta\right)\left(k - k^*\right) - c, \\
\dot{c} &= \text{ constant } + \frac{c^* f''\left(k^*\right)}{\varepsilon_u\left(c^*\right)}\left(k - k^*\right).
\end{aligned}$$

Moreover, from (1.11.23), $f'\left(k^*\right) - \delta = \rho$, so the eigenvalues of this two-equation system are given by the values of ξ that solve the following quadratic form:

$$\det\begin{pmatrix} \rho - n - \xi & -1 \\ \frac{c^* f''(k^*)}{\varepsilon_u(c^*)} & 0 - \xi \end{pmatrix} = 0.$$

It is straightforward to verify that, since $c^* f''\left(k^*\right)/\varepsilon_u\left(c^*\right) < 0$, there are two real eigenvalues,

one negative and one positive. This implies that there exists a one-dimensional stable manifold converging to the steady state, exactly as the saddle path in the figure. Therefore, the local analysis also leads to the same conclusion. However, the local analysis can only establish local stability, whereas the diagrammatic analysis established global stability.

1.11.9 Technological Change and the Canonical Neoclassical Model

The analysis so far was for the neoclassical growth model without any technological change. The neoclassical growth model would not be able to account for long-run growth without some type of exogenous technological change. Therefore, the more interesting version of this model is the one that incorporates technological change. We now analyze the neoclassical model with exogenous technological change.

We extend the production function to:

$$Y(t) = F(K(t), A(t)L(t)), \tag{1.11.25}$$

where

$$A(t) = \exp(gt)A(0).$$

Notice that the production function (1.11.25) imposes purely labor-augmenting (Harrod-neutral) technological change. Only purely labor-augmenting technological change is consistent with balanced growth, as you will show in a problem set.

We continue to adopt Assumptions 1.11.3, 1.11.4 and 1.11.1. Assumption 1.11.2 will be strengthened further in order to ensure finite discounted utility in the presence of sustained economic growth.

The constant returns to scale feature again enables us to work with normalized variables. Now let us define

$$\begin{aligned}\hat{y}(t) &:= \frac{Y(t)}{A(t)L(t)} \\ &= F\left(\frac{K(t)}{A(t)L(t)}, 1\right) \\ &:= f(k(t)),\end{aligned}$$

where

$$k(t) := \frac{K(t)}{A(t)L(t)}$$

is the capital-to-effective-labor ratio, which is defined taking into account that effective labor is increasing because of labor-augmenting technological change.

In addition to the assumptions on technology, we also need to impose a further assumption on preferences in order to ensure balanced growth. We define balanced growth as a pattern of growth consistent with the Kaldor facts of constant capital–output ratio and capital share in national income. These two observations together also imply that the rental rate on capital, $R(t)$, has to be constant. Using (1.11.13), we see that $r(t)$ must then be constant as well.

The Euler equation implies that

$$\frac{\dot{c}(t)}{c(t)} = \frac{1}{\varepsilon_u(c(t))}(r(t) - \rho).$$

If $r(t) \to r^*$, then $\dot{c}(t)/c(t) \to g_c$ is only possible if $\varepsilon_u(c(t)) \to \varepsilon_u$, i.e., if the elasticity of intertemporal substitution is asymptotically constant. Therefore, balanced growth is only consistent with utility functions that have a constant elasticity of intertemporal substitution as $t \to \infty$.

The next example shows the family of utility functions with constant elasticity of intertemporal substitution, which are also those with a constant coefficient of relative risk aversion.

Example 1.11.0.1. (CRRA Utility) Recall that the coefficient of relative risk aversion for a twice-continuously differentiable concave utility function $u(c)$ is

$$\mathcal{R} = -\frac{u''(c)c}{u'(c)}.$$

Constant relative risk aversion (CRRA) utility function satisfies the property that $\mathcal{R}$ is constant. Given the restriction that balanced growth is only possible with preferences featuring a constant elasticity of intertemporal substitution, we start with a utility function that has this

feature throughout. The unique time-separable utility function with this feature is

$$u(c(t)) = \begin{cases} \frac{c(t)^{1-\theta}-1}{1-\theta} & \text{if } \theta \neq 1 \text{ and } \theta \geq 0 \\ \log c(t) & \text{if } \theta = 1 \end{cases} \tag{1.11.26}$$

where the elasticity of marginal utility of consumption, ε_u, is given by the constant θ. When $\theta = 0$, these represent linear preferences, whereas when $\theta = 1$, we have $\log$ preferences. As $\theta \to \infty$, these preferences become infinitely risk-averse, and infinitely unwilling to substitute consumption over time. □

More specifically, we now assume that the economy admits a representative household with CRRA preferences

$$\int_0^\infty \exp(-(\rho - n)t) \frac{\tilde{c}(t)^{1-\theta} - 1}{1-\theta} dt$$

where $\tilde{c}(t) := C(t)/L(t)$ is per capita consumption. We use the notation $\tilde{c}(t)$ in order to preserve $c(t)$ for a further normalization.

We refer to this model, with labor–augmenting technological change and CRRA preferences as given by (1.11.26) as the canonical model, since it is the model used in almost all applications of the neoclassical growth model. The Euler equation in this case takes the simpler form:

$$\frac{\dot{\tilde{c}}(t)}{\tilde{c}(t)} = \frac{1}{\theta}(r(t) - \rho). \tag{1.11.27}$$

Let us first characterize the steady–state equilibrium in this model with technological progress. Since with technological progress there will be growth in per capita income, $\tilde{c}(t)$ will grow. Instead, in analogy with $k(t)$, let us define

$$\begin{aligned} c(t) &:= \frac{C(t)}{A(t)L(t)} \\ &:= \frac{\tilde{c}(t)}{A(t)} \end{aligned}$$

We will see that this normalized consumption level will remain constant along the BGP. In

particular, we have

$$\begin{aligned}\frac{\dot{c}(t)}{c(t)} &:= \frac{\dot{\tilde{c}}(t)}{\tilde{c}(t)} - g \\ &= \frac{1}{\theta}(r(t) - \rho - \theta g).\end{aligned}$$

Moreover, for the accumulation of capital stock, we have

$$\dot{k}(t) = f(k(t)) - c(t) - (n + g + \delta)k(t),$$

where $k(t) := K(t)/A(t)L(t)$. The transversality condition, in turn, can be expressed as

$$\lim_{t\to\infty}\left\{k(t)\exp\left(-\int_0^t [f'(k(s)) - g - \delta - n]\,ds\right)\right\} = 0. \tag{1.11.28}$$

In addition, the equilibrium interest rate, $r(t)$, is still given by (1.11.13), so

$$r(t) = f'(k(t)) - \delta.$$

Since in steady state $c(t)$ must remain constant, we also have

$$r(t) = \rho + \theta g,$$

or

$$f'\left(k^*\right) = \rho + \delta + \theta g, \tag{1.11.29}$$

which pins down the steady–state value of the normalized capital ratio k^* uniquely, in a way similar to the model without technological progress. The level of normalized consumption is then given by

$$c^* = f\left(k^*\right) - (n + g + \delta)k^*,$$

while per capita consumption grows at the rate g. The only additional condition in this case is that because there is growth, we have to make sure that the transversality condition is in fact satisfied. Substituting (1.11.29) into (1.11.28), we have

$$\lim_{t\to\infty}\left\{k(t)\exp\left(-\int_0^t [\rho - (1-\theta)g - n]ds\right)\right\} = 0,$$

which can only hold if the integral within the exponent goes to zero, i.e., if $\rho-(1-\theta)g-n > 0$ or, alternatively, if the following assumption is satisfied:

Assumption 1.11.5. $\rho - n > (1-\theta)g$.

Note that this assumption strengthens Assumption 1.11.2 when $\theta < 1$. In steady state, we have $r = \rho + \theta g$ and the growth rate of output is $g + n$. Therefore, Assumption 1.11.5 is equivalent to requiring that $r > g + n$.

The result that the steady–state capital–labor ratio was independent of preferences is no longer the case, since now k^* given by (1.11.29) depends on the elasticity of marginal utility (or the inverse of the elasticity of intertemporal substitution), θ. The reason for this is that there is now positive growth in output per capita, and thus in consumption per capita. Since individuals face an upward-sloping consumption profile, their willingness to substitute consumption today for consumption tomorrow determines how much they will accumulate and thus the equilibrium effective capital–labor ratio.

While the steady–state effective capital–labor ratio, k^*, is determined endogenously, the steady–state growth rate of the economy is given exogenously and is equal to the rate of labor-augmenting technological progress, g. Therefore, the neoclassical growth model, like the basic Solow growth model, endogenizes the capital–labor ratio, but not the growth rate of the economy. The advantage of the neoclassical growth model is that the capital–labor ratio and the equilibrium level of (normalized) output and consumption are determined by the preferences of the individuals rather than an exogenously fixed saving rate. This also enables us to compare equilibrium and optimal growth (and in this case conclude that the competitive equilibrium is Pareto optimal and any Pareto optimum can be decentralized). But the determination of the rate of growth of the economy is still outside the scope of analysis.

Example 1.11.0.2. Consider the model with CRRA utility and labor-augmenting technological progress at the rate g. Assume that the production function is given by $F(K, AL) = K^\alpha (AL)^{1-\alpha}$, so that

$$f(k) = k^\alpha,$$

and thus $r = \alpha k^{\alpha-1} - \delta$. In this case, suppressing time dependence to simplify notation, the

Euler equation becomes:

$$\frac{\dot{c}}{c} = \frac{1}{\theta}\left(\alpha k^{\alpha-1} - \delta - \rho - \theta g\right),$$

and the accumulation equation can be written as

$$\frac{\dot{k}}{k} = k^{\alpha-1} - \delta - g - n - \frac{c}{k}.$$

Now define $z := c/k$ and $x := k^{\alpha-1}$, which implies that $\dot{x}/x = (\alpha-1)\dot{k}/k$. Therefore, these two equations can be written as

$$\frac{\dot{x}}{x} = -(1-\alpha)(x - \delta - g - n - z) \tag{1.11.30}$$

and

$$\begin{aligned}\frac{\dot{z}}{z} &= \frac{\dot{c}}{c} - \frac{\dot{k}}{k} \\ &= \frac{1}{\theta}(\alpha x - \delta - \rho - \theta g) - x + \delta + g + n + z \\ &= \frac{1}{\theta}((\alpha-\theta)x - (1-\theta)\delta + \theta n) - \frac{\rho}{\theta} + z.\end{aligned} \tag{1.11.31}$$

The two differential equations (1.11.30) and (1.11.31) together with the initial condition $x(0)$ and the transversality condition completely determine the dynamics of the system. This example can be completed for the special case in which $\theta \to 1$ (i.e., $\log$ preferences). □

1.11.10 Comparative Dynamics

We now briefly discuss how comparative dynamics are different in the neoclassical growth model than those in the basic Solow model. Recall that while comparative statics refer to changes in steady state in response to changes in parameters, comparative dynamics look at how the entire equilibrium path of variables changes in response to a change in policy or parameters. Since the purpose here is to give a sense of how these results are different, we will only look at the effect of a change in a single parameter, the discount rate ρ. Imagine a neoclassical growth economy with population growth at the rate n, labor-augmenting technological progress at the rate g and a discount rate ρ that has settled into a steady state represented by (k^*, c^*). Now imagine that the discount rate declines to $\rho' < \rho$. How does the equilibrium path change?

We know from our previous analysis that at the new discount rate $\rho' > 0$, there exists a unique steady state equilibrium that is saddle path stable. Let this steady state be denoted by (k^{**}, c^{**}). Therefore, the equilibrium will ultimately tend to this new steady–state equilibrium. Moreover, since $\rho' < \rho$, we know that the new steady–state effective capital–labor ratio has to be greater than k^*, that is, $k^{**} > k^*$ (while the equilibrium growth rate will remain unchanged). Consumption must drop immediately to reach the new saddle path, so that capital can accumulate towards its new steady–state level. Following this initial reaction, consumption slowly increases along the saddle path to a higher level of consumption.

Comparative dynamics in response to changes in other parameters, including the rate of labor-augmenting technological progress g, the rate of population growth n, and other aspects of the utility function, can also be analyzed similarly. Similar analysis can be applied to work through the comparative dynamics in response to a change in the rate of labor-augmenting technological progress, g, and in response to an anticipated future change in ρ.

1.11.11 The Role of Policy

In this model, the rate of growth of per capita consumption and output per worker are determined exogenously by the growth rate of labor-augmenting technological progress. The level of income, on the other hand, depends on preferences, in particular, on the elasticity of intertemporal substitution, $1/\theta$, the discount rate, ρ, the depreciation rate, δ, the population growth rate, n, and naturally the form of the production function $f(\cdot)$.

This model gives us a way of understanding differences in income per capita across countries in terms of preference and technology parameters. The elasticity of intertemporal substitution and the discount rate can be viewed as potential determinants of economic growth related to cultural or geographic factors. However, an explanation for cross-country and over-time differences in economic growth based on differences or changes in preferences is unlikely to be satisfactory. A more appealing direction may be to link the incentives to accumulate physical capital (and later to accumulate human capital and technology) to the institutional environment of an economy. For now, it is useful to focus on a particularly

simple way in which institutional differences might affect investment decisions, which is through differences in policies. To do this, let us extend the framework in a simple way and introduce linear tax policy. Suppose that returns on capital net of depreciation are taxed at the rate τ and the proceeds of this are redistributed back to the consumers. In that case, the capital accumulation equation, in terms of normalized capital, remains:

$$\dot{k}(t) = f(k(t)) - c(t) - (n + g + \delta)k(t),$$

but the net interest rate faced by households now changes to:

$$r(t) = (1 - \tau)\left(f'(k(t)) - \delta\right),$$

because of the taxation of capital returns. The growth rate of normalized consumption is then obtained from the consumer Euler equation, (1.11.27), as

$$\begin{aligned}\frac{\dot{c}(t)}{c(t)} &= \frac{1}{\theta}(r(t) - \rho - \theta g) \\ &= \frac{1}{\theta}\left((1 - \tau)\left(f'(k(t)) - \delta\right) - \rho - \theta g\right).\end{aligned}$$

An identical argument immediately implies that the steady–state capital–to–effective–labor ratio is given by

$$f'\left(k^*\right) = \delta + \frac{\rho + \theta g}{1 - \tau}. \tag{1.11.32}$$

This equation shows the effects of taxes on steady–state capital–to–effective–labor ratio and output per capita. A higher tax rate τ increases the right-hand side of (1.11.32), and since from Assumption 1.11.4, $f'(\cdot)$ is decreasing, it reduces k^*. Therefore, higher taxes on capital have the effect of depressing capital accumulation and reducing income per capita. This shows one channel through which policy (thus institutional) differences might affect economic performance. We can also note that similar results would be obtained if instead of taxes being imposed on returns from capital, they were imposed on the amount of investment (see next section). Naturally, we have not so far offered a reason why some countries may tax capital at a higher rate than others, which is again a topic that will be discussed later. Before doing this, in the next section we will also discuss how large these effects can be and whether they could account for the differences in cross-country incomes.

1.11.12 Extensions

There are many empirically and theoretically relevant extensions of the neoclassical growth model. These include endogenizing the labor supply decisions on individuals by introducing leisure in the utility function, which corresponds to the version of the neoclassical growth model most often employed in short-run and medium-run macroeconomic analyses. Other important extensions include introducing government expenditures and taxation into the basic model, analyzing the behavior of the basic neoclassical growth model with a free capital account (representing borrowing and lending opportunities for the economy at some exogenously given international interest rate r^*), adding adjustment costs to investment, and exploring versions of the neoclassical model with multiple sectors.

1.11.13 Conclusion

Unlike the Solow-Swan model, the neoclassical growth model explicitly models consumer and firm optimization. Sustained long-term growth can only be generated by technological progress, but technological progress remains exogenous in this model. A major contribution is to endogenize capital accumulation decisions by specifying the preferences of consumers, allowing us to link the saving rates to the instantaneous utility function, to the discount rate, and also to technology and prices in the economy.

Perhaps the most important contribution of this model is that it paves the way for further analysis of capital accumulation, human capital and endogenous technological progress. This model is a conceptually important step towards the elucidation of the ultimate causes of economic growth, providing a useful scaffolding that allows for many generalizations.

For all of its advantages and accomplishments, the neoclassical growth model is unable to generate the observed dispersion in growth rates and capital-income ratios across countries. In addition, the focus is on the more proximate causes of these differences—saving rates, investments in human capital and technology, preferences—which begs the question of what economic factors are the ultimate, more fundamental sources of growth.

1.12 Mathematical Appendix

1.12.1 System of linear ODEs

The goal is to solve the linear system of ODEs given by:

$$\frac{dx(t)}{dt} = B(t) + Ax(t), \tag{1.12.1}$$

where $x(t)$ is an $n \times 1$ vector, A is an $n \times n$ matrix, $B(t)$ is an $n \times 1$ vector that does not depend on x and the operator $\frac{d}{dt}$ takes the derivative of vectors component-wise. Sometimes, the notation

$$\dot{x}(t) := \frac{dx(t)}{dt}$$

is more convenient and we will use both options interchangeably.

To solve (1.12.1), we first solve the homogeneous part:

$$\frac{dx(t)}{dt} = Ax(t). \tag{1.12.2}$$

If A is diagonalizable, we can write

$$A = Q\Lambda Q^{-1},$$

where Λ is a diagonal $n \times n$ matrix with the eigenvalues of A on its diagonal and Q is an $n \times n$ matrix whose columns are the eigenvectors of A. By convention, we will always order the eigenvalues in decreasing order (the largest eigenvalue is Λ_{11} and the corresponding eigenvector is the first column of Q and so on). If A is not diagonalizable, then replace the diagonal matrix Λ by the Jordan normal form J of A, which can be written $J = D + P$ where D is diagonal and P is nilpotent (there exists k such that $P^k = 0$). For these notes, we will stick to diagonalizable matrices.

Multiplying (1.12.2) by Q^{-1} we get:

$$\dot{y}(t) = \Lambda y(t), \tag{1.12.3}$$

where

$$y(t) := Q^{-1}x(t).$$

The solution to (1.12.3) is easy, since all equations are decoupled:

$$y(t) = e^{\Lambda t}y_0, \tag{1.12.4}$$

where y_0 is the $n \times 1$ vector of initial conditions, with

$$y_0 = Q^{-1}x_0.$$

The matrix exponential e^X for a square matrix X can be defined by the usual Taylor expansion of the exponential function

$$e^X := \sum_{k=0}^{\infty} \frac{1}{k!}X^k.$$

Since Λ is diagonal, the i^{th} diagonal entry of $e^{\Lambda t}$ is $e^{\Lambda_{ii}t}$. Multiplying (1.12.4) by Q gives the general solution to (1.12.2):

$$\begin{aligned} Qy(t) &= Qe^{\Lambda t}y_0 \\ x(t) &= Qe^{\Lambda t}Q^{-1}x_0 \\ x(t) &= e^{At}x_0. \end{aligned} \tag{1.12.5}$$

The general solution to the inhomogeneous equation (1.12.1) is the sum of the general solution (1.12.5) and any particular solution to (1.12.1). A straightforward particular solution is:

$$S(t) = \Psi(t)\int_0^t \Psi(s)^{-1}B(s)ds,$$

where $\Psi(t)$ is the fundamental matrix of the homogeneous equation, normalized so that $\Psi(0) = I$:

$$\Psi(t) := e^{At}.$$

If A is diagonalizable with $A = Q\Lambda Q^{-1}$, then $e^{At} = Qe^{\Lambda t}Q^{-1}$. Therefore, the general solution

is:

$$x(t) = S(t) + e^{At}x_0. \tag{1.12.6}$$

We verify that the general solution indeed satisfies (1.12.1). First, we compute

$$\begin{aligned}\dot{S}(t) &= \dot{\Psi}(t)\int_0^t \Psi(s)^{-1}B(s)ds + \Psi(t)\Psi(t)^{-1}B(t)\\ &= A\Psi(t)\int_0^t \Psi(s)^{-1}B(s)ds + B(t)\\ &= AS(t) + B(t).\end{aligned}$$

Then we verify that (1.12.6) satisfies (1.12.1):

$$\begin{aligned}\frac{dx(t)}{dt} &= \dot{S}(t) + \frac{d}{dt}\left(e^{At}x_0\right)\\ &= \dot{S}(t) + Ae^{At}x_0\\ &= AS(t) + B(t) + Ae^{At}x_0\\ &= A\left(S(t) + e^{At}x_0\right) + B(t)\\ &= Ax(t) + B(t).\end{aligned}$$

The 2×2 case

If

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix},$$

a few good relations to remember are:

$$\begin{aligned}\Lambda_{11}, \Lambda_{22} \text{ given by } &\frac{\operatorname{tr}(A) \pm \sqrt{(\operatorname{tr}(A))^2 - 4\det(A)}}{2},\\ \Lambda_{11} + \Lambda_{22} &= \operatorname{tr}(A) = a + d,\\ \Lambda_{11} - \Lambda_{22} &= \sqrt{(\operatorname{tr}(A))^2 - 4\det(A)},\\ \Lambda_{11}\Lambda_{22} &= \det(A) = ad - bc.\end{aligned}$$

The eigenvectors are:

$$\begin{bmatrix} \frac{b}{\Lambda_{11}-a} \\ 1 \end{bmatrix} \text{ and } \begin{bmatrix} \frac{b}{\Lambda_{22}-a} \\ 1 \end{bmatrix}$$

(assuming $b \neq 0$; if $b = 0$ choose the corresponding alternatives, e.g., if $b = 0$ and $c \neq 0$ use $\begin{bmatrix} 1 & (\Lambda_{ii} - d)/c \end{bmatrix}^T$, and if $b = c = 0$ (diagonal A) use the standard basis vectors). If B is constant,

$$\begin{aligned} S(t) &= \Psi(t) \int_{t_{start}}^{t} \Psi(s)^{-1} B \, ds \\ &= e^{At} \int_{t_{start}}^{t} e^{-As} B \, ds \\ &= Q e^{\Lambda t} \int_{t_{start}}^{t} \left[e^{\Lambda s}\right]^{-1} Q^{-1} B \, ds \\ &= Q e^{\Lambda t} \Lambda^{-1} \left(e^{-\Lambda t_{start}} - e^{-\Lambda t}\right) Q^{-1} B \end{aligned}$$

(assuming Λ is invertible; if some $\Lambda_{ii} = 0$, interpret the corresponding diagonal entry as $t - t_{start}$), where we have used that:

$$\begin{aligned} \int_{t_{start}}^{t} \left[e^{\Lambda s}\right]^{-1} ds &= \begin{bmatrix} \int_{t_{start}}^{t} e^{-\Lambda_{11} s} ds & 0 \\ 0 & \int_{t_{start}}^{t} e^{-\Lambda_{22} s} ds \end{bmatrix} \\ &= \begin{bmatrix} \frac{e^{-\Lambda_{11} t_{start}} - e^{-\Lambda_{11} t}}{\Lambda_{11}} & 0 \\ 0 & \frac{e^{-\Lambda_{22} t_{start}} - e^{-\Lambda_{22} t}}{\Lambda_{22}} \end{bmatrix} \\ &= \Lambda^{-1} \left(e^{-\Lambda t_{start}} - e^{-\Lambda t}\right). \end{aligned}$$

Then the solution is:

$$x(t) = Q e^{\Lambda t} \Lambda^{-1} \left(e^{-\Lambda t_{start}} - e^{-\Lambda t}\right) Q^{-1} B + Q e^{\Lambda t} Q^{-1} x_0.$$

We check with $t_{start} = 0$:

$$\begin{aligned}
\dot{x}(t) &= \frac{d}{dt}\left[Qe^{\Lambda t}\Lambda^{-1}\left(e^{-\Lambda t_{start}} - e^{-\Lambda t}\right)Q^{-1}B + Qe^{\Lambda t}Q^{-1}x_0\right] \\
&= \frac{d}{dt}\left[Q\Lambda^{-1}\left(e^{\Lambda t} - I\right)Q^{-1}B + Qe^{\Lambda t}Q^{-1}x_0\right] \\
&= Q\Lambda^{-1}\Lambda e^{\Lambda t}Q^{-1}B + Q\Lambda e^{\Lambda t}Q^{-1}x_0 \\
&= Qe^{\Lambda t}Q^{-1}B + AQe^{\Lambda t}Q^{-1}x_0 \\
&= Qe^{\Lambda t}Q^{-1}B + A\left(Qe^{\Lambda t}Q^{-1}x_0\right) \\
&= Ax(t) + B.
\end{aligned}$$

1.12.2 The maximum principle

Consider the infinite-horizon optimal control problem

$$\max_{u(\cdot)} \int_0^\infty e^{-\rho t} f(x(t), u(t))\, dt$$

subject to

$$\dot{x}(t) = \mu(t, x(t), u(t)), \quad x(0) = x_0, \quad u(t) \in \mathcal{U} \text{ for all } t \geq 0, \quad G(x(t), u(t)) \geq 0,$$

where $x(t) \in \mathbb{R}^n$ are state variables, $u(t) \in \mathbb{R}^s$ are control variables, $u(\cdot)$ is a measurable control function, $\mathcal{U} \subset \mathbb{R}^s$ is the control set, $\rho > 0$ is the discount rate, $f : \mathbb{R}^n \times \mathbb{R}^s \to \mathbb{R}$ is the instantaneous payoff, $\mu : \mathbb{R}_+ \times \mathbb{R}^n \times \mathbb{R}^s \to \mathbb{R}^n$ gives the state dynamics, and $G : \mathbb{R}^n \times \mathbb{R}^s \to \mathbb{R}^q$ is a vector-valued function representing q path constraints.

Define the current-value Hamiltonian

$$\mathcal{H}(t, x, u, \lambda, \phi) = f(x, u) + \lambda^T \mu(t, x, u) + \phi^T G(x, u),$$

with costate $\lambda(t) \in \mathbb{R}^n$ and multipliers $\phi(t) \in \mathbb{R}^q$, where $\phi(t) \geq 0$ componentwise.

If an admissible pair $(x^*(\cdot), u^*(\cdot))$ solves the problem and suitable regularity conditions hold, then there exist an absolutely continuous function $\lambda(\cdot)$ and a measurable function $\phi(\cdot)$ such that, for almost all $t \geq 0$:

1. State equation and feasibility:

$$\dot{x}^*(t) = \mu(t, x^*(t), u^*(t)), \quad x^*(0) = x_0, \quad u^*(t) \in \mathcal{U}, \quad G(x^*(t), u^*(t)) \geq 0.$$

2. Hamiltonian maximization:

$$u^*(t) \in \arg \max_{u \in \mathcal{U},\, G(x^*(t),u) \geq 0} \mathcal{H}\big(t, x^*(t), u, \lambda(t), \phi(t)\big),$$

   and, if $u^*(t)$ is in the interior of the feasible control set, this yields the first-order condition $\partial \mathcal{H}/\partial u = 0$ evaluated at $\big(t, x^*(t), u^*(t), \lambda(t), \phi(t)\big)$.

3. Costate dynamics in current value:

$$\dot{\lambda}(t) = \rho\, \lambda(t) - \frac{\partial \mathcal{H}}{\partial x}\big(t, x^*(t), u^*(t), \lambda(t), \phi(t)\big).$$

4. Complementary slackness and sign:

$$\phi_j(t) \geq 0, \quad G_j(x^*(t), u^*(t)) \geq 0, \quad \phi_j(t)\, G_j(x^*(t), u^*(t)) = 0 \quad \text{for } j = 1, \ldots, q.$$

It is sometimes convenient to work with present-value objects. Define $p(t) = e^{-\rho t}\lambda(t)$ and $\psi(t) = e^{-\rho t}\phi(t)$, and the present-value Hamiltonian

$$\mathcal{H}^{pv}(t, x, u, p, \psi) = e^{-\rho t} f(x, u) + p^T \mu(t, x, u) + \psi^T G(x, u),$$

so that the costate equation becomes

$$\dot{p}(t) = -\frac{\partial \mathcal{H}^{pv}}{\partial x}\big(t, x^*(t), u^*(t), p(t), \psi(t)\big),$$

with the relations $p(t) = e^{-\rho t}\lambda(t)$ and $\psi(t) = e^{-\rho t}\phi(t)$ holding pointwise in time.

Transversality. As a baseline transversality condition for discounted problems with a finite optimal value, one can impose

$$\lim_{t \to \infty} \mathcal{H}^{pv}\big(t, x^*(t), u^*(t), p(t), \psi(t)\big) = 0,$$

that is,

$$\lim_{t\to\infty}\big[e^{-\rho t}f(x^*(t),u^*(t))+p(t)^T\mu\big(t,x^*(t),u^*(t)\big) \\ +\psi(t)^TG\big(x^*(t),u^*(t)\big)\big]=0.$$

On path segments where $G(\cdot)$ is non-binding, the term with ψ vanishes by complementary slackness, and in unconstrained problems this reduces to

$$\lim_{t\to\infty}\big[e^{-\rho t}f(x^*(t),u^*(t))+p(t)^T\mu\big(t,x^*(t),u^*(t)\big)\big]=0.$$

A useful strengthening is obtained under additional structure. Suppose the state space is $\mathbb{R}^n_+$ with $x^*(t)\geq 0$ componentwise for all t, the value function V is concave and differentiable so that $\lambda(t)=\nabla V\big(x^*(t)\big)$, and the problem primitives imply that $\lambda(t)\geq 0$ componentwise (for example, if f and the components of μ are non-decreasing in x). Then the baseline condition implies

$$\lim_{t\to\infty}e^{-\rho t}\lambda(t)^Tx^*(t)=\lim_{t\to\infty}p(t)^Tx^*(t)=0,$$

because concavity of V yields $\lambda(t)^Tx^*(t)\leq V\big(x^*(t)\big)-V(0)$ and hence $0\leq p(t)^Tx^*(t)\leq e^{-\rho t}V\big(x^*(t)\big)-e^{-\rho t}V(0)$, with the right-hand side converging to zero if the total value is finite.

Sufficiency. The necessary conditions above are also sufficient for optimality if the transversality condition holds, and if, for each t, the map $(x,u)\mapsto\mathcal{H}\big(t,x,u,\lambda(t),\phi(t)\big)$ is concave on the set $\{(x,u)\mid u\in\mathcal{U},G(x,u)\geq 0\}$ and this set is convex. In this case, any admissible path satisfying the conditions is globally optimal.

1.12.3 Hamilton-Jacobi-Bellman equation (HJB)

The Hamilton-Jacobi-Bellman (HJB) equation provides a sufficient condition for optimality and is the continuous-time analogue of the Bellman equation. Consider the stochastic

optimal control problem:

$$\begin{aligned} V(t,x) &= \max_{u(\cdot)} \mathbb{E}_t \left[\int_t^\infty e^{-\rho(s-t)} f(s,x(s),u(s)) ds \right] \\ \text{s.t.} \quad dx(s) &= \mu(s,x(s),u(s))ds + \sigma(s,x(s),u(s))dZ_s, \\ u(s) &\in \mathcal{U} \quad \text{for all } s \geq t, \\ x(t) &= x. \end{aligned}$$

Here, $V(t,x)$ is the value function, representing the optimal value of the objective starting from state x at time t. The state variable $x(t) \in \mathbb{R}^n$, the control $u(t) \in \mathcal{U} \subset \mathbb{R}^s$, and Z_t is an m-dimensional standard Brownian motion.

The HJB equation is derived from the Bellman principle of optimality, which states that for a small time interval dt:

$$V(t,x) \approx \max_{u \in \mathcal{U}} \left\{ f(t,x,u)dt + e^{-\rho dt} \mathbb{E}_t[V(t+dt, x+dx)] \right\}.$$

Using the approximation $e^{-\rho dt} \approx 1 - \rho dt$ and applying Itô's lemma to expand $V(t+dt, x+dx)$, we get:

$$\mathbb{E}_t[dV] = \left(\frac{\partial V}{\partial t} + (\nabla_x V)^T \mu + \frac{1}{2} \text{tr}\left(\sigma \sigma^T \nabla_x^2 V \right) \right) dt,$$

where $\nabla_x V$ is the gradient of V with respect to x (an $n \times 1$ column vector) and $\nabla_x^2 V$ is the $n \times n$ Hessian matrix of second partial derivatives.

Substituting this into the Bellman equation, rearranging, dividing by dt, and taking the limit as $dt \to 0$ yields the HJB equation:

$$\rho V(t,x) = \max_{u \in \mathcal{U}} \left\{ f(t,x,u) + \frac{\partial V}{\partial t} + (\nabla_x V)^T \mu(t,x,u) + \frac{1}{2} \text{tr}\left(\sigma(t,x,u) \sigma(t,x,u)^T \nabla_x^2 V \right) \right\}. \tag{1.12.7}$$

An easy way to remember this is through the economic intuition: the required return on the value function (ρV) must equal the flow payoff (f) plus the expected rate of change of the value function ($E[dV]/dt$).

The expression inside the maximization is the Hamiltonian, though definitions can vary. The term inside the trace is the variance-covariance matrix of the stochastic process. Any path

constraints, like $G(x,u) \geq 0$, can be included in the maximization over u.

The Autonomous Case

In many economic models, the functions f, μ, and σ do not depend explicitly on time t. This is known as an autonomous problem. In this case, the value function V will also be independent of time, so $V(t,x) = V(x)$ and its partial derivative with respect to time is zero, $\frac{\partial V}{\partial t} = 0$. The HJB equation (1.12.7) then simplifies to the stationary HJB equation:

$$\rho V(x) = \max_{u \in \mathcal{U}} \left\{ f(x,u) + (\nabla_x V)^T \mu(x,u) + \frac{1}{2} \text{tr}\left(\sigma(x,u)\sigma(x,u)^T \nabla_x^2 V\right) \right\}.$$

This is a non-linear ordinary or partial differential equation for the value function $V(x)$.

Chapter 2

Mathematical Preliminaries

2.1 Probability Spaces, Sigma-Algebras, Filtrations

2.1.1 Discrete time

Times, states, events

Time is indexed by $t \in \{0, 1, \ldots, T\}$ for a positive integer T. One of a finite number of possible states of the world is realized by the terminal time T (either before T or exactly at T). We refer to the state that is realized by T as the true state or the revealed state. The possible states are represented by the elements of a finite set Ω, the state space. We are not yet concerned with the likelihood of any one state occurring, but every contingency represented by a state in Ω is possible and every relevant contingency is represented by a state in Ω. Subsets of Ω are called events.

Information flow, partitions

Time-t information is represented by a partition $\mathcal{F}_t^0$ of Ω. A partition is defined as a set of mutually exclusive nonempty events whose union is Ω. Thus, a partition is a collection of subsets of Ω such that each state belongs to exactly one element of $\mathcal{F}_t^0$. We use partitions to encode what is known at different times. All that is known at time t is what element of $\mathcal{F}_t^0$ contains the true state. We assume that at time 0, it is only known that the true state is an

element of Ω and therefore $\mathcal{F}_0^0 = \{\Omega\}$. At time T the true state is known with certainty and therefore $\mathcal{F}_T^0 = \{\{\omega\} \mid \omega \in \Omega\}$.

Example 2.1.0.1. Suppose Ω is the set of the eight possible outcomes of three coin tosses:

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\},$$

where we encode heads with H and tails with T. Consider the time-1 partition

$$\mathcal{F}_1^0 = \{A_H, A_T\},$$

with

$$A_H = \{HHH, HHT, HTH, HTT\},$$
$$A_T = \{THH, THT, TTH, TTT\}.$$

The partition $\mathcal{F}_1^0$ encodes that at $t = 1$ the true state is either in the set A_H, or in the set A_T. Thus, what is known at $t = 1$ is the outcome of the first coin toss. If the true state is in A_H, the first coin toss is H since all elements of A_H start with H. In addition, this is all that is known, since the outcomes of the last two coin tosses of the elements of A_H are $\{HH, HT, TH, TT\}$, which are all the possible outcomes of tossing the coin two times. By the same logic, if the true state is in A_T, the first coin toss is T.

Conversely, if it is known that the first coin toss is H, the true state is a state of the form HXY with $X, Y \in \{H, T\}$, so the true state is in A_H. Similarly, if the first coin toss is T, the true state is an element of A_T. □

Example 2.1.0.2. Information is generated by observing the outcome of a coin toss at each time $t = 1, 2, \ldots, \mathcal{T}$. A state is a finite sequence $\omega = (\omega_1, \ldots, \omega_{\mathcal{T}})$, where $\omega_t \in \{H, T\}$, and the state space is the Cartesian product $\Omega := \{H, T\}^{\mathcal{T}}$. At time $t > 0$, the first t coin toss outcomes $\bar{\omega}_1, \ldots, \bar{\omega}_t$ with $\bar{\omega}_i \in \{H, T\}$ have been observed and it is therefore known that the true state is an element of the event $\{\omega \in \Omega \mid \omega_1 = \bar{\omega}_1, \ldots, \omega_t = \bar{\omega}_t\}$. The partition $\mathcal{F}_t^0$ is the set of all these events as $(\bar{\omega}_1, \ldots, \bar{\omega}_t)$ ranges over $\{H, T\}^t$. □

As in Example 2.1.0.2, when we think of how information is revealed over time, we assume perfect recall. If at some time the true state is known to belong to a partition element, the

same remains true at all subsequent times. More formally, we assume that if $u > t$, the partition $\mathcal{F}_u^0$ is a refinement of the partition $\mathcal{F}_t^0$, meaning that every event in $\mathcal{F}_t^0$ is the union of events in $\mathcal{F}_u^0$.

Filtrations

Define the sets:

$$\mathcal{F}_t = \{F \mid F \text{ is a union of elements of } \mathcal{F}_t^0\}, \quad t = 0, \ldots, T. \tag{2.1.1}$$

An event F belongs to $\mathcal{F}_t$ if and only if at time t it is known whether F contains the state to be revealed at time T.

Example 2.1.0.3. Let $\Omega := \{1, 2, 3, 4\}$ and $\mathcal{F}_1^0 := \{\{1, 2\}, \{3\}, \{4\}\}$. Then

$$\mathcal{F}_1 = \{\emptyset, \{1, 2\}, \{3\}, \{4\}, \{1, 2, 3\}, \{1, 2, 4\}, \{3, 4\}, \Omega\}.$$

Suppose state 1 is realized at time $T := 2$. At time 1, it is known that the state is either 1 or 2. From that, it can be inferred whether every one of the events in $\mathcal{F}_1$ contains 1 or not, and these are all the events about which such a claim can be made. □

Every $\mathcal{F}_t$ is an algebra of events, meaning that it contains $\emptyset$ and Ω, and it is closed relative to the formation of Boolean set operations: For all $A, B \in \mathcal{F}_t$,

1. the union $A \cup B := \{\omega \mid \omega \in A \text{ or } \omega \in B\}$,
2. the intersection $A \cap B := \{\omega \mid \omega \in A \text{ and } \omega \in B\}$, and
3. the set difference $A \backslash B := \{\omega \mid \omega \in A \text{ and } \omega \notin B\}$,

are all elements of $\mathcal{F}_t$. In particular, for all $F \in \mathcal{F}_t$, the complement $F^c := \Omega \backslash F$ is an element of $\mathcal{F}_t$. This definition of an algebra is of course redundant. For example, since $A \cap B = (A^c \cup B^c)^c$ and $A \backslash B = A \cap B^c$, an algebra (of events) is any nonempty set of events that is closed with respect to the formation of unions and complements. Besides providing convenient notation, algebras are key in generalizing to infinite-dimensional state spaces in which algebras are not generated by partitions.

The intersection of an arbitrary collection of algebras is also an algebra. The union of two

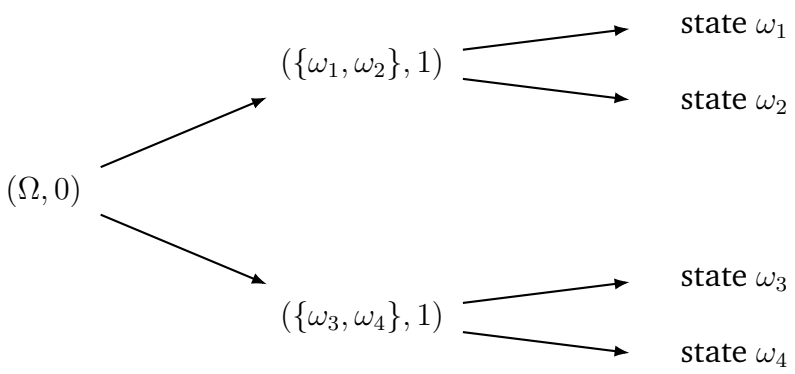

Figure 2.1: Information tree of Example 2.1.0.2 with $T = 2$. Each state ω_i can be identified with the terminal $\mathrm{spot}(\{\omega_i\}, 2)$.

algebras is not necessarily an algebra. The algebra $\sigma(\mathcal{S})$ generated by a set of events $\mathcal{S}$ is the intersection of all algebras that include $\mathcal{S}$. It is straightforward to verify that $\mathcal{F}_t = \sigma(\mathcal{F}_t^0)$ for all t. Conversely, $\mathcal{F}_t^0$ can be recovered from $\mathcal{F}_t$ as the set of nonempty elements of $\mathcal{F}_t$ that do not have a nonempty proper subset in $\mathcal{F}_t$.

Note that $\mathcal{F}_u^0$ is a refinement of $\mathcal{F}_t^0$ if and only if $\mathcal{F}_t \subseteq \mathcal{F}_u$. This motivates the definition of a filtration as a time-indexed sequence $\mathcal{F}_0, \mathcal{F}_1, \ldots, \mathcal{F}_T$ of algebras of events, abbreviated to $\{\mathcal{F}_t\}$, such that $u \geq t$ implies $\mathcal{F}_u \supseteq \mathcal{F}_t$. We can therefore specify the information primitive of our model as a filtration $\{\mathcal{F}_t\}$ satisfying

$$\mathcal{F}_0 = \{\emptyset, \Omega\} \quad \text{and} \quad \mathcal{F}_T = 2^\Omega \text{ (the set of all subsets of } \Omega), \tag{2.1.2}$$

rather than in terms of partitions.

Information tree

Figure 2.1, shows how a filtration $\{\mathcal{F}_t\}$ can be thought of as an information tree, whose nodes correspond to what we call "spots" or "nodes". Formally, a spot of the filtration $\{\mathcal{F}_t\}$ is a pair (F, t) where $F \in \mathcal{F}_t^0$ and $t \in \{0, \ldots, T\}$. The root of the information tree corresponds to the initial spot $(\Omega, 0)$. A terminal spot takes the form $(\{\omega\}, T)$, where $\omega \in \Omega$, and can therefore be identified with the state ω as well as the unique path on the information tree from the initial spot to the given terminal spot. A non-terminal spot $(F, t-1)$, $t \in \{1, \ldots, T\}$, has immediate successor spots $(F_0, t), \ldots, (F_d, t)$, where $F_0, \ldots, F_d$ are the elements of $\mathcal{F}_t^0$

whose union is F. The spot $(F, t-1)$ can be thought of as the set of paths on the information tree from the initial spot to every terminal spot corresponding to a state in F.

Optional

*2.1.2 Resolved sets

The above construction means that even if at some time t we do not precisely know what the true state is, we can make a list of sets that are sure to contain it and that are sure not to contain it. These are the sets that are resolved by the information modeled by the partition $\mathcal{F}_t^0$.

Example 2.2. Suppose Ω is the set of the eight possible outcomes of three coin tosses. If we are told the outcome of the first coin toss only, the sets

$$A_H = \{HHH, HHT, HTH, HTT\}, \quad A_T = \{THH, THT, TTH, TTT\}$$

are resolved. For each of these sets, once we are told the first coin toss, we know if the true state is a member. For example, if we are told the first toss is H, we know that the true event is in A_H and we also know that the true event is not in A_T.

If instead of knowing the outcome of the first coin toss we were told the outcome of the second coin toss, neither A_H nor A_T would be resolved, since they both contain events with H and with T in the second coin toss. □

The empty set $\emptyset$ and the whole space Ω are always resolved, even without any information; the true state never belongs to $\emptyset$ and always belongs to Ω.

Example 2.2. The four sets that are resolved by the first coin toss in Example 2.2 form the algebra

$$\mathcal{F}_1 = \{\emptyset, A_H, A_T, \Omega\}.$$

We think of this algebra as containing the information learned by observing the first coin toss. More precisely, if instead of being told the first coin toss we are told, for each set in $\mathcal{F}_1$, whether or not the true state belongs to the set, we know the outcome of the first coin toss and nothing more.

If we are told the first two coin tosses, we obtain a finer resolution. In particular, the four sets

$$A_{HH} = \{HHH, HHT\}, \qquad A_{HT} = \{HTH, HTT\},$$
$$A_{TH} = \{THH, THT\}, \qquad A_{TT} = \{TTH, TTT\},$$

are resolved. Of course, the sets in $\mathcal{F}_1$ are still resolved. Whenever a set is resolved, so is its complement, which means that A_{HH}^c, A_{HT}^c, A_{TH}^c, and A_{TT}^c are resolved. Whenever two sets are resolved, so is their union, which means that $A_{HH} \cup A_{TH}$, $A_{HH} \cup A_{TT}$, $A_{HT} \cup A_{TH}$, and $A_{HT} \cup A_{TT}$ are resolved. We have already noted that the two other pairwise unions, $A_H = A_{HH} \cup A_{HT}$ and $A_T = A_{TH} \cup A_{TT}$, are resolved. The triple unions are also resolved, and these are the complements already mentioned, e.g.,

$$A_{HH} \cup A_{HT} \cup A_{TH} = A_{TT}^c.$$

In all, we have 16 resolved sets that together form an algebra we call $\mathcal{F}_2$; i.e.,

$$\mathcal{F}_2 = \left\{ \begin{array}{l} \emptyset, A_H, A_T, A_{HH}, A_{HT}, A_{TH}, A_{TT}, A_{HH}^c, A_{HT}^c, A_{TH}^c, A_{TT}^c, \\ A_{HH} \cup A_{TH}, A_{HH} \cup A_{TT}, A_{HT} \cup A_{TH}, A_{HT} \cup A_{TT}, \Omega \end{array} \right\}. \tag{2.1.3}$$

We think of this algebra as containing the information learned by observing the first two coin tosses.

If we are told all three coin tosses, we know the true state and therefore every subset of Ω is resolved. There are 256 subsets of Ω and, taken all together, they constitute the algebra $\mathcal{F}_3$:

$$\mathcal{F}_3 = \text{ The set of all subsets of } \Omega.$$

If we are told nothing about the coin tosses, the only resolved sets are $\emptyset$ and Ω. We form the "trivial" algebra $\mathcal{F}_0$ with these two sets:

$$\mathcal{F}_0 = \{\emptyset, \Omega\}.$$

We have then four algebras, $\mathcal{F}_0, \mathcal{F}_1, \mathcal{F}_2$, and $\mathcal{F}_3$, indexed by time. As time moves forward, we obtain finer resolution. In other words, if $n < m$, then $\mathcal{F}_m$ contains every set in $\mathcal{F}_n$ and also other additional sets. This means that $\mathcal{F}_m$ contains more information than $\mathcal{F}_n$.

The collection of algebras $\mathcal{F}_0$, $\mathcal{F}_1$, $\mathcal{F}_2$, $\mathcal{F}_3$ is an example of a filtration. □

2.1.3 Continuous time

Sigma-algebras

To handle infinite-dimensional state spaces, we formalize the concept of an algebra of events through the concept of σ-algebras. Infinite dimensional state spaces can occur when time is discrete or continuous. Conversely, we can have finite state spaces in both continuous and discrete time. However, the continuous time models and techniques that we are interested in always require an infinite-dimensional state space, which is why we study them.

Definition 2.1.1. A σ-algebra (or σ-field) over a non-empty set Ω is a family $\mathcal{A}$ of subsets of Ω such that:

1. $\Omega \in \mathcal{A}$,
2. $A \in \mathcal{A}$ implies $A^c \in \mathcal{A}$
3. if $(A_n, n \geq 1)$ is a countable family of elements in $\mathcal{A}$, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$.

The definition of σ-algebra is applicable to both finite and infinite-dimensional state spaces.

Definition 2.1.2. Given a collection $\mathcal{A}$ of events, the σ-algebra generated by $\mathcal{A}$ is the intersection of all σ-algebras containing $\mathcal{A}$.

When the state space is finite dimensional, we found that the set $\mathcal{F}_t$ is generated by the partition $\mathcal{F}_t^0$. In fact, it can be shown that any finite σ-algebra can be generated by taking unions of partitions. However, when we move to infinite-dimensional state spaces, restricting ourselves to finite σ-algebras is very restrictive. For example, a finite σ-algebra can only encode the realization of finitely many values of a random outcome. If the random outcome can take values in the real numbers, a finite σ-algebra cannot encode the information contained in it.

Borel Sigma-Algebra

To work with infinite-dimensional state spaces, we have the following:

Definition 2.1.3. The Borel σ-algebra on a topological space X, denoted by $\mathcal{B}(X)$, is the σ-algebra generated by the open sets of X.

The elements of a Borel σ-algebra are called Borel sets. A Borel space is a pair (X, B), where B is the σ-algebra of Borel sets of X.

When we consider infinite-dimensional state spaces, we will only be concerned with the Borel σ-algebra on $\mathbb{R}^d$, $\mathcal{B}(\mathbb{R}^d)$. This σ-algebra contains all open sets, all closed sets, all countable unions of closed sets, all countable intersections of such countable unions, etc.

Probability Measures

We are now ready to discuss the likelihood of events. Roughly speaking, we want to assign a number between 0 and 1 —a probability— to each possible event.

Definition 2.1.4. A pair $(\Omega, \mathcal{F})$ is called a measurable space.

Definition 2.1.5. A probability measure on a measurable space $(\Omega, \mathcal{F})$ is a function $P : \mathcal{F} \to [0, 1]$ such that

1. $P(\emptyset) = 0, P(\Omega) = 1$,
2. for a countable set of disjoint events $A_i \in \mathcal{F}$,
$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

When $\mathcal{F}$ is finite, we can drop the word "measure" and just talk about probabilities. In uncountable probability spaces, we need to be more careful. For example, we would like to say that something that has probability zero cannot happen. Likewise, we may want to assert that if an event has probability one, it must happen with certainty. However, neither of the last two statements are necessarily true. For example, if $\mathcal{F}$ is the Borel σ-algebra in $\Omega = [0, 1]$ and we want all numbers in $[0, 1]$ to be "equally likely", then each number must have a probability measure of zero. If any one number has a probability measure that is

non-zero, then our requirement that they are all "equally likely" means that they all have non-zero probability measure. By the second property in Definition 2.1.5, the probability of the event $\{\omega \mid \omega \text{ is a rational number}\}$ is infinity since there are an infinite number of rational numbers. However, this violates the requirement that the probability measure is between 0 and 1. We conclude that even though drawing any number is possible and, in fact, all numbers are equally likely, the probability measure of any one number is zero.

Filtrations

The extension of filtrations to the continuous-time setting is straightforward.

Definition 2.1.6. Let $\mathcal{T}$ be an indexed set. A filtration $\mathbb{F}$ on Ω is a collection of σ-algebras $\mathcal{F}_t$, for $t \in \mathcal{T}$, such that

1. $\mathcal{F}_0 = \{\emptyset, \Omega\}$
2. $\forall s > t, \quad \mathcal{F}_t \subset \mathcal{F}_s$
3. $\mathcal{F}_T = \mathcal{F}$.

In discrete time $\mathcal{T}$ is a countable set: $\mathcal{T} = \{0, 1, 2, ..., T\}$ with T finite or infinite. In continuous time $\mathcal{T}$ is uncountable: $\mathcal{T} = [0, T]$ with T finite or infinite.

Definition 2.1.7. A filtered probability space is a probability space with a filtration $\mathbb{F}$.

2.2 Stochastic Processes

A random variable is a function of the form $x : \Omega \to \mathbb{R}^d$ with d a positive integer. A stochastic process, or simply a process, is a time-indexed sequence of random variables. More formally, a stochastic process is defined as a time-indexed sequence of random variables $(x_t, t \in \mathcal{T})$ or as a function of the form $x : \Omega \times \mathcal{T} \to \mathbb{R}^d$, where $x_t(\omega) = x(\omega, t)$ for $\omega \in \Omega$, $t \in \mathcal{T}$, and d a positive integer. Sometimes we write $x(t)$ instead of x_t. The function $x(\omega, \cdot) : \mathcal{T} \to \mathbb{R}^d$, for any fixed $\omega \in \Omega$, is a path of the process x.

Measurable Random Variables

Suppose information is represented by a given underlying filtration $\{\mathcal{F}_t\}$ on Ω satisfying (2.1.2). If a process is to represent an observed quantity, it cannot reveal more information than implied by the postulated filtration. For example, if $T = 1$ and $\Omega = \{0, 1\}$, the process $x_0(\omega) = x_1(\omega) = \omega$ is not consistent with the information structure, because observation of the realization of x_0 at time zero reveals the state ω. To formalize this type of informational constraint, we introduce the notion of measurability with respect to an algebra.

Definition 2.2.1. Let $(\Omega, \mathcal{F}, P, \mathbb{F})$ be a filtered probability space. A random variable $X : \Omega \to \mathbb{R}^d$ is measurable with respect to $\mathcal{F}$ or $\mathcal{F}$-measurable if and only if for any set $A \in \mathcal{B}(\Omega)$, the set

$$\{\omega \in \Omega : X(\omega) \in A\}$$

belongs to $\mathcal{F}$.

We denote by $L(\mathcal{F})$ (or simply L when clear) the set of random variables that are $\mathcal{F}$-measurable, by $L^1(\mathcal{F})$ (or L^1 when clear) the set of $\mathcal{F}$-measurable random variables that are also integrable, and by $L^2(\mathcal{F})$ (or L^2 when clear) the set of $\mathcal{F}$-measurable random variables whose square is integrable.

Adapted Processes

We can now formalize the requirement that a process respects the given information structure with the notion of adaptedness. The process x is said to be adapted to the underlying filtration $\{\mathcal{F}_t\}$ if $x_t \in L(\mathcal{F}_t)$ for every time t. Since $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $\mathcal{F}_T = 2^\Omega$, the initial value of an adapted process is constant, while its terminal value can be any $\mathcal{F}_T$-measurable random variable.

We denote the space of all processes adapted to the filtration $\{\mathcal{F}_t\}$ by $\mathcal{L}(\mathcal{F}_t)$ (or simply $\mathcal{L}$ when clear) and the set of all strictly positive processes adapted to the filtration $\{\mathcal{F}_t\}$ by $\mathcal{L}_{++}(\mathcal{F}_t)$ (or $\mathcal{L}_{++}$ when clear).

σ-algebra Generated by Random Variables

In applications, it is common to specify the filtration $\{\mathcal{F}_t\}$ as the information revealed by given processes representing observable quantities. To formally define this way of constructing the information stream, we first define a notion of information revealed by a given set of random variables.

The σ-algebra generated by a set S of random variables is the intersection of all σ-algebras relative to which every $x \in S$ is measurable, and is denoted by $\sigma(S)$. If $S = \{x_1, \ldots, x_n\}$, $\sigma(x_1, \ldots, x_n) := \sigma(S)$ is the same as the σ-algebra generated by the partition of all nonempty events of the form

$$\{(x_1, \ldots, x_n) \mid x_i = \alpha \text{ for all } i = 1, \ldots, n \text{ and } \alpha \in \mathbb{R}\}.$$

We interpret $\sigma(x_1, \ldots, x_n)$ as the information that can be inferred by observing the realization of the random variables $x_1, \ldots, x_n$. Any other variable whose realization is revealed by this information must be determined as a function of the realization of $(x_1, \ldots, x_n)$:

Proposition 2.2.1. *Given any random variables x_1, ..., x_n, a random variable y is $\sigma(x_1, \ldots, x_n)$-measurable if and only if there exists a function $f : \mathbb{R}^n \to \mathbb{R}$ such that $y = f(x_1, \ldots, x_n)$.*

Proof. That $f(x_1, \ldots, x_n)$ is $\sigma(x_1, \ldots, x_n)$-measurable follows from Proposition 1.1.3. Conversely, suppose y is $\sigma(x_1, \ldots, x_n)$-measurable and let $\{(x_1(\omega), \ldots, x_n(\omega)) \mid \omega \in \Omega\} = \{\alpha_1, \ldots, \alpha_m\} \subseteq \mathbb{R}^n$. The σ-algebra $\sigma(x_1, \ldots, x_n)$ is generated by the partition $\{A_1, \ldots, A_m\}$, where $A_i := \{(x_1, \ldots, x_n) = \alpha_i\}$. Let $y = \sum_{j=1}^m \beta_i 1_{A_i}$, where β_i is the constant value of y on A_i. Selecting any function $f : \mathbb{R}^n \to \mathbb{R}$ such that $\beta_i = f(\alpha_i)$ results in $y = f(x_1, \ldots, x_n)$. □

Filtration Generated by Stochastic Processes

A filtration is said to be generated by the processes $B_t^1, \ldots, B_t^d$, where each time-0 value B_0^i is a constant, if for all t, $\mathcal{F}_t = \sigma(\{B_s^1, \ldots, B_s^d \mid s \leq t\})$. Writing $B = (B^1, \ldots, B^d)$, it follows from Proposition 2.2.1 that in this case a process x is adapted if and only if x_0 is constant

and for every time $t > 0$, there exists a function $f(t, \cdot) : \mathbb{R}^{d \times t} \to \mathbb{R}$ such that

$$x(\omega, t) = f(t, B(\omega, 1), \ldots, B(\omega, t)), \quad \omega \in \Omega.$$

In other words, x_t is a function of the path of B up to time t.

*2.2.1 Stopping Time

Indicator Function

We define the indicator function 1_A of a set A as the function that takes the value 1 on A and 0 on the complement of A in the implied domain. For example, if A is an event, then 1_A is the random variable

$$1_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A \\ 0, & \text{if } \omega \notin A \end{cases}$$

If $A \subseteq \Omega \times \{0, \ldots, T\}$, then 1_A denotes the process defined as above, but with (ω, t) in place of ω.

Stopping Time

Related to the notion of an adapted process is that of a stopping time, defined as a function of the form $\tau : \Omega \to \mathcal{T} \cup \{\infty\}$ for some time index $\mathcal{T}$, provided that $\{\tau \leq t\} \in \mathcal{F}_t$ for every time t. The last restriction is equivalent to the adaptedness of the indicator process $1_{\{\tau \leq t\}}$, which takes the value zero prior to the (random) time τ, and the value one from time τ on (which on the event $\{\tau = \infty\}$ is never). A stopping time, or corresponding indicator process, announces the (first) arrival of an event which is consistent with the information stream encoded in the underlying filtration. For example, if x is an adapted process, then the first time that $x_t \geq 1$ defines a stopping time (with the value ∞ being assigned on the event that x always remains below one). On the other hand, the first time that x reaches its path maximum is not generally a stopping time. For any process x and stopping time τ, the random variable x_τ or $x(\tau)$ is defined by letting $x_\tau(\omega) = x(\omega, \tau(\omega))$, with the convention $x(\omega, \infty) = 0$.

2.3 An Informal Introduction to Stochastic Calculus

Brownian motion is a continuous-time scalar stochastic process such that, given the initial value x_0 at time $t = 0$, the random variable x_t for any $t > 0$ is normally distributed with mean $x_0 + \mu_t$ and variance σ_t^2. The parameter μ measures the trend, and σ the volatility, of the process. This process was first formulated to represent the motion of small particles suspended in a liquid. We shall sometimes refer to a "particle" performing the Brownian motion, x_t as its "position", and a graph of x_t against t as its "path".

We can think of Brownian motion as the cumulation of independent identically distributed normal increments, the infinitesimal random increment dx over the infinitesimal time dt having mean μdt and variance $\sigma^2 dt$. Just as we would write a general normal (μ, σ) variable as $\mu + \sigma w$ where w is a standard normal random variable of zero mean and unit variance, we can write

$$dx = \mu dt + \sigma dw \tag{2.3.1}$$

where w is a standardized Brownian motion (Wiener process) whose increment dw has zero mean and variance dt. This is the usual shorthand notation for Brownian motion. Figure 2.2 gives a visual summary of the last two paragraphs.

2.3.1 Random Walk Representation

We approximate Brownian motion by a discrete random walk. Then the normal distribution arises as the limit of a sum of independent binary variables Δx over discrete time intervals Δt, when these go to zero in a particular way.

Divide time into discrete periods of length Δt, and let space consist of discrete points along a line, Δh being the step-length or the distance between successive points. Let Δx be a random variable that follows a random walk: in one time period it moves up one step in space with probability p, and one step down with probability $q = 1 - p$. Note: Δh is a given positive number, and Δx is a random variable that takes values $\pm \Delta h$. Figure 2.3 shows all

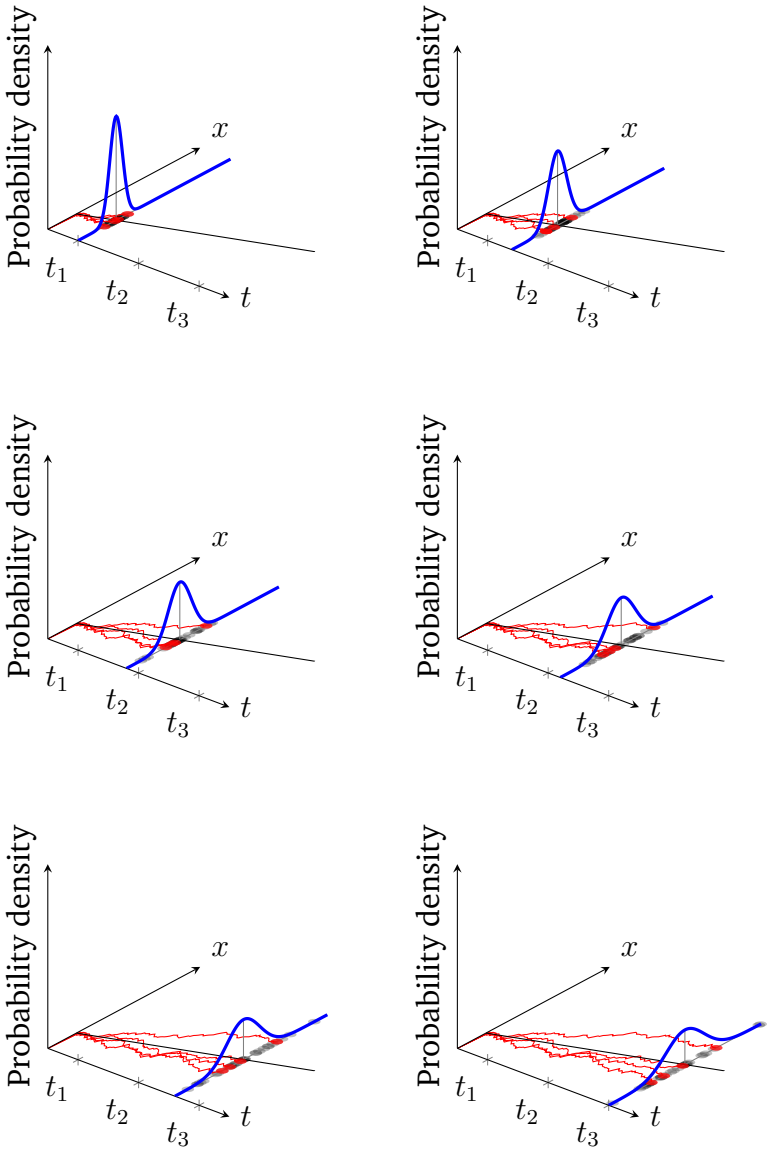

Figure 2.2: Paths of Brownian motion x_t are plotted in red. The different panels show the evolution of paths and of the normal distribution of x_t in blue as time goes by. The mean of the distribution is on black straight line. The variance of the distribution increases with t.

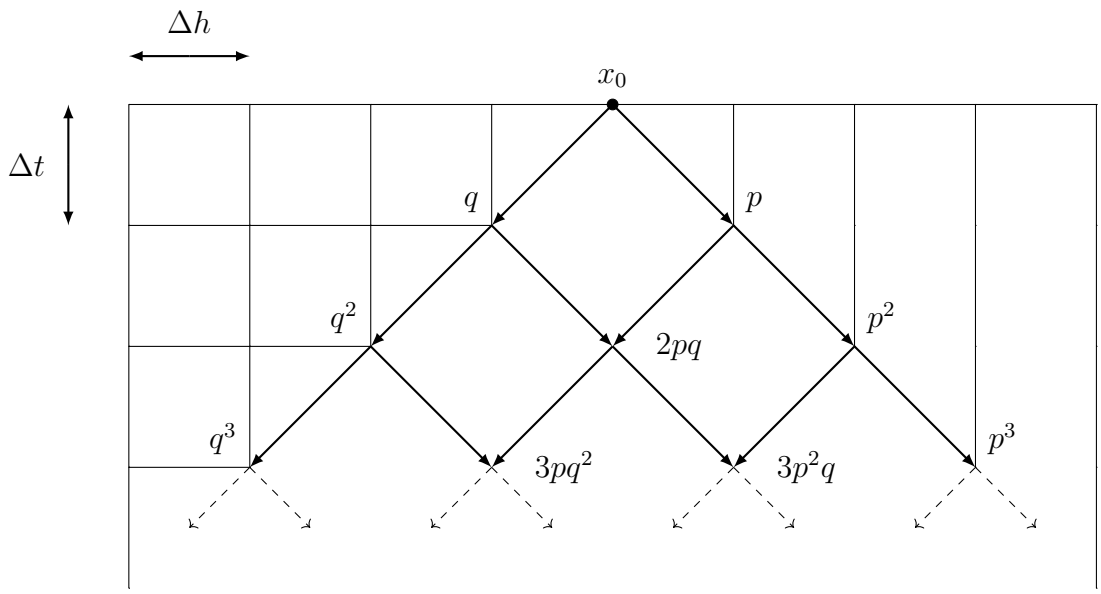

Figure 2.3: Random walk representation.

this compactly. We see various possible paths, with time marching downward and position shown horizontally. At each point in time and space, the probability of reaching it is also shown.

The mean of Δx is

$$E[\Delta x] = p\Delta h + q(-\Delta h) = (p-q)\Delta h. \tag{2.3.2}$$

Also,

$$E\left[(\Delta x)^2\right] = p(\Delta h)^2 + q(-\Delta h)^2 = (\Delta h)^2,$$

so the variance of Δx is

$$\begin{aligned} \mathrm{Var}[\Delta x] &= E\left[(\Delta x)^2\right] - (E[\Delta x])^2 && (2.3.3)\\ &= \left[1-(p-q)^2\right](\Delta h)^2 = 4pq(\Delta h)^2 && (2.3.4)\end{aligned}$$

A time interval of length t has $n = t/\Delta t$ such discrete steps. Since the successive steps of the random walk are independent, the cumulated change $x_t - x_0$ is a binomial random variable with mean

$$n(p-q)\Delta h = t(p-q)\Delta h/\Delta t$$

and variance

$$4npq(\Delta h)^2 = 4tpq(\Delta h)^2/\Delta t$$

These expressions are minor modifications of the very familiar and elementary binomial distribution. There a 'success' in any one trial counts as 1 and occurs with probability p, while a failure counts as 0 and occurs with probability $q = 1 - p$. The (random) number of successes in n independent trials has expectation np and variance npq. The expressions above are perfectly analogous. Now success counts as Δh and failure as $-\Delta h$; for example the variance is $4(\Delta h)^2$ times that of the usual binomial expression.

Now set

$$\Delta h = \sigma\sqrt{\Delta t} \tag{2.3.5}$$

and

$$p = \frac{1}{2}\left[1 + \frac{\mu}{\sigma}\sqrt{\Delta t}\right], \quad q = \frac{1}{2}\left[1 - \frac{\mu}{\sigma}\sqrt{\Delta t}\right] \tag{2.3.6}$$

or

$$p = \frac{1}{2}\left[1 + \frac{\mu}{\sigma^2}\Delta h\right], \quad q = \frac{1}{2}\left[1 - \frac{\mu}{\sigma^2}\Delta h\right]. \tag{2.3.7}$$

Then

$$4pq = 1 - \left(\frac{\mu}{\sigma}\right)^2 \Delta t.$$

Substitute these into the above expressions, and let Δt go to zero. For given t, the number of steps goes to infinity. Then the binomial distribution converges to the normal, with mean

$$t\frac{\mu}{\sigma^2}\Delta h\frac{\Delta h}{\Delta t} = \mu t$$

and variance

$$t\left[1 - \left(\frac{\mu}{\sigma}\right)^2 \Delta t\right]\frac{\sigma^2\Delta t}{\Delta t} \to \sigma^2 t$$

These are exactly the values we need for Brownian motion. Thus we can regard Brownian motion as the limit of the random walk, when the time interval and the space step-length go to zero together, while preserving the relation (2.3.5) between them. Figure 2.4 shows

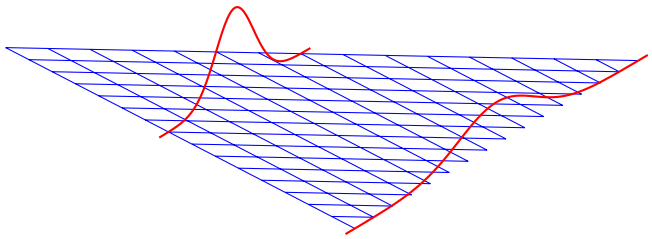

Figure 2.4: Binomial approximation of normal distribution.

the binomial process from Figure 2.3 after many steps and the normal distribution that it approximates at two different times.

The mean of $x_t - x_0$ is μt, and its standard deviation is $\sigma\sqrt{t}$. For large t, we have $\sqrt{t} \ll t$; in the long run, the trend is the dominant determinant of Brownian motion. But for small t, we have $t \ll \sqrt{t}$, so volatility dominates in the short run.

Another manifestation of this volatility is seen by calculating the expected length of a path. We have

$$E(|\Delta x|) = \Delta h,$$

so the total expected length of the path over the time interval from 0 to t is

$$t\Delta h/\Delta t = t\sigma/\sqrt{\Delta t} \to \infty$$

as Δt goes to zero. For small but finite Δt, the total length of almost all sample paths is very large. Therefore each path must have many ups and downs and look very jagged. Most such sample paths are not differentiable. When discussing the expected rate of change, therefore, we must write $E[dx]/dt$, not $E[dx/dt]$.

2.3.2 A First Pass at Ito's Lemma

Suppose x follows Brownian motion with parameters μ, σ. Consider a stochastic process y that is related to x by $y = f(x)$ where f is a given non-random function. We want to relate changes in y to those in x. The rules of conventional calculus suggest writing $dy = f'(x)dx$. But this turns out to be wrong. Starting at $y_0 = f(x_0)$, consider the position a small amount

of time t later.

$$y_t - y_0 = f'(x_0)(x_t - x_0) + \frac{1}{2}f''(x_0)(x_t - x_0)^2 + \dots$$

Hence

$$\begin{aligned} E[y_t - y_0] &= f'(x_0)E[x_t - x_0] + \frac{1}{2}f''(x_0)E\left[(x_t - x_0)^2\right] + \dots \\ &= f'(x_0)\mu t + \frac{1}{2}f''(x_0)\left[\sigma^2 t + \mu^2 t^2\right] + \dots \\ &= \left[\mu f'(x_0) + \frac{1}{2}\sigma^2 f''(x_0)\right]t + \dots, \end{aligned}$$

where in each case the dots represent terms in higher powers of t that can be ignored when t is small. But note that the second order term in the Taylor expansion of $f(x)$ contributes a term that is of the first order in t. The reason is that the variance of the increments of x is linear in t. This is the feature that makes the calculus of Brownian motion so different from the usual calculus of non-random variables.

A similar calculation will show that

$$\mathrm{Var}[y_t - y_0] = f'(x_0)^2\sigma^2 t + \dots$$

Let x denote a general starting position and $y = f(x)$. Consider the infinitesimal increment dy over the next infinitesimal time interval dt. We can use the above expressions replacing t bt dt and ignoring higher order terms in dt. Therefore dy has mean

$$E[dy] = \left[f'(x)\mu + \frac{1}{2}f''(x)\sigma^2\right]dt$$

and variance

$$\mathrm{Var}[dy] = f'(x)^2\sigma^2 dt$$

So y follows the general process defined by

$$dy = \left[f'(x)\mu + \frac{1}{2}f''(x)\sigma^2\right]dt + f'(x)\sigma dw. \tag{2.3.8}$$

This is Ito's Lemma for the special case we are considering. A slight generalization is easily

available: if $y = f(x,t)$, the Taylor expansion has an addition term in f_t, and

$$dy = \left[f_x(x,t)\mu + \frac{1}{2} f_{xx}(x,t)\sigma^2 + f_t(x,t) \right] dt + f_x(x,t)\sigma dw \tag{2.3.9}$$

For a simple intuition, return to the discrete random walk formulation, and suppose x has zero trend, so $p = q = 1/2$. Now $E[\Delta x] = 0$, and as time passes the distribution of x merely spreads out with linearly increasing variance around an unchanging mean. From the standard intuition of risk-aversion, or from Jensen's inequality, we know that the sign of $E[\Delta y]$ depends on the curvature of the function f. A risk-averse person will dislike the increase in risk, and a risk-lover will like it. Therefore $E[\Delta y]$ will be negative if f is concave (f'' is negative) and positive if f is convex (f'' is positive). We have

$$\begin{aligned} E[\Delta y] =& \frac{1}{2} f(x + \Delta h) + \frac{1}{2} f(x - \Delta h) - f(x) \\ =& \frac{1}{2} \left[f'(x)\Delta h + \frac{1}{2} f''(x)(\Delta h)^2 + \ldots \right] \\ & + \frac{1}{2} \left[-f'(x)\Delta h + \frac{1}{2} f''(x)(\Delta h)^2 + \ldots \right] \\ =& \frac{1}{2} f''(x)(\Delta h)^2 + \ldots = \frac{1}{2} f''(x)\sigma^2 \Delta t + \ldots \end{aligned}$$

Regarding Δt as an infinitesimal dt, this is exactly the same as the additional term in (2.3.8) that ordinary calculus would not have led us to expect. Thus Ito's Lemma is basically a consequence of Jensen's inequality when we take into account the particular relation between the space steps and the time intervals of Brownian motion. Readers can now do the slightly messier case where $\mu \neq 0$ and get an exact correspondence with (2.3.8).

2.3.3 Geometric Brownian Motion

Now suppose x follows the Brownian motion (2.3.1), and let $X = e^x$. Ito's Lemma gives

$$E[dX] = \left(e^x \mu + \frac{1}{2} e^x \sigma^2 \right) dt = X \left(\mu + \frac{1}{2} \sigma^2 \right) dt$$

and

$$\mathrm{Var}[dX] = \left(e^x\right)^2 \sigma^2 dt = X^2 \sigma^2 dt$$

Therefore the process of X can be written

$$dX/X = \left(\mu + \frac{1}{2}\sigma^2\right) dt + \sigma dw.$$

This is called a geometric Brownian motion or a proportional Brownian motion. It is particularly useful in economics because it is always a positive process (if X_0 is positive), so it provides a good way to model asset prices. When we need to emphasize the difference between this geometric Brownian motion and the "standard" Brownian motion (2.3.1) we refer to the latter as arithmetic Brownian motion or absolute Brownian motion.

If X follows the geometric Brownian motion

$$dX/X = \nu dt + \sigma dz \tag{2.3.10}$$

then using Ito's Lemma we find that $x = \ln X$ follows the ordinary Brownian motion

$$dx = \left(\nu - \frac{1}{2}\sigma^2\right) dt + \sigma dw$$

Note that when $x = \ln X$, the trend coefficients on the right hand sides of the expressions (2.3.1) for dx and (2.3.10) for dX/X differ by $\frac{1}{2}\sigma^2$. Thus $d\ln X \neq dX/X$; this is the Jensen-Ito effect discussed above. The logarithm is a concave function, and therefore $d\ln X < dX/X$, and a calculation shows that $\frac{1}{2}\sigma^2 dt$ is just the right difference.

Suppose X follows the geometric Brownian motion (2.3.10), starting at $t = 0$ in the known position X_0. Let $x = \ln X$ and $x_0 = \ln X_0$. We know that x follows the arithmetic Brownian motion (2.3.1) with $\mu = \nu - \frac{1}{2}\sigma^2$. Then at any positive time t, $x_t = \ln X_t$ is normally distributed with mean $x_0 + \mu t$ and variance $\sigma^2 t$. In other words, X_t has a lognormal distribution with parameters $x_0 + \mu t$ and $\sigma\sqrt{t}$. For such a distribution, it can be shown that

$$E[X_t] = \exp\left(x_0 + \mu t + \frac{1}{2}\sigma^2 t\right) = X_0 e^{\nu t}$$

This is a special case of the formula for a general exponential. We see once again the Jensen-Ito effect at work; the exponential is a convex function, therefore

$$E[X_t] = E[\exp x] > \exp E[x_t] = \exp(x_0 + \mu t)$$

You can get further practice by finding the stochastic process for $1/X$ when X follows the

geometric Brownian motion (2.3.10).

2.3.4 Some Generalizations

An obvious generalization of the Brownian motion (2.3.1) is obtained by letting the trend coefficient μ and the volatility coefficient σ depend on the current state x and also on time; thus

$$dx = \mu(x,t)dt + \sigma(x,t)dw. \tag{2.3.11}$$

A process whose trend and volatility coefficients are functions of the current state is often called a diffusion process; when they are functions of time as well, it is sometimes called an Ito process.

The geometric Brownian motion is an important special case. We can cast (2.3.10) in the diffusion process from (2.3.11) by writing

$$dX = \mu X dt + \sigma X dw$$

so the coefficients are both proportional to the current state.

In some economic applications, we need processes that revert toward some central level $\bar{x}$ of the state variable. For example, there may be an equilibrating force on prices. Now the trend coefficient has a sign opposite to that of $x - \bar{x}$. When the mean reversion is linear, we have

$$dx = -\theta(x - \bar{x})dt + \sigma dw \tag{2.3.12}$$

where θ is some positive constant. The process in (2.3.12) is the continuous-time analog of the discrete-time $AR(1)$ process.

Finally, several variables x_i for $i = 1, 2, \ldots m$, may follow Brownian motion, their volatility components being linear combinations of independent standard Wiener processes w_j for $j = 1, 2, \ldots n$:

$$dx_i = \mu_i dt + \sum_{j=1}^{n} a_{ij} dw_j$$

Then

$$E\left[dx_i\right] = \mu_i dt$$

and

$$\mathrm{Var}\left[dx_i\right] = \sum_{j=1}^{n} \left(a_{ij}\right)^2 dt, \quad \mathrm{Cov}\left[dx_i, dx_k\right] = \sum_{j=1}^{n} a_{ij} a_{kj} dt.$$

This allows quite general correlation between the different dx_i.

2.4 A More Formal Introduction to Stochastic Calculus

We now define Brownian motion, Ito integrals, and Ito processes more formally. This will allow us to develop further intuition for why we need them and how they work.

2.4.1 Brownian Motion

We now give a definition of Brownian motion.

Definition 2.4.1. A Brownian motion is a stochastic process Z in $\mathbb{R}^d$ such that

1. $Z_0 = 0$ a.s.,
2. for all s and t, $Z_s - Z_t$ is normal with mean 0 and covariance matrix $(s-t)I$,
3. for all $0 \leq t_0 < t_1 < .. < t_n \leq T$, the random variables $Z_{t_0}, Z_{t_1} - Z_{t_0}, .., Z_{t_n} - Z_{t_{n-1}}$, are independent.

Brownian motion is the basic building block for continuous stochastic processes. The independence property in the last item of the definition makes Brownian motion a Markov process.

The standard filtration $\mathbb{F}$ of a Brownian motion Z is defined as follows. Consider the σ-algebra $\mathcal{F}_t^Z$ generated by the sets

$$\{\omega \in \Omega : Z(\omega, s) \in A\},$$

where A is a set in the Borel σ-algebra on $\mathbb{R}^d$, and $s \leq t$. This is the σ-algebra generated by the random variables Z_s for $s \leq t$, and contains all the history of the Brownian motion up to

time t. Consider next the σ-algebra $\mathcal{F}_t$ generated by the sets in $\mathcal{F}_t^Z$ and the subsets of all zero probability events in $\mathcal{F}$. The filtration $\mathbb{F}$ is the collection of the σ-algebras $\mathcal{F}_t$ for $t \in \mathcal{T}$. We will use Brownian filtration to model information flow in financial markets.

Definition 2.4.2 (Martingale)**.** Consider an adapted stochastic process X such that X_t is integrable for all t. X is a martingale iff $X_t = E(X_s \mid \mathcal{F}_t)$ for all $s > t$.

A Brownian motion is a martingale w.r.t. its standard filtration. Indeed, for $s > t$

$$E(Z_s \mid \mathcal{F}_t) = E(Z_t + (Z_s - Z_t) \mid \mathcal{F}_t) = Z_t + E(Z_s - Z_t \mid \mathcal{F}_t) = Z_t$$

We now formalize the idea mentioned earlier that the "total expected length" of the path of a Brownian motion is infinite. We denote a partition $0 = t_0 < t_1 < .. < t_n = T$ of $[0, T]$ by Π, and the set of all partitions by $\mathcal{P}$.

Definition 2.4.3 (Variation of a function)**.** The variation of a function $f : \mathcal{T} \to \mathbb{R}$ is

$$V(f) = \sup_{\Pi \in \mathcal{P}} \sum_i \left| f\left(t_{i+1}\right) - f\left(t_i\right) \right|.$$

The sample paths of a Brownian motion have infinite total variation, almost surely (with probability 1). This implies that the paths are nowhere differentiable (with respect to t). This formalizes the idea that each path of the Brownian motion "has many ups and downs and look very jagged" mentioned earlier.

Even though Brownian motion is nowhere differentiable and has unbounded total variation, it turns out that it has bounded *quadratic* variation. This observation is the cornerstone of the stochastic Ito integral we study below.

The length of a partition is defined by

$$\ell(\Pi) := \max_i \left| t_{i+1} - t_i \right|.$$

Definition 2.4.4 (Quadratic variation of Brownian motion)**.** A quadratic variation of a one-dimensional Brownian motion process W on $[0, T]$ is

$$[W, W]_T := \lim_{\ell(\Pi) \to 0} \sum_i \left| W(t_{i+1}) - W(t_i) \right|^2$$

where the limit is taken in probability over all $\Pi \in \mathcal{P}$.

We can explicitly compute the quadratic variation of a one-dimensional Brownian motion as $[Z]_T = T$, so it is finite.

2.4.2 Ito's Integral for Simple Integrands

Let $W(t)$ for $t \geq 0$ be a Brownian motion and $\mathcal{F}(t)$ the filtration generated by it. We fix a positive number T and seek to make sense of

$$\int_0^T \Delta(t) dW(t) \tag{2.4.1}$$

where $\Delta(t)$ is an adapted stochastic process. Anything that depends on the path of a random process is itself random. Requiring $\Delta(t)$ to be adapted means that we require $\Delta(t)$ to be $\mathcal{F}(t)$-measurable for each $t \geq 0$. In other words, the information available at time t is sufficient to evaluate $\Delta(t)$ at that time. When we are standing at time 0 and t is strictly positive, $\Delta(t)$ is unknown to us. It is a random variable. When we get to time t, we have sufficient information to evaluate $\Delta(t)$; its randomness has been resolved. We focus on adapted processes because, in economic applications, $\Delta(t)$ is usually be a decision variable — such as the position we take in an asset at time t — or a state variable — such as inflation — that are known at t.

Increments of the Brownian motion after time t are independent of $\mathcal{F}(t)$, and since $\Delta(t)$ is $\mathcal{F}(t)$-measurable, it must also be independent of these future Brownian increments.

The problem we face when trying to assign meaning to the Ito integral (2.4.1) is that Brownian motion paths cannot be differentiated with respect to time. If $g(t)$ is a differentiable function, then we can define

$$\int_0^T \Delta(t) dg(t) = \int_0^T \Delta(t) g'(t) dt$$

where the right-hand side is an ordinary (Lebesgue) integral with respect to time. This will not work for Brownian motion.

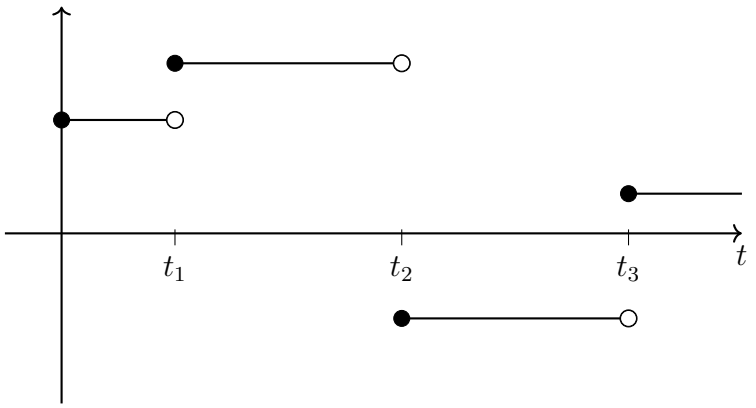

Figure 2.5: A path of a simple process.

2.4.3 Construction of the Integral

To define the integral (2.4.1), Ito devised the following way around the nondifferentiability of the Brownian paths. We first define the Ito integral for simple integrands $\Delta(t)$ and then extend it to nonsimple integrands as a limit of the integral of simple integrands. We describe this procedure.

Let $\Pi = \{t_0, t_1, \ldots, t_n\}$ be a partition of $[0, T]$. Assume that $\Delta(t)$ is constant in t on each subinterval $[t_j, t_{j+1})$. Such a process $\Delta(t)$ is a simple process.

Figure 2.5 shows a single path of a simple process $\Delta(t)$. We shall always choose these simple processes, as shown in this figure, to take a value at a partition time t_j and then remain constant up to, but not including, the next partition time t_{j+1}. Although it is not apparent from Figure 2.5, the path shown depends on the same ω on which the path of the Brownian motion $W(t)$ depends (neither ω nor $W(t)$ are shown in the figure). If one were to choose a different ω, there would be a different path of the Brownian motion and possibly a different path of $\Delta(t)$. However, the value of $\Delta(t)$ can depend only on the information available at time t. Since there is no information at time 0, the value of $\Delta(0)$ must be the same for all paths, and hence the first piece of $\Delta(t)$, for $0 \leq t < t_1$, does not really depend on ω. The value of $\Delta(t)$ on the second interval, $[t_1, t_2)$, can depend on observations made during the first time interval $[0, t_1)$.

The Ito integral of a simple process $\Delta(t)$ is defined by:

$$I(t) := \sum_{j=0}^{k-1} \Delta(t_j)[W(t_{j+1}) - W(t_j)] + \Delta(t_k)[W(t) - W(t_k)] \tag{2.4.2}$$

which we write as

$$I(t) = \int_0^t \Delta(u)dW(u).$$

In particular, we can take $t = t_n = T$, and (2.4.2) provides a definition for the Ito integral (2.4.1). We have managed to define this integral not only for the upper limit of integration T but also for every upper limit of integration t between 0 and T.

2.4.4 Properties of the Integral

The Ito integral (2.4.2) is defined over the martingale $W(t)$. A martingale has no tendency to rise or fall, and hence it is to be expected that $I(t)$, thought of as a process in its upper limit of integration t, also has no tendency to rise or fall. We formalize this observation by the next theorem.

Theorem 2.4.1. *The Ito integral defined by (2.4.2) is a martingale.*

Proof. The proof is in Section *2.6.1. □

Because $I(t)$ is a martingale and $I(0) = 0$, we have $\mathbb{E}I(t) = 0$ for all $t \geq 0$. It follows that $\mathrm{Var}(I(t)) = \mathbb{E}I^2(t)$, a quantity that can be evaluated by the formula in the next theorem.

Theorem 2.4.2 (Ito isometry)**.** *The Ito integral defined by (2.4.2) satisfies*

$$\mathbb{E}I^2(t) = \mathbb{E}\int_0^t \Delta^2(u)du.$$

Proof. The proof is in Section *2.6.1. □

We turn to the quadratic variation of the Ito integral $I(t)$.

Theorem 2.4.3. *The quadratic variation accumulated up to time t by the Ito integral (2.4.2) is*

$$[I, I](t) = \int_0^t \Delta^2(u)du. \tag{2.4.3}$$

Proof. The proof is in Section *2.6.1. □

In Theorems 2.4.2 and 2.4.3, we finally see how the quadratic variation and the variance of a process can differ. The quadratic variation is computed path-by-path, and the result can depend on the path. If along one path of the Brownian motion we have large values of $\Delta(u)$, the Ito integral will have a large quadratic variation. Along a different path, we could have small positions $\Delta(u)$ and the Ito integral would have a small quadratic variation. The quadratic variation can be regarded as a measure of risk, and it depends on the size of the positions we take. The variance of $I(t)$ is an average over all possible paths of the quadratic variation. Because it is the expectation of something, it cannot be random. We emphasize here that what we are calling variance is not the empirical variance. Empirical (or sample) variance is computed from a realized path and is an estimator of the theoretical variance we are discussing. The empirical variance is sometimes carelessly called variance, which creates the possibility of confusion.

The Ito integral formula $I(t) = \int_0^t \Delta(u)dW(u)$ can be written in differential form as $dI(t) = \Delta(t)dW(t)$. Similarly, the quadratic variation formula $[W, W](t) = t$, can be written in differential form as $dW(t)dW(t) = dt$. Thus, at least informally, we can square $dI(t)$:

$$dI(t)dI(t) = \Delta^2(t)dW(t)dW(t) = \Delta^2(t)dt \tag{2.4.4}$$

This equation says that the Ito integral $I(t)$ accumulates quadratic variation at rate $\Delta^2(t)$ per unit time.

Remark 2.4.1. Equations

$$I(t) = \int_0^t \Delta(u)dW(u) \tag{2.4.5}$$

and

$$dI(t) = \Delta(t)dW(t) \tag{2.4.6}$$

mean almost the same thing. The first equation has the precise meaning given by (2.4.2). The second equation has the imprecise meaning that when we move forward a little bit in time from time t, the change in the Ito integral $I(t)$ is $\Delta(t)$ times the change in the Brownian

motion $dW(t)$. It also has a precise meaning, which one obtains by integrating both sides, remembering to put in a constant of integration $I(0)$:

$$I(t) = I(0) + \int_0^t \Delta(u)dW(u) \tag{2.4.7}$$

We say that the second is the differential form of the third and that the third is the integral form of the second. These two equations mean exactly the same thing.

The only difference between the second and third equations, and hence the only difference between the first and second, is that the first specifies the initial condition $I(0) = 0$, whereas (2.4.6) and (2.4.7) permit $I(0)$ to be any arbitrary constant.

2.4.5 Ito's Integral for More General Integrands

In this section, we define the Ito integral $\int_0^T \Delta(t)dW(t)$ for integrands $\Delta(t)$ that are allowed to vary continuously with time and also to jump. In particular, we no longer assume that $\Delta(t)$ is a simple process.

We do assume that $\Delta(t), t \geq 0$, is adapted to the filtration $\mathcal{F}(t), t \geq 0$. We also assume the mean-square-integrability condition

$$\mathbb{E}\int_0^T \Delta^2(t)dt < \infty \tag{2.4.8}$$

In order to define $\int_0^T \Delta(t)dW(t)$, we approximate $\Delta(t)$ by simple processes. Figure 2.6 suggests how this can be done. In that figure, the continuously varying $\Delta(t)$ is shown as a solid red line and the approximating simple integrand is dashed. Notice that $\Delta(t)$ is allowed to jump. The approximating simple integrand is constructed by choosing a partition $0 = t_0 < t_1 < t_2 < t_3 < t_4 < t_5$, setting the approximating simple process equal to $\Delta(t_j)$ at each t_j, and then holding the simple process constant over the subinterval $[t_j, t_{j+1})$. As the maximal step size of the partition approaches zero, the approximating integrand will become a better and better approximation of the continuously varying one.

In general, then, it is possible to choose a sequence $\Delta_n(t)$ of simple processes such that as $n \to \infty$ these processes converge to the continuously varying $\Delta(t)$. By "converge", we mean

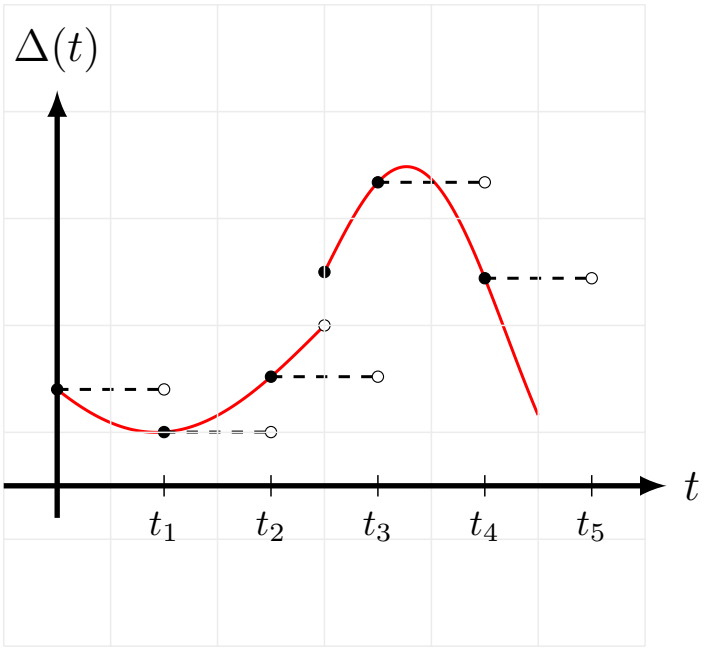

Figure 2.6: Approximating a continuously varying integrand.

that

$$\lim_{n\to\infty} \mathbb{E}\int_0^T |\Delta_n(t) - \Delta(t)|^2\, dt = 0.$$

For each $\Delta_n(t)$, the Ito integral $\int_0^t \Delta_n(u)dW(u)$ has already been defined for $0 \le t \le T$. We define the Ito integral for the continuously varying integrand $\Delta(t)$ by the formula

$$\int_0^t \Delta(u)dW(u) = \lim_{n\to\infty} \int_0^t \Delta_n(u)dW(u), \quad 0 \le t \le T$$

This integral inherits the properties of Ito integrals of simple processes:

(i) (Continuity) As a function of the upper limit of integration t, the paths of $I(t)$ are continuous.

(ii) (Adaptivity) For each t, $I(t)$ is $\mathcal{F}(t)$-measurable.

(iii) (Linearity) If $I(t) = \int_0^t \Delta(u)dW(u)$ and $J(t) = \int_0^t \Gamma(u)dW(u)$, then $I(t) \pm J(t) = \int_0^t (\Delta(u) \pm \Gamma(u))dW(u)$; furthermore, for every constant c, $cI(t) = \int_0^t c\Delta(u)dW(u)$.

(iv) (Martingale) $I(t)$ is a martingale.

(v) (Ito isometry) $\mathbb{E}I^2(t) = \mathbb{E}\int_0^t \Delta^2(u)du$.

(vi) (Quadratic variation) $[I, I](t) = \int_0^t \Delta^2(u)du$.

2.4.6 Square Integrable Functions

We have defined the Ito integral $\int_0^T \Delta^2(t)dW(t)$ under the square-integrability condition. It turns out that many economically interesting trading strategies do not fall into the class. The integral can be defined under the weaker condition (without the expectation)

$$\int_0^T \Delta^2(t)dt < \infty, \quad \text{almost surely ,}$$

i.e., $\Delta(t) \in L^2(\mathcal{F})$, but then the integral is not guaranteed to be a martingale, but a *local martingale*, a much weaker property. We usually impose restrictions on $\Delta(t)$ so that it is still in $L^2(\mathcal{F})$ but so that the Ito integral is a martingale.

2.4.7 Ito Processes

Now we can define the process in equation (2.3.11) more precisely. A diffusion process or Ito process is a process S in $\mathbb{R}$ such that

$$S_t = S_0 + \int_0^t \mu_s ds + \int_0^t \sigma_s dZ_s \tag{2.4.9}$$

where the process $\mu \in \mathcal{L}^1$ and the process $\sigma \in \mathcal{L}^2$.

An Ito process is the sum of a "normal" (non-stochastic) integral and a stochastic integral. A Brownian motion is obviously an Ito process.

We frequently write an Ito process in "differential" form, as in equation (2.3.11),

$$dS_t = \mu_t dt + \sigma_t dZ_t, \tag{2.4.10}$$

and refer to dS_t as the increment of S at time t. If μ, σ are mean-square integrable then

$$\frac{d}{d\tau} E_t\left(S_\tau\right)\Big|_{\tau=t} = \mu_t, \quad \text{a.s.} \tag{2.4.11}$$

and

$$\frac{d}{d\tau} \operatorname{Var}_t\left(S_\tau\right)\Big|_{\tau=t} = \sigma_t^2, \quad \text{a.s.} \tag{2.4.12}$$

We frequently write these equations as $E(dS_t) = \mu_t dt$ and $\operatorname{Var}(dS_t) = \sigma_t^2 dt$, respectively, and

refer to $\mu_t dt$ as the mean of dS_t, and $\sigma_t^2 dt$ as its variance. We refer to the process μ as the drift process and to the process σ as the diffusion process.

2.4.8 Ito's Lemma

We now restate Ito's Lemma more formally:

Theorem 2.4.4. *[Ito's Lemma - One-Dimensional Case] Suppose that S is an Ito process with*

$$dS_t = \mu_t dt + \sigma_t dZ_t$$

and $f : \mathbb{R}^2 \to \mathbb{R}$ is twice continuously differentiable. Then the process C defined by $C_t = f(t, S_t)$ is an Ito process with

$$dC_t = \left(\mathcal{D}_S f(t, S_t) + f_t(t, S_t)\right) dt + f_S(t, S_t)\, \sigma_t dZ_t$$

where

$$\mathcal{D}_S f(t, S_t) = f_S(t, S_t)\, \mu_t + \frac{1}{2} f_{SS}(t, S_t)\, \sigma_t^2$$

As mentioned earlier, Ito's lemma extends the chain rule of calculus for stochastic integrals. If the process S was just a non-stochastic integral

$$dS_t = \mu_t dt$$

then the chain rule of calculus would imply that

$$dC_t = f_S(t, S_t) dS_t + f_t(t, S_t) dt = [f_S(t, S_t)\mu_t + f_t(t, S_t)] dt.$$

When, however, S includes a stochastic integral, it is not true that

$$dC_t = f_S(t, S_t) dS_t + f_t(t, S_t) dt$$

because this would omit the term

$$\frac{1}{2} f_{SS}(t, S_t)\, \sigma_t^2 dt$$

2.4.9 Ito's Lemma for Two Variables

Suppose X and Y are Ito processes with differentials

$$dX(t) = \Theta_1(t)dt + \sigma_{11}(t)dW_1(t) + \sigma_{12}(t)dW_2(t) \tag{2.4.13}$$

$$dY(t) = \Theta_2(t)dt + \sigma_{21}(t)dW_1(t) + \sigma_{22}(t)dW_2(t) \tag{2.4.14}$$

where W_1 and W_2 are independent Brownian motions. Then

$$dX(t)dX(t) = \left(\sigma_{11}^2(t) + \sigma_{12}^2(t)\right) dt \tag{2.4.15}$$

$$dX(t)dY(t) = \left(\sigma_{11}(t)\sigma_{21}(t) + \sigma_{12}(t)\sigma_{22}(t)\right) dt \tag{2.4.16}$$

$$dY(t)dY(t) = \left(\sigma_{21}^2(t) + \sigma_{22}^2(t)\right) dt \tag{2.4.17}$$

Equations (2.4.15)-(2.4.17) can be obtained by multiplying the equations (2.4.13) and (2.4.14) for $dX(t)$ and $dY(t)$ and using the multiplication table

$$\begin{aligned} dW_i(t)dW_i(t) &= dt, \\ dW_i(t)dt &= dtdW_i(t) = 0, \\ dtdt &= 0, \end{aligned}$$

and

$$dW_1(t)dW_2(t) = 0. \tag{2.4.18}$$

Equation (2.4.18) holds for independent Brownian motions. If instead we had

$$dW_1(t)dW_2(t) = \rho dt$$

for a constant $\rho \in [-1, 1]$, then ρ would be the correlation between $W_1(t)$ and $W_2(t)$ (i.e., $\mathbb{E}\left[W_1(t)W_2(t)\right] = \rho t$).

Now suppose $f(t, x, y)$ is a function of the time variable t and two dummy variables x and y.

The Ito-Doeblin formula is now

$$\begin{aligned} & df(t, X(t), Y(t)) \\ & = f_t(t, X(t), Y(t))dt + f_x(t, X(t), Y(t))dX(t) + f_y(t, X(t), Y(t))dY(t) \\ & \quad \frac{1}{2} f_{xx}(t, X(t), Y(t))dX(t)dX(t) + f_{xy}(t, X(t), Y(t))dX(t)dY(t) \\ & \quad + \frac{1}{2} f_{yy}(t, X(t), Y(t))dY(t)dY(t). \end{aligned} \tag{2.4.19}$$

Replacing all the differentials on the right-hand side of (2.4.19) by their formulas (2.4.13)-(2.4.17) and integrating, one obtains a formula for the stochastic process $f(t, X(t), Y(t))$ as the sum of $f(0, X(0), Y(0))$, an ordinary integral with respect to time, an Ito integral with respect to dW_1, and an Ito integral with respect to dW_2.

Equation (2.4.19) gives us Ito's product rule:

$$d(X(t)Y(t)) = X(t)dY(t) + Y(t)dX(t) + dX(t)dY(t)$$

The analogous rule for non-stochastic differentials, $d(X(t)Y(t)) = X(t)dY(t) + Y(t)dX(t)$, omits the term $dX(t)dY(t)$. When $dX(t)$ and $dY(t)$ are not stochastic, the term $dX(t)dY(t)$ is of order dt^2 and can be ignored. Instead, when $dX(t)$ and $dY(t)$ are stochastic the term $dX(t)dY(t)$ is of order dt and cannot be ignored.

2.5 Change of Measure

Theorem 2.5.1 (Girsanov's Theorem)**.** *Consider a process $\eta \in (\mathcal{L}^2)^d$ such that ξ^η is a martingale. Then the process Z^η defined by*

$$Z_t^\eta = Z_t + \int_0^t \eta_s ds \tag{2.5.1}$$

is a Brownian motion under Q^η. Z^η has a martingale representation property under Q^η : for any local Q^η-martingale M_t, adapted to $\mathbb{F}$, there exists a process $\phi \in (\mathcal{L}^2)^d$, such that

$$M_t = M_0 + \int_0^t \phi_s dZ_s^\eta$$

Girsanov's theorem says that, by adding a drift to a process that is a Brownian motion under the probability measure P, we can obtain a process that is a Brownian motion under the

probability measure Q^η. To understand intuitively why the process Z^η has no drift, we consider the increment

$$dZ_t^\eta = dZ_t + \eta_t dt$$

In addition, we think of the increment dZ_t as taking the values $\sqrt{dt}$ and $-\sqrt{dt}$. Under P, the probability of each value (conditional on time t) is $1/2$. Under Q^η, the probability of $\sqrt{dt}$ is

$$\frac{1}{2}\frac{\xi_{t+dt}^\eta}{\xi_t^\eta}\bigg|_{dZ_t=\sqrt{dt}} = \frac{1}{2}\exp\left(-\eta_t\sqrt{dt}\right) + o(\sqrt{dt})$$

and the probability of $-\sqrt{dt}$ is

$$\frac{1}{2}\exp\left(\eta_t\sqrt{dt}\right) + o(\sqrt{dt})$$

Therefore, the expectation of dZ_t^η is

$$\begin{aligned}&\frac{1}{2}\exp\left(-\eta_t\sqrt{dt}\right)\left(\sqrt{dt}+\eta_t dt\right) + \frac{1}{2}\exp\left(\eta_t\sqrt{dt}\right)\left(-\sqrt{dt}+\eta_t dt\right) + o(dt)\\ =&\frac{1}{2}\left(1-\eta_t\sqrt{dt}\right)\left(\sqrt{dt}+\eta_t dt\right) + \frac{1}{2}\left(1+\eta_t\sqrt{dt}\right)\left(-\sqrt{dt}+\eta_t dt\right) + o(dt).\end{aligned}$$

This is equal to 0 in order dt.

Girsanov's theorem has the following important implication.

Proposition 2.5.1. *Consider an Ito process S in $\mathbb{R}^N$,*

$$S_t = S_0 + \int_0^t \mu_s ds + \int_0^t \sigma_s dZ_s$$

and processes $\nu \in (\mathcal{L}^1)^N$ and $\eta \in (\mathcal{L}^2)^d$ such that

$$\sigma_t\eta_t = \mu_t - \nu_t$$

If the process ξ^η is a martingale, then S is an Ito process under Q^η, and

$$S_t = S_0 + \int_0^t \nu_s ds + \int_0^t \sigma_s dZ_s^\eta$$

We can thus write an Ito process S under P as an Ito process under Q^η. The drift changes, but the diffusion stays the same.

In fact, the reverse is true: we can show that under any equivalent probability measure w.r.t.

which S is a martingale, the diffusion part of the process S_t stays the same, and only the drift changes.

2.6 Summary

Let $W(t)$ be a Brownian motion and $\Delta(t)$ a stochastic process adapted to the filtration of the Brownian motion. The Ito integral

$$I(t) = \int_0^t \Delta(u)dW(u)$$

is a martingale. Because it is zero at time $t = 0$, its expectation is zero for all t.

Its variance is given by Ito's isometry

$$\mathbb{E}I^2(t) = \mathbb{E}\int_0^t \Delta^2(u)du$$

An Ito process is a process of the form

$$X(t) = X(0) + \int_0^t \Delta(u)dW(u) + \int_0^t \Theta(u)du$$

where $X(0)$ is nonrandom and $\Delta(u)$ and $\Theta(u)$ are adapted stochastic processes. In differential notation, we write

$$dX(t) = \Delta(t)dW(t) + \Theta(t)dt$$

To compute expressions in differential form, we use the multiplication table

$$dW(t)dW(t) = dt, \quad dW(t)dt = dtdW(t) = 0, \quad dtdt = 0$$

For example,

$$\begin{aligned} dX(t)dX(t) &= (\Delta(t)dW(t) + \Theta(t)dt)^2 \\ &= \Delta^2(t)dW(t)dW(t) + 2\Delta(t)\Theta(t)dW(t)dt + \Theta^2(t)dtdt \\ &= \Delta^2(t)dt \end{aligned}$$

Let $f(t, x)$ is a function of the time variable t and a stochastic process x. The Ito-Doeblin

formula is

$$df(t, X(t)) = f_t(t, X(t))dt + f_x(t, X(t))dX(t) + \frac{1}{2} f_{xx}(t, X(t))dX(t)dX(t).$$

For two stochastic processes $X(t)$ and $Y(t)$, Ito's product rule is

$$d(X(t)Y(t)) = X(t)dY(t) + Y(t)dX(t) + dX(t)dY(t).$$

Optional

*2.6.1 Proofs

Proof of Theorem 2.4.1. Let $0 \leq s \leq t \leq T$ be given. We shall assume that s and t are in different subintervals of the partition Π (i.e., there are partition points t_ℓ and t_k such that $t_\ell < t_k, s \in [t_\ell, t_{\ell+1})$, and $t \in [t_k, t_{k+1})$). If s and t are in the same subinterval, the following proof simplifies. The Ito integral may be rewritten as

$$I(t) = \sum_{j=0}^{\ell-1} \Delta(t_j) \left[W(t_{j+1}) - W(t_j)\right] + \Delta(t_\ell) \left[W(t_{\ell+1}) - W(t_\ell)\right] \tag{*2.6.1}$$

$$+ \sum_{j=\ell+1}^{k-1} \Delta(t_j) \left[W(t_{j+1}) - W(t_j)\right] + \Delta(t_k) \left[W(t) - W(t_k)\right] \tag{*2.6.2}$$

We must show that $\mathbb{E}[I(t) \mid \mathcal{F}(s)] = I(s)$. We take the conditional expectation of each of the four terms on the right-hand side. Every random variable in the first sum $\sum_{j=0}^{\ell-1} \Delta(t_j) \left[W(t_{j+1}) - W(t_j)\right]$ is $\mathcal{F}(s)$-measurable because the latest time appearing in this sum is t_ℓ and $t_\ell \leq s$. Therefore,

$$\mathbb{E}\left[\sum_{j=0}^{\ell-1} \Delta(t_j) \left[W(t_{j+1}) - W(t_j)\right] \mid \mathcal{F}(s)\right] = \sum_{j=0}^{\ell-1} \Delta(t_j) \left[W(t_{j+1}) - W(t_j)\right]. \tag{*2.6.3}$$

For the second term on the right-hand side, we use the properties of expectations and the martingale property of W to write

$$\mathbb{E}\left[\Delta(t_\ell) \left(W(t_{\ell+1}) - W(t_\ell)\right) \mid \mathcal{F}(s)\right] = \Delta(t_\ell) \left(\mathbb{E}\left[W(t_{\ell+1}) \mid \mathcal{F}(s)\right] - W(t_\ell)\right) \tag{*2.6.4}$$

$$= \Delta(t_\ell) \left(W(s) - W(t_\ell)\right) \tag{*2.6.5}$$

Adding these results, we obtain $I(s)$.

It remains to show that the conditional expectations of the third and fourth terms on the right-hand side are zero. We will then have $\mathbb{E}[I(t) \mid \mathcal{F}(s)] = I(s)$.

The summands in the third term are of the form $\Delta(t_j)\,[W(t_{j+1}) - W(t_j)]$, where $t_j \geq t_{\ell+1} > s$. This permits us to use the law of iterated expectations

$$\begin{aligned}
&\mathbb{E}\left\{\Delta(t_j)\ (W(t_{j+1}) - W(t_j)) \mid \mathcal{F}(s)\right\} \\
&\qquad = \mathbb{E}\left\{\mathbb{E}\left[\Delta(t_j)\,(W(t_{j+1}) - W(t_j)) \mid \mathcal{F}(t_j)\right] \mid \mathcal{F}(s)\right\} \\
&\qquad = \mathbb{E}\left\{\Delta(t_j)\,(\mathbb{E}\left[W(t_{j+1}) \mid \mathcal{F}(t_j)\right] - W(t_j)) \mid \mathcal{F}(s)\right\} \\
&\qquad = \mathbb{E}\left\{\Delta(t_j)\,(W(t_j) - W(t_j)) \mid \mathcal{F}(s)\right\} = 0
\end{aligned}$$

At the end, we have used the fact that W is a martingale. Because the conditional expectation of each of the summands in the third term on the right-hand side is zero, the conditional expectation of the whole term is zero:

$$\mathbb{E}\left\{\sum_{j=\ell+1}^{k-1} \Delta(t_j)\,[W(t_{j+1}) - W(t_j)] \mid \mathcal{F}(s)\right\} = 0$$

The fourth term on the right-hand side is treated like the summands in the third term, with the result that

$$\begin{aligned}
&\mathbb{E}\left\{\Delta(t_k)\ (W(t) - W(t_k)) \mid \mathcal{F}(s)\right\} \\
&\qquad = \mathbb{E}\left\{\mathbb{E}\left[\Delta(t_k)\,(W(t) - W(t_k)) \mid \mathcal{F}(t_k)\right] \mid \mathcal{F}(s)\right\} \\
&\qquad = \mathbb{E}\left\{\Delta(t_k)\,(\mathbb{E}\left[W(t) \mid \mathcal{F}(t_k)\right] - W(t_k)) \mid \mathcal{F}(s)\right\} \\
&\qquad = \mathbb{E}\left\{\Delta(t_k)\,(W(t_k) - W(t_k)) \mid \mathcal{F}(s)\right\} = 0.
\end{aligned}$$

□

Proof of Theorem 2.4.2. To simplify the notation, we set $D_j = W(t_{j+1}) - W(t_j)$ for $j = 0, \ldots, k-1$ and $D_k = W(t) - W(t_k)$ so that the Ito integral may be written as $I(t) = \sum_{j=0}^{k} \Delta(t_j)\, D_j$ and

$$I^2(t) = \sum_{j=0}^{k} \Delta^2(t_j)\, D_j^2 + 2 \sum_{0 \leq i < j \leq k} \Delta(t_i)\, \Delta(t_j)\, D_i D_j$$

We first show that the expected value of each of the cross terms is zero. For $i < j$, the random variable $\Delta(t_i)\,\Delta(t_j)\,D_i$ is $\mathcal{F}(t_j)$-measurable, while the Brownian increment D_j is independent of $\mathcal{F}(t_j)$. Furthermore, $\mathbb{E}D_j = 0$. Therefore,

$$\begin{aligned}\mathbb{E}\left[\Delta(t_i)\,\Delta(t_j)\,D_iD_j\right] &= \mathbb{E}\left[\Delta(t_i)\,\Delta(t_j)\,D_i\right]\cdot\mathbb{E}D_j\\ &= \mathbb{E}\left[\Delta(t_i)\,\Delta(t_j)\,D_i\right]\cdot 0\\ &= 0\end{aligned}$$

We next consider the square terms $\Delta^2(t_j)\,D_j^2$. The random variable $\Delta^2(t_j)$ is $\mathcal{F}(t_j)$- measurable, and the squared Brownian increment D_j^2 is independent of $\mathcal{F}(t_j)$. Furthermore, $\mathbb{E}D_j^2 = t_{j+1} - t_j$ for $j = 0, \ldots, k-1$ and $\mathbb{E}D_k^2 = t - t_k$. Therefore,

$$\mathbb{E}I^2(t) = \sum_{j=0}^{k}\mathbb{E}\left[\Delta^2(t_j)\,D_j^2\right] \tag{*2.6.6}$$

$$= \sum_{j=1}^{k}\mathbb{E}\Delta^2(t_j)\cdot\mathbb{E}D_j^2 \tag{*2.6.7}$$

$$= \sum_{j=0}^{k-1}\mathbb{E}\Delta^2(t_j)\,(t_{j+1}-t_j) + \mathbb{E}\Delta^2(t_k)\,(t-t_k) \tag{*2.6.8}$$

But $\Delta(t_j)$ is constant on $[t_j, t_{j+1})$, so

$$\Delta^2(t_j)(t_{j+1}-t_j) = \int_{t_j}^{t_{j+1}}\Delta^2(u)\,du.$$

Similarly,

$$\Delta^2(t_k)(t-t_k) = \int_{t_k}^{t}\Delta^2(u)\,du.$$

We may thus continue to obtain

$$\begin{aligned}\mathbb{E}I^2(t) &= \sum_{j=0}^{k-1}\mathbb{E}\int_{t_j}^{t_{j+1}}\Delta^2(u)du + \mathbb{E}\int_{t_k}^{t}\Delta^2(u)du\\ &= \mathbb{E}\left[\sum_{j=0}^{k-1}\int_{t_j}^{t_{j+1}}\Delta^2(u)du + \int_{t_k}^{t}\Delta^2(u)du\right] = \mathbb{E}\int_0^t\Delta^2(u)du.\end{aligned}$$

□

Proof of Theorem 2.4.3. We first compute the quadratic variation accumulated by the

Ito integral on one of the subintervals $[t_j, t_{j+1}]$ on which $\Delta(u)$ is constant. For this, we choose partition points

$$t_j = s_0 < s_1 < \cdots < s_m = t_{j+1}$$

and consider

$$\sum_{i=0}^{m-1} \left[I\left(s_{i+1}\right) - I\left(s_i\right)\right]^2 = \sum_{i=0}^{m-1} \left[\Delta\left(t_j\right)\left(W\left(s_{i+1}\right) - W\left(s_i\right)\right)\right]^2 \quad \text{(*2.6.9)}$$

$$= \Delta^2\left(t_j\right) \sum_{i=0}^{m-1} \left(W\left(s_{i+1}\right) - W\left(s_i\right)\right)^2 \quad \text{(*2.6.10)}$$

As $m \to \infty$ and the step size $\max_{i=0,\ldots,m-1}\left(s_{i+1} - s_i\right)$ approaches zero, the term

$$\sum_{i=0}^{m-1} \left(W(s_{i+1}) - (s_i)\right)^2$$

converges to the quadratic variation accumulated by Brownian motion between times t_j and t_{j+1}, which is $t_{j+1} - t_j$. Therefore, the quadratic variation accumulated by the Ito integral between times t_j and t_{j+1} is

$$\Delta^2\left(t_j\right)\left(t_{j+1} - t_j\right) = \int_{t_j}^{t_{j+1}} \Delta^2(u)du$$

where again we have used the fact that $\Delta(u)$ is constant for $t_j \leq u < t_{j+1}$. Analogously, the quadratic variation accumulated by the Ito integral between times t_k and t is $\int_{t_k}^{t} \Delta^2(u)du$. Adding up all these pieces, we obtain the result. □

Chapter 3

Asset Pricing: Discrete Time

3.1 Financial Assets

3.1.1 Security Markets

A financial asset is defined by the payoffs it provides to its holder. We refer to financial assets with exogenously given payoffs as securities.

There are two dates, 0 and 1. Securities are traded at date 0, and their payoffs are realized at date 1. Date 0, the present, has no uncertainty. On date 1, one of S states can occur.

Security j is defined by its payoff x_j, an element of $\mathbb{R}^S$, where x_{js} denotes the payoff that the holder of one unit (or one share) of security j receives in state s at date 1. Payoffs are in units of the single consumption good available in this economy. The consumption good is therefore the numéraire. Payoffs may be positive, zero, or negative. There exists a finite number J of securities with payoffs $x_1, \ldots, x_J$, with $x_j \in \mathbb{R}^S$, which are exogenously given.

The $J \times S$ matrix X of payoffs of all securities

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_J \end{bmatrix}$$

is called the payoff matrix. The vectors x_j are understood to be row vectors. In general, vectors are understood to be either row vectors or column vectors, as the context requires.

A portfolio consists of holdings in the J securities. These holdings may be positive, zero, or negative. A positive holding of a security is called a long position, and a negative holding is called a short position.

A portfolio is denoted by a J-dimensional vector h, where h_j represents the holdings of security j. The portfolio payoff is $\sum_j h_j x_j$ and can be represented as hX.

The set of payoffs available via trades of securities is the asset span and is denoted by $\mathcal{M}$:

$$\mathcal{M} = \left\{ z \in \mathbb{R}^S \colon z = hX \text{ for some } h \in \mathbb{R}^J \right\}$$

Thus $\mathcal{M}$ is the subspace of $\mathbb{R}^S$ spanned by the security payoffs, that is, the row span of the payoff matrix X. If $\mathcal{M} = \mathbb{R}^S$, then markets are complete. If $\mathcal{M}$ is a proper subspace of $\mathbb{R}^S$, then markets are incomplete. When markets are complete, any element of $\mathbb{R}^S$ can be obtained as a portfolio payoff.

Theorem 3.1.1. *Markets are complete iff the payoff matrix X has rank S*[1]

Proof. Asset span $\mathcal{M}$ equals the whole space $\mathbb{R}^S$ iff the equation $z = hX$, with J unknowns h_j, has a solution for every $z \in \mathbb{R}^S$. A necessary and sufficient condition for this is that X has rank S. □

A security is called redundant if its payoff can be generated as the payoff of a portfolio of other securities. There are no redundant securities iff the payoff matrix X has rank J.

The prices of securities at date 0 are denoted by a J-dimensional vector $p = (p_1, \ldots, p_J)$. The price of portfolio h at security prices p is $ph = \sum_j p_j h_j$. The return r_j on security j is its payoff x_j divided by its price p_j (assumed to be nonzero; the return on a payoff with zero

1. We use "A iff B" as an abbreviation for "A if and only if B". It has the same meaning as "A is equivalent to B" and as "A is a necessary and sufficient condition for B". We sometimes use $A \iff B$ to denote "A iff B".

Proving necessity in "A iff B" means proving "B implies A" ($B \implies A$) whereas proving sufficiency means proving "A implies B" ($A \implies B$).

price is undefined):

$$r_j = \frac{x_j}{p_j}$$

Thus, "return" means gross return ("net return" equals gross return minus one). Throughout, we work with gross returns.

The asset span can be specified using the returns on the securities rather than their payoffs. In this case, the asset span is the subspace of $\mathbb{R}^S$ spanned by the returns on the securities.

The following example illustrates the concepts introduced earlier.

Example 3.1.0.1. Let there be three states and two securities. Security 1 is risk free and has payoff $x_1 = (1, 1, 1)$. Security 2 is risky with $x_2 = (1, 2, 2)$. The payoff matrix is

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \end{bmatrix}$$

The asset span is

$$\mathcal{M} = \{(z_1, z_2, z_3) : z_1 = h_1 + h_2, z_2 = h_1 + 2h_2, z_3 = h_1 + 2h_2, \text{ for some } (h_1, h_2)\},$$

which is a two-dimensional subspace of $\mathbb{R}^3$. By inspection, $\mathcal{M} = \{(z_1, z_2, z_3) : z_2 = z_3\}$. At prices $p_1 = 0.8$ and $p_2 = 1.25$, security returns are $r_1 = (1.25, 1.25, 1.25)$ and $r_2 = (0.8, 1.6, 1.6)$. □

3.2 Pricing

3.2.1 The Law of One Price

The law of one price says that all portfolios with the same payoff have the same price:

$$\text{if } hX = h'X, \text{ then } ph = ph'$$

for any two portfolios h and h'.

Because the payoff matrix X is given exogenously, the law of one price is properly interpreted as a restriction on security prices as opposed to a restriction on payoffs. If there are no redundant securities, only one portfolio can generate any given payoff, and thus the law of

one price is trivially satisfied. If there are redundant securities, the law of one price may or may not be satisfied depending on security prices.

A necessary and sufficient condition for the law of one price to hold is that every portfolio with zero payoff has zero price. If the law of one price does not hold, then any payoff (that is, any contingent claim in the asset span) can be purchased at any price.

3.2.2 The Payoff Pricing Functional

For any security prices p we define a mapping $q : \mathcal{M} \to \mathbb{R}$ that assigns to each payoff the price(s) of the portfolio(s) that generate(s) that payoff. Formally,

$$q(z) := \{w : w = ph \text{ for some } h \text{ such that } z = hX\}. \tag{3.2.1}$$

In general, the mapping q is a correspondence rather than a single-valued function.

If the law of one price holds, then q is single valued. Further, it is a linear functional:

Theorem 3.2.1. *The law of one price holds iff q is a linear functional on the asset span $\mathcal{M}$.*

Proof. If the law of one price holds, then, as just noted, q is single valued. To prove linearity, consider payoffs $z, z' \in \mathcal{M}$ such that $z = hX$ and $z' = h'X$ for some portfolios h and h'. For arbitrary $\lambda, \mu \in \mathbb{R}$, the payoff $\lambda z + \mu z'$ can be generated by the portfolio $\lambda h + \mu h'$ with price $\lambda ph + \mu ph'$. Because q is single valued, Definition 3.2.1 implies that

$$q\left(\lambda z + \mu z'\right) = \lambda ph + \mu ph'.$$

The right-hand side of the above equation equals $\lambda q(z) + \mu q\left(z'\right)$, and thus q is linear. Conversely, if q is a functional, then the law of one price holds by definition. □

Whenever the law of one price holds, we call q the payoff pricing functional.

3.2.3 State Prices in Complete Markets

Let e_s denote the sth basis vector in the space $\mathbb{R}^S$ of contingent claims, with 1 in the sth place and zeros elsewhere. Vector e_s is the state claim or the Arrow-Debreu security of state s. It is the claim to one unit of consumption contingent on the occurrence of state s. If

markets are complete and if the law of one price holds, then the payoff pricing functional assigns a unique price to each state claim. Let

$$q_s := q(e_s)$$

denote the price of the state claim of state s. We call q_s the state price of state s.

Because any linear functional on $\mathbb{R}^S$ can be identified by its values on the basis vectors of $\mathbb{R}^S$, the payoff pricing functional q can be represented as

$$q(z) = qz$$

for every $z \in \mathbb{R}^S$, where q on the right-hand side of the above equation is an S-dimensional vector of state prices. With some abuse of notation, we use the same notation for the functional and the vector that represents it.

Because the price of each security equals the value of its payoff under the payoff pricing functional, we have

$$p_j = qx_j$$

or, in matrix notation,

$$p = Xq \tag{3.2.2}$$

Equation (3.2.2) is a system of linear equations that associates state prices with given security prices.

A left inverse of the payoff matrix X is a matrix L that satisfies $LX = I_S$, where I_S is the $S \times S$ identity matrix. A left inverse exists iff X is of rank S, which occurs if $J \geq S$ and the columns of X are linearly independent. Markets are complete iff a left inverse of X exists.

A matrix L defined by

$$L = (X'X)^{-1} X'$$

where the prime indicates transposition, is a left inverse of X. Thus, L exists iff $X'X$ is invertible.

Premultiplying both sides of equation (3.2.2) by a left inverse of the payoff matrix, it follows that

$$q = Lp$$

(all left inverses give the same value for q).

The results of this section depend on the assumption of market completeness because otherwise L does not exist (or, alternatively, because state claim e_s may not be in the asset span $\mathcal{M}$, and thus $q(e_s)$ may not be defined).

3.3 Arbitrage

3.3.1 Notation

Our convention on inequalities is as follows: for two vectors $x, y \in \mathbb{R}^n$,

$x \geq y$ means that $x_i \geq y_i \quad \forall i$;	x is greater than y,
$x > y$ means that $x \geq y$ and $x \neq y$;	x is greater than but not equal to y,
$x \gg y$ means that $x_i > y_i \quad \forall i$;	x is strictly greater than y.

For a vector x, positive means $x \geq 0$, positive and nonzero means $x > 0$, and strictly positive means $x \gg 0$. These definitions apply to scalars as well. For scalars, "positive and nonzero" is equivalent to "strictly positive".

3.3.2 Arbitrage and Strong Arbitrage

A strong arbitrage is a portfolio that has a positive payoff and a strictly negative price. An arbitrage is a portfolio that is either a strong arbitrage or has a positive and nonzero payoff and zero price. Formally, a strong arbitrage is a portfolio h that satisfies $hX \geq 0$ and $ph < 0$, and an arbitrage is a portfolio h that satisfies $hX \geq 0$ and $ph \leq 0$ with either $hX \neq 0$ or $ph \neq 0$ (or both).

It is possible for a portfolio to be an arbitrage but not a strong arbitrage:

Example 3.3.0.1. Let there be two securities with payoffs $x_1 = (1, 1)$ and $x_2 = (1, 2)$ and

prices $p_1 = p_2 = 1$. Then, portfolio $h = (-1, 1)$ is an arbitrage but not a strong arbitrage. In fact, there is no strong arbitrage. □

If there exists no portfolio with a positive and nonzero payoff, then any arbitrage is a strong arbitrage. Further, the law of one price does not hold iff there exists a portfolio with zero payoff and strictly negative price. Such a portfolio is a strong arbitrage.

Example 3.3.0.2. Suppose that two securities have payoffs $x_1 = (-1, 2, 0)$ and $x_2 = (2, 2, -1)$. A portfolio $h = (h_1, h_2)$ has a positive payoff if

$$-h_1 + 2h_2 \geq 0$$

$$h_1 + h_2 \geq 0$$

and

$$-h_2 \geq 0$$

These inequalities are satisfied by the zero portfolio alone. Therefore, there exists no portfolio with positive and nonzero payoff, implying further that any arbitrage is a strong arbitrage. Because there are no redundant securities, the law of one price holds for any security prices, so there is no strong arbitrage.

Consequently, there is no arbitrage for any security prices. □

3.3.3 Visual Representation

It is helpful to have a visual representation of the set of security prices that exclude arbitrage. Suppose that there are two securities with payoffs x_1 and x_2, and consider the payoff pairs (x_{1s}, x_{2s}) in each state. These pairs are denoted $x_{\cdot 1}, \ldots, x_{\cdot S}$. Figure 3.1 is drawn on the assumption that $x_{js} > 0$ for all j and s, but the analysis does not depend on this restriction.

Now interpret the coordinate axes as portfolio weights h_1 and h_2 so that any point in the diagram is associated with a portfolio (h_1, h_2). For each $x_{\cdot s}$, construct a line perpendicular to $x_{\cdot s}$ through the origin. The set of portfolios h with positive payoff in state s is the set of points northeast of this line. If this construction is performed in each state, the intersection of the indicated portfolio sets gives the set of portfolios with positive payoffs in all states.

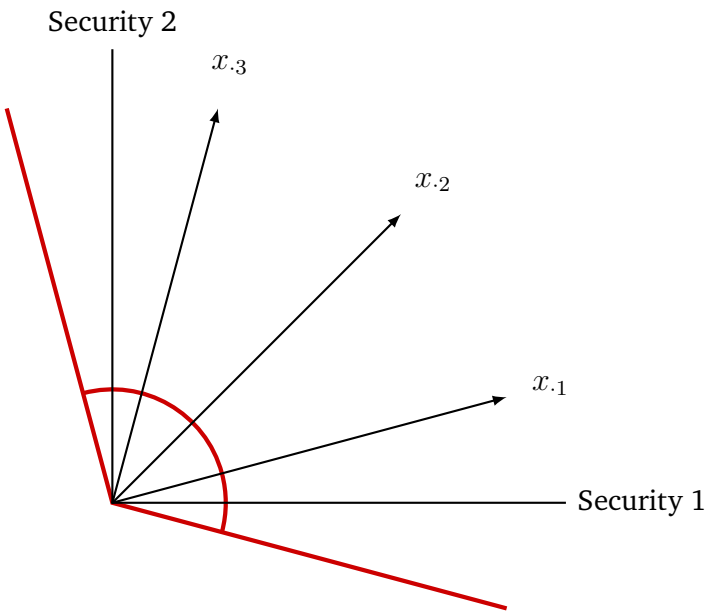

Figure 3.1: The rays labeled $x_{\cdot 1}, x_{\cdot 2}$, and $x_{\cdot 3}$ show payoffs of securities 1 and 2 in states 1, 2, and 3. The cone indicated by the red arc and the red rays shows portfolios that have positive payoffs in all states.

The indicated portfolios are those for which the ray through the point h intersects the arc.

Suppose that security prices are given by $p = (p_1, p_2)$, as shown in figure 3.2. Then the set of zero-price portfolios consists of the line through the origin perpendicular to p. Figure 3.3, which combines figures 3.1 and 3.2, shows that the set of positive-payoff portfolios intersects the set of negative-price portfolios only at the origin, and thus there is no arbitrage.

This conclusion is a consequence of the fact that p lies in the interior of the cone defined by the $x_{\cdot s}$. If p lies on the boundary of the cone, then there is an arbitrage, but not a strong one (figure 3.4), whereas if p lies outside the cone, then there exists strong arbitrage (figure 3.5).

The preceding construction, being two-dimensional, is necessarily restricted to the case in which agents take nonzero positions in at most two securities. It is worth noting that, if there are more than two securities, nonexistence of an arbitrage if portfolios are restricted to contain at most two securities is consistent with existence of an arbitrage if portfolios are unrestricted. This is illustrated by the following example.

Example 3.3.0.3. Consider three securities with payoffs $x_1 = (1, 1, 0), x_2 = (0, 1, 1)$, and

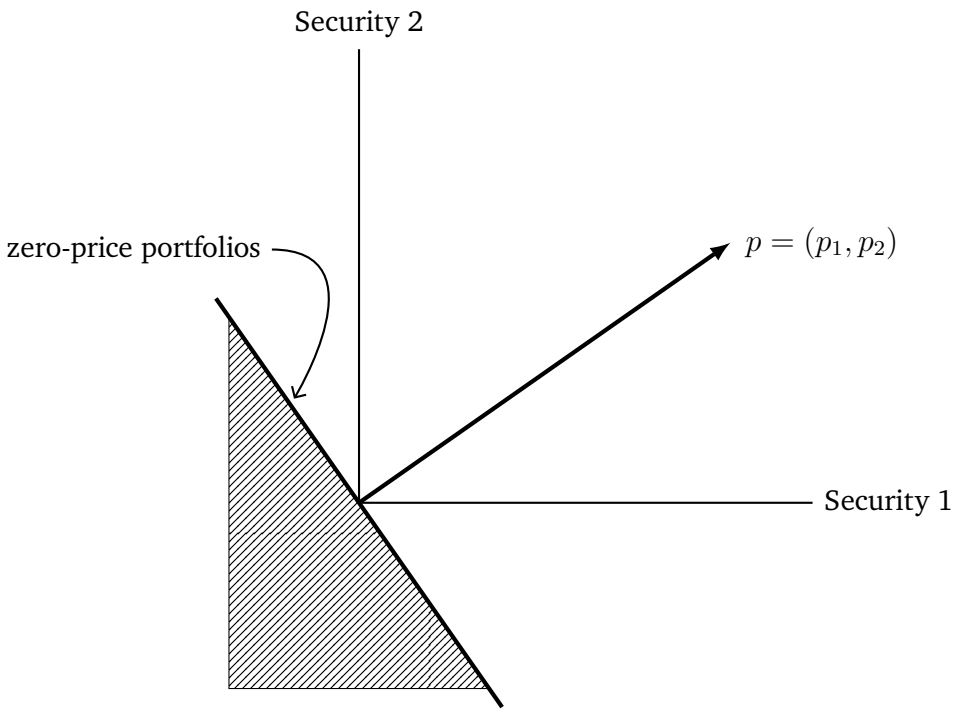

Figure 3.2: Portfolios in the shaded region have a negative price.

$x_3 = (1, 0, 1)$ and with prices $p_1 = 1$, and $p_2 = p_3 = 1/2$. No arbitrage exists with nonzero positions in any two of these securities, but portfolio $h = (-1, 1, 1)$ is an arbitrage. □

3.3.4 Positivity of the Payoff Pricing Functional

A functional is positive if it assigns positive value to every positive element of its domain. It is strictly positive if it assigns strictly positive value to every positive and nonzero element of its domain. Note that if there is no positive (positive and nonzero) element in the domain of a functional, then the functional is trivially positive (strictly positive). Our terminology of positive and strictly positive functionals is consistent with the terminology of positive and strictly positive vectors in the following sense: a linear functional $F : \mathbb{R}^l \to \mathbb{R}$ has a representation in the form of a scalar product $F(x) = fx$ for some vector $f \in \mathbb{R}^l$. Functional F is strictly positive (positive) iff the corresponding vector f is strictly positive (positive).

Absence of arbitrage or strong arbitrage at given security prices corresponds to the payoff pricing functional's being strictly positive or positive.

Theorem 3.3.1. *The payoff pricing functional is linear and strictly positive iff there is no*

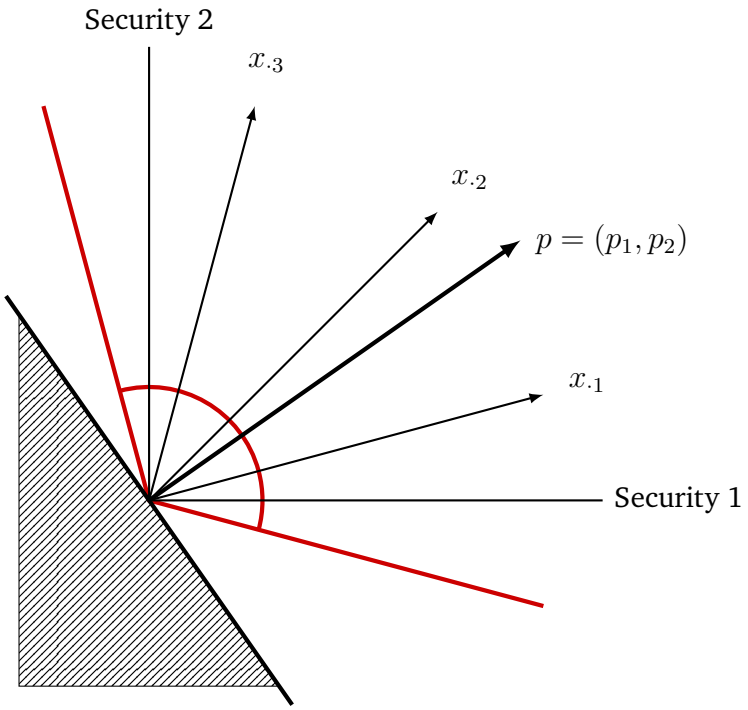

Figure 3.3: The portfolios in the cone determined by the red rays have positive payoffs; the portfolios in the shaded region have a negative price. These regions intersect only at the origin, indicating the absence of arbitrage. This conclusion follows from the fact that p lies in the cone generated by the security payoffs.

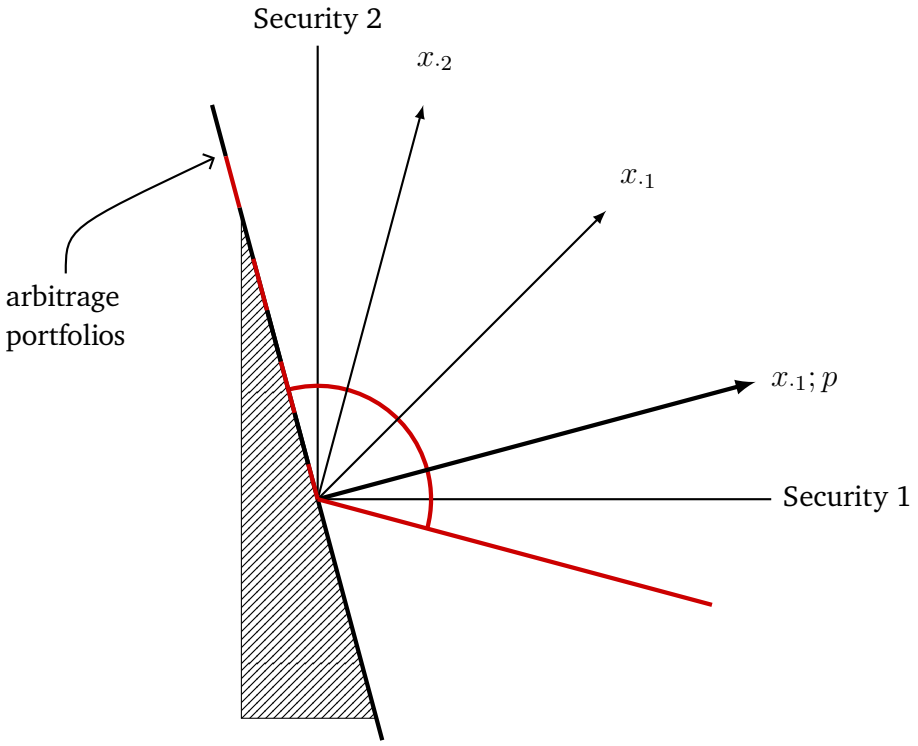

Figure 3.4: The ray p coincides with one of the boundaries of the cone generated by security payoffs. The interpretation is that there exists arbitrage, but not strong arbitrage.

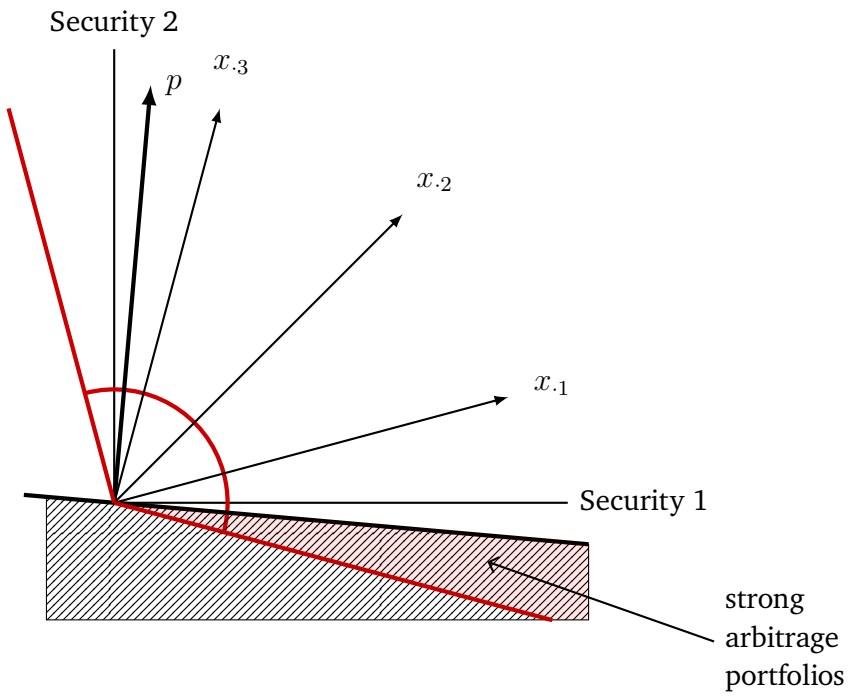

Figure 3.5: The ray p lies outside the cone generated by the security payoffs. The portfolios in the indicated region are arbitrages.

arbitrage.

Proof. The necessity of the absence of arbitrage if the payoff pricing functional is linear and strictly positive is obvious. To prove sufficiency, note that exclusion of arbitrage implies satisfaction of the law of one price, which in turn implies that q is a linear functional. If $z \in \mathcal{M}$, then $q(z) = ph$ for h such that $hX = z$. Exclusion of arbitrage implies that $q(z) > 0$, if $z > 0$, and thus q is strictly positive. □

We also have the following theorem:

Theorem 3.3.2. *The payoff pricing functional is linear and positive iff there is no strong arbitrage.*

The proof is similar to that of Theorem 3.2.1.

3.3.5 Positive State Prices

If markets are complete, then the law of one price implies the existence of a state price vector q such that

$$p = Xq \tag{3.3.1}$$

Because the payoff matrix X is left invertible under complete markets, the vector q that solves equation (3.3.1) is unique. In view of

$$q(z) = qz$$

the absence of arbitrage is equivalent to state prices being strictly positive $(q \gg 0)$, and the absence of strong arbitrage is equivalent to those prices being positive $(q \geq 0)$.

3.4 Valuation

We have already established that security prices can be characterized by a payoff pricing functional mapping the asset span into the reals. The payoff pricing functional is linear and strictly positive (positive) iff security prices exclude arbitrage (strong arbitrage). With complete markets, the payoff pricing functional is uniquely determined, implying the uniqueness of state prices.

To study incomplete markets, we use a valuation functional, an extension of the payoff pricing functional from the asset span $\mathcal{M}$ to the entire contingent claim space $\mathbb{R}^S$. Thus, a valuation functional is a linear functional

$$Q : \mathbb{R}^S \to \mathbb{R}$$

that coincides with the payoff pricing functional on the asset $\operatorname{span} \mathcal{M}$; that is,

$$Q(z) = q(z) \quad \text{for every } z \in \mathcal{M}.$$

A valuation functional assigns values to all contingent claims, not just to payoffs. Of special interest is a valuation functional that is strictly positive (positive) because, in the case of complete markets, this property is equivalent to the absence of arbitrage (strong arbitrage).

The following simple example illustrates a positive valuation functional.

Example 3.4.0.1. Suppose that there are two states and a single security with payoff $x_1 = (1, 2)$ and price $p_1 = 1$. The asset span is $\mathcal{M} = \text{span}\{(1, 2)\} = \{(\alpha, 2\alpha) \colon \alpha \in \mathbb{R}\}$, and the payoff pricing functional is given by $q(\alpha, 2\alpha) = \alpha$. Each functional $Q \colon \mathbb{R}^2 \to \mathbb{R}$ defined by $Q(z) = q_1 z_1 + q_2 z_2$, where $q_1, q_2 \geq 0$ and $q_1 + 2q_2 = 1$ is a positive valuation functional. □

3.4.1 The Fundamental Theorem of Finance

If we consider an arbitrary vector of security prices, can we be assured that a strictly positive (positive) valuation functional exists? It cannot exist if security prices permit arbitrage (strong arbitrage) because then either the payoff pricing functional does not exist or it is not strictly positive (positive).

We come now to a critical question: If security prices are such as to exclude arbitrage, does a strictly positive valuation functional exist? The answer is provided in the following theorem.

Theorem 3.4.1 (Fundamental Theorem of Finance)**.** *Security prices exclude arbitrage iff there exists a strictly positive valuation functional.*

Suppose now only that security prices exclude strong arbitrage. This weakening of the condition implies a weakening of the conclusion:

Theorem 3.4.2 (Fundamental Theorem of Finance, Weak Form)**.** *Security prices exclude strong arbitrage iff there exists a positive valuation functional.*

For both theorems, necessity follows from Theorems 3.2.1 and 3.3.2 because existence of a strictly positive (positive) valuation functional implies existence of a strictly positive (positive) payoff pricing functional, the payoff pricing functional being a restriction of the valuation functional.

The proof of sufficiency is more involved. The extension of the payoff pricing functional q from the asset span to the entire commodity space is achieved by extending q one dimension at a time. In the first step we choose a contingent claim $\hat{z}$ not in the asset span $\mathcal{M}$ and extend q to the subspace spanned by $\mathcal{M}$ and $\hat{z}$. This extended subspace has dimension

equal to the dimension of $\mathcal{M}$ plus one. The extension of the payoff pricing functional is achieved by specifying a value π for the contingent claim $\hat{z}$. For the extension to remain strictly positive (positive), the chosen value π must be such that all payoffs greater than $\hat{z}$ have prices that are strictly greater (greater) than π, and all payoffs less than $\hat{z}$ have prices that are strictly less (less) than π. These restrictions define an interval in which π must lie. The extension is the payoff pricing functional for security markets consisting of J securities with payoffs $\{x_1, \ldots, x_J\}$ and prices $\{p_1, \ldots, p_J\}$ and a security with payoff $\hat{z}$ and price π.

In the second step, we choose a contingent claim not in the span of the $J + 1$ securities of step 1 and extend the payoff pricing functional to the subspace spanned by the $J + 1$ securities of step 1 and the new contingent claim. After $S - J$ steps we achieve an extension to the entire commodity space.

The formal proof is provided below as optional.

Optional

*3.4.2 Proof of Sufficiency in the Fundamental Theorem of Finance

Because all of the steps in the construction of the extension are the same, we present only the first one.

Bounds on the Values of Contingent Claims

We now define the upper and lower bounds on the value of a contingent claim $z \in \mathbb{R}^S$ that can be inferred from the prices of the payoffs in $\mathcal{M}$. The upper bound

$$q_u(z) := \min_h \{ph : hX \geq z\} \tag{*3.4.1}$$

is the lowest price of a portfolio, the payoff of which dominates the contingent claim. If such a portfolio does not exist, we set $q_u(z) = +\infty$. For example, if $\mathcal{M} = \text{span}\{(1, 0)\}$ and $z = (1, 1)$, then $q_u(z) = +\infty$.

The lower bound

$$q_\ell(z) := \max_h \{ph : hX \leq z\} \tag{*3.4.2}$$

is the highest price of a portfolio, the payoff of which is dominated by the contingent claim. If such a portfolio does not exist, we set $q_\ell(z) = -\infty$. For example, if $\mathcal{M} = \text{span}\{(1,0)\}$ and $z = (-1,-1)$, then $q_\ell(z) = -\infty$.

For a contingent claim in the asset span, the lower and the upper bounds coincide with the value under the payoff pricing functional as long as there exists no strong arbitrage:

Proposition *3.4.1. *If security prices exclude strong arbitrage, then* $q_u(z) = q_\ell(z) = q(z)$ *for every* $z \in \mathcal{M}$.

Proof. By the definitions of the bounds we have $q_u(z) \leq q(z)$ and $q_\ell(z) \geq q(z)$ for $z \in \mathcal{M}$. Suppose that $q_u(z) < q(z)$ for some $z \in \mathcal{M}$. Then $q_u(z) < +\infty$ and there exists a portfolio h' such that

$$h'X \geq z$$

and

$$ph' < q(z).$$

Let h be a portfolio such that $hX = z$ and $ph = q(z)$. Then portfolio $h' - h$ is a strong arbitrage. This contradicts the assumption. The proof that $q_\ell(z) = q(z)$ is similar. □

The following two examples illustrate the bounds on the values of contingent claims that are not in the asset span.

Example 3.2. In Example 3.4.0.1, the contingent claim $z = (1,1)$ is not in the asset span. We have

$$q_u(z) = \min\{h : (h, 2h) \geq (1,1)\} = 1 \tag{*3.4.3}$$

$$q_\ell(z) = \max\{h : (h, 2h) \leq (1,1)\} = \frac{1}{2}. \tag{*3.4.4}$$

Thus the bounds on the value of z are $1/2$ and 1. □

Example 3.2. Let there be two securities: security 1, a bond with risk-free payoff $x_1 = (1,1,1)$, and security 2 , a stock with payoff $x_2 = (1,2,4)$. The prices of the bond and

stock are, respectively, $p_1 = 1/2$ and $p_2 = 1$. A nontraded call option on the stock with a strike price of 3 has the payoff $z = (0, 0, 1)$. That payoff is not in the span of the payoffs on the stock and the bond and hence cannot be priced using the payoff pricing functional.

A lower bound on the value of the call is determined by solving

$$\max_{h_1,h_2} (p_1 h_1 + p_2 h_2) \tag{*3.4.5}$$

subject to

$$h_1 x_1 + h_2 x_2 \leq z$$

The constraint implies that h_1 and h_2 satisfy

$$h_1 + h_2 \leq 0, \tag{*3.4.6}$$

$$h_1 + 2h_2 \leq 0, \tag{*3.4.7}$$

$$h_1 + 4h_2 \leq 1. \tag{*3.4.8}$$

The linear program (*3.4.5) can easily be solved graphically.

One can also argue as follows: because there are two choice variables, it is permissible to assume that at the solution at least two of the constraints are satisfied with equality. Constraints (*3.4.6) and (*3.4.7) are satisfied at equality by $h_1 = h_2 = 0$, at which point constraint (*3.4.8) is satisfied. Constraints (*3.4.6) and (*3.4.8) are satisfied at equality by $h_1 = -1/3, h_2 = 1/3$, at which point constraint (*3.4.7) is violated. Constraints (*3.4.7) and (*3.4.8) are satisfied at equality by $h_1 = -1, h_2 = 1/2$, at which point constraint (*3.4.6) is satisfied.

The two points at which two of the constraints are satisfied as equalities and the third constraint is satisfied both give portfolios with zero price, and thus zero is the lower bound for the value of the call.

The upper bound on the value of the call is determined by solving

$$\min_{h_1,h_2} (p_1 h_1 + p_2 h_2)$$

subject to

$$h_1 + h_2 \geq 0, \tag{*3.4.9}$$

$$h_1 + 2h_2 \geq 0, \tag{*3.4.10}$$

$$h_1 + 4h_2 \geq 1. \tag{*3.4.11}$$

As earlier, the minimum is attained at a point at which at least two of the constraints are satisfied with equality. Because constraints (*3.4.9)-(*3.4.11) are the reverse inequalities to (*3.4.6)-(*3.4.8), the only point that satisfies two of the constraints with equality is $h_1 = -1/3, h_2 = 1/3$. The price of this portfolio is $1/6$. Thus, the bounds on the value of the call option are zero and $1/6$. □

Important properties of the bounds q_ℓ and q_u are given in the following propositions.

Proposition *3.4.2. *If security prices exclude strong arbitrage, then $q_u(z) \geq q_\ell(z)$ for every contingent claim $z \in \mathbb{R}^S$. Further, $q_u(z) > -\infty$ and $q_\ell(z) < +\infty$ for every $z \in \mathbb{R}^S$.*

Proof. Suppose that $q_u(z) < q_\ell(z)$ for some $z \in \mathbb{R}^S$. Then $q_u(z) < +\infty$ and $q_\ell(z) > -\infty$, and there exist portfolios h' and h'' such that

$$h'X \leq z \leq h''X$$

and

$$ph' > ph''.$$

But then the portfolio $h'' - h'$ satisfies $(h'' - h')\, X \geq 0$ and $p\,(h'' - h') < 0$, and thus it is a strong arbitrage. This contradicts the assumption.

As to the second part, we can assume that there are no redundant securities. If there are redundant securities, the absence of arbitrage implies that the law of one price holds, and the payoff pricing functional and the upper and lower bounds in the markets with a

smaller subset of nonredundant securities are the same as with the full set of securities.

Suppose by contradiction that $q_u(z) = -\infty$ for some $z \in \mathbb{R}^S$. Then there exists a sequence of portfolios $\{h^n\}$ such that

$$h^n X \geq z \tag{*3.4.12}$$

and

$$\lim_{n\to\infty} ph^n = -\infty. \tag{*3.4.13}$$

Equation (*3.4.13) implies that sequence $\{h^n\}$ is unbounded, that is, $\lim \|h^n\| = +\infty$ where $\|h^n\|$ denotes the Euclidean norm of h^n. Consider the bounded sequence $\{h^n / \|h^n\|\}$ and its nonzero limit $\hat{h}$. Dividing both sides of inequality (*3.4.12) by $\|h^n\|$ and taking limits as n goes to infinity, we obtain

$$\hat{h}X \geq 0. \tag{*3.4.14}$$

Further, equation (*3.4.13) implies that

$$p\hat{h} \leq 0. \tag{*3.4.15}$$

Because portfolio $\hat{h}$ is nonzero and there are no redundant securities, its payoff is nonzero and inequalities (*3.4.14) and (*3.4.15) imply that $\hat{h}$ is an arbitrage. This is a contradiction.

The proof that $q_\ell(z) < +\infty$ is similar. □

Also we have the following proposition.

Proposition *3.4.3. *If security prices exclude arbitrage, then $q_u(z) > q_\ell(z)$ for every contingent claim z not in the asset span.*

Proof. In view of Proposition *3.4.2, we only have to prove that $q_u(z) \neq q_\ell(z)$ for every $z \notin \mathcal{M}$. Suppose that $q_u(z) = q_\ell(z)$ for some $z \notin \mathcal{M}$. It follows from the second part of Proposition *3.4.2 that $q_u(z) < +\infty$ and $q_\ell(z) > -\infty$. Therefore there exist portfolios h'

and h'' such that

$$h'X \leq z \leq h''X \tag{*3.4.16}$$

and

$$ph' = ph'' = q_u(z)$$

Neither of the weak inequalities in expression (*3.4.16) can be an equality because z is not in the asset span; that is, it cannot be generated by a portfolio. Consequently, $(h'' - h')\,X > 0$, and $p\,(h'' - h') = 0$, and thus the portfolio $h'' - h'$ is an arbitrage. This is a contradiction. □

The Extension

Having derived upper and lower bounds on the value of any contingent claim, we turn now to how these bounds are used to extend the payoff pricing functional.

Fix a contingent $\text{claim}\ \hat{z} \notin \mathcal{M}$. Define $\mathcal{N}$ by

$$\mathcal{N} = \{z + \lambda\hat{z} : z \in \mathcal{M} \text{ and } \lambda \in \mathbb{R}\}$$

Thus $\mathcal{N}$ is the subspace of $\mathbb{R}^S$ that has dimension equal to the dimension of $\mathcal{M}$ plus one and contains $\mathcal{M}$ and $\hat{z}$. It is the asset span of $J + 1$ securities with payoffs $\{x_1, \ldots, x_J\}$ and $\hat{z}$.

If there is no strong arbitrage – equivalently, if the payoff pricing functional q is positive – then Proposition *3.4.2 implies that a finite value π can be chosen to satisfy

$$q_\ell(\hat{z}) \leq \pi \leq q_u(\hat{z}).$$

We extend q to a linear functional on $\mathcal{N}$ in that we define $Q : \mathcal{N} \to \mathbb{R}$ by

$$Q(z + \lambda\hat{z}) := q(z) + \lambda\pi. \tag{*3.4.17}$$

We now prove that Q, as just defined, is the desired positive extension of q.

Proposition *3.4.4. *If $q : \mathcal{M} \to \mathbb{R}$ is positive, so is $Q : \mathcal{N} \to \mathbb{R}$.*

Proof. Let $y \in \mathcal{N}$. Then

$$y = z + \lambda\hat{z}$$

for some $z \in \mathcal{M}$ and some $\lambda \in \mathbb{R}$. Of the three possibilities for λ, suppose first that $\lambda > 0$. Then $y \geq 0$ implies

$$\hat{z} \geq -\frac{z}{\lambda}. \tag{*3.4.18}$$

If we apply q_ℓ to both sides of inequality (*3.4.18) and use the implication of definition (*3.4.2) that q_ℓ is an increasing function, the result is

$$q_\ell(\hat{z}) \geq q_\ell\left(-\frac{z}{\lambda}\right) \tag{*3.4.19}$$

By Proposition *3.4.1, the functions q and q_ℓ coincide on $\mathcal{M}$. Because $-z/\lambda \in \mathcal{M}$, we have $q_\ell(-z/\lambda) = q(-z/\lambda)$. Therefore, inequality (*3.4.19) becomes

$$q_\ell(\hat{z}) \geq q\left(-\frac{z}{\lambda}\right) \tag{*3.4.20}$$

Because $\pi \geq q_\ell(\hat{z})$, inequality (*3.4.20) implies that

$$\pi \geq q\left(-\frac{z}{\lambda}\right)$$

or alternatively that

$$q(z) + \lambda\pi \geq 0. \tag{*3.4.21}$$

Because the left-hand side of inequality (*3.4.21) equals $Q(y)$, we obtain that $Q(y) \geq 0$.

If $\lambda < 0$, a similar argument, but using q_u and the fact that $\pi \leq q_u(\hat{z})$, also gives $Q(y) \geq 0$. Finally, if $\lambda = 0$, then $y = z$ and $Q(y) = q(z)$. The positivity of q implies that if $y \geq 0$, then $Q(y) \geq 0$. □

If there is no arbitrage — equivalently, if q is strictly positive — then Proposition *3.4.3 implies that π can be chosen to satisfy

$$q_\ell(\hat{z}) < \pi < q_u(\hat{z}).$$

Then the following holds true.

Proposition *3.4.5. *If $q : \mathcal{M} \to \mathbb{R}$ is strictly positive, so is $Q : \mathcal{N} \to \mathbb{R}$.*

The proof is essentially the same as the proof of Proposition *3.4.4.

For the prices $\{p_1, \ldots, p_J\}$ and π, functional Q, as defined in equation (*3.4.17), is the payoff pricing functional on $\mathcal{N}$. Therefore Q is strictly positive (positive) on $\mathcal{N}$ iff the indicated prices exclude arbitrage (strong arbitrage) in $J+1$ securities markets with payoffs $\{x_1, \ldots, x_J\}$ and $\hat{z}$.

Example 3.2. In example 3.2, define

$$\mathcal{N} = \{z + \lambda\hat{z} : z \in \mathcal{M}, \lambda \in \mathbb{R}\}$$

where $\mathcal{M} = \text{span}\{(1, 2)\}$, and $\hat{z} = (1, 1)$. Thus $\mathcal{N} = \mathbb{R}^2$. We have the following bounds on the value π of $\hat{z}$ (see equations (*3.4.3) and (*3.4.4)):

$$\frac{1}{2} \leq \pi \leq 1$$

We choose $\pi = 3/4$ and define $Q : \mathcal{N} \to \mathbb{R}$ by

$$Q(z + \lambda\hat{z}) = q(z) + \frac{3}{4}\lambda$$

for $z \in \mathcal{M}$ and $\lambda \in \mathbb{R}$. Recall that $q(z) = \alpha$ for $z = (\alpha, 2\alpha)$. One can easily check that

$$Q(1, 0) = \frac{1}{2} \quad \text{and} \quad Q(0, 1) = \frac{1}{4}$$

and hence that

$$Q(y_1, y_2) = \frac{1}{2}y_1 + \frac{1}{4}y_2$$

Thus, Q is strictly positive. □

3.4.3 Uniqueness of the Valuation Functional

Extending the payoff pricing functional into a valuation functional does not result in a unique valuation functional when markets are incomplete. When markets are incomplete, if security prices exclude arbitrage, then there exists a continuum of values of π that define strictly positive extensions.

When markets are complete, the asset span $\mathcal{M}$ equals the contingent claim space $\mathbb{R}^S$, and the payoff pricing functional is the valuation functional. It turns out that this is the only case of a unique strictly positive valuation functional.

Theorem 3.4.3. *Suppose that security prices exclude arbitrage. Then security markets are complete iff there exists a unique strictly positive valuation functional.*

Proof. The sufficiency of market completeness for the uniqueness of the valuation operator is obvious. Necessity follows from Proposition *3.4.3. If markets are not complete, so that there exists a contingent claim not in the asset span, then there exists a nondegenerate interval of values of that contingent claim that gives rise to different strictly positive valuation functionals. □

Theorem 3.4.3 does not extend to security prices that exclude strong arbitrage, but do not exclude arbitrage.

Example 3.4.0.5. Suppose that there are two states and a single risk-free security with payoff $x_1 = (1, 1)$ and price $p_1 = 0$. Markets are incomplete in this example, and security prices exclude strong arbitrage, but they permit arbitrage. The payoff pricing functional is the zero functional on the asset span $\mathcal{M} = \{(\alpha, \alpha) \colon \alpha \in \mathbb{R}\}$. Further, the upper and the lower bounds on the value of any contingent claim in $\mathbb{R}^2$ are zero. Therefore, the only positive extension of the payoff pricing functional is the zero functional on $\mathbb{R}^2$. Thus we have uniqueness of the valuation operator without complete markets. □

3.5 State Prices and Risk-Neutral Probabilities

3.5.1 Introduction

By the fundamental theorem of finance, the payoff pricing functional can be extended to a strictly positive (positive) valuation functional iff security prices exclude arbitrage (strong arbitrage). We now show that each strictly positive (positive) valuation functional can be represented by a vector of strictly positive (positive) state prices. State prices can easily be calculated as a strictly positive (positive) solution to a system of linear equations relating

security prices and their payoffs. An implication of the existence of strictly positive (positive) state prices is the absence of arbitrage (strong arbitrage). An implication of the uniqueness of state prices is that markets are complete.

The valuation functional can also be represented by strictly positive (positive) probabilities of the states. These probabilities, known as risk-neutral probabilities, are simple transforms of the state prices and therefore are just as useful as those prices. Under the risk-neutral probabilities representation, the price of each security equals its expected payoff discounted by the risk-free return.

3.5.2 State Prices

If markets are complete, the payoff pricing functional q is defined on the entire contingent claim space $\mathbb{R}^S$, and the state price vector $q = (q_1, \ldots, q_S)$ provides a representation of the functional q as $q(z) = qz$ for every payoff $z \in \mathbb{R}^S$. We now extended the same kind of representation to incomplete markets using the valuation functional rather than the payoff pricing functional.

A valuation functional, being a linear functional on $\mathbb{R}^S$, can be identified by its values on the basis vectors of that space. Let

$$q_s := Q(e_s) \tag{3.5.1}$$

for every s, where e_s is the state claim for state s. The value q_s is the state price of state s. If Q is strictly positive (positive), then each state price q_s is strictly positive (positive).

Because every contingent claim $z \in \mathbb{R}^S$ can be written as $z = \sum_s z_s e_s$, we have

$$Q(z) = \sum_s z_s Q(e_s) = \sum_s z_s q_s \tag{3.5.2}$$

or

$$Q(z) = qz. \tag{3.5.3}$$

Equation (3.5.3) is the state-price representation of the valuation functional Q. It defines a one-to-one relation between valuation functionals and state-price vectors. Because the

valuation functional in incomplete markets is not unique (Theorem 3.4.3), state prices are not unique either.

Equation (3.5.3) provides a simple method for pricing payoffs without determining a portfolio that generates the payoff under consideration. Once state prices are known, the price of every payoff can be obtained. Equation (3.5.3) can also be applied to contingent claims not in the asset span, although for any such claim the derived value will depend on the state-price vector used. It follows from the proof of the fundamental theorem of finance that the derived value is independent of the state-price vector iff the contingent claim lies in the asset span.

State prices can be characterized as solutions to a system of linear equations just as under complete markets. To see this we apply equation (3.5.3) to the payoff x_j of security j. Because $Q(x_j) = p_j$, we obtain

$$p_j = qx_j, \tag{3.5.4}$$

or in vector-matrix notation

$$p = Xq \tag{3.5.5}$$

State prices are a solution to the system of J equations (3.5.5) with S unknowns q_s. Strictly positive state prices are a strictly positive solution; positive state prices are a positive solution. If markets are incomplete, then the payoff matrix X has rank less than S, and the independent equations of (3.5.5) are fewer in number than the number of unknowns. If markets are complete, then state prices are unique. Of course, if markets are incomplete there are also nonpositive solutions to equation (3.5.5), but they do not qualify as state prices.

We have the following:

Theorem 3.5.1. *There exists a strictly positive valuation functional iff there exists a strictly positive solution to equation (3.5.5). Each strictly positive solution q defines a strictly positive valuation functional Q satisfying $Q(z) = qz$ for every $z \in \mathbb{R}^S$.*

Proof. It was proven in equations (3.5.1)-(3.5.5) that state prices associated with a strictly

positive valuation functional are a solution to equation (3.5.5). Existence of a valuation functional follows from the fact that, if q is a strictly positive solution to equation (3.5.5), then the functional Q defined by $Q(z) = qz$ is linear and strictly positive. Whenever $z \in \mathcal{M}$, then $z = hX$ for some portfolio h, and $Q(z) = qz = hXq = ph$ (that is, Q coincides with the payoff pricing functional on $\mathcal{M}$). Thus, Q is a strictly positive valuation functional. □

Similarly, the following theorem holds.

Theorem 3.5.2. *There exists a positive valuation functional iff there exists a positive solution to equation (3.5.5). Each positive solution q defines a positive valuation functional Q satisfying $Q(z) = qz$ for every $z \in \mathbb{R}^S$.*

Theorems 3.5.1 and 3.5.2 say that state-price vectors can be defined either as the values of the state claims under valuation functionals, as in equation (3.5.1), or as a strictly positive (positive) solution to equation (3.5.5). The fundamental theorem of finance can be restated to say that security prices exclude arbitrage (strong arbitrage) iff there exists a strictly positive (positive) state-price vector.

Example 3.5.0.1. In example 3.2, there were two securities: a risk-free bond with payoff $x_1 = (1, 1, 1)$ and price $p_1 = 1/2$ and a risky stock with payoff $x_2 = (1, 2, 4)$ and price $p_2 = 1$. Positive state prices q_1, q_2, q_3 are a positive solution to the system of two equations

$$q_1 + q_2 + q_3 = \frac{1}{2}$$

and

$$q_1 + 2q_2 + 4q_3 = 1$$

Using q_3 as a parameter (we have two equations and three unknowns), the solution is

$$q_1 = 2q_3, \quad q_2 = \frac{1}{2} - 3q_3$$

For state prices to be positive, we must have $0 \leq q_3 \leq 1/6$. If $0 < q_3 < 1/6$, then state prices are strictly positive. The existence of a strictly positive solution verifies that security prices $p_1 = 1/2$ and $p_2 = 1$ exclude arbitrage. □

3.5.3 Visual Representation

In Chapter 3 we presented a visual analysis of security prices for two securities. It was shown that security prices exclude strong arbitrage whenever the price vector lies in the convex cone generated by the vectors of payoffs of the two securities in each state. Security prices exclude arbitrage whenever the vector of security prices lies in the interior of that cone. That is precisely the diagrammatic interpretation of the existence of strictly positive (positive) state prices. Equation (3.5.5) with positive state prices q_s means that the vector of security prices p lies in the cone generated by vectors $x_{\cdot s} = (x_{1s}, \ldots, x_{Js})$ in $\mathbb{R}^J$. If the state prices are strictly positive, then vector p lies in the interior of that cone.

3.5.4 Risk-Free Payoffs

A contingent claim that does not depend on the state is risk free. If markets are complete, risk-free claims are necessarily in the asset span. If markets are incomplete, it may or may not be possible to construct a portfolio with a nonzero risk-free payoff.

If a nonzero risk-free payoff lies in the asset span, then all risk-free payoffs lie in the asset span, and as long as the law of one price holds, they all have the same return. We denote that risk-free return by $\bar{r}$. It follows from equation (3.5.4) that $\bar{r}$ satisfies

$$\bar{r} = \frac{1}{\sum_s q_s}$$

3.5.5 Risk-Neutral Probabilities

Suppose that security prices exclude arbitrage (strong arbitrage) and that a risk-free payoff with strictly positive return $\bar{r}$ lies in the asset span. Let q be a strictly positive (positive) state price vector. Define

$$\pi_s^* := \bar{r} q_s = \frac{q_s}{\sum_s q_s}$$

for every s. So defined, the π_s^*'s are strictly positive (positive) and sum to one. It is natural to interpret them as probabilities. We call them risk-neutral probabilities.

When equipped with risk-neutral probabilities, the set of states S can be regarded as a

probability space. Date-1 consumption plans, security payoffs, contingent claims, and others, which we have thus far regarded as vectors with S components, can now be regarded as random variables on the probability space S. Here and throughout these notes we make no distinction in notation between a random variable and the vector of values the random variables take on.

Let E^* denote the expectation with respect to the probabilities π^*. Then $E^*(z) = \sum_s \pi_s^* z_s$ for a contingent claim z. We have

$$qz = \sum_s q_s z_s = \frac{1}{\bar{r}} \sum_s \pi_s^* z_s = \frac{1}{\bar{r}} E^*(z) \tag{3.5.6}$$

Applying equation (3.5.6) to $z = x_j$ and using equation (3.5.5), we obtain

$$p_j = \frac{1}{\bar{r}} E^*\left(x_j\right) \tag{3.5.7}$$

for every security j.

Equation (3.5.7) says that the price of each security equals the expectation of its payoff with respect to probabilities π^* discounted by the risk-free return. We emphasize that the expectation is taken with respect to probabilities π^* derived from state prices, rather than from agents' subjective probabilities.

Equation (3.5.7) can also be written in terms of returns as

$$\bar{r} = E^*(r_j).$$

Using equations (3.5.6) and (3.5.3), we obtain

$$Q(z) = \frac{1}{\bar{r}} E^*(z) \tag{3.5.8}$$

for every $z \in \mathbb{R}^S$. Equation (3.5.8) is the representation of the valuation functional Q by risk-neutral probabilities. The value of each contingent claim equals the discounted expectation of the claim with respect to risk-neutral probabilities.

Because risk-neutral probabilities are rescaled state prices, they have all the properties of those prices. They are characterized as strictly positive (positive) solutions to equation (3.5.7). Their existence and strict positivity (positivity) are equivalent to the absence of

arbitrage (strong arbitrage); their uniqueness is equivalent to market completeness.

Example 3.5.0.2. The risk-neutral probabilities of Example 3.5.0.1 can be derived by multiplying state prices by the risk-free return $\bar{r}$. Because $\bar{r} = 2$, we have

$$\pi_1^* = 2\pi_3^*, \quad \pi_2^* = 1 - 3\pi_3^*, \quad \text{and} \quad 0 \leq \pi_3^* \leq \frac{1}{3}.$$

Because state prices are not unique, neither are risk-neutral probabilities.

Risk-neutral probabilities can also by derived directly from the system of equations (3.5.7); that is,

$$1 = \pi_1^* + \pi_2^* + \pi_3^*,$$

and

$$2 = \pi_1^* + 2\pi_2^* + 4\pi_3^*.$$

□

3.6 Optimal Portfolios with One Risky Security

An agent's willingness to invest in a risky security depends on, among other things, the expected return of that security relative to the return on a risk-free investment. In this section, we analyze agents' optimal portfolios in a simple setting of two securities: a single risky security and a risk-free security.

We assume agents' utility functions have an expected utility representation with strictly increasing and twice differentiable utility functions. We also assume that date-0 consumption does not enter agents' utility functions. Furthermore, we assume that their endowments at date 1 are in the asset span. When the endowments are in the asset span, we refer to this economy as a securities market economy. Finally, we assume that optimal portfolios exist, except where otherwise indicated.

3.6.1 Portfolio Choice and Wealth

The consumption-portfolio choice problem of an agent with a strictly increasing expected utility function that depends only on date-1 consumption can be written as

$$\max_{c_1,h} E[v(c_1)] \tag{3.6.1}$$

subject to

$$ph = w_0 \tag{3.6.2}$$

and

$$c_1 = w_1 + hX, \tag{3.6.3}$$

with an additional restriction on consumption if such is imposed (for example, consumption must be positive when working with some utility functions). The date-1 consumption plan c_1 and the date-1 endowment w_1 in (3.6.2)-(3.6.4) are understood as scalar random variables on the set of states S with probability measure π. Security payoffs X are understood as a J-dimensional random variable.

A change in notation will facilitate the analysis of optimal portfolios. If, as assumed, the agent's date-1 endowment lies in the asset span, then we have $w_1 = \hat{h}X$ for some portfolio $\hat{h}$. Date-1 budget constraint (3.6.4) can be written as

$$c_1 = (h + \hat{h})X. \tag{3.6.4}$$

The agent's wealth is defined as the sum of his date-0 endowment plus the price of the portfolio generating his date-1 endowment:

$$w := w_0 + p\hat{h}. \tag{3.6.5}$$

Note that the price of portfolio $\hat{h}$ equals the value of the date-1 endowment w_1 under the payoff pricing functional, that is, $p\hat{h} = q\,(w_1)$. Unless the agent's date-1 endowment is zero, wealth w depends on security prices.

The date-0 budget constraint (3.6.3) can be written as

$$p(h + \hat{h}) = w \tag{3.6.6}$$

Let a_j denote the amount of wealth invested in security j, that is,

$$a_j = p_j(\hat{h}_j + h_j)$$

equation (3.6.6) can be written as

$$c_1 = \sum_{j=1}^{J} \frac{a_j x_j}{p_j} = \sum_{j=1}^{J} a_j r_j$$

where r_j is the return on security j.

Summing up, we obtain the following portfolio choice problem

$$\max_{\{a_j\}} E\left[v(\sum_{j=1}^{J} a_j r_j)\right] \tag{3.6.7}$$

subject to

$$\sum_{j=1}^{J} a_j = w. \tag{3.6.8}$$

If the agent is strictly risk averse, then the optimal consumption plan $c_1 = \sum_j a_j r_j$ is unique. Two distinct consumption plans cannot both be optimal, because any strictly convex combination of the two would also be budget feasible and would yield strictly higher expected utility. If, in addition, there are no redundant securities, then the agent's optimal portfolio is also unique.

If one of the securities, say security 1, is risk free with return $\bar{r}$, then it is convenient to solve the budget constraint (3.6.8) for

$$a_1 = w - \sum_{j=2}^{J} a_j \tag{3.6.9}$$

and substitute (3.6.9) in the objective (3.6.7). Thus the agent's portfolio choice problem consists of solving

$$\max_{a_2 \ldots a_J} E\left[v\left(w\bar{r} + \sum_{j=2}^{J} a_j(r_j - \bar{r})\right)\right] \tag{3.6.10}$$

The maximization is constrained only by the requirement that consumption lie in the specified consumption set.

3.6.2 Optimal Portfolios with One Risky Security

Let there be one risky security with return r. The difference $r - \bar{r}$ between r and the return on the risk-free security, which is assumed to be nonzero, is the excess return on the risky security. The agent's optimal investment in the risky security, denoted by a^*, is a solution to the problem

$$\max_a E[v(w\bar{r} + (r - \bar{r})a)], \tag{3.6.11}$$

which, as noted, may involve an additional restriction that consumption be positive: $w\bar{r} + (r - \bar{r})a > 0$. The agent's wealth w is assumed to be strictly positive.

If security prices exclude arbitrage and if consumption is restricted to be positive, then maximization problem (3.6.11) has a solution (which we do not prove here). In the present context of two securities, one of which is risk free, the condition that there be no arbitrage has a simple characterization in terms of securities' returns. The risky return r must be lower than the risk-free return $\bar{r}$ in some states and higher in other states. Otherwise, if r is uniformly above $\bar{r}$, then $r - \bar{r}$ is an arbitrage, and if r is uniformly below $\bar{r}$, then $\bar{r} - r$ is an arbitrage.

If the agent is strictly risk averse, the optimal investment is unique because in the present setting neither security is redundant. The optimal investment a^* is then a function of the agent's wealth w, the risk-free return $\bar{r}$, and the (distribution of the) risky return r. Further, because utility function v is twice differentiable, a^* is a differentiable function of its arguments whenever the consumption plan generated by a^* is interior.

The interior optimal investment a^* satisfies the first-order condition

$$E\left[v'\left(w\bar{r} + a^*(r - \bar{r})\right)(r - \bar{r})\right] = 0.$$

3.6.3 Risk Premium and Optimal Portfolios

The risk premium on a security is defined as its expected excess return; that is, its expected return less the risk-free return. If the risk premium is zero, then the security is said to be priced fairly, meaning that the excess return on the security is a fair game; that is, a random variable with zero expectation. Of course, there is no suggestion that anything is unfair about nonzero risk premia.

A risk-neutral agent is indifferent among all portfolios if the risk premium on the risky security is zero. If the risk premium is nonzero and there are no restrictions on consumption, then her optimal investment does not exist. If her consumption is restricted to be positive, then the agent will hold long the security with high expected return and sell short the security with low expected return until the positivity restriction becomes binding.

Whether a strictly risk-averse agent chooses a positive or a negative investment in the risky security depends on the risk premium on the risky security.

Theorem 3.6.1. *If an agent is strictly risk averse, then the optimal investment in the risky security is strictly positive, zero, or strictly negative iff the risk premium on the risky security is strictly positive, zero, or strictly negative.*

Proof. Because w is strictly positive, zero investment in the risky security results in a strictly positive risk-free consumption. Therefore $a = 0$ is an interior point of the interval of the investment choices whether or not consumption is restricted to be positive. The derivative of expected utility in maximization problem (3.6.11) with respect to a at $a = 0$ is $v'(w\bar{r})(\mu - \bar{r})$. Because $v'(w\bar{r})$ is strictly positive, the derivative is strictly positive, zero, or strictly negative iff $\mu - \bar{r}$ is strictly positive, zero, or strictly negative. Because expected utility is strictly concave in a, the sign of the derivative at zero investment determines whether the optimal investment is positive, zero, or negative. □

It is important to keep in mind that the optimal investment a^* characterized in Theorem 3.6.1 is the part of total wealth w invested in the risky security. Because w consists of the date-0 endowment plus the price of the portfolio the agent is endowed with (see equation

(3.6.5)), zero investment a^* means that the agent sells all of the shares of the risky security that he is endowed with and invests the proceeds in the risk-free security.

3.7 The Expectations and Pricing Kernels

The payoff pricing functional — and also its extension, the valuation functional — can be represented either by state prices or by risk-neutral probabilities. We now derive another representation of the payoff pricing functional, the pricing kernel. The existence of the pricing kernel is a consequence of the Riesz representation theorem, which, in the present context, says that any linear functional on a vector space can be represented by a vector in that space.

The Riesz representation theorem is a powerful tool in many areas, including in machine learning and econometrics. For the finite-dimensional contingent claims space $\mathbb{R}^S$, the Riesz representation theorem is not really needed since we can use standard methods of Euclidean geometry (mostly, projections). However, when we later consider state spaces that are infinite-dimensional, we will model them as Hilbert spaces and be required to use the Riesz representation theorem. Hilbert spaces are generalizations of $\mathbb{R}^S$ that have enough structure to "behave like" $\mathbb{R}^S$ in many ways (especially in the way projections work).

3.7.1 Hilbert Spaces and Inner Products

An inner product or scalar product on a vector space $\mathcal{H}$ is a function from $\mathcal{H} \times \mathcal{H}$ to $\mathbb{R}$. It is usually indicated by a dot, and therefore is often termed a dot product. Inner products obey the following properties for all $x, y, z \in \mathcal{H}$ and all $a, b \in \mathbb{R}$:

- symmetry: $x \cdot y = y \cdot x$,
- linearity: $x \cdot (ay + bz) = a(x \cdot y) + b(x \cdot z)$,
- strict positivity: $x \cdot x > 0$ when $x \neq 0$.

The inner product defines a norm of a vector in the vector space $\mathcal{H}$ as

$$\|x\| := \sqrt{x \cdot x}$$

The norm satisfies the following important properties for all $x, y \in \mathcal{H}$:

- triangle inequality: $\|x + y\| \leq \|x\| + \|y\|$,
- Cauchy-Schwarz inequality: $|x \cdot y| \leq \|x\|\|y\|$.

Further, the norm defines the convergence of a sequence of vectors in $\mathcal{H}$ and therefore the continuity of functionals on $\mathcal{H}$.

A Hilbert space is a vector space $\mathcal{H}$ that is equipped with an inner product and is complete with respect to the norm induced by its inner product. In this context, completeness means that any Cauchy sequence of elements of the vector space $\mathcal{H}$ converges to an element of that space.

3.7.2 The Expectations Inner Product

The space $\mathbb{R}^S$ of state-contingent date-1 consumption plans is a Hilbert space. The most familiar inner product in that space is the Euclidean inner product:

$$x \cdot y = \sum_s x_s y_s$$

Another inner product, important in the derivation of the capital asset pricing model, is the expectations inner product,

$$x \cdot y = E(xy)$$

where, as usual, $E(xy) = \sum_s \pi_s x_s y_s$ for a probability measure π on S. The norm induced by the expectations inner product is

$$\|x\| = \sqrt{E\left(x^2\right)} = \sqrt{\text{var}(x) + [E(x)]^2}.$$

The Cauchy-Schwarz inequality for the expectations inner product is

$$|E(xy)| \leq \sqrt{E\left(x^2\right) E\left(y^2\right)},$$

and implies, when applied to $x - E(x)$ and $y - E(y)$, that the correlation $\rho(x, y)$ between x and y is less than one in absolute value.

3.7.3 Orthogonal Vectors

Two vectors $x, y \in \mathcal{H}$ are orthogonal, denoted by $x \perp y$, iff their inner product is zero:

$$x \perp y \text{ iff } x \cdot y = 0.$$

A collection of vectors $\{z_1, \ldots, z_n\}$ in a Hilbert space $\mathcal{H}$ is an orthogonal system if $z_i \perp z_j$ for all $i \neq j$. If in addition $\|z_i\| = 1$ for every i, then the collection $\{z_1, \ldots, z_n\}$ is an orthonormal system. An orthonormal system is an orthonormal basis for its linear span.

Theorem 3.7.1 (Pythagorean Theorem)**.** *If $\{z_1, \ldots, z_n\}$ is an orthogonal system in a Hilbert space $\mathcal{H}$, then*

$$\left\| \sum_{i=1}^{n} z_i \right\|^2 = \sum_{i=1}^{n} \|z_i\|^2 \tag{3.7.1}$$

Proof. Write the left-hand side using the inner product and apply the definition of orthogonality. □

A useful implication of the Pythagorean theorem is the following:

Theorem 3.7.2. *Any orthogonal system of nonzero vectors is linearly independent.*

Proof. Let $\{z_1, \ldots, z_n\}$ be an orthogonal system with $z_i \neq 0$ for each i. Suppose that

$$\sum_{i=1}^{n} \lambda_i z_i = 0 \tag{3.7.2}$$

for some $\lambda_i \in \mathbb{R}$. Because $\{\lambda_1 z_1, \ldots, \lambda_n z_n\}$ is also an orthogonal system, it follows from equations (3.7.1) and (3.7.2) that

$$\sum_{i=1}^{n} \lambda_i^2 \|z_i\|^2 = \left\| \sum_{i=1}^{n} \lambda_i z_i \right\|^2 = 0$$

This implies that $\lambda_i = 0$ for every i, and thus that the vectors $z_1, \ldots, z_n$ are linearly independent. □

3.7.4 Orthogonal Projections

A vector $x \in \mathcal{H}$ is orthogonal to a linear subspace $\mathcal{Z} \subset \mathcal{H}$ iff it is orthogonal to every vector in $z \in \mathcal{Z}$:

$$x \perp \mathcal{Z} \text{ iff } x \cdot z = 0 \quad \forall z \in \mathcal{Z}.$$

If the subspace $\mathcal{Z}$ is the linear span of vectors $z_1, \ldots, z_n$, then a vector x is orthogonal to $\mathcal{Z}$ iff it is orthogonal to every z_i for $i = 1, \ldots, n$. The set of all vectors orthogonal to a subspace $\mathcal{Z}$ is the orthogonal complement of $\mathcal{Z}$ and is denoted $\mathcal{Z}^\perp$. It is a linear subspace of $\mathcal{H}$.

Theorem 3.7.3. *For any finite-dimensional subspace $\mathcal{Z}$ of a Hilbert space $\mathcal{H}$ and any vector $x \in \mathcal{H}$, there exist unique vectors $x^{\mathcal{Z}} \in \mathcal{Z}$ and $y \in \mathcal{Z}^\perp$ such that $x = x^{\mathcal{Z}} + y$.*

Proof. We only prove the theorem in the finite-dimensional case. Let $\{z_1, \ldots, z_n\}$ be an orthogonal basis for $\mathcal{Z}$, and define

$$x^{\mathcal{Z}} = \sum_{i=1}^{n} \frac{x \cdot z_i}{z_i \cdot z_i} z_i \tag{3.7.3}$$

and

$$y = x - x^{\mathcal{Z}}.$$

The vector $x^{\mathcal{Z}}$ so defined is in $\mathcal{Z}$. We have

$$y \cdot z_j = \left(x - \sum_{i=1}^{n} \frac{x \cdot z_i}{z_i \cdot z_i} z_i \right) \cdot z_j \tag{3.7.4}$$

$$= \left(x - \frac{x \cdot z_j}{z_j \cdot z_j} z_j \right) \cdot z_j = 0. \tag{3.7.5}$$

Therefore $y \perp z_j$ for every $j = 1, \ldots, n$. Hence, $y \in \mathcal{Z}^\perp$.

To see that $x^{\mathcal{Z}}$ is unique, suppose that $x = x_1^{\mathcal{Z}} + y_1 = x_2^{\mathcal{Z}} + y_2$ for some $x_1^{\mathcal{Z}}, x_2^{\mathcal{Z}} \in \mathcal{Z}$ and $y_1, y_2 \in \mathcal{Z}^\perp$. The Pythagorean Theorem implies

$$\|y_2\|^2 = \left\| x_1^{\mathcal{Z}} - x_2^{\mathcal{Z}} \right\|^2 + \|y_1\|^2, \tag{3.7.6}$$

and

$$\|y_1\|^2 = \left\|x_1^{\mathcal{Z}} - x_2^{\mathcal{Z}}\right\|^2 + \|y_2\|^2 . \tag{3.7.7}$$

Equation (3.7.6) implies that $\|y_2\| \geq \|y_1\|$, and equation (3.7.7) implies that $\|y_2\| \leq \|y_1\|$. It follows that $\|y_1\| = \|y_2\|$, and therefore also that

$$\left\|x_1^{\mathcal{Z}} - x_2^{\mathcal{Z}}\right\|^2 = 0$$

thus, by the strict positivity of inner products, $x_1^{\mathcal{Z}} = x_2^{\mathcal{Z}}$. □

If $\mathcal{Z}$ is a (finite-dimensional) subspace of a Hilbert space $\mathcal{H}$, then Theorem 3.7.3 implies that $\mathcal{H}$ can be decomposed as $\mathcal{H} = \mathcal{Z} + \mathcal{Z}^{\perp}$, with $\mathcal{Z} \cap \mathcal{Z}^{\perp} = \{0\}$.

Vector $x^{\mathcal{Z}}$ of the unique decomposition of Theorem 3.7.3 is the orthogonal projection of x on $\mathcal{Z}$. If the projection is taken with respect to the expectations inner product, then the coefficients of the representation (3.7.3) of the orthogonal projection are

$$\frac{x \cdot z_i}{z_i \cdot z_i} = \frac{E(xz_i)}{E(z_i^2)},$$

and we have

$$x^{\mathcal{Z}} = \sum_{i=1}^{n} \frac{E(xz_i)}{E(z_i^2)} z_i. \tag{3.7.8}$$

Thus, the projection with respect to the expectations inner product is the same as the linear regression of x on the z_i's. Equation (3.7.8) is the equation for the predicted value of the dependent variable for given values of the independent variables.

Example 3.7.0.1. In the Hilbert space $\mathbb{R}^2$ with the expectations inner product given by probabilities $(1/4, 3/4)$, let $\mathcal{Z} = \text{span}\{(1,1)\}$ and $x = (1,2)$. The orthogonal projection $x^{\mathcal{Z}}$ is

$$x^{\mathcal{Z}} = \frac{(1,2)\cdot(1,1)}{(1,1)\cdot(1,1)}(1,1) = \frac{7}{4}(1,1) = (7/4, 7/4)$$

□

3.7.5 Diagrammatic Methods in Hilbert Spaces

One of the most appealing features of Hilbert spaces is that they lend themselves well to diagrammatic representations. To see this, consider a two-dimensional Hilbert space in which coordinates are expressed in terms of an orthonormal basis ϵ_1, ϵ_2. The inner product of two vectors x and y is given by

$$x \cdot y = (x_1\epsilon_1 + x_2\epsilon_2) \cdot (y_1\epsilon_1 + y_2\epsilon_2)\,.$$

Because ϵ_1 and ϵ_2 are orthonormal, we have

$$x \cdot y = x_1y_1 + x_2y_2,$$

and thus we can represent the Hilbert space by the Euclidean plane of ordered pairs of real numbers with the "natural basis" $(1,0),(0,1)$ and in which the inner product is the Euclidean inner product. Therefore x and y are orthogonal if they are perpendicular; that is, if $x_1y_1 + x_2y_2 = 0$.

In finance applications the basis vectors are the state claims. Although they are orthogonal under the expectations inner product, they do not constitute an orthonormal basis because they do not have unit norm:

$$e_s \cdot e_s = E(e_s^2) = \pi_s \neq 1.$$

If we use state claims as the basis in a diagrammatic representation, then orthogonal payoffs are perpendicular only if the probabilities of all states are the same. Otherwise orthogonal projections are skewed. For instance, the orthogonal projection $x^{\mathcal{Z}} = (7/4, 7/4)$ of vector $x = (1,2)$ on $\mathcal{Z} = \text{span}\{(1,1)\}$ in Example 3.7.0.1 differs from the perpendicular projection $(3/2, 3/2)$.

3.7.6 Riesz Representation Theorem

A linear and (norm) continuous functional on a Hilbert space has a simple form; it is the inner product with a vector in that space.

Theorem 3.7.4 (Riesz Representation Theorem)**.** *If $F : \mathcal{H} \to \mathbb{R}$ is a continuous linear*

functional on a Hilbert space $\mathcal{H}$, then there exists a unique vector k_f in $\mathcal{H}$ such that

$$F(x) = k_f \cdot x \quad \forall x \in \mathcal{H}. \tag{3.7.9}$$

Proof. If F is the zero functional, then we take $k_f = 0$. Suppose that F is a nonzero functional. Let $\mathcal{N} = \{x \in \mathcal{H} : F(x) = 0\}$ be the null space of F and $\mathcal{N}^\perp$ the orthogonal complement of $\mathcal{N}$. We have $\mathcal{H} = \mathcal{N} + \mathcal{N}^\perp$, and $\mathcal{N}^\perp \neq \{0\}$.

Choose a nonzero vector z in $\mathcal{N}^\perp$. By multiplying z by a scalar we can have $F(z) = 1$. Any vector $x \in \mathcal{H}$ can be written as

$$x = [x - F(x)z] + F(x)z.$$

Note that $[x - F(x)z] \in \mathcal{N}$. Because $z \in \mathcal{N}^\perp$, it follows that

$$z \cdot x = z \cdot [F(x)z]. \tag{3.7.10}$$

Now set

$$k_f = \frac{z}{(z \cdot z)}.$$

Then equation (3.7.10) implies

$$k_f \cdot x = \frac{F(x)(z \cdot z)}{z \cdot z} = F(x)$$

and thus k_f satisfies equation (3.7.9).

It remains to show that k_f is unique. If there are k_f and k_f' satisfying equation (3.7.9), then

$$\left(k_f - k_f'\right) \cdot x = 0$$

holds for every $x \in \mathcal{H}$; hence, $\left(k_f - k_f'\right) = 0$. □

The vector k_f in the representation (3.7.9) is called the Riesz kernel corresponding to F.

Optional

*3.7.7 Construction of the Riesz Kernel

Finding the Riesz kernel for a linear functional on the Hilbert space $\mathbb{R}^S$ with the Euclidean inner product is easy. The kernel is obtained from $k_{fs} = F(e_s)$, which implies by linearity that $F(x) = \sum_s k_{fs} x_s$. Obtaining the kernel for the expectations inner product is equally easy. The functional F can first be written $F(x) = \sum_s k_s x_s$. Then $k_{fs} = k_s/\pi_s$ gives the desired representation $F(x) = \sum_s \pi_s k_{fs} x_s = E(k_f x)$.

Any complete subspace of a Hilbert space is a Hilbert space in its own right under the same inner product. The Riesz Representation Theorem can therefore be applied to linear functionals on complete subspaces of a Hilbert space. Thus if $\mathcal{Z}$ is a complete subspace of a Hilbert space $\mathcal{H}$ and F is a continuous linear functional on $\mathcal{Z}$, then there exists a unique kernel k_f in $\mathcal{Z}$ such that $F(z) = k_f \cdot z$ holds for every $z \in \mathcal{Z}$.

If the subspace $\mathcal{Z}$ is a linear span of a finite collection of vectors $\{z_1, \ldots, z_n\}$, then kernel k_f of a linear functional $F : \mathcal{Z} \to \mathbb{R}$ can be constructed as follows: Let

$$w_i = F(z_i)$$

for $i = 1, \ldots, n$ be the values of F on the basis vectors of $\mathcal{Z}$. The kernel k_f has to satisfy n equations

$$w_i = k_f \cdot z_i \quad i = 1, \ldots, n \tag{*3.7.11}$$

Because $k_f \in \mathcal{Z}$, we have $k_f = \sum_{j=1}^n a_j z_j$. Substituting in equation (*3.7.11), we obtain n equations

$$w_i = \sum_{j=1}^n a_j z_j \cdot z_i \quad i = 1, \ldots, n$$

with n unknowns a_j that can be solved using standard methods.

The following example illustrates the preceding construction:

Example 3.7. Let $\mathcal{Z} = \operatorname{span}\{(1,1)\} \subset \mathbb{R}^2$, and let the inner product be the expectations

inner product given by probabilities $(1/4, 3/4)$. Let $F : \mathcal{Z} \to \mathbb{R}$ be given by

$$F(z) = 2z_1$$

for $z = (z_1, z_2) \in \mathcal{Z}$.

Vector $(1, 1)$ constitutes a basis of $\mathcal{Z}$. The kernel k_f has to satisfy $k_f = a(1, 1)$ for some scalar a. Because $F(1, 1) = 2$, we can solve for a from the single equation

$$2 = a(1, 1) \cdot (1, 1) = a(1/4 + 3/4).$$

Thus $a = 2$ and

$$k_f = (2, 2).$$

□

3.7.8 The Expectations Kernel

The asset span is a subspace of the Hilbert space $\mathbb{R}^S$ with the expectations inner product; hence it is a Hilbert space in its own right. Consequently the Riesz Representation Theorem applies to linear functionals defined on the asset span. Two linear functionals on the asset span $\mathcal{M}$ are of particular interest: the expectations functional, discussed in this section, and the payoff pricing functional, discussed in the next section. The probability measure π defining the expectations inner product is taken to be agents' subjective probability measure. If agents' preferences have expected utility representations, then π is the probability measure (assumed common to all agents) of the expected utility.

The expectations functional E maps every payoff $z \in \mathcal{M}$ into its expectation $E(z)$. The Riesz kernel k_e associated with the expectations functional is the unique payoff that satisfies

$$E(z) = E(k_e z), \quad \forall z \in \mathcal{M}. \tag{3.7.12}$$

We emphasize that equation (3.7.12) need not be valid for contingent claims outside the asset span.

If the risk-free payoff is in the asset span $\mathcal{M}$, then the expectations kernel k_e is risk free and

equal to one in every state. If the risk-free payoff is not in the asset span, then the kernel k_e is the orthogonal projection of the risk-free payoff on $\mathcal{M}$. To see this, observe that

$$E[(e - k_e)z] = 0$$

for every z in $\mathcal{M}$, where e denotes the payoff of one in every state. Therefore $e - k_e$ is orthogonal to $\mathcal{M}$. Because $e = (e - k_e) + k_e$, it follows that k_e is the projection of e onto $\mathcal{M}$.

Example 3.7.0.3. Assume that there are three states and two securities with payoffs $x_1 = (1, 1, 0)$ and $x_2 = (0, 1, 1)$. The probability of each state is $1/3$. To find the expectations kernel we consider the following two equations for expected payoffs:

$$\frac{2}{3} = E\left(k_e x_1\right) \tag{3.7.13}$$

and

$$\frac{2}{3} = E\left(k_e x_2\right). \tag{3.7.14}$$

Because the expectations kernel k_e lies in the asset span, we have

$$k_e = h_1 x_1 + h_2 x_2 = (h_1, h_1 + h_2, h_2) \tag{3.7.15}$$

for some portfolio (h_1, h_2). Substituting equation (3.7.15) in equations (3.7.13) and (3.7.14) we obtain

$$\frac{2}{3} = \frac{1}{3}h_1 + \frac{1}{3}\left(h_1 + h_2\right)$$

and

$$\frac{2}{3} = \frac{1}{3}\left(h_1 + h_2\right) + \frac{1}{3}h_2.$$

The solution is $h_1 = h_2 = 2/3$, which gives

$$k_e = \left(\frac{2}{3}, \frac{4}{3}, \frac{2}{3}\right).$$

Note that k_e is not the risk-free payoff because the risk-free payoff is not in the asset span. □

3.7.9 The Pricing Kernel

The Riesz kernel associated with the payoff pricing functional q on the asset span $\mathcal{M}$ is the pricing kernel k_q. It is the unique payoff in $\mathcal{M}$ that satisfies

$$q(z) = E(k_q z), \quad \forall z \in \mathcal{M}. \tag{3.7.16}$$

The expectation $E(k_q z)$ is well defined for contingent claims z not in the asset span, but it does not in general define a positive valuation functional on $\mathbb{R}^S$. This is so because the pricing kernel need not be positive (or strictly positive) even if there is no strong arbitrage (arbitrage). For example, if there is no portfolio with a strictly positive payoff, then the pricing kernel cannot be strictly positive.

If there is no arbitrage (strong arbitrage), then there exists a strictly positive (positive) state price vector $q = (q_1, \ldots, q_S)$ such that

$$q(z) = \sum_s q_s z_s \tag{3.7.17}$$

for every $z \in \mathcal{M}$. Consider the vector of state prices rescaled by the probabilities of states, denoted by $q/\pi = (q_1/\pi_1, \ldots, q_s/\pi_S)$. We can rewrite equation (3.7.17) as

$$q(z) = E(\frac{q}{\pi} z). \tag{3.7.18}$$

equations (3.7.16) and (3.7.18) imply that

$$E\left[(\frac{q}{\pi} - k_q)z\right] = 0 \tag{3.7.19}$$

for every $z \in \mathcal{M}$, and hence that $q/\pi - k_q$ is orthogonal to $\mathcal{M}$. Because $q/\pi = (q/\pi - k_q) + k_q$, it follows that the pricing kernel k_q is the projection of q/π on $\mathcal{M}$.

The pricing kernel is unique regardless of whether markets are complete or incomplete. If markets are incomplete, then there exist multiple state price vectors. When rescaled by probabilities, all these vectors have the same projection on the asset span, and that projection is the pricing kernel k_q. If markets are complete, then there exists a unique state price vector q and the pricing kernel k_q equals q/π.

If q is an equilibrium payoff pricing functional, then

$$q(z) = E\left(\frac{\partial_1 v}{E\left(\partial_0 v\right)} z\right) \tag{3.7.20}$$

for every $z \in \mathcal{M}$, where $\partial_1 v / E\left(\partial_0 v\right)$ is the vector of marginal rates of substitution of an agent whose utility function has an expected utility representation $E[v(c)]$ and whose equilibrium consumption is interior. The projection of the vector $\partial_1 v / E(\partial_0 v)$ on the asset span $\mathcal{M}$ equals the pricing kernel k_q. If markets are complete, the vector of marginal rates of substitution equals k_q, and this holds for all agents with interior consumption.

If the risk-free payoff is in the asset span, then

$$E(k_q) = E(k_q k_e) = \frac{1}{\bar{r}}$$

Example 3.7.0.4. In example 3.7.0.3, assume that security prices are $p_1 = 1, p_2 = 4/3$. To find the pricing kernel, we consider the equations for the prices of securities

$$1 = E(k_q x_1)$$

and

$$4/3 = E(k_q x_2).$$

The pricing kernel k_q lies in the asset span, and thus we have

$$k_q = h_1 x_1 + h_2 x_2 = (h_1, h_1 + h_2, h_2)$$

for some portfolio (h_1, h_2). The solution is $h_1 = 2/3$, $h_2 = 5/3$, which gives

$$k_q = (\frac{2}{3}, \frac{7}{3}, \frac{5}{3})$$

□

3.7.10 Stochastic Discount Factors

A representation of the payoff pricing functional that is closely related to the Riesz representation by the pricing kernel is the stochastic discount factor. A stochastic discount factor is

any contingent claim $m \in \mathbb{R}^s$ that satisfies

$$q(z) = E(mz), \quad \forall z \in \mathcal{M} \tag{3.7.21}$$

Of course, the pricing kernel k_q is a stochastic discount factor. If markets are incomplete, there exist stochastic discount factors other than the pricing kernel. Examples include the vector of state prices rescaled by the probabilities of states as in equation (3.7.18), and the vector of marginal rates of substitution as in equation (3.7.20). All stochastic discount factors have the same projection onto the asset span, and that projection is the pricing kernel. This is so because, in analogy to equation (3.7.19), the equality

$$0 = E[(m - k_q)z]$$

holds for every $z \in \mathcal{M}$.

3.8 The Mean-Variance Frontier Payoffs

Although variance does not in general provide an accurate measure of risk, the analysis of expected returns and variances of returns plays an important role in the theory and applications of finance. This analysis leads to identification of returns that have minimal variance for a given expected return.

The analysis relies on the Hilbert space methods — in particular, on the representations of the payoff pricing functional by the pricing kernel and of the expectations functional by the expectations kernel. The returns that attain minimum variance for a given expected return lie on a line passing through the returns on the pricing kernel and the expectations kernel.

3.8.1 Mean-Variance Frontier Payoffs

A payoff is a mean-variance frontier payoff if there is no other payoff with the same price and the same expectation but a smaller variance. In other words, the mean-variance frontier payoffs minimize variance subject to constraints on price and expectation.

Let $\mathcal{E}$ be the span of the expectations kernel k_e and the pricing kernel k_q. It is a subspace of the asset span $\mathcal{M}$. The central result of this chapter is the following:

Theorem 3.8.1. *A payoff is a mean-variance frontier payoff iff it lies in the span of the expectations kernel and the pricing kernel.*

Proof. Taking the orthogonal projection (with respect to the expectations inner product) of an arbitrary payoff $z \in \mathcal{M}$ onto $\mathcal{E}$ results in

$$z = z^{\mathcal{E}} + \epsilon$$

with $z^{\mathcal{E}} \in \mathcal{E}$ and $\epsilon \in \mathcal{E}^{\perp}$. In particular, ϵ is orthogonal to both k_e and k_q. Therefore ϵ has zero expectation and zero price, implying that z and $z^{\mathcal{E}}$ have the same expectation and the same price. Further, because ϵ is orthogonal to $z^{\mathcal{E}}$ and $E(\epsilon) = 0$, it follows that $\operatorname{cov}\left(\epsilon, z^{\mathcal{E}}\right) = E\left(\epsilon z^{\mathcal{E}}\right) - E(\epsilon)E\left(z^{\mathcal{E}}\right) = 0$. Consequently, $\operatorname{var}(z) = \operatorname{var}\left(z^{\mathcal{E}}\right) + \operatorname{var}(\epsilon)$, and thus $\operatorname{var}\left(z^{\mathcal{E}}\right) \leq \operatorname{var}(z)$ with strict inequality if $\epsilon \neq 0$. This implies that every mean-variance frontier payoff lies in $\mathcal{E}$.

For the converse, we have to show that every payoff in $\mathcal{E}$ is a mean-variance frontier payoff. Suppose, on the contrary, that there exists a payoff z in $\mathcal{E}$ that is not a mean-variance frontier payoff. Then there must exist another payoff z' with the same price and the same expectation but strictly lower variance than z. Using the argument of the first part of the proof, we can assume that $z' \in \mathcal{E}$. Because z and z' have the same price and the same expectation, we have $E\left[k_q\left(z - z'\right)\right] = 0$ and $E\left[k_e(z - z')\right] = 0$. This implies that $z - z' \in \mathcal{E}^{\perp}$. Because also $z - z' \in \mathcal{E}$, it follows that $z = z'$. This contradicts the assumption that z' has lower variance than z. □

If the expectations kernel and the pricing kernel are collinear (that is, $k_q = \gamma k_e$ for some $\gamma \neq 0$), then the set of mean-variance frontier payoffs $\mathcal{E}$ is a line. The expectations kernel and the pricing kernel are collinear iff all portfolios have the same expected return (equal to $1/\gamma$). If the risk-free payoff lies in the asset span, then k_e and k_q are collinear iff fair pricing holds. Under fair pricing — that is, when $E\left(r_j\right) = \bar{r}$ for every security j — the kernels are $k_e = e$ and $k_q = (1/\bar{r})e$, where e is the risk-free unit payoff.

When k_e and k_q are not collinear, the set of mean-variance frontier payoffs $\mathcal{E}$ is a plane.

If there are only two nonredundant securities, then the asset span is a plane. Further, if the

expectations and pricing kernels are not collinear, then the asset span coincides with the set of mean-variance frontier payoffs. Thus, every payoff is a mean-variance frontier payoff if there are two securities. Note that the number of states is irrelevant.

For brevity, the term "frontier payoff" is often used in place of "mean-variance frontier payoff".

3.8.2 Frontier Returns

The return associated with any payoff having a nonzero price equals that payoff divided by its price. Frontier returns are the returns on the frontier payoffs. Equivalently, frontier returns are the frontier payoffs that have unit price.

It follows from Theorem 3.8.1 that the return r_q on the pricing kernel and the return r_e on the expectations kernel are frontier returns. They are

$$r_e = \frac{k_e}{E\left(k_e k_q\right)} = \frac{k_e}{E\left(k_q\right)} \quad \text{and} \quad r_q = \frac{k_q}{E\left(k_q^2\right)}$$

where the pricing kernel was used to represent the prices of k_q and k_e.

If the expectations kernel and the pricing kernel are collinear, then the returns r_e and r_q are the same. The set of frontier returns consists of the single return r_e. If the risk-free payoff lies in the asset span, that single return equals the risk-free return $\bar{r}$.

We assume throughout that the expectations kernel and the pricing kernel are not collinear. If k_e and k_q are not collinear, then the set of frontier returns is the line passing through the return r_q and the return r_e. This line can be indexed by a single parameter λ, and thus

$$r_\lambda = r_e + \lambda\left(r_q - r_e\right),$$

where $-\infty < \lambda < \infty$.

Example 3.8.0.1. Suppose that there are three equally likely states and that three securities

are traded. The security returns are

$$r_1 = (3, 0, 0)$$
$$r_2 = (0, 6, 0)$$
$$r_3 = \left(\frac{6}{7}, \frac{3}{7}, \frac{9}{7}\right).$$

We wish to know which, if any, of these returns are on the mean-variance frontier.

To determine whether any of the security returns is a mean-variance frontier return, we locate the set of frontier returns. We first find the returns on the expectations and pricing kernels. Because markets are complete, the expectations kernel is the risk-free payoff $(1, 1, 1)$, and the pricing kernel is the state-price vector q rescaled by the probabilities of states. The state-price vector is the unique solution to the equations

$$1 = 3q_1$$
$$1 = 6q_2$$
$$1 = \frac{6}{7}q_1 + \frac{3}{7}q_2 + \frac{9}{7}q_3$$

The solution is $q_1 = 1/3, q_2 = 1/6, q_3 = 1/2$. The pricing kernel equals q/π, that is $(1, 1/2, 3/2)$. The returns of the expectations and pricing kernels are obtained using the pricing kernel. Because markets are complete the expectations kernel is $(1, 1, 1)$, and multiplying by the pricing kernel shows that its price is 1. Therefore its return r_e is $(1, 1, 1)$. The price of the pricing kernel $(1, 1/2, 3/2)$ is $7/6$, and the return r_q equals r_3. Return r_3 is therefore a frontier return. Returns r_1 and r_2 are not, because they are not on the line connecting r_e and r_q. □

Chapter 4

Asset Pricing: Continuous Time

4.1 Asset Pricing in Continuous Time

4.1.1 The Model

The securities market model consists of

1. a probability space $(\Omega, \mathcal{F}, P)$,
2. a time inverval $\mathcal{T} = [0, T]$,
3. a Brownian motion $Z = (Z_1, .., Z_d)$ on $(\Omega, \mathcal{F}, P)$,
4. the standard filtration $\mathbb{F}$ of Z,
5. $N + 1$ securities indexed by $n = 0, .., N$.

(a) Security 0 is the "riskless" security, and can be interpreted as the value of a bank account. Its price at time 0 is $B_0 \equiv S_{0,0} = 1$, and its price at time t is given by

$$dB_t = r_t B_t dt$$

where r is the short rate process.

(b) Securities $1, \ldots, N$ are the "risky" securities. Their prices are given by an Ito process

$$d \begin{pmatrix} S_{1,t} \\ .. \\ S_{N,t} \end{pmatrix} = \mu_t dt + \sigma_t dZ_t$$

where $\mu \in (\mathcal{L}^1)^N$ and $\sigma \in (\mathcal{L}^2)^{N \times d}$.

We set $S = (S_0, .., S_N)$ and $S_{1N} = (S_1, .., S_N)$.

Implicit in this formulation is that securities do not pay dividends. We also assume that there is no intermediate consumption. We will introduce dividends and intermediate consumption later on.

Definition 4.1.1. A trading strategy is a process $\theta \in \mathcal{L}(S)$.

For a trading strategy θ, we use $\theta_{1N} = (\theta_1, .., \theta_N)$ to denote the investment in the risky assets. We define the stochastic integral $\int_0^t \theta_s dS_s$ as the Ito process

$$\int_0^t \theta_s dS_s = \int_0^t \left(\theta_{0,s} r_s B_s + \theta_{1N,s} \mu_s\right) ds + \int_0^t \theta_{1N,s} \sigma_s dZ_s$$

Definition 4.1.2. A trading strategy is self-financing iff

$$\theta_t S_t = \theta_0 S_0 + \int_0^t \theta_s dS_s \tag{4.1.1}$$

Equation (4.1.1) is the dynamic budget constraint in continuous time. In discrete time, and without dividends and intermediate consumption, the dynamic budget constraint is

$$0 = \theta_{t-1} S_t - \theta_t S_t$$

Equation 5.1.5 implies that

$$\theta_t S_t = \theta_{t-1} S_{t-1} + \theta_{t-1} (S_t - S_{t-1})$$

Therefore,

$$\theta_t S_t = \theta_0 S_0 + \sum_{s=0}^{t-1} \theta_s (S_{s+1} - S_s)$$

In continuous time, the latter sum is the stochastic integral $\int_0^t \theta_t dS_t$.

Definition 4.1.3. A cash flow is a pair (C_0, C_T) where $C_0 \in \mathbb{R}$ and C_T is $\mathcal{F}_T$-measurable.

Definition 4.1.4. A self-financing trading strategy θ finances a cash flow (C_0, C_T) iff $C_0 = -\theta_0 S_0$, and $C_T = \theta_T S_T$.

Definition 4.1.5. A cash flow is marketable iff it is financed by a trading strategy θ.

We denote by M the set of marketable cash flows.

4.1.2 SPD and EMM

In discrete time we showed that absence of arbitrage is equivalent to the existence of strictly positive state prices. Strictly positive state prices are in turn equivalent to a (strictly positive) state-price density (SPD), and to an equivalent martingale measure (EMM).

In continuous time things are more complicated. Existence of a SPD or an EMM does not preclude arbitrage. Arbitrages can be constructed by following doubling strategies. We will eliminate doubling strategies by imposing plausible restrictions on the set of trading strategies. Under these restrictions, existence of a SPD or an EMM does preclude arbitrage. Absence of arbitrage, however, does not imply existence of a SPD or an EMM.

In this section we define SPD and EMM, and show that existence of one is equivalent to existence of the other.

Definition 4.1.6. An arbitrage is a marketable cash flow such that $C_0 \geq 0, C_T \geq 0$, and $C_0 > 0$ or $\text{prob}(C_T > 0) > 0$.

Definition 4.1.7. A SPD is a strictly positive Ito process π such that πS is a martingale, and $\pi_0 = 1$.

Definition 4.1.8. An EMM is a probability measure Q such that $\hat{S} \equiv S/B$ is a martingale under Q, and Q is equivalent to P.

Proposition 4.1.1. *A SPD exists iff an EMM exists.*

Proof. Suppose that a SPD exists. Denote the SPD by π and set $\xi = \pi B$. The function ξ_T is a Radon-Nikodym derivative and thus defines a probability measure Q. Indeed, $\xi_T > 0$ and, since ξ is a martingale, $E(\xi_T) = \xi_0 = 1$. The probability measure Q is equivalent to P since $\xi_T > 0$ a.s. Therefore, Q is an EMM if $\hat{S} = S/B$ is a martingale under Q. We have

$$E_t^Q\left(\hat{S}_s\right) = \frac{E_t\left(\hat{S}_s\xi_T\right)}{E_t(\xi_T)} = \frac{E_t\left(\hat{S}_s\xi_s\right)}{\xi_t} = \frac{E_t(\pi_s S_s)}{\xi_t}$$

Since πS is a martingale under P, this is equal to

$$\frac{S_t}{B_t}\frac{\pi_t}{\pi_t} = \hat{S}_t$$

Therefore, $\hat{S}$ is a martingale under Q.

Suppose that an EMM exists. Denote the EMM by Q, its Radon-Nikodym derivative w.r.t. P by ξ_T, and set $\xi_t = E_t(\xi_T)$ and $\pi = \xi/B$. Since Q is equivalent to $P, \xi_T > 0$ a.s. and thus $\pi_T > 0$ a.s. By the martingale representation theorem, ξ is an Ito process, and thus π is also an Ito process. Therefore, π is an SPD if πS is a martingale under P. We have

$$E_t(\pi_s S_s) = E_t\left(\frac{\xi_s}{B_s}S_s\right) = E_t\left(\xi_T\frac{S_s}{B_s}\right) = E_t^Q\left(\frac{S_s}{B_s}\right)E_t(\xi_T) = E_t^Q\left(\frac{S_s}{B_s}\right)\xi_t$$

Since S/B is a martingale under Q, this is equal to

$$\frac{S_t}{B_t}\xi_t = \pi_t S_t$$

Therefore, πS is a martingale under P. □

Notice that if Q is an EMM, we have

$$\hat{S}_t = E_t^Q\left(\hat{S_T}\right)$$

Since,

$$B_t = \exp\left(\int_0^t r_s ds\right)$$

we get

$$S_t = E_t^Q\left[\exp\left(-\int_t^T r_s ds\right) S_T\right]$$

The price at time t is equal to the expectation under Q of the price at time T, discounted at the riskless rate.

4.1.3 Doubling Strategies

In continuous time, existence of a SPD or an EMM does not preclude arbitrage. Arbitrages can be constructed by following doubling strategies. The idea behind doubling strategies can be understood in the following example. Suppose that $T = 1$ and there are two securities. The first is a riskless bond whose price is equal to 1. (The short rate is thus equal to 0 .) The second is a risky stock. At $t = 0$ the stock price is equal to 1 . The stock price is then constant, except for times $t = 1 - 1/2^n$, when it jumps. The jump at time $t = 1 - 1/2^n$ is either $1/2^n$ or $-1/2^n$ with equal probabilities. The stock price is not a Brownian motion, but is a martingale.

The trading strategy starts with one share in the stock and -1 in the bond. The wealth at $t = 0$ is thus 0 . If at $t = 1/2$ the stock price goes up, the stock is sold, and the proceeds are invested in the bond until $t = 1$. The wealth at $t = 1$ is thus $1/2$. If the stock price goes down, 3 shares of the stock are bought. If at $t = 3/4$ the stock price goes up, the stock is sold and the proceeds are invested in the bond until $t = 1$. The wealth at $t = 1$ is thus

$$-1 - \frac{1}{2} \times 3 + \frac{3}{4} \times (3 + 1) = \frac{1}{2}$$

If the stock price goes down, 21 shares of the stock are bought, and so on. The wealth at $t = 1$ is $1/2$ with probability 1 . This is an arbitrage.

In this example the stock price is not a Brownian motion. Duffie 6.C. presents a doubling strategy in a Brownian motion context.

Mathematically, the important feature of a doubling strategy θ is that, even when the price S is a martingale, the stochastic integral $\int_0^t \theta_s dS_s$ is not a martingale. The stochastic integral is in fact only a local martingale.

A doubling strategy can be criticized on two grounds. First, trading can take place an arbitrarily large number of times. Second, wealth can become arbitrarily negative, and there is thus no bankruptcy. This suggests two plausible restrictions on the set of trading strategies.

The first restriction is that trading can take place only a finite number of times. Trading strategies thus have to be simple. With simple strategies, the stochastic integral $\int_0^t \theta_s d\hat{S}_s$ is a martingale. However, with simple strategies, the continuous-time model loses much of its power. The replicating strategies in the Black-Scholes model, for instance, involve continuous trading.

The second restriction is that (discounted) wealth has to stay above a threshold. Under this restriction, the stochastic integral $\int_0^t \theta_s d\hat{S}_s$ turns out to be a super-martingale, and this suffices for absence of arbitrage.

A third restriction is motivated by the mathematics of the stochastic integral. The stochastic integral $\int_0^t \theta_s d\hat{S}_s$ is a local martingale because $\theta \in \mathcal{L}(\hat{S})$. If, however, $\theta \in \mathcal{H}^2(\hat{S})$, then $\int_0^t \theta_s d\hat{S}_s$ is a martingale.

We will focus on the second and third restrictions. Moreover, we will discount prices and wealth by the riskless rate. The second restriction is that trading strategies belong to the set

$$\underline{\Theta}(\hat{S}) \equiv \left\{\theta : \exists k, \text{ s.t.} \quad \theta_t \hat{S}_t \geq k \forall t\right\}$$

Discounted wealth thus has to be greater than a threshold k. The third restriction is that trading strategies belong to the set $\theta \in \mathcal{H}^2(\hat{S})$. Notice that when the short rate is bounded, we have $\underline{\Theta}(\hat{S}) = \underline{\Theta}(S)$ and $\mathcal{H}^2(\hat{S}) = \mathcal{H}^2(S)$.

4.1.4 From SPD/EMM to No Arbitrage

Theorem 4.1.1 provides conditions under which existence of an EMM implies absence of arbitrage.

Theorem 4.1.1. *Suppose that an EMM Q exists. If trading strategies belong to $\mathcal{H}^2(\hat{S})$, and the Radon-Nikodym derivative $dQ/dP \in L^2$, then there is no arbitrage. Alternatively, if trading strategies belong to $\underline{\Theta}(\hat{S})$, then there is no arbitrage.*

Proof. The idea of the first part of the proof is that the discounted stock prices are martingales, therefore the discounted gain process is a martingale, and hence as long as C_T is positive, C_0 must be negative, because $\hat{C}_0 = E^Q\left[\hat{C}_T\right]$.

Consider a cash flow (C_0, C_T) financed by a trading strategy θ. We have

$$d\left(\theta_t S_t / B_t\right) = \frac{d\left(\theta_t S_t\right)}{B_t} - \frac{\theta_t S_t r_t}{B_t} dt = \frac{\theta_t dS_t}{B_t} - \frac{\theta_t S_t r_t}{B_t} dt = \theta_t d\left(\frac{S_t}{B_t}\right)$$

where we used first Ito's lemma, then the self-financing constraint (equation 5.1.4), and then Ito's lemma. Integrating, we get

$$\theta_t \hat{S}_t = \theta_0 \hat{S}_0 + \int_0^t \theta_s d\hat{S}_s$$

Equation (5.4.1) implies that the strategy θ is self-financing w.r.t. the discounted price process $\hat{S}$. It also implies that

$$\hat{C}_T \equiv \frac{C_T}{B_T} = \theta_T \hat{S}_T = \theta_0 \hat{S}_0 + \int_0^T \theta_t d\hat{S}_t = -\hat{C}_0 + \int_0^T \theta_t d\hat{S}_t$$

Suppose that $\theta \in \mathcal{H}^2(\hat{S})$. Below we will show that the stochastic integral $\int_0^t \theta_s d\left(\hat{S}_s\right)$ is a martingale under Q. This will imply that

$$E^Q \left[\int_0^T \theta_t d\hat{S}_t \right] = 0$$

and thus that the cash flow (C_0, C_T) is not an arbitrage, since

$$E^Q \hat{C}_T + \hat{C}_0 = 0$$

The rest is purely technical. To show that $\int_0^t \theta_s d\hat{S}_s$ is a martingale under Q, we write it in terms of a Brownian motion. Ito's lemma implies that

$$d\hat{S}_{1N,t} = \frac{\mu_t - r_t S_{1N,t}}{B_t} dt + \frac{\sigma_t}{B_t} dZ_t \equiv \hat{\mu}_t dt + \hat{\sigma}_t dZ_t$$

Since $\hat{S}_{1N}$ is a martingale under Q, the diffusion invariance principle implies that there exists a Brownian motion Z^Q under Q such that

$$d\hat{S}_{1N,t} = \hat{\sigma}_t dZ_t^Q$$

Therefore,

$$\int_0^t \theta_s d\hat{S}_s = \int_0^t \theta_{1N,s} d\hat{S}_{1N,s} = \int_0^t \theta_{1N,s}\hat{\sigma}_s dZ_s^Q$$

This is a martingale under Q if

$$E^Q\left[\left(\int_0^T \left(\theta_{1N,t}\hat{\sigma}_t\right)^2 dt\right)^{\frac{1}{2}}\right] < \infty$$

We have

$$\begin{aligned} E^Q\left[\left(\int_0^T \left(\theta_{1N,t}\hat{\sigma}_t\right)^2 dt\right)^{\frac{1}{2}}\right] &= E\left[\frac{dQ}{dP}\left(\int_0^T \left(\theta_{1N,t}\hat{\sigma}_t\right)^2 dt\right)^{\frac{1}{2}}\right] \\ &\leq \left[E\left(\frac{dQ}{dP}\right)^2\right]^{\frac{1}{2}}\left[E\left[\int_0^T \left(\theta_{1N,t}\hat{\sigma}_t\right)^2 dt\right]\right]^{\frac{1}{2}}, \end{aligned}$$

where we used first the fact that dQ/dP is the Radon-Nikodym derivative of Q w.r.t. P, and then the Cauchy-Schwarz inequality. Since $dQ/dP \in L^2$ and $\theta \in \mathcal{H}^2(\hat{S})$, both terms are finite.

Suppose next that $\theta \in \underline{\Theta}(\hat{S})$. The stochastic integral $\int_0^t \theta_s d\left(\hat{S}_s\right)$ is a local martingale under Q. This stochastic integral is bounded below since

$$\int_0^t \theta_s d\hat{S}_s = \theta_t\hat{S}_t - \theta_0\hat{S}_0 \geq k - \theta_0\hat{S}_0$$

Since a local martingale that is bounded below is a super-martingale, we have

$$E^Q\left[\int_0^T \theta_t d\hat{S}_t\right] \leq 0$$

Equation (5.4.2) implies that

$$E^Q\left[\hat{C}_T\right] + C_0 \leq 0$$

and thus the cash flow (C_0, C_T) is not an arbitrage. □

Combining Proposition 4.1.1 and Theorem 4.1.1, we get the following corollary:

Corollary 4.1.1. *Suppose that an SPD, π, exists. If trading strategies belong to $\mathcal{H}^2(\hat{S})$, and $\pi_T B_T \in L^2$, then there is no arbitrage. Alternatively, if trading strategies belong to $\underline{\Theta}(\hat{S})$, then there is no arbitrage.*

4.1.5 From Security Prices to EMM

Whether an EMM exists or not, depends on security prices. We now determine conditions on security prices so that an EMM exists.

To write security prices under the EMM, we will use Girsanov's theorem. Therefore we look for an EMM whose Radon-Nikodym derivative w.r.t. P has the form required in that theorem. More precisely, we consider a process $\eta \in (\mathcal{L}^2)^d$ that satisfies Novikov's condition. We also consider the process

$$\xi_t^\eta = \exp\left(-\int_0^t \eta_s' dZ_s - \frac{1}{2}\int_0^t \eta_s^2 ds\right)$$

and the probability measure Q^η whose Radon-Nikodym derivative w.r.t. P is equal to ξ_T^η.

Discounted security prices follow the process

$$d\left(\frac{S_{1N,t}}{B_t}\right) = \frac{\mu_t - r_t S_{1N,t}}{B_t} dt + \frac{\sigma_t}{B_t} dZ_t = \hat{\mu}_t dt + \hat{\sigma}_t dZ_t$$

Girsanov's theorem implies that, under Q^η, discounted security prices follow the process

$$d\left(\frac{S_{1N,t}}{B_t}\right) = (\hat{\mu}_t - \hat{\sigma}_t\eta_t)\,dt + \hat{\sigma}_t dZ_t^{\eta}$$

Discounted security prices are thus a martingale under Q^{η} if

$$\hat{\mu}_t - \hat{\sigma}_t\eta_t = 0 \tag{4.1.2}$$

and $\hat{\sigma}_t \in (\mathcal{H}^2)^{N\times d}$.

Therefore, the conditions for an EMM to exist are

1. equation (5.5.1) has a solution η_t.
2. $\hat{\sigma}_t \in (\mathcal{H}^2)^{N\times d}$.
3. η satisfies Novikov's condition.

Equation 5.5.1 has an economic interpretation. We can write the n 'th "component" of this equation as

$$\frac{\mu_{n,t}}{S_{n,t}} - r_t = \sum_{i=1}^{d} \frac{\sigma_{n,i,t}}{S_{n,t}}\eta_{i,t}$$

The LHS of equation 5.5.2 is equal to the instantaneous expected return on security n minus the riskless rate. This is the risk premium of security n. The term $\sigma_{n,i,t}/S_{n,t}$ is the loading of the instantaneous return on security n on the i 'th Brownian motion. Finally, $\eta_{i,t}$ is the risk premium of the i 'th Brownian motion. Equation 5.5.2 thus states that the risk premium of security n is the sum over Brownian motions of the security's loading on each Brownian motion, times the Brownian motion's risk premium.

The condition that equation 5.5.1 has a solution, means that the risk premia of the securities are "consistent", in that they can be derived from risk premia of the underlying Brownian motions. If equation 5.5.1 did not have a solution, there would exist two securities, or

portfolios of securities that would have the same loadings on the Brownian motions, but different risk premia. Therefore, there would be an arbitrage.

Proposition 4.1.2. *Suppose that equation* (4.1.2) *does not have a solution. Then there are arbitrages, both for trading strategies in* $\mathcal{H}^2(\hat{S})$ *and in* $\underline{\Theta}(\hat{S})$.

Proof. See Duffie 6.G. □

For an EMM to exist, the risk premia of the securities must be derived from risk premia of the underlying Brownian motions. Moreover, the risk premia of the Brownian motions must satisfy Novikov's condition. This means that they cannot get too large. If, in particular, the risk premia are bounded, then they satisfy Novikov's condition.

Notice that the risk premia of the Brownian motions (and of the securities) can be positive or negative. A positive risk premium means that the Brownian motion goes down when marginal utility is high, and up when it is low.

The conditions for an EMM to exist thus have an economic interpretation. The EMM also has an economic interpretation. Ito's lemma implies that

$$d\xi_t^\eta = -\xi_t^\eta \eta_t' dZ_t$$

ξ_t^η is the ratio of the probabilities, under Q and under P, of the Brownian motion path up to time t. Suppose that the risk premium of a Brownian motion is positive. Then, the ratio ξ^η will increase at time $t + dt$ if the Brownian motion goes down. Therefore, Q will put more weight than P on high marginal utility states. This is because Q incorprorates risk-aversion.

The SPD corresponding to the EMM also has an economic interpretation. The SPD is given by $\pi = \xi^\eta / B$. Ito's lemma and equation 5.5.3 imply that

$$d\pi_t = -\pi_t r_t dt - \pi_t \eta_t' dZ_t$$

The interpretation of the drift term $-\pi_t r_t dt$, is that the SPD decreases over time, since it incorporates discounting. The interpretation of the diffusion term $-\pi_t \eta_t' dZ_t$, is the same as for the EMM: the SPD puts more weight on high marginal utility states, since it incorporates risk-aversion.

4.1.6 From No Arbitrage to Security Prices

We have shown that, under restrictions on trading strategies, existence of a SPD or an EMM implies absence of arbitrage. We now examine the converse. Under the same restrictions on trading strategies, does absence of arbitrage imply existence of a SPD or an EMM?

When time is discrete and the number of states finite, this converse result is true, and can be shown by applying the separating hyperplane theorem to the marketed subspace and the positive orthant. When time is continuous, however, the separating hyperplane theorem can no longer be applied because the positive orthant has empty interior.

When time is continuous, absence of arbitrage does not imply existence of a SPD or an EMM. Although absence of arbitrage does not imply existence of a SPD or an EMM, absence of approximate arbitrage, a stronger concept than arbitrage does imply existence of a SPD or an EMM. See Duffie 6.K.

4.2 Applications of FTAP, Continuous Time

4.2.1 Redundant Securities

We consider a securities market model consisting of

1. a probability space $(\Omega, \mathcal{F}, P)$,
2. a time inverval $\mathcal{T} = [0, T]$,
3. a Brownian motion $Z = (Z_1, .., Z_d)$ on $(\Omega, \mathcal{F}, P)$,
4. the standard filtration $\mathbb{F}$ of Z,
5. $N + 1$ securities indexed by $n = 0, .., N$.
   (a) Security 0 is the "riskless" security, and its price is given by

$$dB_t = r_t B_t dt.$$

(b) Securities $1, \ldots, N$ are the "risky" securities. Their prices are given by an Ito process

$$d \begin{pmatrix} S_{1,t} \\ . \\ S_{N,t} \end{pmatrix} = \mu_t dt + \sigma_t dZ_t$$

where $\mu \in (\mathcal{L}^1)^N$ and $\sigma \in (\mathcal{L}^2)^{N \times d}$.

We set $S = (S_0, .., S_N)$ and $S_{1N} = (S_1, .., S_N)$. We assume that securities pay no dividends, and there is no intermediate consumption. We denote S_0 by B.

Discounted prices follow the process

$$d\hat{S}_{1N,t} = d\left(\frac{S_{1N,t}}{B_t}\right) = \frac{\mu_t - r_t S_{1N,t}}{B_t} dt + \frac{\sigma_t}{B_t} dZ_t \equiv \hat{\mu}_t dt + \hat{\sigma}_t dZ_t$$

We assume that $\hat{\mu}_t$ and $\hat{\sigma}_t$ satisfy the conditions that guarantee existence of an EMM. We denote by η_t the risk premia associated to the Brownian motions. These are the solutions to

$$\hat{\mu}_t = \hat{\sigma}_t \eta_t.$$

We denote the EMM by Q, and the Brownian motion under Q, obtained from Girsanov's theorem, by Z^Q. We can write the dynamics of discounted prices as

$$d\hat{S}_{1N,t} = d\left(\frac{S_{1N,t}}{B_t}\right) = \hat{\sigma}_t dZ_t^Q$$

We assume that trading strategies are in $\mathcal{L}(S)$ and are such that the stochastic integral $\int_0^t \theta_s d\left(S_s/B_s\right)$ is a martingale under Q.

Definition 4.2.1. A self-financing trading strategy θ replicates an $\mathcal{F}_T$-measurable random variable C_T iff $C_T = \theta_T S_T$.

Definition 4.2.2. A new security with time T payoff $\tilde{S}_T$ is redundant iff there exists an admissible trading strategy $\tilde{\theta}$ that replicates $\tilde{S}_T$.

Proposition 4.2.1. *For a redundant security,* $\tilde{S}_t = \tilde{\theta}_t S_t$.

Proof. If

$$\tilde{\theta}_t S_t < \tilde{S}_t$$

then there is an arbitrage obtained by buying the replicating strategy and shorting the redundant security. A similar arbitrage exists if the opposite inequality holds. □

Proposition 4.2.2. *For a redundant security,*

$$\tilde{S}_t = E_t^Q \left[\exp\left(- \int_t^T r_s ds \right) \tilde{S}_T \right]$$

Proof. We denote by $\tilde{\theta}$ the replicating strategy of the redundant security. Since $\tilde{\theta}$ is selffinancing, it is also self-financing w.r.t. the discounted price process S/B. Therefore,

$$d\left(\frac{\tilde{\theta}_t S_t}{B_t} \right) = \tilde{\theta}_t d\left(\frac{S_t}{B_t} \right)$$

Integrating, we get,

$$\frac{\tilde{S}_T}{B_T} = \frac{\tilde{\theta}_T S_T}{B_T} = \tilde{\theta}_t \hat{S}_t + \int_t^T \tilde{\theta}_s d\left(\hat{S}_s \right)$$

Since the stochastic integral $\int_0^t \theta_s d\left(S_s / B_s \right)$ is a martingale under Q, we have

$$\frac{\tilde{\theta}_t S_t}{B_t} = E_t^Q \left[\frac{\tilde{S}_T}{B_T} \right]$$

Proposition 4.2.1 implies then that

$$\frac{\tilde{S}_t}{B_t} = E_t^Q \left[\frac{\tilde{S}_T}{B_T} \right]$$

and thus

$$\tilde{S}_t = E_t^Q \left[\exp\left(- \int_t^T r_s ds \right) \tilde{S}_T \right]$$

□

The redundant security is said to be priced by arbitrage.

4.2.2 Complete Markets

Consider a market in which there exists a bounded market price of risk process η_t : $\hat{\sigma}_t \hat{\eta}_t = \mu_t$, an therefore there exists an EMM Q and there is no arbitrage. To apply arbitrage pricing, we need to characterize the set of cash flows that can be replicated. As always, we limit the admissible trading strategies to belong to $\underline{\Theta}(\hat{S})$ or $\underline{\Theta}(\hat{S})$. It would be easy to demonstrate replication if we allowed for a larger class of trading strategies, e.g., $\mathcal{L}^2$, in fact, it would appear that all markets are complete. But such strategies can produce arbitrage gains.

We denote by $L^2(Q)$ the set of random variables that are square-integrable w.r.t. Q.

Definition 4.2.3. Markets are complete iff all $\mathcal{F}_T$-measurable random variables C_T such that $C_T/B_T \in L^2(Q)$ can be replicated by a trading strategy in $\mathcal{H}^2(\hat{S})$ or (if $C_T \geq 0$) by a trading strategy in $\underline{\Theta}(\hat{S})$.

Theorem 4.2.1. *Markets are complete iff the rank of σ_t is equal to d a.s.*

Proof. Here we outline the main ideas of the proof.

Suppose that the rank of σ_t is equal to d a.s. Consider an $\mathcal{F}_T$-measurable random variable C_T such that $C_T/B_T \in L^2(Q)$. By the law of iterative expectations, the process $E_t^Q\left(C_T/B_T\right)$ is a martingale under Q. Using the martingale representation part of the Girsanov's theorem, there exists a process $\zeta \in \left(\mathcal{L}^2\right)^d$ such that

$$E_t^Q\left[\frac{C_T}{B_T}\right] = E^Q\left[\frac{C_T}{B_T}\right] + \int_0^t \zeta_s dZ_s^Q \tag{4.2.1}$$

In fact, since $C_T/B_T \in L^2(Q), \zeta \in \left(\mathcal{H}^2\right)^d$ (standard result, e.g., Protter (2004, Thm. 27, Corollary 3), which says that a local martingale is in fact a square integrable martingale iff the isometry relation holds).

Since the rank of σ_t is equal to d, we can find a process θ_{1N} such that

$$\theta_{1N,t}\hat{\sigma}_t = \zeta_t. \tag{4.2.2}$$

The process θ_{1N} represents the investment in the risky assets. We define the investment in the riskless asset by

$$E_t^Q\left[\hat{C}_T\right] = \theta_t \hat{S}_t \tag{4.2.3}$$

Since $\theta_{1N}\hat{\sigma} = \zeta \in \left(\mathcal{H}^2\right)^d$, we know that $\theta_{1N}\hat{\sigma} \in \mathcal{L}^2$. Also, since $\theta^{1N}\hat{\mu} = \theta^{1N}\hat{\sigma}\eta = \zeta\eta$, and $\zeta, \eta \in \left(\mathcal{L}^2\right)^d$, Cauchy-Schwarz inequality implies that $\theta_{1N}\hat{\mu} \in \mathcal{L}^1$. Thus, θ_t is a mathematically valid trading strategy, i.e., the corresponding gain process is well defined.

We now need to show that the strategy θ is self-financing, replicates C_T, and in $\mathcal{H}^2(\hat{S})$ and in $\underline{\Theta}(\hat{S})$. To show that θ is self-financing, we plug equation (6.2.2) into equation (6.2.1) and get

$$\begin{aligned} E_t^Q\left[\hat{C}_T\right] &= E^Q\left[\hat{C}_T\right] + \int_0^t \theta_{1N,s}\hat{\sigma}_s dZ_s^Q \\ &= E^Q\left[\hat{C}_T\right] + \int_0^t \theta_{1N,s} d\hat{S}_{1N,s} \\ &= E^Q\left[\hat{C}_T\right] + \int_0^t \theta_s d\hat{S}_s. \end{aligned} \tag{4.2.4}$$

Combining with equation (6.2.3), we get

$$\theta_t \hat{S}_t = \theta_0 S_0 + \int_0^t \theta_s d\hat{S}_s \tag{4.2.5}$$

Therefore, θ is self-financing w.r.t. the discounted price process $\hat{S} \equiv S/B$. This implies that θ is self-financing w.r.t. to the undiscounted price process S.

Equation 6.2.3 for $t = T$ implies that θ replicates C_T.

Finally, we need to show that θ is in $\mathcal{H}^2(\hat{S})$ and in $\underline{\Theta}(\hat{S})$, if $C_T \geq 0$. We will not prove the first result, it can be found in Duffie Thm 6.I., p. 118.

Note: While we didn't prove that $\theta \in \mathcal{H}^2(\hat{S})$, remember that such condition itself was only sufficient for absence of doubling strategies (or for existence of EMM to imply absence of arbitrage), it was not necessary. In fact, a weaker condition would suffice, that θ must be martingale-generating under Q, i.e., that $\int \theta_t d\hat{S}_t$ is a martingale under Q. It is easy to see that in our case we obtained such a $\theta : \int_0^t \theta_s d\hat{S}_s$ is a martingale under Q since

$$\int_0^t \theta_s d\hat{S}_s = E_t^Q\left[\hat{C}_T\right] - E^Q\left[\hat{C}_T\right] \tag{4.2.6}$$

The results that $\theta \in \underline{\Theta}(\hat{S})$ is easier to show. Eq. (6.2.3) implies that $\theta_t \hat{S}_t = E_t^Q\left[\hat{C}_T\right] \geq 0$, since $C_T \geq 0$.

For the converse implication, see Duffie 6.I. □

The condition for market completeness has a similar flavor to the condition in discrete time.

In discrete time, markets are complete iff the number of linearly independent securities is equal to the number of nodes one period ahead. In continuous time, the number of nodes one period ahead is replaced by the number of Brownian motions. Moreover, linear independence of securities is measured by the rank of the diffusion matrix σ_t, and thus corresponds to an infinitesimal time interval

In continuous time, we can replicate an infinite-dimensional set of cash flows, with a finite number of securities. To some extent, this is not surprising, since trading can take place infinitely often.

As in discrete time, markets are complete iff the EMM is unique.

Theorem 4.2.2. *Markets are complete iff the EMM is unique.*

Proof. Suppose that markets are complete. Then the rank of σ_t is equal to d a.s. Consider an EMM Q, denote its Radon-Nikodym derivative w.r.t. P by ξ_T, and set $\xi_t = E_t(\xi_T)$. By the law of iterative expectations, the process ξ is a martingale under P. The martingale representation theorem implies that there exists a process $\rho \in (\mathcal{L}^2)^d$ such that

$$\xi_t = \xi_0 - \int_0^t \rho_s' dZ_s = 1 - \int_0^t \rho_s' dZ_s \tag{4.2.7}$$

Therefore, $d\xi_t = -\rho_t' dZ_t$, and Ito's lemma implies that

$$d(\log(\xi_t)) = -\frac{\rho_t'}{\xi_t} dZ_t - \frac{1}{2}\frac{\|\rho_t\|^2}{\xi_t^2} dt \tag{4.2.8}$$

Integrating, we get

$$\xi_t = \exp\left(-\int_0^t \frac{\rho_s'}{\xi_s} dZ_s - \frac{1}{2}\int_0^t \frac{\|\rho_s\|^2}{\xi_s^2} ds\right) \tag{4.2.9}$$

Since ξ_t is a martingale, Girsanov's theorem implies that, under Q, discounted security prices follow the process

$$d\left(\frac{S_{1N,t}}{B_t}\right) = \left(\hat{\mu}_t - \hat{\sigma}_t \frac{\rho_t}{\xi_t}\right) dt + \hat{\sigma}_t dZ_t^Q \tag{4.2.10}$$

where Z^Q is a Brownian motion under Q. Since Q is an EMM, discounted security prices are a martingale, and thus

$$\hat{\mu}_t - \hat{\sigma}_t \frac{\rho_t}{\xi_t} = 0 \tag{4.2.11}$$

Since the rank of σ_t is equal to d, this equation uniquely determines ρ_t/ξ_t. Therefore, Q is unique.

The converse is left as an exercise. □

4.2.3 The Black-Scholes model

The Model

We assume that there are 2 securities. The price of the first security is given by $B_0 = 1$ and

$$dB_t = rB_t dt \tag{4.2.12}$$

The short rate is thus constant. We refer to the first security as the bond. The price of the second security is given by

$$dS_t^1 = \mu S_{1,t} dt + \sigma S_{1,t} dZ_t \tag{4.2.13}$$

where Z_t is an one-dimensional Brownian motion. The drift and the diffusion are thus proportional to the price. Such an Ito process is called a Geometric Brownian Motion. We refer to the second security as a stock, and denote its price by S_t. Ito's lemma implies that

$$d\left(\log\left(S_t\right)\right) = \left(\mu - \frac{1}{2}\sigma^2\right)dt + \sigma dZ_t \tag{4.2.14}$$

Therefore,

$$S_t = S_0 \exp\left[\left(\mu - \frac{1}{2}\sigma^2\right)t + \sigma Z_t\right]. \tag{4.2.15}$$

The discounted stock price follows the process

$$d\hat{S}_t = (\mu - r)\hat{S}_t dt + \sigma \hat{S}_t dZ_t \tag{4.2.16}$$

This is also a Geometric Brownian Motion.

The conditions that guarantee existence of an EMM are satisfied. The risk-premium for the Brownian motion is

$$\eta = \frac{\mu - r}{\sigma}. \tag{4.2.17}$$

The Radon-Nikodym derivative of the EMM Q w.r.t. P is

$$\xi_T = \exp\left(-\eta Z_T - \frac{1}{2}\eta^2 T\right) \tag{4.2.18}$$

The Brownian motion under Q, obtained from Girsanov's theorem is

$$Z_t^Q = Z_t + \eta t \tag{4.2.19}$$

Since $\sigma > 0$, markets are complete, and the EMM is unique.

Pricing: The Martingale Approach

Consider now a new security with a time T payoff $\tilde{S}_T \in L^1(Q)$. This security is redundant, and its time t price is given by

$$\tilde{S}_t = E_t^Q \left[\exp(-r(T-t)) \tilde{S}_T \right]. \tag{4.2.20}$$

If the time T payoff is a function of S_T, i.e. $\tilde{S}_T = G(S_T)$, we have

$$\tilde{S}_t = E_t^Q \left[\exp(-r(T-t)) G(S_T) \right] \tag{4.2.21}$$

To compute the expectation (6.3.3), we note that the discounted stock price follows the process

$$d\hat{S}_t = \sigma \hat{S}_t dZ_t^Q \tag{4.2.22}$$

Therefore,

$$S_T = S_t \exp\left[\left(r - \frac{1}{2}\sigma^2 \right)(T-t) + \sigma \left(Z_T^Q - Z_t^Q \right) \right]$$

Since Z_t^Q is a Brownian motion under Q, the expectation (6.3.3) is

$$\exp(-r(T-t)) E\left[G\left\{ S_t \exp\left[\left(r - \frac{1}{2}\sigma^2 \right)(T-t) + (\sigma\sqrt{T-t})\tilde{\epsilon} \right] \right\} \right]$$

where $\tilde{\epsilon}$ is a standardized normal variable, i.e. a normal variable with mean 0 and variance 1. Notice that the expectation (6.3.4) depends only on S_t and t.

When the new security is a European call or a European put, we get the Black-Scholes prices. We denote the price of a European call by $C(S_t, t)$, and that of a European put by $P(S_t, t)$.

Proposition 4.2.3. *We have*

$$C\left(S_t, t\right) = S_t N\left(z_1\right) - \exp(-r(T-t))KN\left(z_2\right), \tag{4.2.23}$$

and

$$P\left(S_t, t\right) = \exp(-r(T-t))KN\left(-z_2\right) - S_t N\left(-z_1\right)$$

where $N(.\ )is the cumulative distribution function of the standard normal distribu-$ *tion,*

$$z_1 = \frac{\log\left(\frac{S_t}{K}\right) + \left(r + \frac{1}{2}\sigma^2\right)(T-t)}{\sigma\sqrt{T-t}}$$

and

$$z_2 = z_1 - \sigma\sqrt{T-t}.$$

Proof. We compute the price of the call. The price of the put follows by a similar argument, or by put-call parity

$$P\left(S_t, t\right) = C\left(S_t, t\right) + K\exp(-r(T-t)) - S_t. \tag{4.2.24}$$

The time T payoff of the call is

$$\begin{aligned} G\left(S_T\right) &= \max\left(S_T - K, 0\right) \\ &= \max\left[S_t \exp\left[\left(r - \frac{1}{2}\sigma^2\right)(T-t) + (\sigma\sqrt{T-t})\tilde{\epsilon}\right] - K, 0\right]. \end{aligned} \tag{4.2.25}$$

This is equal to 0 for

$$S_t \exp\left[\left(r - \frac{1}{2}\sigma^2\right)(T-t) + (\sigma\sqrt{T-t})\tilde{\epsilon}\right] \leq K \tag{4.2.26}$$

i.e.,

$$z_2 \equiv \frac{\log\left(\frac{S_t}{K}\right) + \left(r - \frac{1}{2}\sigma^2\right)(T-t)}{\sigma\sqrt{T-t}} \leq -\tilde{\epsilon}, \tag{4.2.27}$$

and to

$$S_t \exp\left[\left(r - \frac{1}{2}\sigma^2\right)(T-t) + (\sigma\sqrt{T-t})\tilde{\epsilon}\right] - K \tag{4.2.28}$$

otherwise.

The call price is

$$\begin{aligned} C(S_t, t) &= \exp(-r(T-t))E\left[\max\left[S_t \exp\left[\left(r - \frac{1}{2}\sigma^2\right)(T-t) + (\sigma\sqrt{T-t})\tilde{\epsilon}\right] - K, 0\right]\right] \\ &= C_1 + C_2, \end{aligned} \tag{4.2.29}$$

where

$$C_1 = \exp(-r(T-t))\frac{1}{\sqrt{2\pi}}\int_{-z_2}^{\infty} S_t \exp\left[\left(r - \frac{1}{2}\sigma^2\right)(T-t) + (\sigma\sqrt{T-t})y\right]\exp\left(-\frac{1}{2}y^2\right)dy \tag{4.2.30}$$

and

$$C_2 = -\exp(-r(T-t))\frac{1}{\sqrt{2\pi}}\int_{-z_2}^{\infty} K \exp\left(-\frac{1}{2}y^2\right)dy \tag{4.2.31}$$

The term C_2 is equal to

$$C_2 = -\exp(-r(T-t))K\left(1-N\left(-z_2\right)\right) = -\exp(-r(T-t))KN\left(z_2\right). \tag{4.2.32}$$

Therefore, it is equal to the second term in the Black-Scholes equation (6.3.5).

The term C_1 is equal to

$$\begin{aligned} C_1 &= \exp(-r(T-t))S_t \exp(r(T-t))\frac{1}{\sqrt{2\pi}}\int_{-z_2}^{\infty}\exp\left[-\frac{1}{2}(y-\sigma\sqrt{T-t})^2\right]dy \\ &= S_t\frac{1}{\sqrt{2\pi}}\int_{-z_2-\sigma\sqrt{T-t}}^{\infty}\exp\left(-\frac{1}{2}y^2\right)dy \end{aligned} \tag{4.2.33}$$

Setting $z_1 = z_2 + \sigma\sqrt{T-t}$, we can write C_1 as

$$C_1 = S_t\left(1-N\left(-z_1\right)\right) = S_t N\left(z_1\right). \tag{4.2.34}$$

Therefore, C_1 is equal to the first term in the Black-Scholes equation (4.2.23). □

Pricing: The PDE Approach

So far, we priced redundant securities by following the martingale approach, i.e. by using the fact that the price of a redundant security is the expectation, under the EMM, of the security's discounted payoff. The original approach to arbitrage pricing was the PDE approach. It consists in deriving a Partial Differential Equation (PDE) for the price of the redundant security. We now present the PDE approach, in the context of the Black-Scholes model. We also explain the relation between the martingale and the PDE approaches.

We assume that the time T payoff of the redundant security is $G\left(S_T\right)$, and denote the time t price of the security by $g\left(S_t, t\right)$. The martingale approach implies that

$$g\left(S_t, t\right) = E_t^Q\left[\exp(-r(T-t))G\left(S_T\right)\right]. \tag{4.2.35}$$

We can write this as

$$\exp(-rt)g\left(S_t,t\right)=E_t^Q\left[\exp(-rT)G\left(S_T\right)\right] \tag{4.2.36}$$

By the law of iterative expectations, the process $\exp(-rt)g\left(S_t,t\right)$ is a martingale under Q.

Suppose that the function $g(S,t)$ is twice continuously differentiable. Noting that

$$dS_t=rS_tdt+\sigma S_tdZ_t^Q \tag{4.2.37}$$

and applying Ito's lemma, we get

$$\begin{aligned} d\big(\exp(-rt)g(S_t,t)\big)=&\exp(-rt)\Big[-rg(S_t,t)+\mathcal{D}_Sg(S_t,t)+g_t(S_t,t)\Big]dt\\ &+\exp(-rt)g_S(S_t,t)\sigma S_tdZ_t^Q. \end{aligned} \tag{4.2.38}$$

where

$$\mathcal{D}_Sg\left(S_t,t\right)=g_S\left(S_t,t\right)rS_t+\frac{1}{2}g_{SS}\left(S_t,t\right)\sigma^2S_t^2 \tag{4.2.39}$$

Since the process $\exp(-rt)g\left(S_t,t\right)$ is a martingale under Q, the drift is equal to 0 . Therefore, the function $g(S,t)$ solves the PDE

$$-g(S,t)r+\mathcal{D}_Sg(S,t)+g_t(S,t)=0 \tag{4.2.40}$$

The PDE approach consists in solving the PDE (6.3.12), with the terminal condition $g(S,T)=G(S)$.

Conversely, suppose that the function $g(S,t)$ solves the PDE (6.3.12), with the terminal condition $g(S,T)=G(S)$. Then, equation (6.3.9) follows from the Feynman-Kac theorem.

The Feynman-Kac theorem concerns the general PDE

$$h(S,t) - g(S,t)r(S,t) + \left[g_S(S,t)\mu(S,t) + \frac{1}{2} g_{SS}(S,t)\sigma(S,t)^2 + g_t(S,t) \right] = 0, \qquad (4.2.41)$$

with the terminal condition $g(S,T) = G(S)$. To this PDE, is associated a stochastic differential equation (SDE)

$$dS_t = \mu\left(S_t, t\right) dt + \sigma\left(S_t, t\right) dZ_t, \qquad (4.2.42)$$

Under regularity conditions, there exists an Ito process S^x that solves the SDE (6.3.14) with the initial condition $S_t = x$. The Feynman-Kac theorem is that, under regularity conditions, a solution to the PDE (6.3.13) is given by

$$g(x,t) = E\left[\int_t^T \phi(t,s) h\left(S_s^x, s\right) ds + \phi(t,T) G\left(S_T^x\right) \right] \qquad (4.2.43)$$

where

$$\phi(t,s) = \exp\left(\int_t^s r\left(S_u^x, u\right) du \right) \qquad (4.2.44)$$

4.2.4 Portfolio Choice: Martingale Approach

We assume that trading strategies are in $\mathcal{L}(S)$ and are such that the stochastic integral $\int_0^t \theta_s d\left(S_s/B_s\right)$ is a martingale under Q. (Alternatively, we could assume that trading strategies are in $\mathcal{L}(S) \cap \underline{\Theta}(S/B)$.)

Definition 4.2.4. A consumption plan is a pair (c, C_T) where $c \in \mathcal{L}, C_T$ is $\mathcal{F}_T$-measurable, $c \geq 0$, and $C_T \geq 0$.

We consider an investor who consumes over time. The investor has wealth W in period 0 . The investor's preferences are given by a time-additive expected utility function

$$U(c, C_T) = E\left[\int_0^T u_t(c_t)\, dt + U_T(C_T)\right] \tag{4.2.45}$$

where the functions u_t and U_T are strictly increasing and concave. The integral and the expectation in (6.5.15) are well-defined if both u_t and U_T are bounded. If this is not the case (for instance $u_t(c) = \log(c_t)$) then we consider only those consumption plans for which the integral and the expectation can be defined.

Definition 4.2.5. A consumption plan is feasible iff the cash flow $(-W, c, C_T)$ is marketable.

We denote by $\mathcal{C}$ the set of feasible consumption plans. The investor's problem, $\mathcal{P}$, is

$$\begin{gathered}\max_{c, C_T} U(c, C_T) \\ (c, C_T) \in \mathcal{C}\end{gathered} \tag{4.2.46}$$

Definition 4.2.6. A consumption plan (c, C_T) is optimal iff it solves $\mathcal{P}$. A trading strategy is optimal iff it finances $(-W, c, C_T)$, where (c, C_T) is an optimal consumption plan.

As in discrete time, we can solve $\mathcal{P}$ using the martingale approach or the dynamic programming approach.

As in discrete time, the martingale approach is most powerful when there are complete markets. This is because, when markets are complete, the optimization problem $\mathcal{P}$ is equivalent to a simple static problem. We first assume complete markets (the rank of the diffusion matrix σ_t is equal to d a.s.) and proceed in four steps. First, we show the equivalence between the problem $\mathcal{P}$ and the static problem. Second, we solve the static problem, and determine the optimal consumption plan. Third, we determine the optimal trading strategy and fourth, we solve the problem $\mathcal{P}$ in a special case. We next assume incomplete markets and explain how the martingale approach extends.

The Static Problem

Proposition 4.2.4. *When markets are complete, the problem $\mathcal{P}$ is equivalent to the problem $\mathcal{P}_Q$ given by*

$$\max_{c,C_T} U(c, C_T), \quad \textit{subject to } E^Q\left[\int_0^T \frac{c_t}{B_t}dt + \frac{C_T}{B_T}\right] = W, c \geq 0, \quad C_T \geq 0. \tag{4.2.47}$$

Proof. We first show that a feasible consumption plan (c, C_T) satisfies equation (6.5.16). Denote by θ the trading strategy that finances the cash flow $(-W, c, C_T)$. Since θ finances the discounted cash flow $(-W, c/B, C_T/B_T)$ under the discounted price process S/B, we have

$$\theta_t S_t/B_t = W + \int_0^t \theta_s d\left(\frac{S_s}{B_s}\right) - \int_0^t c_s/B_s ds \tag{4.2.48}$$

Since the stochastic integral $\int_0^t \theta_s d(S_s/B_s)$ is a martingale under Q, we have

$$E^Q\left[\int_0^t \theta_s d\left(\frac{S_s}{B_s}\right)\right] = 0 \tag{4.2.49}$$

Equations (6.5.17) and (6.5.18) imply that

$$E^Q\left[\int_0^t c_s/B_s\, ds\right] = W - E^Q[\theta_t S_t/B_t]. \tag{4.2.50}$$

Setting $t = T$ in equation (6.5.19), and noting that $\theta_T S_T = C_T$, we get equation (6.5.16).

We next show the converse, i.e., that if a consumption plan (c, C_T) satisfies equation (6.5.16), then it is feasible. We use the martingale representation theorem and construct a strategy that finances the cash flow $(-W, c, C_T)$. The random variable

$$X = \int_0^T \frac{c_t}{B_t} dt + \frac{C_T}{B_T} \tag{4.2.51}$$

belongs to $L^1(Q)$ since it is positive and has finite expectation (and equal to W). By the law of iterative expectations, the process $E_t^Q(X)$ is a martingale under Q. The martingale representation theorem implies that there exists a process $\zeta \in (\mathcal{L}^2)^d$ such that

$$E_t^Q(X) = E^Q(X) + \int_0^t \zeta_s dZ_s^Q \tag{4.2.52}$$

Since the rank of σ_t is equal to d, we can find a process θ_{1N} such that

$$\theta_{1N,t}\hat{\sigma}_t = \zeta_t \tag{4.2.53}$$

The process $\theta_{1N,t}$ represents the investment in the risky assets. We define the investment in the riskless asset by

$$E_t^Q \left[\int_t^T c_s/B_s ds + \frac{C_T}{B_T} \right] = \theta_t S_t / B_t \tag{4.2.54}$$

We need to show that the strategy θ finances the cash flow $(-W, c, C_T)$, satisfies $\theta_T S_T = C_T$, is in $\mathcal{L}(S)$, and is such that the stochastic integral $\int_0^t \theta_s d\left(S_s/B_s\right)$ is a martingale under Q. To show that θ finances the cash flow (W, c, C_T), we note that equations (6.1.4, 6.5.16, 6.5.20, and 6.5.21), imply that

$$E_t^Q(X) = W + \int_0^t \theta_s d\left(\frac{S_s}{B_s}\right) \tag{4.2.55}$$

Equations (6.5.22) and (6.5.23) imply equation (6.5.17). Therefore, θ finances the discounted cash flow $(W, c/B, C_T/B_T)$ under the discounted price process S/B. It thus also finances the undiscounted cash flow $(-W, c, C_T)$ under the undiscounted price process S.

The proof of the remaining properties of θ is as in the theorem that markets are complete iff the rank of σ_t is equal to d.

□

Proposition 4.2.4 implies that when markets are complete, the problem $\mathcal{P}$ is equivalent to the simple "static" problem $\mathcal{P}_Q$. This static problem consists in maximizing utility subject to the static budget constraint (6.5.16).

Note: When the horizon is infinite, a similar static formulation can be obtained by setting $T = \infty$. Huang and Pages (1992) show under what conditions the infinite-horizon problem has a solution and when this solution can be recovered as a limit $T \to \infty$ of solutions for finite-horizon economies. Note one important issue for infinite-horizon problems: the measure Q is no longer equivalent to P on $\mathcal{F}_\infty$. In fact, the two measures are typically mutually singular, i.e., there exists a set $K \in \mathcal{F}_\infty$, such $E^P\left[1_{\{K\}}\right] = 1$ while $E^Q\left[1_{\{K\}}\right] = 0$. This is because when $\int_0^\infty \eta_t^2 dt = \infty$ with P-probability $1, \lim_{t\to\infty} \xi_t^\eta = 0$ almost surely as well (see Huang and Pages 1992, Lemma 1).

Optimal Consumption

To solve the static problem $\mathcal{P}_Q$, we write the static budget constraint (6.5.16) in terms of the probability measure P rather than Q. We denote the Radon-Nikodym derivative of Q w.r.t. P by ξ_T, and set $\xi_t = E_t(\xi_T)$. We have

$$\begin{aligned} &E^Q\left[\int_0^T \frac{c_t}{B_t}dt + \frac{C_T}{B_T}\right] = E\left[\xi_T\left(\int_0^T \frac{c_t}{B_t}dt + \frac{C_T}{B_T}\right)\right] \\ =&E\left[\int_0^T \frac{\xi_t}{B_t}c_t dt + \frac{\xi_T}{B_T}C_T\right] = E\left[\int_0^T \pi_t c_t dt + \pi_T C_T\right]. \end{aligned} \tag{4.2.56}$$

(For the second equality we used Fubini's theorem, and for the third the relation between the EMM and the SPD.) Therefore, we can write the static problem $\mathcal{P}_Q$ as

$$\max_{c, C_T} E\left[\int_0^T u_t(c_t)\,dt + U_T(C_T)\right] \tag{4.2.57}$$

subject to

$$E\left[\int_0^T \pi_t c_t dt + \pi_T C_T\right] = W \tag{4.2.58}$$

and

$$c_t \geq 0, \quad C_T \geq 0 \tag{4.2.59}$$

Let's proceed heuristically, and write the Lagrangian

$$L = E\left[\int_0^T u_t\left(c_t\right) dt + U_T\left(C_T\right)\right] + \lambda\left[W - E\left[\int_0^T \pi_t c_t dt + \pi_T C_T\right]\right] \tag{4.2.60}$$

where λ is the Lagrange multiplier of the static budget constraint (6.5.24). The first-order conditions are

$$u_t'\left(c_t^*\right) = \lambda \pi_t \tag{4.2.61}$$

and

$$U_T'\left(C_T^*\right) = \lambda \pi_T \tag{4.2.62}$$

and are equalities as long as $c_t^*, C_T^* > 0$. To ensure that $c_t^*, C_T^* > 0$, we assume from now on that the functions u_t and U_T satisfy the Inada conditions. These are

$$\lim_{c_t \to 0} u_t'\left(c_t\right) = \infty \quad \lim_{c_t \to \infty} u_t'\left(c_t\right) = 0$$

and similarly for U_T. Notice that equations (6.5.25) and (6.5.26) are the same as in discrete

time.

Denoting by $i_t(y)$ the inverse of u_t' and by $I_t(y)$ the inverse of U_T', we can write the first-order conditions as

$$c_t^* = i_t\left(\lambda\pi_t\right) \tag{4.2.63}$$

and

$$C_T^* = I_T\left(\lambda\pi_T\right) \tag{4.2.64}$$

To determine c_t^* and C_T^*, we need to determine the Lagrange multiplier $\lambda.\lambda$ is determined by the static budget constraint (6.5.24)

$$E\left[\int_0^T \pi_t i_t\left(\lambda\pi_t\right)dt + \pi_T I_T\left(\lambda\pi_T\right)\right] = W \tag{4.2.65}$$

Proposition 4.2.5 makes the above heuristic analysis rigorous.

Proposition 4.2.5. *Suppose that there exists λ solving equation* (4.2.65)*. Then the solution to $\mathcal{P}_Q$ is given by equations* (4.2.63) *and* (4.2.64)*.*

Proof. Consider a consumption plan (c, C_T) that satisfies the static budget constraint (6.5.24). We will show that expected utility is smaller than under the consumption plan (c^*, C_T^*). Since u_t is concave, we have

$$\begin{aligned} u_t\left(c_t\right) &\leq u_t\left(c_t^*\right) + u_t'\left(c_t^*\right)\left(c_t - c_t^*\right) \\ &\leq u_t\left(c_t^*\right) + \lambda\pi_t\left(c_t - c_t^*\right) \end{aligned}$$

and similarly for U_T. Integrating and taking expectations, we get

$$U(c, C_T) \leq U(c^*, C_T^*) + \lambda \left[F(c, C_T) - F(c^*, C_T^*) \right],$$

where

$$F(c, C_T) = E\left[\int_0^T \pi_t c_t dt + \pi_T C_T \right]$$

Since (c, C_T) satisfies the static budget constraint, we have $F(c, C_T) = W$. Since λ solves equation (4.2.65), we have $F(c^*, C_T^*) = W$. Therefore, (c^*, C_T^*) dominates (c, C_T). □

Optimal Trading Strategy

In Proposition 4.2.4 we constructed a trading strategy that finances a cash flow (W, c, C_T). This general construction applies to the cash flow (W, c^*, C_T^*) that corresponds to the optimal consumption plan. However, the construction is based on the martingale representation theorem, and does not explicitly determine the investment in the risky assets. We now determine this investment.

We consider the wealth W_t^* required at time t to finance the optimal consumption plan from time t on. This is equal to the expectation under Q of the discounted value of the consumption plan at time t, i.e.

$$W_t^* = B_t E_t^Q \left[\int_t^T \frac{c_s^*}{B_s} ds + \frac{C_T^*}{B_T} \right] = \frac{1}{\pi_t} E_t \left[\int_t^T \pi_s i_s(\lambda \pi_s) ds + \pi_T I_T(\lambda \pi_T) \right]. \quad (4.2.66)$$

We want to write the wealth W_t^* as function of some "state" variables. For notational simplicity, we set $S = (S_1, .., S_N)$ from now on. We assume that the short rate r_t, and the drift μ_t and diffusion σ_t of the Ito process S, depend only on the value of some state variables, X, at time t, and on t. For simplicity, we further assume that the state variables are the prices, i.e. $X = S$. We also recall that the SPD π evolves according to

$$d\pi_t = -\pi_t r_t dt - \pi_t \eta_t' dZ_t. \tag{4.2.67}$$

Since the risk premia η_t depend only on S_t and t, the wealth W_t^* depends only on π_t, S_t, and t. We can thus set $W_t^* = F(\pi_t, S_t, t)$. The optimal trading strategy can be derived from the function F. The function F can be computed in two equivalent ways. First, directly, as an expectation, and second, indirectly, as a solution of a PDE. We will derive the PDE, and then show how the optimal trading strategy can be derived from F.

4.3 The PDE

The derivation of the PDE is very similar to that of the Black-Scholes PDE. The differences are that there is intermediate consumption, and that there are $N+2$ variables (π, S, and t), instead of just 2.

The process

$$\int_0^t \frac{c_s^*}{B_s} ds + \frac{F(\pi_t, S_t, t)}{B_t} \tag{4.3.1}$$

is a martingale under Q since it is equal to $E_t^Q(X)$, where

$$X = \int_0^T \frac{c_t^*}{B_t} dt + \frac{C_T}{B_T} \tag{4.3.2}$$

Equation (6.5.31) implies that

$$d\pi_t = \pi_t \left(-r_t + \eta_t^2\right) dt - \pi_t \eta_t' dZ_t^Q \tag{4.3.3}$$

and equation (6.1.4) implies that

$$dS_t = r_t S_t dt + \sigma_t dZ_t^Q \tag{4.3.4}$$

Ito's lemma implies that the drift term of the process $E_t^Q(X)$ is

$$\frac{1}{B_t}\left(c_t^* - r_t F\left(\pi_t, S_t, t\right) + \mathcal{D}_{\pi S} F\left(\pi_t, S_t, t\right) + F_t\left(\pi_t, S_t, t\right)\right) \tag{4.3.5}$$

where

$$\mathcal{D}_{\pi S} F = F_\pi \pi_t \left(-r_t + \eta_t^2\right) + F_S r_t S_t + \frac{1}{2}\left(\pi_t^2 \eta_t^2 F_{\pi\pi} - 2\pi_t F'_{\pi S}\sigma_t \eta_t + \operatorname{tr}\left(\sigma_t \sigma_t' F_{SS}\right)\right) \tag{4.3.6}$$

and the diffusion term (the coefficient of dZ_t^Q) is

$$\frac{1}{B_t}\left(-F_\pi\left(\pi_t, S_t, t\right)\pi_t \eta_t' + F_S'\left(\pi_t, S_t, t\right)\sigma_t\right) \tag{4.3.7}$$

Since the process $E_t^Q(X)$ is a martingale under Q, the drift term is equal to 0 . Therefore, the function $F(\pi, S, t)$ solves the PDE

$$c_t^* - r_t F(\pi, S, t) + \mathcal{D}_{\pi S} F(\pi, S, t) + F_t(\pi, S, t) = 0 \tag{4.3.8}$$

with the terminal condition $F(\pi, S, T) = C_T^*$. We have thus shown that the function $F(\pi, S, t)$ defined by

$$F\left(\pi_t, S_t, t\right) = B_t E_t^Q\left[\int_t^T \frac{c_s^*}{B_s} dt + \frac{C_T^*}{B_T}\right] \tag{4.3.9}$$

solves the PDE (6.5.32) with the terminal condition $F(\pi, S, T) = C_T^*$. The converse follows from the Feynman-Kac theorem.

4.4 Optimal Trading Strategy

Since the drift term of the process $E_t^Q(X)$ is equal to 0 , we have

$$E_t^Q(X) = E^Q(X) + \int_0^t \frac{-F_\pi\left(\pi_s, S_s, s\right)\pi_s\eta_s' + F_S'\left(\pi_s, S_s, s\right)\sigma_s}{B_s} dZ_s^Q \tag{4.4.1}$$

Comparing this equation with equations (6.5.20 and 6.5.21), we define the investment in the risky assets by

$$\left(\theta_{1N,t}^*\right)\sigma_t = -F_\pi\left(\pi_t, S_t, t\right)\pi_t\eta_t' + F_S'\left(\pi_t, S_t, t\right)\sigma_t \tag{4.4.2}$$

Equation (6.5.33) has a unique solution iff $N = d$. To obtain the solution for $N = d$, we multiply equation (6.5.33) from the left by σ_t', and use equation (6.1.3). We get

$$\left[\theta_{1N,t}^*\right]' = -F_\pi\left(\pi_t, S_t, t\right)\pi_t\left(\sigma_t\sigma_t'\right)^{-1}\left(\mu_t - r_t S_t\right) + F_S\left(\pi_t, S_t, t\right) \tag{4.4.3}$$

We define the investment in the riskless asset by

$$F\left(\pi_t, S_t, t\right) = \theta_{0,t}^* B_t + \theta_{1N,t}^* S_t. \tag{4.4.4}$$

The optimal trading strategy can thus be derived from the function F.

Incomplete Markets

When markets are incomplete, things are more complicated. The problem $\mathcal{P}$ is no longer equivalent to a simple static problem of maximizing utility subject to a single budget constraint 6.5.16. In fact, since there are many EMMs, there is a budget constraint associated to each EMM.

As in discrete time, we can show a duality result. This is that the problem $\mathcal{P}$ is equivalent to

a dual problem, which consists of (i) solving the complete markets problem for each EMM and (ii) minimizing the maximum value over all EMMs.

4.5 Equilibrium: Continuous-Time Models

4.5.1 The Model

We consider a probability space $(\Omega, \mathcal{F}, P)$, a time inverval $\mathcal{T} = [0, T]$, a Brownian motion $Z = (Z_1, .., Z_d)$ on $(\Omega, \mathcal{F}, P)$, and the standard filtration $\mathbb{F}$ of Z. We assume that there are N securities that pay dividends at a rate $\delta = (\delta_{1,t}, .., \delta_{N,t}) \in (\mathcal{L}^1)^N$, and have a time T price equal to $S_T = (S_{1,T}, .., S_{N,T})$. The supply of the securities is $x = (x_1, .., x_N)$. Abusing notation, we denote by x the process that is always equal to x. There are I agents. Agent i 's preferences are given by a time-additive expected utility function

$$U_i(c_i, C_{i,T}) = E\left[\int_0^T u_{i,t}(c_{i,t})\, dt + U_{i,T}(C_{i,T})\right] \tag{4.5.1}$$

where the functions $u_{i,t}$ and $U_{i,T}$ are strictly increasing and concave. Agent i receives an endowment of the consumption good at a rate $e_i \in \mathcal{L}^1$. He also receives an endowment of the securities at time 0 . Denoting the latter endowment by $\bar{\theta}_{i,0}$, we have

$$\sum_{i=1}^{I} \bar{\theta}_{i,0} = x \tag{4.5.2}$$

We consider security price processes that are equal to S_T at time T, and are of the form

$$dS_t = \mu_t dt + \sigma_t dZ_t, \tag{4.5.3}$$

where $\mu \in (\mathcal{L}^1)^N$ and $\sigma \in (\mathcal{L}^2)^{N \times d}$. We set $\mu_t = I_{S_t} \bar{\mu}_t$ and $I_{S_t} \bar{\sigma}_t$, where I_{S_t} is a $N \times N$ diagonal matrix with n 'th diagonal element equal to $S_{n,t}$. We assume that trading strategies

are in $\mathcal{L}(S)$ and are such that the stochastic integral $\int_0^t \theta_s d\left(S_s/B_s\right)$ is a martingale under Q. (Alternatively, we could assume that trading strategies are in $\mathcal{L}(S) \cap \underline{\Theta}(S/B)$.)

To a price process S, we associate a set of marketable cash flows. A consumption plan $(c_i, C_{i,T})$ is feasible for agent i iff the cash flow $\left(-\bar{\theta}_{i,0}S_0, c_i - e_i, C_{i,T}\right)$ is marketable. We denote by $\mathcal{C}_i$ the set of feasible cash flows for agent i. Agent i 's problem $\mathcal{P}_i$ is

$$\begin{aligned} &\max_{c_i, C_{i,T}} U\left(c_i, C_{i,T}\right) \\ &\left(c_i, C_{i,T}\right) \in \mathcal{C}_i \end{aligned} \tag{4.5.4}$$

A consumption plan $(c_i, C_{i,T})$ is optimal iff it solves $\mathcal{P}_i$. A trading strategy θ_i is optimal iff it finances the cash flow $\left(-\bar{\theta}_{i,0}S_0, c_i - e_i, C_{i,T}\right)$.

Definition 8.1.1 A securities market (SM) equilibrium is a price process S, a vector of trading strategies $(\theta_1, .., \theta_I)$, and consumption policies $(c_1, .., c_I)$, such that

1. (optimization) (c_i, θ_i) is optimal for agent i
2. (market-clearing)

$$\begin{aligned} \sum_{i=1}^{I} \theta_i &= x \\ \sum_{i=1}^{I} e_i - c_i &= 0 \end{aligned} \tag{4.5.5}$$

4.5.2 CAPMs

We assume that an equilibrium exists. We also assume that there exists an equilibrium short rate process r_t. This process is such that the instantaneous demand for riskless borrowing and lending is 0 . For simplicity, we will assume that risky assets do not pay dividends.

Since in equilibrium there is no arbitrage, the equation

$$\bar{\mu}_t - r_t 1 = \bar{\sigma}_t \eta_t \tag{4.5.6}$$

has a solution η. This solution is unique only when markets are complete.

We will derive two CAPM-type equations that link the expected return on a security to its covariance with aggregate variables. To derive these equations, we start from equation (8.2.1). The n 'th "component" of this equation is

$$\bar{\mu}_{n,t} - r_t = \sum_{j=1}^{d} \bar{\sigma}_{n,j,t} \eta_{j,t} \tag{4.5.7}$$

The LHS of equation (8.2.2) is equal to the instantaneous expected return on security n, $\bar{\mu}_{n,t}$, minus the riskless rate, r_t. This is the risk premium of security n. Equation (8.2.2) links this risk premium to the loadings, $\bar{\sigma}_{n,j,t}$, of security n on the Brownian motions, and to the risk premia, $\eta_{j,t}$, of the Brownian motions.

We next write equation (8.2.2) in terms of the instantaneous covariance of security n with the SPD π that corresponds to a solution η of equation (8.2.1). π evolves according to

$$d\pi_t = -\pi_t r_t dt - \pi_t \eta_t' dZ_t. \tag{4.5.8}$$

Using the notation

$$E_t\left(\frac{dS_{n,t}}{S_{n,t}}\right) = \frac{E_t\left(dS_{n,t}\right)}{S_{n,t}} = \frac{\mu_{n,t} dt}{S_{n,t}} = \bar{\mu}_{n,t} dt \tag{4.5.9}$$

and

$$\operatorname{Cov}_t\left(\frac{dS_{n,t}}{S_{n,t}},\frac{d\pi_t}{\pi_t}\right)=\frac{\operatorname{Cov}_t\left(dS_{n,t},d\pi_t\right)}{S_{n,t}\pi_t}=-\frac{\sum_{j=1}^{d}\sigma_{n,j,t}\left(\pi_t\eta_{j,t}\right)dt}{S_{n,t}\pi_t}=-\sum_{j=1}^{d}\bar{\sigma}_{n,j,t}\eta_{j,t}dt \tag{4.5.10}$$

we can write equation (8.2.2) as

$$\frac{E_t\left(dS_{n,t}\right)}{S_{n,t}}-r_tdt=-\operatorname{Cov}_t\left(\frac{dS_{n,t}}{S_{n,t}},\frac{d\pi_t}{\pi_t}\right) \tag{4.5.11}$$

The risk premium of security n is thus proportional to minus the instantaneous covariance of the security with the SPD. When markets are complete, equation (8.2.3) holds for the unique SPD. When markets are incomplete, it holds for any SPD. The intuition behind equation (8.2.3) is that a security is valuable, and thus has a low risk premium, if it has a high payoff in high SPD states, i.e. in states where consumption is valuable.

An important implication of (8.2.3) is that the instantaneous Sharpe ratio of stock returns is bounded from above by the volatility of the state-price density:

$$\left|\frac{\bar{\mu}_n-r}{\bar{\sigma}_n}\right|\leq\|\eta_t\| \tag{4.5.12}$$

Thus, for any model to generate high Sharpe ratios, it needs to have a sufficiently high volatility of the state-price density. This is a very important result, which is known in discrete time as the Hansen-Jaganathan bound.

CCAPM

The first CAPM equation is the consumption CAPM (CCAPM). When markets are complete, the optimal consumption of agent i is given by

$$u'_{i,t}\left(c_{i,t}\right)=\lambda_i\pi_t \tag{4.5.13}$$

where π is the unique SPD. When markets are incomplete, equation (8.2.5) holds for some SPD π, because of the duality result. Equation (8.2.5) implies that

$$d\pi_t = \frac{1}{\lambda_i} d\left[u'_{i,t}\left(c_{i,t}\right)\right] = \frac{1}{\lambda_i}\left[u''_{i,t}\left(c_{i,t}\right) dc_{i,t} + \text{ terms in } dt\right] \tag{4.5.14}$$

Equations (8.2.3, 8.2.5, and 8.2.6) imply that

$$\frac{E_t\left(dS_{n,t}\right)}{S_{n,t}} - r_t dt = \left[-\frac{u''_{i,t}\left(c_{i,t}\right)}{u'_{i,t}\left(c_{i,t}\right)}\right] \mathrm{Cov}_t\left(\frac{dS_{n,t}}{S_{n,t}}, dc_{i,t}\right) \tag{4.5.15}$$

Equation (8.2.7) links the risk premium of security n to the instantaneous covariance of the security with the consumption of agent i. Note that, when markets are complete, (8.2.5) implies that consumption growth is perfectly instantaneously correlated across agents, since

$$dc_{i,t} = [\ldots]dt + \lambda_i i'_{i,t}\left(\lambda_i \pi_t\right) d\pi_t \tag{4.5.16}$$

where $i_{i,t}(\cdot)$ denotes the inverse of $u'_{i,t}(\cdot)$.

To obtain the covariance with the aggregate consumption, we divide (8.2.7) by the term in brackets and sum across agents. Denoting the aggregate consumption by c_t, we get

$$\frac{E_t\left(dS_{n,t}\right)}{S_{n,t}} - r_t dt = A_t \,\mathrm{Cov}_t\left(\frac{dS_{n,t}}{S_{n,t}}, dc_t\right) \tag{4.5.17}$$

where

$$A_t = -\frac{1}{\sum_{i=1}^{I} \frac{u'_{i,t}(c_{i,t})}{u''_{i,t}(c_{i,t})}} > 0 \tag{4.5.18}$$

This is the CCAPM. It states that the instantaneous risk premium of a security is proportional to the instantaneous covariance of the security with aggregate consumption.

The intuition behind the CCAPM is that a security is valuable, and thus has a low risk premium, if it has a high payoff in states where consumption is low. This intuition is the same as in discrete time. In continuous time, however, there are two advantages. First, the CCAPM holds even when markets are incomplete. Second, the CCAPM involves the covariance with consumption and not with the marginal utility of consumption. The intuition is that in a small time interval, changes are small. Therefore, agents' utility is approximately quadratic, and thus mean-variance analysis can apply. Mean-variance analysis does not require complete markets, and involves the covariance with consumption.

If we are considering a representative-agent economy, or a complete-market economy in which a representative agent can be constructed by aggregation of individual utilities, then (8.2.4) implies that

$$\left|\frac{\bar{\mu}_n - r}{\bar{\sigma}_n}\right| \leq \gamma_t \sigma_{c,t} \tag{4.5.19}$$

where γ_t is the curvature of the utility function (of the representative agent) at the current consumption level, $\gamma_t = c_t u_t'' (c_t) / u_t' (c_t)$, and $\sigma_{c,t}$ is the instantaneous volatility of consumption growth. The above constraint has important implications for building models with empirically realistic quantitative implications.

The CCAPM is often associated with complete financial markets. This is too restrictive. We do not need market to be dynamically complete to obtain the CCAPM relation, as shown above. In fact, one can generalize the CCAPM even further. Our analysis below is somewhat heuristic and emphasizes the intuition behind the results. Proving the formal statements can be quite difficult in continuous-time models.

ICAPM

The second CAPM equation is the intertemporal CAPM (ICAPM). Suppose that the drift μ_t and diffusion σ_t of the Ito process S depend only on the value of some state variables, X, at time t, and on t. Then the value function of agent i at time t depends on the agent's wealth $W_{i,t}$, on X_t, and on t. We denote this value function by $V_i (W_{i,t}, X_t, t)$. The first order

condition for consumption is

$$u'_{i,t}\left(c_{i,t}\right) = V_{i,W}\left(W_{i,t}, X_t, t\right) \tag{4.5.20}$$

Equations (8.2.5) and (8.2.11) imply that

$$d\pi_t = \frac{1}{\lambda_i} d\left[V_{i,W}\left(W_{i,t}, X_t, t\right)\right] = \frac{1}{\lambda_i}\left[V_{i,WW} dW_{i,t} + \left(V_{i,WX}\right)' dX_t + \text{ terms in } dt\right]. \tag{4.5.21}$$

Equations (8.2.3, 8.2.5, 8.2.11, and 8.2.12) imply that

$$\frac{E_t\left(dS_{n,t}\right)}{S_{n,t}} - r_t dt = \left(-\frac{V_{i,WW}}{V_{i,W}}\right) \mathrm{Cov}_t\left(\frac{dS_{n,t}}{S_{n,t}}, dW_{i,t}\right) + \left(-\frac{\left(V_{i,WX}\right)'}{V_{i,W}}\right) \mathrm{Cov}_t\left(\frac{dS_{n,t}}{S_{n,t}}, dX_t\right). \tag{4.5.22}$$

Equation (8.2.13) links the risk premium of security n to the instantaneous covariance of the security with the wealth of agent i and with the state variables X. To obtain the covariance with the aggregate wealth W_t, we proceed as for the CCAPM, and get

$$\frac{E_t\left(dS_{n,t}\right)}{S_{n,t}} - r_t dt = A_t^W \mathrm{Cov}_t\left(\frac{dS_{n,t}}{S_{n,t}}, dW_t\right) + A_t^X \mathrm{Cov}_t\left(\frac{dS_{n,t}}{S_{n,t}}, dX_t\right) \tag{4.5.23}$$

for two processes $A_t^W > 0$ and A_t^X. This is the ICAPM. It states that the risk premium of a security depends on the instantaneous covariance of the security with aggregate wealth and with the state variables X. Note that instead of the covariance with changes in the state variable X, one can equivalently use the covariance with returns on the hedging portfolios, which are portfolios of assets with maximum correlation with changes in X, i.e. portfolios given by

$$\left(\bar{\sigma}_t \bar{\sigma}_t'\right)^{-1} \bar{\sigma}_t \sigma'_{X,t} \tag{4.5.24}$$

The intuition behind the ICAPM is the same as in discrete time. The risk premium of a security increases in the covariance of the security with aggregate wealth. It also depends on the covariance of the security with the state variables, since the security might be used to hedge changes in the state variables. In continuous time there are three advantages. First, the ICAPM holds even when markets are incomplete. Second, the ICAPM involves the covariance with wealth and not with the marginal utility of wealth. Third, the effects of the covariance with wealth and with the state variables can be very neatly separated.

When the investment opportunity set is constant, agents' value functions depend only on their wealth and on time. The ICAPM thus becomes

$$\frac{E_t\left(dS_{n,t}\right)}{S_{n,t}} - r_t dt = A_t^W \operatorname{Cov}_t\left(\frac{dS_{n,t}}{S_{n,t}}, dW_t\right) \tag{4.5.25}$$

This is the standard CAPM.

4.6 Empirical Facts and Puzzles

Unconditional properties (average behavior).

- Real stock returns have high mean and high variance: in the post-war sample, the average aggregate stock returns are approximately 8% per year, and volatility is around $15 - 16\%$.
- The average risk-free rate is low. Using 3-month T-bills as a proxy, obtain an average real rate of 1% per year. (Caveat: T-bills are nominal, so in real terms returns are not exactly risk-free). Volatility of interest rates in low, estimated at approximately 1% per year.
- Real consumption growth has very low volatility. For non-durables, annual volatility is approximately 1%. The mean of the growth rate is approximately 2% per year. In the longer sample, including the pre-war period, consumption growth is more volatile, up to 4% annual standard deviation.

- Real consumption growth and stock returns have correlation of approximately 0.3 (at one-year horizon).
- Real dividend growth is more volatile than that of consumption, but much less volatile than stock returns. Annual standard deviation is approximately 6%.
- Correlation between consumption and dividend growth is approximately 25%.

4.7 Conditional properties (predictability).

- Excess stock returns are forecastable. The log price-dividend ratio forecasts 10% of the variance at a 1 -year horizon and almost 40% at a 4 -year horizon. Other variables can predict stock returns, e.g., the term spread.
- Consumption growth is not well forecast by it own history or by the price-dividend ratio. The R^2 of the predictive regression (with the price-dividend ratio) is about 4% at horizons of 1 to 4 years.
- Dividend growth is not well forecasted by the $\log$ of the price-dividend ratio. The R^2 is about 8% at horizons of 1 to 4 years.

4.7.1 A Benchmark Model

We now analyze the asset pricing implications of a very simple model.

Consider an economy with a representative agent, who has CRRA preferences:

$$E_0\left[\int_0^T e^{-\rho t}\frac{c_t^{1-\gamma}}{1-\gamma}dt\right], \quad \gamma > 0 \tag{4.7.1}$$

It is common to assume that T is very large and treat it as infinite.

There are two assets in the economy, a stock and a bond. The stock pays a stream of dividends δ_t, given by

$$\frac{d\delta_t}{\delta_t} = \mu_\delta dt + \sigma_\delta dZ_t \tag{4.7.2}$$

The risk-free interest rate is denoted by r_t. The agent is endowed with a single share of the stock.

Because of the market-clearing conditions, the optimal consumption policy is given by

$$c_t^* = \delta_t \tag{4.7.3}$$

and therefore the state-price density is given by

$$\pi_t = e^{-\rho t} \left(\delta_t/\delta_0\right)^{-\gamma} \tag{4.7.4}$$

We can now compute the risk-free interest rate (using Ito's Lemma):

$$r_t = r = \frac{-E_t\left[d\pi_t\right]/\pi_t}{dt} = \rho + \gamma\mu - \frac{1}{2}\gamma(\gamma+1)\sigma^2 \tag{4.7.5}$$

The price of the stock, S_t, is given by the standard formula, $S_t = E_t\left[\int_t^T \left(\pi_s/\pi_t\right)\delta_s ds\right]$, which now takes form

$$S_t = E_t\left[\int_t^T e^{-\rho(s-t)}\left(\delta_s/\delta_t\right)^{-\gamma}\delta_s ds\right] = \delta_t E_t\left[\int_t^T e^{-\rho(s-t)}\left(\delta_s/\delta_t\right)^{1-\gamma} ds\right] = A_t\delta_t, \tag{4.7.6}$$

where

$$A_t \equiv E_t\left[\int_t^T e^{-\rho(s-t)}\left(\delta_s/\delta_t\right)^{1-\gamma} ds\right] \tag{4.7.7}$$

A_t is a deterministic function of time, and when $T \to \infty$, A_t becomes a constant. Note that

the diffusion part of dS_t/S_t equals the diffusion part of $d\delta_t/\delta_t$, for that we don't need infinite T. We can immediately compute the moments of stock returns:

$$\sigma_R = \sigma_\delta \tag{4.7.8}$$

$$\mu_R - r = E_t\left[\frac{dS_t + \delta_t dt}{S_t dt} - r\right] = -\frac{\mathrm{Cov}\left(\frac{dS_t}{S_t}, \frac{d\pi_t}{\pi_t}\right)}{dt} = \gamma\sigma_R\sigma_\delta = \gamma\sigma_\delta^2 \tag{4.7.9}$$

4.7.2 Puzzles

We can identify three problems with the benchmark model. These problems came to be known as puzzles, although the fact that a simple model like the one above does not match the data is hardly surprising.

First, note that the volatility of stock returns equals the volatility of consumption growth. In our model, consumption is the same as dividends, so we cannot distinguish the two, but in either case, we cannot match the level of stock return volatility in the data.

Second, the model cannot simultaneously match the level of the interest rate and the level of the equity premium (in a robust way and with realistic parameter values). Suppose we use $\gamma = 5$. Then, with $\sigma_\delta = 1\%$, we obtain equity premium of 0.0005 . This is the so-called equity-premium puzzle. At the same time, the interest rate equals approximately

$$r = \rho + 0.1 > 10\%, \tag{4.7.10}$$

which is too high. This is known as the risk-free rate puzzle.

One could object that our model equates consumption with dividends, which is counterfactual. Indeed, a model could do a better job if we allowed dividends to be more volatile than consumption. But note that the fundamental puzzles still remain. The risk-free rate is determined entirely by consumption growth. So, we still have the same expression. The instantaneous Sharpe ratio of stock returns is bounded from above by the volatility of the growth rate of the state-price density, which equals $\gamma\sigma_{c^*}$, where σ_{c^*} denotes the volatility of

growth of equilibrium consumption. To get the right equity premium, Sharpe ratio must be approximately $0.06/0.15 = 0.4$. That means $\gamma \geq 40$. Most economists would like to see a risk-aversion parameter below 5 or at most 10. Note also that at such high values of γ the model has ridiculous implications for the risk-free rate. A tiny change in parameters of the consumption growth process implies huge swings in interest rates.

Finally, note that we could test the plausibility of our state-price density process without solving the equilibrium model, but rather by directly estimating the Euler equations. For example, the observed stock returns should satisfy

$$E_t \left[d \left(S_t \cdot \pi_t + \int_0^t \pi_s \delta_s ds \right) \right] = 0 \tag{4.7.11}$$

where

$$\pi_t = e^{-\rho t} \left(c_t^* \right)^{-\gamma} \tag{4.7.12}$$

Of course, empirically one would test a discrete-time version of the Euler equations,

$$E_t \left[\frac{S_{t+1} + \delta_{t+1}}{S_t} \frac{\pi_{t+1}}{\pi_t} - 1 \right] = 0 \tag{4.7.13}$$

Such empirical tests (e.g., Hansen and Singleton (1982)) reject the simple "CRRA" model.

4.7.3 First Attempt: Recursive Preferences

One obvious limitation of our time-separable CRRA model of preferences is that the same parameter γ controls investor's aversion to risk and willingness to substitute over time. In fact, the elasticity of intertemporal substitution ψ is equal to $1/\gamma$ for such preferences. It is natural to investigate more general forms of preferences in order to generate a more realistic state-price density process. One such extension is to allow for the utility function to be non-separable over time. In particular, we consider recursive preferences.

A recursive utility function is defined as a solution to

$$U_t = G\left(c_t, m\left(U_{t+1}\right)\right) \tag{4.7.14}$$

where $m\left(U_{t+1}\right)$ denotes the distribution of continuation utility values U_{t+1}. A common specialization of this relation is

$$U_t = G\left(c_t, v^{-1}E_t\left[v\left(U_{t+1}\right)\right]\right), \tag{4.7.15}$$

where v is another utility function. The standard time-separable utility function is a special case. There are many ways to recover it. For example, let $v(x) = x$ and $G(x, y) = u(x) + \delta y$. We obtain a familiar recursive relation on the expected utility:

$$U_t = u\left(c_t\right) + E_t\left[U_{t+1}\right]. \tag{4.7.16}$$

There are other possibilities, however. For example, let

$$G(x, y) = \left(x^\theta + \delta y^\theta\right)^{1/\theta} \tag{4.7.17}$$

Also, let $v(x) = x^\theta$. Then,

$$U_t^\theta = c_t^\theta + \delta E_t\left[U_{t+1}^\theta\right] \tag{4.7.18}$$

We see that U_t^θ is our standard time-separable utility function, so U_t is obtained by a nonlinear transformation, which preserves ranking of all consumption streams, i.e., is ordinally equivalent. The widely used Epstein-Zin formulation assumes that $v(x) = x^\alpha$, where α is generally different from θ. As we discussed earlier, this model of preferences disentangles risk aversion from elasticity of intertemporal substitution, which helps match

quantitative properties of asset prices and consumption data and is often a useful feature to build into a model. Tractability is an issue, however.

Duffie and Epstein show that the continuous-time analog of recursive utility, known as stochastic differential utility, can be expressed as a solution of

$$U_t = E_t \left[\int_t^T f\left(c_s, U_s\right) ds \right] \tag{4.7.19}$$

where $f(\cdot, \cdot)$ is called an intertemporal aggregator.

Exercise. Use Ito's lemma to check that the standard time-separable utility function

$$E\left[\int_0^T e^{-\rho t} u(c_t)\, dt \right]$$

satisfies the recursive formulation with $f(c, U) = u(c) - \rho U$.

We consider a particular functional form of the aggregator, which makes preferences homothetic:

$$f(c, U) = \frac{1}{1 - \psi^{-1}} \left\{ \frac{\rho c^{1-\psi^{-1}}}{\left((1-\gamma)U\right)^{\frac{\gamma-\psi^{-1}}{1-\gamma}}} - \rho(1-\gamma)U \right\}. \tag{4.7.20}$$

Here ρ will play the role of the time-preference parameter, γ controls risk aversion, and ψ the elasticity of intertemporal substitution. The standard Bellman equation must be modified only slightly. In particular, if the investment opportunity set is constant, the value function equals $e^{-\rho t} V\left(W_t, t\right)$, where V solves

$$\max_{c,\phi} f(c, V) + V_t + V_W \left[\left(r + \left(\mu_R - r\right)\phi\right) W - c \right] + \frac{1}{2} V_{WW} \sigma_R^2 \phi^2 W^2 = 0, \tag{4.7.21}$$

where μ_R and σ_R denote the mean and volatility of stock returns. The guess $V\left(W_t, t\right) = A(t) W_t^{1-\gamma}$ still works and the portfolio composition is the same as in the Merton's model,

$$\phi^* = \frac{\mu_R - r}{\sigma_R^2} \tag{4.7.22}$$

What differs is the consumption policy, which now depends on ψ and γ separately. We can solve for equilibrium in our simple model by conjecturing that the moments of stock returns and the risk-free rate are constant and then imposing market clearing conditions to pin down the precise values. Skipping the algebra, we find that

$$\sigma_R = \sigma_\delta \tag{4.7.23}$$

and

$$r = \rho + \psi^{-1}\mu_\delta - \frac{1}{2}\left(1 + \psi^{-1}\right)\gamma\sigma_\delta^2 \tag{4.7.24}$$

The volatility puzzle still remains. However, we see now that high risk aversion does not need to imply low interest rates! This is a useful property of recursive preferences.

So far we said nothing about time-series predictability of stock returns. In our simple model, expected stock returns are constant, therefore there is no predictability. We will see models with time-varying expected stock returns later in the course.

Bibliography

Acemoglu, Daron. 2002. “Directed Technical Change.” The Review of Economic Studies 69, no. 4 (October): 781–809. https://doi.org/10.1111/1467-937X.00226. https://academic.oup.com/restud/article-abstract/69/4/781/1556072.

———. 2003. “Labor- and Capital-Augmenting Technical Change.” Journal of the European Economic Association 1, no. 1 (March): 1–37. https://doi.org/10.1162/154247603322256756. https://academic.oup.com/jeea/article-abstract/1/1/1/2294416.

———. 2009. Introduction to Modern Economic Growth. Princeton, NJ: Princeton University Press, January. ISBN: 9780691132921. https://press.princeton.edu/books/hardcover/9780691132921/introduction-to-modern-economic-growth.

Acemoglu, Daron, and Pascual Restrepo. 2018. “The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment.” American Economic Review 108, no. 6 (June): 1488–1542. https://doi.org/10.1257/aer.20160696. https://www.aeaweb.org/articles?id=10.1257/aer.20160696.

Aghion, Philippe, Peter Howitt, and Ross Levine. 2018. “Financial Development and Innovation-Led Growth.” In Handbook of Finance and Development, edited by Thorsten Beck and Ross Levine, 3–30. Cheltenham, UK: Edward Elgar. ISBN: 978-1-78536-050-3. https://doi.org/10.4337/9781785360510.00007.

Autor, David, David Dorn, Lawrence F. Katz, Christina Patterson, and John Van Reenen. 2020. "The Fall of the Labor Share and the Rise of Superstar Firms." The Quarterly Journal of Economics 135, no. 2 (May): 645–709. https://doi.org/10.1093/qje/qjaa004. https://academic.oup.com/qje/article/135/2/645/5721414.

Autor, David H., David Dorn, and Gordon H. Hanson. 2013. "The China Syndrome: Local Labor Market Effects of Import Competition in the United States." American Economic Review 103, no. 6 (October): 2121–2168. https://doi.org/10.1257/aer.103.6.2121. https://www.aeaweb.org/articles?id=10.1257/aer.103.6.2121.

Baldwin, Richard. 2016. The Great Convergence: Information Technology and the New Globalization. Cambridge, MA: Harvard University Press, November. ISBN: 9780674660489. https://www.hup.harvard.edu/books/9780674660489.

Bank, World. 1993. The East Asian Miracle: Economic Growth and Public Policy. New York, NY: Oxford University Press / World Bank, September. ISBN: 9780195209938. https://documents.worldbank.org/en/publication/documents-reports/documentdetail/975081468244550798.

———. 2025. (World Development Indicators). https://databank.worldbank.org/source/world-development-indicators.

Barkai, Simcha. 2020. "Declining Labor and Capital Shares." The Journal of Finance 75, no. 5 (October): 2421–2463. https://doi.org/10.1111/jofi.12842. https://onlinelibrary.wiley.com/doi/10.1111/jofi.12842.

Barro, Robert J. 1991. "Economic Growth in a Cross Section of Countries." The Quarterly Journal of Economics 106, no. 2 (May): 407–443. https://doi.org/10.2307/2937943. https://academic.oup.com/qje/article-abstract/106/2/407/1905452.

Barro, Robert J., and Xavier Sala-i-Martin. 2003. Economic Growth. 2nd ed. Cambridge, MA: MIT Press, October. ISBN: 9780262025539.

Bloom, David E., and Jeffrey G. Williamson. 1998. "Demographic Transitions and Economic Miracles in Emerging Asia." The World Bank Economic Review 12, no. 3 (September): 419–455. https://doi.org/10.1093/wber/12.3.419. https://academic.oup.com/wber/article-abstract/12/3/419/1632238.

Bolt, Jutta, Robert Inklaar, Herman de Jong, and Jan Luiten van Zanden. 2018. Rebasing 'Maddison': New Income Comparisons and the Shape of Long-Run Economic Development. GGDC Research Memorandum GD-174. Groningen Growth and Development Centre, January. https://www.rug.nl/ggdc/historicaldevelopment/maddison/releases/maddison-project-database-2018.

Byrne, David M., John G. Fernald, and Marshall B. Reinsdorf. 2016. "Does the United States Have a Productivity Slowdown or a Measurement Problem?" Brookings Papers on Economic Activity 2016, no. 1 (March): 109–182. https://doi.org/10.1353/eca.2016.0014. https://www.brookings.edu/articles/does-the-united-states-have-a-productivity-slowdown-or-a-measurement-problem/.

Byrne, David M., Stephen D. Oliner, and Daniel E. Sichel. 2017. "How Fast Are Semiconductor Prices Falling?" Brookings Papers on Economic Activity 2017, no. 1 (March): 247–300. https://www.brookings.edu/articles/how-fast-are-semiconductor-prices-falling/.

Caselli, Francesco. 2005. "Accounting for Cross-Country Income Differences." In Handbook of Economic Growth, Volume 1A, edited by Philippe Aghion and Steven N. Durlauf, 679–741. Amsterdam: Elsevier. https://doi.org/10.1016/S1574-0684(05)01009-9. https://www.sciencedirect.com/science/article/pii/S1574068405010099.

Collins, Susan M., and Barry P. Bosworth. 1996. "Economic Growth in East Asia: Accumulation versus Assimilation?" Brookings Papers on Economic Activity 1996, no. 2 (September): 135–204. https://doi.org/10.2307/2534621. https://www.brookings.edu/wp-content/uploads/1997/06/1996b_bpea_collins_bosworth_rodrik.pdf.

Comin, Diego, and Bart Hobijn. 2010. "An Exploration of Technology Diffusion." American Economic Review 100, no. 5 (December): 2031–2059. https://doi.org/10.1257/aer.100.5.2031. https://www.aeaweb.org/articles?id=10.1257/aer.100.5.2031.

Comin, Diego A., and Martí Mestieri. 2018. "If Technology Has Arrived Everywhere, Why Has Income Diverged?" American Economic Journal: Macroeconomics 10, no. 3 (July): 137–178. https://doi.org/10.1257/mac.20150175. https://www.aeaweb.org/articles?id=10.1257/mac.20150175.

Corrado, Carol, Charles Hulten, and Daniel Sichel. 2005. "Measuring Capital and Technology: An Expanded Framework." In Measuring Capital in the New Economy, edited by Carol Corrado, John Haltiwanger, and Daniel Sichel, 11–46. Chicago: University of Chicago Press. ISBN: 978-0-226-11604-4. https://doi.org/10.7208/chicago/9780226116174.003.0002. https://www.nber.org/books-and-chapters/measuring-capital-new-economy/measuring-capital-and-technology-expanded-framework.

Crouzet, Nicolas, and Janice C. Eberly. 2019. Understanding Weak Capital Investment: The Role of Market Concentration and Intangibles. NBER Working Paper 25869. National Bureau of Economic Research, May. https://www.nber.org/papers/w25869.

Dao, Mai Chi, Mitali Das, Zsoka Koczan, and Weicheng Lian. 2017. Why Is Labor Receiving a Smaller Share of Global Income? Theory and Empirical Evidence. IMF Working Paper WP/17/169. International Monetary Fund, July. https://www.elibrary.imf.org/downloadpdf/view/journals/001/2017/169/001.2017.issue-169-en.pdf.

Duernecker, Georg, Berthold Herrendorf, and Ákos Valentinyi. 2021. "The Productivity Growth Slowdown and Kaldor's Growth Facts." Journal of Economic Dynamics and Control 130:104200. ISSN: 0165-1889. https://doi.org/10.1016/j.jedc.2021.104200.

Eggertsson, Gauti B., Jacob A. Robbins, and Ella Getz Wold. 2018. Kaldor and Piketty's Facts: The Rise of Monopoly Power in the United States. NBER Working Paper 24287. National Bureau of Economic Research, February. https://doi.org/10.3386/w24287. https://www.nber.org/papers/w24287.

Elsby, Michael W. L., Bart Hobijn, and Aysegul Sahin. 2013. "The Decline of the U.S. Labor Share." Brookings Papers on Economic Activity 2013, no. 1 (March): 1–63. https://www.brookings.edu/articles/the-decline-of-the-u-s-labor-share/.

Feenstra, Robert C., Robert Inklaar, and Marcel P. Timmer. 2015. "The Next Generation of the Penn World Table." American Economic Review 105, no. 10 (October): 3150–3182. https://doi.org/10.1257/aer.20130954. https://www.aeaweb.org/articles?id=10.1257/aer.20130954.

Fernald, John G. 2015. Productivity and Potential Output Before, During, and After the Great Recession. Working Paper 2014-15. Federal Reserve Bank of San Francisco, June. https://www.frbsf.org/research-and-insights/publications/working-papers/2014/15/.

Fund, International Monetary. 2017. World Economic Outlook, April 2017: Chapter 3. Understanding the Downward Trend in Labor Income Shares. April 2017. International Monetary Fund, April. https://www.imf.org/en/Publications/WEO/Issues/2017/04/04/world-economic-outlook-april-2017.

Gollin, Douglas. 2002. "Getting Income Shares Right." Journal of Political Economy 110, no. 2 (April): 458–474. https://doi.org/10.1086/338747. https://www.journals.uchicago.edu/doi/10.1086/338747.

Gordon, Robert J. 2016. The Rise and Fall of American Growth: The U.S. Standard of Living since the Civil War. Princeton, NJ: Princeton University Press, January. ISBN: 9780691147727. https://press.princeton.edu/books/hardcover/9780691147727/the-rise-and-fall-of-american-growth.

Greenwood, Jeremy, Zvi Hercowitz, and Per Krusell. 1997. "Long-Run Implications of Investment-Specific Technological Change." American Economic Review 87, no. 3 (June): 342–362. https://www.jstor.org/stable/2951349.

Gutiérrez, Germán, and Thomas Philippon. 2017. "Investment-less Growth: An Empirical Investigation." Brookings Papers on Economic Activity 2017, no. 2 (September): 67–140. https://www.brookings.edu/bpea-articles/investment-less-growth-an-empirical-investigation/.

Hall, Robert E., and Charles I. Jones. 1999. "Why Do Some Countries Produce So Much More Output Per Worker Than Others?" The Quarterly Journal of Economics 114, no. 1 (February): 83–116. https://doi.org/10.1162/003355399555954. https://academic.oup.com/qje/article-abstract/114/1/83/1921741.

Herrendorf, Berthold, Richard Rogerson, and Akos Valentinyi. 2014. "Growth and Structural Transformation." In Handbook of Economic Growth, Volume 2B, edited by Philippe Aghion and Steven N. Durlauf, 855–941. Amsterdam: Elsevier. https://doi.org/10.1016/B978-0-444-53540-5.00006-9. https://www.sciencedirect.com/science/article/pii/B9780444535405000069.

———. 2019. "Growth and the Kaldor Facts." Federal Reserve Bank of St. Louis Review 101, no. 2 (February): 85–99. https://research.stlouisfed.org/publications/review/2019/02/13/growth-and-the-kaldor-facts.

Inada, Ken-Ichi. 1963. "On a Two-Sector Model of Economic Growth: Comments and a Generalization." Review of Economic Studies 30 (2): 119–127.

Jones, Charles I. 2016. "The Facts of Economic Growth." In Handbook of Macroeconomics, Volume 2A, edited by John B. Taylor and Harald Uhlig, 3–69. Amsterdam: Elsevier. https://doi.org/10.1016/bs.hesmac.2016.03.002. https://www.sciencedirect.com/science/article/pii/S1574004816300024.

Jones, Charles I., and Paul M. Romer. 2010. "The New Kaldor Facts: Ideas, Institutions, Population, and Human Capital." American Economic Journal: Macroeconomics 2, no. 1 (January): 224–245. https://doi.org/10.1257/mac.2.1.224. https://www.aeaweb.org/articles?id=10.1257/mac.2.1.224.

Jordà, Òscar, Katharina Knoll, Dmitry Kuvshinov, Moritz Schularick, and Alan M. Taylor. 2019. "The Rate of Return on Everything, 1870–2015." The Quarterly Journal of Economics 134, no. 3 (August): 1225–1298. https://doi.org/10.1093/qje/qjz012. https://academic.oup.com/qje/article-abstract/134/3/1225/5435538.

Jorgenson, Dale W., Mun S. Ho, and Kevin J. Stiroh. 2008. "A Retrospective Look at the U.S. Productivity Growth Resurgence." Journal of Economic Perspectives 22, no. 1 (January): 3–24. https://doi.org/10.1257/jep.22.1.3. https://www.aeaweb.org/articles?id=10.1257/jep.22.1.3.

Jorgenson, Dale W., and Kevin J. Stiroh. 2000. "Raising the Speed Limit: U.S. Economic Growth in the Information Age." Brookings Papers on Economic Activity 2000, no. 1 (June): 125–235. https://doi.org/10.1353/eca.2000.0018. https://www.brookings.edu/bpea-articles/raising-the-speed-limit-u-s-economic-growth-in-the-information-age/.

Kaldor, Nicholas. 1957. "A Model of Economic Growth." The Economic Journal 67, no. 268 (December): 591–624. https://doi.org/10.2307/2227704. https://academic.oup.com/ej/article-abstract/67/268/591/5248725.

———. 1961. "Capital Accumulation and Economic Growth." In The Theory of Capital, edited by F. A. Lutz and D. C. Hague, 177–222. London: Palgrave Macmillan. ISBN: 978-1-349-08452-4. https://doi.org/10.1007/978-1-349-08452-4_10. https://ideas.repec.org/h/pal/intecp/978-1-349-08452-4_10.html.

Karabarbounis, Loukas, and Brent Neiman. 2014. "The Global Decline of the Labor Share." The Quarterly Journal of Economics 129, no. 1 (February): 61–103. https://doi.org/10.1093/qje/qjt032. https://academic.oup.com/qje/article/129/1/61/1890096.

Keller, Wolfgang. 2004. "International Technology Diffusion." Journal of Economic Literature 42, no. 3 (September): 752–782. https://doi.org/10.1257/0022051042177685. https://www.aeaweb.org/articles?id=10.1257/0022051042177685.

King, Robert G., and Sergio T. Rebelo. 1993. "Transitional Dynamics and Economic Growth in the Neoclassical Model." The American Economic Review 83, no. 4 (September): 908–931.

Koh, Dongya, Raul Santaeulalia-Llopis, and Yu Zheng. 2024. "Labor Share Decline and the Capitalization of Intellectual Property Products." Review of Economic Dynamics 54 (January): 181–207. https://doi.org/10.1016/j.red.2023.11.003. https://www.sciencedirect.com/science/article/pii/S1094202523000730.

Loecker, Jan De, Jan Eeckhout, and Gabriel Unger. 2020. "The Rise of Market Power and the Macroeconomic Implications." The Quarterly Journal of Economics 135, no. 2 (May): 561–644. https://doi.org/10.1093/qje/qjz041. https://academic.oup.com/qje/article/135/2/561/5687353.

OECD. 2012. OECD Employment Outlook 2012, Chapter 3: Labour Losing to Capital: What Explains the Declining Labour Share? OECD Publishing, July. https://www.oecd.org/en/publications/reports/2012/07/oecd-employment-outlook-2012_g1g1dcdb.html.

———. 2015. The Future of Productivity. Paris: OECD Publishing, July. ISBN: 978-92-64-24853-3. https://www.oecd.org/en/publications/reports/2015/12/the-future-of-productivity_9789264248533-en.html.

Oliner, Stephen D., and Daniel E. Sichel. 2000. "The Resurgence of Growth in the Late 1990s: Is Information Technology the Story?" Journal of Economic Perspectives 14, no. 4 (December): 3–22. https://doi.org/10.1257/jep.14.4.3. https://www.aeaweb.org/articles?id=10.1257/jep.14.4.3.

Piketty, Thomas. 2014. Capital in the Twenty-First Century. Cambridge, MA: Harvard University Press, March. ISBN: 9780674430006. https://www.hup.harvard.edu/books/9780674430006.

Pritchett, Lant. 1997. "Divergence, Big Time." Journal of Economic Perspectives 11, no. 3 (June): 3–17. https://doi.org/10.1257/jep.11.3.3. https://www.aeaweb.org/articles?id=10.1257/jep.11.3.3.

Reinhart, Carmen M., and Kenneth S. Rogoff. 2009. This Time Is Different: Eight Centuries of Financial Folly. Princeton, NJ: Princeton University Press, October. ISBN: 9780691142166. https://press.princeton.edu/books/hardcover/9780691142166/this-time-is-different.

Rognlie, Matthew. 2015. "Deciphering the Fall and Rise of the Net Capital Share." Brookings Papers on Economic Activity 2015, no. 1 (March): 1–69. https://www.brookings.edu/wp-content/uploads/2016/06/2015a_rognlie.pdf.

Solow, Robert M. 1956. "A Contribution to the Theory of Economic Growth." The Quarterly Journal of Economics 70, no. 1 (February): 65–94. https://doi.org/10.2307/1884513. https://academic.oup.com/qje/article-abstract/70/1/65/1885041.

Syverson, Chad. 2017. "Challenges to Mismeasurement Explanations for the US Productivity Slowdown." Journal of Economic Perspectives 31, no. 2 (May): 165–186. https://doi.org/10.1257/jep.31.2.165. https://www.aeaweb.org/articles?id=10.1257/jep.31.2.165.

Vollrath, Dietrich. 2020. Fully Grown: Why a Stagnant Economy Is a Sign of Success. Chicago: University of Chicago Press, January. ISBN: 9780226666808. https://press.uchicago.edu/ucp/books/book/chicago/F/bo44800441.html.

Young, Alwyn. 1995. "The Tyranny of Numbers: Confronting the Statistical Realities of the East Asian Growth Experience." The Quarterly Journal of Economics 110, no. 3 (August): 641–680. https://doi.org/10.2307/2946695. https://academic.oup.com/qje/article-abstract/110/3/641/1859236.