INTRODUCTION TO THE SERIES

The aim of the *Handbooks in Economics* series is to produce Handbooks for various branches of economics, each of which is a definitive source, reference, and teaching supplement for use by professional researchers and advanced graduate students. Each Handbook provides self-contained surveys of the current state of a branch of economics in the form of chapters prepared by leading specialists on various aspects of this branch of economics. These surveys summarize not only received results but also newer developments, from recent journal articles and discussion papers. Some original material is also included, but the main goal is to provide comprehensive and accessible surveys. The Handbooks are intended to provide not only useful reference volumes for professional collections but also possible supplementary readings for advanced courses for graduate students in economics.

CONTENTS OF THE HANDBOOK

VOLUME I

VOLUME II

Part 2 – MATHEMATICAL APPROACHES TO MICROECONOMIC THEORY

Part 3 – MATHEMATICAL APPROACHES TO COMPETITIVE EQUILIBRIUM

PREFACE TO THE HANDBOOK

The field of mathematical economics

Mathematical economics includes various applications of mathematical concepts and techniques to economics, particularly economic theory. This branch of economics traces its origins back to the early nineteenth century, as noted in the historical introduction, but it has developed extremely rapidly in recent decades and is continuing to do so. Many economists have discovered that the language and tools of mathematics are extremely productive in the further development of economic theory. Simultaneously, many mathematicians have discovered that mathematical economic theory provides an important and interesting area of application of their mathematical skills and that economics has given rise to some important new mathematical problems, such as game theory.

Purpose

The *Handbook of Mathematical Economics* aims to provide a definitive source, reference, and teaching supplement for the field of mathematical economics. It surveys, as of the late 1970's, the state of the art of mathematical economics. Bearing in mind that this field is constantly developing, the Editors believe that now is an opportune time to take stock, summarizing both received results and newer developments. Thus all authors were invited to review and to appraise the current status and recent developments in their presentations. In addition to its use as a reference, the Editors hope that this Handbook will assist researchers and students working in one branch of mathematical economics to become acquainted with other branches of this field. Each of the chapters can be read independently.

Organization

The Handbook includes 29 chapters on various topics in mathematical economics, arranged into five parts: *Part 1* treats *Mathematical Methods in Economics*, including reviews of the concepts and techniques that have been most useful for the mathematical development of economic theory. *Part 2* elaborates on *Mathematical Approaches to Microeconomic Theory*, including consumer, pro-

ducer, oligopoly, and duality theory. *Part 3* treats *Mathematical Approaches to Competitive Equilibrium*, including such aspects of competitive equilibrium as existence, stability, uncertainty, the computation of equilibrium prices, and the core of an economy. *Part 4* covers *Mathematical Approaches to Welfare Economics*, including social choice theory, optimal taxation, and optimal economic growth. *Part 5* treats *Mathematical Approaches to Economic Organization and Planning*, including organization design and decentralization.

Level

All of the topics presented are treated at an advanced level, suitable for use by economists and mathematicians working in the field or by advanced graduate students in both economics and mathematics.

Acknowledgements

Our principal acknowledgements are to the authors of chapters in the *Handbook of Mathematical Economics*, who not only prepared their own chapters but also provided advice on the organization and content of the Handbook and reviewed other chapters.

KENNETH J. ARROW
Stanford University

MICHAEL D. INTRILIGATOR
University of California, Los Angeles

HISTORICAL INTRODUCTION

KENNETH J. ARROW and MICHAEL D. INTRILIGATOR*

Stanford University and University of California, Los Angeles

Much of the field of mathematical economics is presented in the chapters in this Handbook. Indeed, while the field of "mathematical economics" could be defined, as in the Preface, as one that "includes various applications of mathematical concepts and techniques to economics, particularly economic theory", an alternative approach to defining the field would be to enumerate all its parts. Pragmatically, our definition of the field in this sense is provided by the Table of Contents of the Handbook. We recognize, however, that this definition is not truly complete; considerations of space limitations and priorities have caused the omission of some very active fields of mathematical economics.

A historical perspective will provide the reader with a sharper sense of the background of and interrelationships among the various chapters. We conclude with a list of eleven important developments in mathematical economics over the period since 1961.

This introduction divides the history of mathematical economics into three broad and somewhat overlapping periods: the calculus-based marginalist period (1838–1947), the set-theoretic/linear models period (1948–1960), and the current period of integration (1961–present). These dates are only suggestive. Calculus-based marginalist analysis has never ceased; the set-theoretic/linear models analysis was begun by 1933 and still continues to be significant.

1. The calculus-based marginalist period: 1838–1947

The early period of mathematical economics was one in which economics borrowed methodologies from the physical sciences and related mathematics to develop a formal theory based largely on calculus. By assuming sufficiently smooth functions (e.g., utility and production functions) and maximizing behavior, a reasonably complete theory of the behavior of microeconomic agents and of general equilibrium was developed. The basic mathematical tool was the calculus, particularly the use of total and partial derivatives and Lagrange

*We are indebted to several of the Handbook authors, but especially to Lionel McKenzie for useful suggestions.

multipliers to characterize maxima. The mathematical foundations of the modern theories of the consumer, the producer, oligopoly, and general equilibrium were developed in this period.

A seminal work, which may be treated as the starting point of mathematical economics,[1] was Cournot (1838). Cournot's contributions may be sorted under two general headings: theory of the firm and the interaction of firms and consumers in single markets. As to the theory of the firm, Cournot's basic hypothesis was that firms choose output levels to maximize profits. He studied and rigorously defined both the cases of perfect competition and of monopoly. As to the interaction of firms and consumers in single markets, Cournot developed both the equating of supply and demand in (single) competitive markets and the problem of oligopoly, where sellers' competition is limited. The "Cournot solution" to oligopoly is still a standard approach, and a suitable generalization plays an important role in the development of game theory. In the Handbook, oligopoly theory is developed in Chapter 11 by Friedman, while game theory is discussed in Chapter 7 by Shubik.

Theory of the firm: Cournot's profit-maximizing hypothesis was extended primarily through the development of the production function concept in the last quarter of the nineteenth century, so that a full theory dealing with demands for inputs as well as supply of output appeared. The development was shared by many authors, such as Walras (1874) [but the production function and the marginal productivity theory did not appear until the third edition (1896)], Wicksteed (1894), Wicksell (1893), and J. B. Clark (1889). Hotelling (1932) gave perhaps the first fully coherent account. In the Handbook, the theory of the firm is surveyed in Chapter 10 by Nadiri.

Theory of the consumer: The development of the theory of consumer demand from maximization of a utility function under a budget constraint was first begun by Gossen (1854), Jevons (1871), and Walras (1874) and elaborated by Marshall (1890). A full deduction of the properties of utility-maximizing demand functions was achieved by Slutsky (1915) and further studied by Hicks and Allen (1934), Hotelling (1935), Georgescu-Roegen (1936), and Wold (1943–44, 1953). The foundations of utility were deepened in several ways: the replacement of cardinal utility by ordinal was due to Fisher (1892) and Pareto (1909); axiomatizations of cardinal utility were due to Frisch (1926, 1932) and Alt (1936); and the revealed preference approach was initiated by Samuelson (1938), and further developed by Houthakker (1950) and Uzawa (1960). Chapter 9 of the Handbook, by Barten and Böhm, surveys the theory of consumer demand.

[1]There are always predecessors. For a study of the prehistory of mathematical economics, see Theocharis (1961).

General equilibrium: The fundamental concept that markets are interrelated and therefore the equilibrium of the economy is characterized by simultaneous equality of supply and demand on all markets is due to Walras (1874). The concept was further developed and expounded by Pareto (1896, 1909). The case that an equilibrium exists was made plausible by showing that the number of equations equalled the number of unknowns [see also Marshall (1890)]. The optimality of the competitive equilibrium was argued by both Walras and Pareto.

Stability of equilibrium: In the case of equilibrium on a single market, the conditions for stability had been discussed by Cournot (1838) and Marshall (1890). The question of stability of general equilibrium was discussed extensively in Walras (1874), though not very rigorously. The first discussions from a rigorous viewpoint appeared in Hicks (1939a), and Samuelson (1941). Important later papers on stability included Arrow and Hurwicz (1958), Hahn (1958), (1962), Arrow, Block and Hurwicz (1959), Uzawa (1961, 1962), and Hahn and Negishi (1962), building not only on Hicks and Samuelson but also on Mosak (1944) and Metzler (1945). In the Handbook, stability is treated in Chapter 16 by Hahn.

Optimal resource allocation: The first systematic calculation of benefits and costs, essentially using the modern concepts of consumers' and producers' surplus, was due to Dupuit (1844). A clear definition of optimality in the presence of many individuals was given by Pareto (1909). The characterization of optimal and sub-optimal states became known as the field of welfare economics; a synthesis of all earlier work was achieved by Hotelling (1938), Bergson (1938), and Hicks (1939b, 1941).

The particular problems of optimization over time were first studied by Ramsey (1928) and, with special reference to exhaustible resources, by Hotelling (1931). The problem of optimization when the range of possible taxes is limited was first studied by Ramsey (1927). None of these papers had much immediate impact but led to very considerable amounts of research in the postwar period.

Generalized bargaining: Edgeworth (1881) first studied the outcomes of an economy in which all kinds of commodity bargains could be made, not merely those possible in a price system. The set of possible outcomes was called the *contract curve*. A generalized version of this concept, known as the *core*, has been further developed in game theory in general and specifically with reference to economic systems; see Chapter 18 of this Handbook, by Hildenbrand.

The culmination of the calculus-based marginalist school, which combined many previous results with newer developments, is found in two classic books which continue to be highly influential: Hicks (1946) and Samuelson (1947).

Each both summarized received theory and developed newer concepts. One new concept in Hicks (1946) was that of temporary equilibrium, which was developed extensively later; in the Handbook, it is the subject of Chapter 19, by Grandmont. Samuelson (1947) incorporated the work on revealed preference and on stability previously referred to.

2. The set-theoretic/linear models period: 1948–1960

The set-theoretic/linear models period was primarily a post-World War II phenomenon in which the earlier calculus basis for mathematical economics was replaced by a set-theoretic basis and by linear models. Using set theory meant greater generality in that the classical assumption of smooth functions could be replaced by more general functions. Using linear models also meant treatment of phenomena that could not be represented by smooth functions e.g. vertices of polyhedral figures. The basic mathematical tools of the set-theoretic approach, including mathematical analysis, convexity, and elements of topology, are summarized in the Handbook in Chapter 1 by Green and Heller.

The new approach had already been set forth in the context of economic growth in an important paper of von Neumann (1937), of which the methodology was even more important than the context. Another work that played an important role in developing the set-theoretic approach was Arrow (1951a). This book was concerned with the axiomatization of social choice theory, but in the process of doing so, it used set-theoretic techniques, which provided a framework for studying the problems of general equilibrium theory. In the Handbook, social choice theory is developed in Chapter 22 by Sen, while mathematical approaches to competitive equilibrium are treated in Part 3, Chapters 15–21.

Two highly influential papers in the development of the theory of general equilibrium were Wald (1933–34) and Arrow and Debreu (1954). Wald (1933–34, 1936) provided the first rigorous analysis of general equilibrium, building upon earlier developments of Zeuthen (1932), Neisser (1932), von Stackelberg (1933), and Schlesinger (1933–34). Arrow and Debreu (1954) and, independently, McKenzie (1954) made extensive use of set-theoretic approaches in formulating the problem of the existence of a competitive equilibrium and proving existence under appropriate conditions.

The existence problem was further analyzed in McKenzie (1955, 1959, 1961), Gale (1955), Nikaidô (1956), and Debreu (1962). An important tool in this analysis was the Kakutani fixed point theorem, in Kakutani (1941) — a generalization of the Brouwer fixed point theorem.

The optimality of competitive equilibrium (welfare economics) was restudied by set-theoretic and convex-set methods by Arrow (1951b), and Debreu (1951, 1954a). The subject of welfare economics is treated in the Handbook in Part 4, Chapters 22–26.

In the theory of the consumer, further axiomatic developments in the utility function, especially in relation to the ordinalist hypothesis were presented in Debreu (1954, 1964) and Rader (1963). This subject is included in the development of consumer theory by Barten and Böhm in Chapter 9. There was also an axiomatization of utility theory for choice among uncertain options. The early paper of Ramsey (1926) was neglected, and the influential contributors were von Neumann and Morgenstern (1947), Marschak (1950), and Herstein and Milnor (1953). Ramsey (1926) had also axiomatized the related concept of subjective probability; this was subsequently developed, largely independently of Ramsey's work, by Savage (1954), building upon the earlier work of de Finetti (1937).

In many respects the applications in this period of set-theoretic concepts to the theory of economic equilibrium culminated in Debreu (1959), a classic book which has been extremely influential and one which has played a role relative to the modern set-theoretic period comparable to that played by Hicks (1946) and Samuelson (1947) relative to the classical calculus-based period. As in the case of the earlier books, Debreu (1959) both summarized the state of the theory and developed extensions, in particular to equilibrium under uncertainty, building upon Arrow (1953). The topic of equilibrium under uncertainty is treated in the Handbook in Chapter 20 by Radner. A book which summarized later developments in applying both set-theoretic and calculus-based concepts to the theory of economic equilibrium was Arrow and Hahn (1971).

This period from 1948 to 1960 was also one that witnessed the development of linear models, with many areas of application and related developments. Essentially systems of linear equations and systems of linear inequalities replaced the use of partial derivatives of the calculus-based marginalist period. The input–output model, a linear model of interindustry relations, had been developed both before and during this period in Leontief (1941, 1966). The related activity analysis model of production was developed in Koopmans, ed. (1951), Morgenstern, ed. (1954), Koopmans (1957), and, in the Soviet Union, by Kantorovich (1942, 1959). The von Neumann multisector growth model (1937) was the subject of attention in this period, in particular, in Kemeny, Morgenstern, and Thompson (1956), and Gale (1956). This model has played an important role in both general equilibrium theory and growth theory.

Linear programming was developed in this period, stemming from the work of Dantzig (1949, 1951, 1963), although there had been earlier results on systems of linear inequalities. This approach culminated in Dorfman, Samuelson and Solow (1958) and Gale (1960). These books treated not only linear programming, but also linear models of general equilibrium and linear growth models. Of fundamental importance was the development during this period of a related model of capital accumulation in Malinvaud (1953). Dorfman, Samuelson and Solow (1958) presented the initial formulation of the turnpike theorem, which was later proved in Radner (1961), Morishima (1961, 1964), McKenzie (1963), and Nikaidô (1964). In the Handbook, linear programming is treated in Chapter 2 by

Intriligator, and the theory of growth and turnpike theorems is treated in Chapter 26 by McKenzie.

Game theory was also in the process of development in this period, based, in part, on the analysis of linear models. Its origins dated back to von Neumann (1928) but the fundamental developments appeared in von Neumann and Morgenstern (1947) and Nash (1950). The developments of game theory over this period are summarized in Luce and Raiffa (1957). Game theory is treated in the Handbook in Chapter 7 by Shubik.

3. The current period of integration: 1961–present

The current period is one of integration, in which modern mathematical economics combines elements of calculus, set theory, and linear models. It is also a period in which mathematical ideas have been extended to virtually all areas of economics. There are many topics in mathematical economics under development in the current period, which has been and continues to be an extremely fruitful one for mathematical economics. This section presents eleven important topics under development in this period from 1961 to the late 1970's.

(1) Uncertainty and information:[2] Included are the theory of risk aversion, as developed in Pratt (1964) and in Arrow (1970); equilibrium under uncertainty, in Diamond (1967) and Radner (1968); microeconomic applications, in McCall (1971); insurance, in Borch (1968); search behavior, in Rothschild (1974) and Lucas and Prescott (1974); and market signalling, in Spence (1974). In the Handbook the economics of uncertainty is treated in Chapter 6 by Lippman and McCall, information is treated in Chapter 23 by Arrow, the microeconomic theory of investment under uncertainty is treated in Chapter 13 by Merton, and equilibrium under uncertainty is treated in Chapter 20 by Radner.

(2) Global analysis: Mathematical methods which combine calculus and topology are used to study properties of economic equilibria and their variation with respect to changes in the underlying economy. Debreu (1970) pioneered with a study of the conditions under which there are only a finite set of equilibria. In the Handbook, the mathematics of global analysis is the subject of Chapter 8, by Smale, while the applications to economics are surveyed in Chapter 17 by Dierker.

[2]While the analysis of uncertainty is based on the theory of probability and statistics, this analysis should not be confused with econometrics, which refers rather to the inductive study of empirical data by statistical methods in order to estimate economics relationships and to test economic hypotheses, as opposed to the deductive study of formal theories in mathematical economics.

(3) Duality theory: This is an approach to many aspects of economic theory that combines set-theoretic and calculus techniques. Important works in this area include Hotelling (1932, 1935), Roy (1947), McKenzie (1956–57), Shephard (1953, 1970), Samuelson (1953–54), Uzawa (1964a), Chipman (1966), Diewert (1974), and Fuss and McFadden, eds. (1978). In the Handbook, Chapter 12, by Diewert, discusses duality approaches to microeconomic theory.

(4) Aggregate demand functions: The theory of the consumer shows that demand functions of utility-maximizing individuals must satisfy some restrictive conditions. To what extent, if any, are these or similar conditions necessarily true of aggregate demand functions? Sonnenschein (1973) first gave arguments suggesting that aggregated demand functions are not restricted by the condition that the individual demand functions arise from utility maximization. Subsequent important papers are those of Mantel (1974) and Debreu (1974). This topic is discussed in the Handbook in Chapter 14 by Shafer and Sonnenschein.

(5) Core of economy and markets with a continuum of traders: The intuitive concept of a "large" number of traders basic to the hypothesis of perfect competition has been formalized in recent work as either a limit as the number of traders goes to infinity or as a continuum of traders. In large economies, as Edgeworth (1881) had already stated, the core (or contract curve) tends to coincide with the set of competitive equilibria. This theory combines elements of game theory, general equilibrium theory, and measure theory. This analysis was developed in Shubik (1959), Scarf (1962), Debreu and Scarf (1962), Aumann (1964, 1966), Vind (1964, 1965), and in Hildenbrand (1968, 1970a, 1970b). The core of an economy is treated in the Handbook in Chapter 18 by Hildenbrand. Measure theory is the subject of Chapter 5 by Kirman.

(6) Temporary equilibrium: The concept of temporary equilibrium was introduced by Hicks (1939). In such an equilibrium trade takes place sequentially, with each agent forecasting his or her future endowments on the basis of current and past states of the economy. The equilibrium can involve all prices moving fast enough to clear all markets or, alternatively, allow for quantity rationing. This subject is treated in the Handbook in Chapter 19 by Grandmont.

(7) Computation of equilibrium prices: This is a particular case of the computation of fixed points of mappings in which the fixed point is interpreted as an equilibrium price vector, the implied allocation being a feasible one that clears all markets. The major work in this area is Scarf (1967, 1973). This topic is covered in the Handbook in Chapter 21 by Scarf.

(8) Social choice theory: Social choice theory is concerned with the aggregation of individual preferences into social choices. The modern literature on this

subject stems largely from Arrow (1951a), a book that developed the framework for analyzing this problem and that introduced the possibility and impossibility theorems. According to the possibility theorem majority rule satisfies certain axioms of social choice when there are only two alternatives for the society. According to the impossibility theorem, if there are three or more alternatives for the society then no system of aggregation, including majority rule, can satisfy the axioms of social choice. Much of the literature on this subject up to the 1960's is treated in Sen (1970). Social choice is discussed in the Handbook in Chapter 22 by Sen.

(9) Optimal taxation: Early work in this area included that of Ramsey (1927) and Hotelling (1938), while important recent articles include Boiteux (1956), Mirrlees (1971), and Diamond and Mirrlees (1971). This topic is treated in the Handbook in Chapter 24 by Mirrlees, dealing with optimal taxation as an element of normative second-best theory. Chapter 25, by Sheshinski, discusses positive second-best theory.

(10) Optimal growth theory: This area has been developed in Samuelson and Solow (1956), Samuelson (1965), Uzawa (1964b), Koopmans (1965, 1967), Cass (1965, 1966), von Weizsäcker (1965), Gale (1967), Shell, ed. (1967), and Cass and Shell, eds. (1976). In fact, the problem was initially formulated as the problem of optimal savings in an article that was decades ahead of its time, that of Ramsey (1928). The problem was then addressed using more modern tools of analysis and combining this theory with that of multisector growth models in the 1960's. Growth theory and turnpike theorems are treated in the Handbook in Chapter 26 by McKenzie. The mathematical basis of optimal growth theory includes the theory of dynamical systems, as discussed in Chapter 3 by Varian, and control theory, as discussed in Chapter 4 by Kendrick.

(11) Organization theory: This area includes team theory, decentralization, the problem of incentives, and planning. Important earlier works in this area include Simon (1957), Hurwicz (1960), and Marschak and Radner (1972). This topic is represented in the Handbook in Chapter 27 by Marschak, Chapter 28 by Hurwicz, and Chapter 29 by Heal.

By way of summary, eleven important topics in mathematical economics since 1961 have been:

1. *Uncertainty and information* (Chapters 6, 13, 20, 23)
2. *Global analysis* (Chapters 8, 17)
3. *Duality theory* (Chapter 12)
4. *Aggregate demand functions* (Chapter 14)

5. *Core of an economy and markets with a continuum of traders* (Chapters 5, 7, 18)
6. *Temporary equilibrium* (Chapter 19)
7. *Computation of equilibrium prices* (Chapter 21)
8. *Social choice theory* (Chapter 22)
9. *Optimal taxation* (Chapters 24, 25)
10. *Optimal growth theory* (Chapters 3, 4, 26)
11. *Organization theory* (Chapters 27, 28, 29)

References

Alt, F. (1936), "Über die Messbarkeit des Nutzens", Zeitschrift für Nationalökonomie, 7:161–169. Translated as: "On the measurability of utility", in: J. S. Chipman, L. Hurwicz, M. K. Richter and H. F. Sonnenschein, eds., Preferences, utility and demand. New York: Harcourt Brace Jovanovich.

Arrow, K. J. (1951a), Social choice and individual values. New York: Wiley. In 1963, 2nd ed.

Arrow, K. J. (1951b), "An extension of the basic theorems of welfare economics", in: J. Neyman, ed., Proceedings of the 2nd Berkeley symposium on mathematical statistics. Berkeley, CA: University of California Press.

Arrow, K. J. (1953), "Le rôle des valeurs boursières pour la répartition la meilleure des risques", Econometrie, 41–48. In 1964 translated as: "The role of securities in the optimal allocation of risk-bearing", Review of Economic Studies, 31:91–96.

Arrow, K. J. (1970), Essays in the theory of risk-bearing. Amsterdam: North-Holland.

Arrow, K. J., D. Block and L. Hurwicz (1959), "On the stability of the competitive equilibrium, II", Econometrica, 27:82–109.

Arrow, K. J. and G. Debreu (1954), "Existence of equilibrium for a competitive economy", Econometrica, 22:265–290.

Arrow, K. J. and F. Hahn (1971), General competitive analysis. San Francisco, CA: Holden-Day.

Arrow, K. J. and L. Hurwicz (1958), "On the stability of the competitive equilibrium, I", Econometrica, 26:522–552.

Aumann, R. J. (1964), "Markets with a continuum of traders", Econometrica, 32:39–50.

Aumann, R. J. (1966), "Existence of competitive equilibria in markets with a continuum of traders", Econometrica, 34:1–17.

Bergson, A. (1938), "A reformulation of certain aspects of welfare economics", Quarterly Journal of Economics, 53:310–334.

Boiteux, M. (1956), "Sur la gestion des monopoles publics astreints à l'équilibre budgétaire", Econometrica, 24:22–40.

Borch, K. H. (1968), The economics of uncertainty. Princeton, NJ: Princeton University Press.

Cass, D. (1965), "Optimum growth in an aggregative model of capital accumulation", Review of Economic Studies, 32:233–240.

Cass, D. (1966), "Optimum growth in an aggregative model of capital accumulation: A turnpike theorem", Econometrica, 34:833–850.

Cass, D. and K. Shell, eds. (1976), The Hamiltonian approach to dynamic economics. New York: Academic Press.

Chipman, J. S. (1966), "A survey of the theory of international trade: Part 3, The modern theory", Econometrica, 34:18–76.

Chipman, J. S., L. Hurwicz, M. K. Richter and H. F. Sonnenschein, eds. (1971), Preferences, utility and demand. New York: Harcourt Brace Jovanovich.

Clark, J. B. (1889), "The possibility of a scientific law of wages", Publications of the American Economic Association, 4:37–69.

Cournot, A. (1838), Recherches sur les principes mathématiques de la théorie des richesses. In 1929 translated as: Researches into the mathematical principles of the theory of wealth. New York: Macmillan.
Dantzig, G. B. (1949), "Programming of interdependent activities, II: Mathematical model", Econometrica, 17:200–211.
Dantzig, G. B. (1951), "Maximization of a linear function of variables subject to linear inequalities", in: T. C. Koopmans, eds., Activity analysis of production and allocation. New York: Wiley.
Dantzig, G. B. (1963), Linear programming and extensions. Princeton, NJ: Princeton University Press.
Debreu, G. (1951), "The coefficient of resource utilization", Econometrica, 19: 273–292.
Debreu, G. (1954a), "Valuation equilibrium and Pareto optimum", Proceedings of the National Academy of Sciences, 40:588–592.
Debreu, G. (1954b), "Representation of a preference ordering by a numerical function", in: R. M. Thrall, C. H. Coombs and R. L. Davis, eds., Decision processes. New York: Wiley.
Debreu, G. (1959), Theory of value. New York: Wiley.
Debreu, G. (1962), "New concepts and techniques for equilibrium analysis", International Economic Review, 3:257–273.
Debreu, G. (1964), "Continuity properties of Paretian utility", International Economic Review, 5:285–293.
Debreu, G. (1970), "Economies with a finite set of equilibria", Econometrica, 38:387–392.
Debreu, G. (1974), "Excess demand functions", Journal of Mathematical Economics, 1:15–23.
Debreu, G. and H. Scarf (1963), "A limit theorem on the core of an economy", International Economic Review, 4:235–246.
de Finetti, B. (1937), "La prevision: Ses lois logiques, ses sources subjectives", Annales de l'Institut Henri Poincaré, 7:1–68.
Diamond, P. A. (1967), "The role of the stock market in a general equilibrium model with technological uncertainty", American Economic Review, 57:759–776.
Diamond, P. A. and J. Mirrlees (1971), "Optimal taxation and public production – I, II", American Economic Review, 61:8–27, 261–278.
Diewert, E. (1974), "Applications of duality theory", in: M. D. Intriligator and D. A. Kendrick, eds., Frontiers of quantitative economics, Vol. II. Amsterdam: North-Holland.
Dorfman, R., P. A. Samuelson and R. M. Solow (1958), Linear programming and economic analysis. New York: McGraw-Hill.
Dupuit, J. (1844), "De la mesure de l'utilité des travaux publics", Annales des Ponts et Chausées, 2nd Series, 8:332–375. In 1952 translated as: "On the measurement of the utility of public works", International Economic Papers, 2:83–110.
Edgeworth, F. Y. (1881), Mathematical psychics. London: Routledge & Kegan Paul.
Fisher, I. (1892), "Mathematical investigations in the theory of value and prices", in: Transactions of the Connecticut Academy of Arts and Sciences, Vol. 9. New Haven, CT: Connecticut Academy of Art and Sciences.
Frisch, R. (1926), "Sur un problème d'économie pure". Norsk Matematisk Forenings Skrifter, 16:1–40. In 1957 reprinted in: Metroeconomica, 9:79–111. Translated as: "On a problem in pure economics", in: J. S. Chipman, L. Hurwicz, M. K. Richter and H. F. Sonnenschein, eds., Preferences, utility and demand. New York: Harcourt Brace Jovanovich.
Frisch, R. (1932), New methods of measuring marginal utility. Tübingen: Mohr.
Fuss, M. and D. McFadden, eds. (1978), Production economics: A dual approach to theory and applications. Amsterdam: North-Holland.
Gale, D. (1955), "The law of supply and demand", Mathematica Scandinavica, 3:155–169.
Gale, D. (1956), "The closed linear model of production", in: H. W. Kuhn and A. W. Tucker, eds., Linear inequalities and related systems. Princeton, NJ: Princeton University Press.
Gale, D. (1960), The theory of linear economic models. New York: McGraw-Hill.
Gale, D. (1967), "On optimal development in a multi-sector economy", Review of Economic Studies, 34:1–18.
Georgescu-Roegen, N. (1936), "The pure theory of consumer's behavior", Quarterly Journal of Economics, 50:545–593.
Gossen, H. H. (1854), Entwickelung der Gesetze des menschlichen Verkehrs und der daraus fliessenden Regeln für menschliches Handeln. Braunschweig: Fr. Vieweg and Sohn.

Hahn, F. H. (1958), "Gross substitutes and the dynamic stability of general equilibrium", Econometrica, 26:169–170.
Hahn, F. H. (1962), "On the stability of pure exchange equilibrium", International Economic Review, 3:206–213.
Hahn, F. H. and T. Negishi (1962), "A theorem on non-tatônnement stability", Econometrica, 30:463–469.
Herstein, I. N. and J. Milnor (1953), "An axiomatic approach to measurable utility", Econometrica, 21:291–297.
Hicks, J. R. (1939a), Value and capital. New York: Oxford University Press.
Hicks, J. R. (1939b), "The foundations of welfare economics", Economic Journal, 49:696–712.
Hicks, J. R. (1941), "The rehabilitation of consumers' surplus", Review of Economic Studies, 8:108–116.
Hicks, J. R. (1946), Value and capital, 2nd ed. New York: Oxford University Press.
Hicks, J. R. and R. G. D. Allen (1934), "A reconsideration of the theory of value", Economica, 1:52–76.
Hildenbrand, W. (1968), "The core of an economy with a measure space of economic agents", Review of Economic Studies, 35:443–452.
Hildenbrand, W. (1970a), "Existence of equilibria for economies with production and a measure space of consumers", Econometrica, 38:608–623.
Hildenbrand, W. (1970b), "On economies with many agents", Journal of Economic Theory, 2:161–188.
Hotelling, H. (1931), "The economics of exhaustible resources", Journal of Political Economy, 39:137–175.
Hotelling, H. (1932), "Edgeworth's taxation paradox and the nature of demand and supply functions", Journal of Political Economy, 40:577–616.
Hotelling, H. (1935), "Demand functions with limited budgets", Econometrica, 3:66–78.
Hotelling, H. (1938), "The general welfare in relation to problems of taxation and of railway and utility rates", Econometrica, 6:242–269.
Houthakker, H. (1950), "Revealed preference and the utility function", Economica, 17:159–174.
Hurwicz, L. (1960), "Optimality and informational efficiency in resource allocation processes", in: K. J. Arrow, S. Karlin and P. Suppes, eds., Mathematical methods in the social sciences, 1959. Stanford, CA: Stanford University Press.
Jevons, W. S. (1871), The theory of political economy. London and New York: Macmillan. In 1965, 5th ed. New York: A. M. Kelley.
Kakutani, S. (1941), "A generalization of Brouwer's fixed point theorem", Duke Mathematical Journal, 8:451–459.
Kantorovich, L. V. (1942), "On the translocation of masses" (in Russian), Dokl. Akad. Nauk U.S.S.R., 37:199–201.
Kantorovich, L. V. (1959), Economic calculation of optimal utilization of resources (in Russian). Moscow: Publishing House of the Academy of Sciences of the U.S.S.R. Translated as: The best uses of economic resources. Oxford: Pergamon Press.
Kemeny, J. G., O. Morgenstern and G. L. Thompson (1956), "A generalization of the von Neumann model of an expanding economy", Econometrica, 24:115–135.
Koopmans, T. C., ed. (1951), Activity analysis of production and allocation. New York: Wiley.
Koopmans, T. C. (1957), Three essays on the state of economic science. New York: McGraw-Hill.
Koopmans, T. C. (1965), "On the concept of optimal economic growth", in: The econometric approach to development planning. Amsterdam: North-Holland.
Koopmans, T. C. (1967), "Objectives, constraints, and outcomes in optimal growth models", Econometrica 35:1–15.
Leontief, W. W. (1941), The structure of the American economy, 1919–1939. New York: Oxford University Press. In 1951, 2nd ed.
Leontief, W. W. (1966), Input–output economics. New York: Oxford University Press.
Lucas, R., Jr. and E. Prescott (1974), "Equilibrium search and unemployment", Journal of Economic Theory, 7:188–209.
Luce, R. D. and H. Raiffa (1957), Games and decisions. New York: Wiley.
Malinvaud, E. (1953), "Capital accumulation and the efficient allocation of resources", Econometrica, 21:233–268.

Mantel, R. (1974), "On the characterization of aggregate excess demand", Journal of Economic Theory, 7:348–353.
Marschak, J. (1950), "Rational behavior, uncertain prospects, and measurable utility", Econometrica, 18:111–141.
Marschak, J. and R. Radner (1972), Economic theory of teams. New Haven, CT: Yale University Press.
Marshall, A. (1890), Principles of economics. London and New York: Macmillan.
McCall, J. (1971), "Probabilistic microeconomics", The Bell Journal of Economics and Management Science, 2:403–433.
McKenzie, L. (1954), "On equilibrium in Graham's model of world trade and other competitive systems", Econometrica, 22:147–161.
McKenzie, L. (1955), "Competitive equilibrium with dependent consumer preferences", in: H. A. Antosiewiez, ed., Proceedings of the 2nd symposium on linear programming. Washington, DC: National Bureau of Standards.
McKenzie, L. (1956–1957), "Demand theory without a utility index", Review of Economic Studies, 24:185–189.
McKenzie, L. (1959), "On the existence of general equilibrium for a competitive market", Econometrica, 27:54–71.
McKenzie, L. (1961), "On the existence of general equilibrium: Some corrections", Econometrica, 29:247–248.
McKenzie, L. (1963), "Turnpike theorems for a generalized Leontief model", Econometrica, 31:165–180.
Metzler, L. (1945), "The stability of multiple markets: The Hicks conditions", Econometrica, 13:277–292.
Mirrlees, J. (1971), "An exploration in the theory of optimal income taxation", Review of Economic Studies, 38:175–208.
Morgenstern, O., ed. (1954), Economic activity analysis. New York: Wiley.
Morishima, M. (1961), "Proof of a turnpike theorem: The 'no joint production' case", Review of Economic Studies, 28:89–97.
Morishima, M. (1964), Equilibrium, stability, and growth. New York: Oxford University Press.
Mosak, J. L. (1944), General equilibrium theory in international trade. Bloomington, IN: Principia.
Nash, J. F., Jr. (1950), "Equilibrium in n-person games", Proceedings of the National Academy of Sciences, 36:48–49.
Neisser, H. (1932), "Lohnhöhe und Beschäftigungsgrad im Marktgleichgewicht", Weltwirtschaftliches Archiv, 36:415–455.
Nikaidô, H. (1956), "On the classical multilateral exchange problem", Metroeconomica, 8:135–145.
Nikaidô, H. (1964), "Persistence of continual growth near the von Neumann ray: A strong version of the Radner turnpike theorem", Econometrica, 32:151–162.
Pareto, V. (1896), Cours d'économie politique. Lausanne: Rouge.
Pareto, V. (1909), Manuel d'économie politique. Paris: Giard.
Pratt, J. W. (1964), "Risk aversion in the small and in the large", Econometrica, 32:122–136.
Rader, J. T. (1963), "The existence of a utility function to represent preferences", Review of Economic Studies, 30:229–232.
Radner, R. (1961), "Paths of economic growth that are optimal with regard only to final states: A turnpike theorem", Review of Economic Studies, 28:98–104.
Radner, R. (1968), "Competitive equilibrium under uncertainty", Econometrica, 36:31–58.
Ramsey, F. P. (1926), "Truth and probability". In 1931 published in: F. P. Ramsey, The foundations of mathematics and other logical essays. London: K. Paul, Trench, Trubner, & Co.
Ramsey, F. P. (1927), "A contribution to the theory of taxation", Economic Journal, 37:47–61.
Ramsey, F. P. (1928), "A mathematical theory of saving", Economic Journal, 38:543–559.
Rothschild, M. (1974), "Searching for the lowest price when the distribution of prices is unknown", Journal of Political Economy, 82:689–711.
Roy, R. (1942), De l'utilité. Paris: Hermann.
Roy, R. (1947), "La distribution du revenu entre les divers biens", Econometrica, 15:205–225.
Samuelson, P. A. (1938), "A note on the pure theory of consumer's behavior", Economica N.S., 5:61–71.

Samuelson, P. A. (1941), "The stability of equilibrium: Comparative statics and dynamics", Econometrica 9:97–120.
Samuelson, P. A. (1947), Foundations of economic analysis. Cambridge: Harvard University Press.
Samuelson, P. A. (1953–54), "Prices of factors and goods in general equilibrium", Review of Economic Studies, 21:1–20.
Samuelson, P. A. (1965), "A catenary turnpike theorem involving consumption and the golden rule", American Economic Review, 55:486–496.
Samuelson, P. A. and R. M. Solow (1956), "A complete capital model involving heterogeneous capital goods", Quarterly Journal of Economics, 70:537–562.
Savage, L. J. (1954), The foundation of statistics. New York: Wiley.
Scarf, H. E. (1962), "An analysis of markets with a large number of participants", in: M. Maschler, ed., Recent advances in game theory. Princeton, NJ: Princeton University Press.
Scarf, H. E. (1967), "On the computation of equilibrium prices", in: Ten economic studies in the tradition of Irving Fisher. New York: Wiley.
Scarf, H. E. (1973), The computation of economic equilibria. New Haven, CT: Yale University Press.
Schlesinger, K. (1933–34), "Über die Produktiongleichungen der ökonomischen Wertlehre", Ergebnisse eines Mathematischen Kolloquiums, 6: 10–11.
Sen, A. K. (1970), Collective choice and social welfare. San Francisco, CA: Holden-Day.
Shell, K., ed. (1967), Essays on the theory of optimal economic growth. Cambridge, MA: MIT Press.
Shepard, R. W. (1953), Cost and production functions. Princeton, NJ: Princeton University Press.
Shepard, R. W. (1970), Theory of cost and production functions. Princeton, NJ: Princeton University Press.
Shubik, M. (1959), "Edgeworth market games", in: A. W. Tucker and R. D. Luce, eds., Contributions to the theory of games, IV. Princeton, NJ: Princeton University Press.
Simon, H. (1957), Models of man. New York: Wiley.
Slutsky, E. (1915), "Sulla teoria del bilancio del consumatore", Giornale degli Economisti, 51:19–23. In 1952 translated as: "On the theory of the budget of the consumer", in: G. Stigler and K. Boulding, eds., Readings in price theory. Homewood, IL: Richard D. Irwin.
Sonnenschein, H. (1965), "A study of the relation between transitive preference and the structure of choice", Econometrica, 33:642–735.
Sonnenschein, H. (1973), "Do Walras' identity and continuity characterize the class of community excess demand functions?", Journal of Economic Theory, 6:345–354.
Spence, A. M. (1974), Market signaling. Cambridge, MA: Harvard University Press.
Theocharis, R. (1961), Early developments in mathematical economics. London: Macmillan.
Uzawa, H. (1960), "Preference and rational choice in the theory of consumption", in: K. J. Arrow, S. Karlin and P. Suppes, eds., Mathematical methods in the social sciences, 1959. Stanford, CA: Stanford University Press.
Uzawa, H. (1961), "The stability of dynamic processes", Econometrica, 29:617–631.
Uzawa, H. (1962), "On the stability of Edgeworth's barter process", International Economic Review, 3:218–232.
Uzawa, H. (1964a), "Duality principles in the theory of cost and production", International Economic Review, 5:216–220.
Uzawa, H. (1964b), "Optimal growth in a two-sector model of capital accumulation", Review of Economic Studies, 31:1–24.
Vind, K. (1964), "Edgeworth allocations in an exchange economy with many traders", International Economic Review, 5:165–177.
Vind, K. (1965), "A theorem on the core of an economy", Review of Economic Studies, 32:47–48.
von Neumann, J. (1928), "Zur Theorie der Gesellschaftsspiele", Mathematische Annalen, 100:295–320. In 1959 translated in: A. W. Tucker and R. D. Luce, eds., Contributions to the theory of games. Princeton: Princeton University Press.
von Neumann, J. (1937), "Uber ein ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes", Ergebnisse eines Mathematischen Kolloquiums, 8:73–83. In 1945 translated as: "A model of general economic equilibrium", Review of Economic Studies, 13:1–9.

von Neumann, J. and O. Morgenstern (1947), The theory of games and economic behavior, 2nd ed. Princeton, NJ: Princeton University Press.
von Stackelberg, H. (1933), "Zwei kritische Bemerkungen zur Preistheorie Gustav Cassels", Zeitschrift für Nationalökonomie, 4:456–472.
von Weizsäcker, C. C. (1965), "Existence of optimal programs of accumulation for an infinite time horizon", Review of Economic Studies, 32:85–104.
Wald, A. (1933–34), "Über die eindeutige positive Lösbarkeit der neuen Produktionsgleichungen", Ergebnisse eines Mathematischen Kolloquiums, 6:12–20.
Wald, A. (1934–35), "Uber die Produktionsgleichungen der ökonomische Wertlehre", Ergebnisse eines Mathematischen Kolloquiums, 7:1–6.
Wald, A. (1936), "Uber einige Gleichungssysteme der mathematischen Ökonomie", Zeitschrift für Nationalökonomie, 7:637–670. In 1951 translated as: "On some systems of equations of mathematical economics", Econometrica, 19:368–403.
Walras, L. (1874), Elements d'economie politique pure. Lausanne: L. Corbaz. In 1954 translated by William Jaffé as: Elements of pure economics. Homewood, IL: Richard D. Irwin.
Wicksell, K. (1893), Über Wert, Kapital und Rente nach den neuen nationalökonomischen Theorien. In 1954 translated as: Value, capital and rent. London: Allen & Unwin.
Wicksteed, P. H. (1894), An essay on the co-ordination of the laws of production. London: Macmillan. In 1932 reprinted. London: London School of Economics and Political Science.
Wold, H. (1943–44), "A synthesis of pure demand analysis", Scandinavisk Aktuarietidskrift, 26:85–118 and 200–263, 27:69–120.
Wold, H., in association with L. Jureen (1953), Demand analysis. New York: Wiley.
Zeuthen, F. (1932), "Das Prinzip der Knappheit, technische Kombination und ökonomische Qualität", Zeitschrift für Nationalökonomie, 4:1–24.

Chapter 1

MATHEMATICAL ANALYSIS AND CONVEXITY WITH APPLICATIONS TO ECONOMICS

JERRY GREEN and WALTER P. HELLER*

Harvard University and University of California, San Diego

The following is intended as a guide through the basic mathematical concepts commonly used in economic theory. We have not aimed at either completeness of coverage or generality. Rather, the goal has been to provide statements of the most basic propositions in the areas of mathematics usually referred to as point-set topology and convex analysis. The reader is referred to Rudin (1964, ch. 1) and Simmons (1963, ch. 1) for notions of ordering, the real number system, inf and sup, and other concepts from elementary set theory, as these are not presented here. Occasionally we shall delve somewhat more deeply into specialized material where a result of special usefulness in economic theory is not readily available in the literature or where the proof we provide conveys insight of relevance to the economic contexts in which the result is often used. In the text of each section, propositions are proved only if they are of this latter nature. Many proofs are gathered at the ends of the sections. Still others are omitted if they are completely straightforward or if they are easily available in the literature. Suggestions for references for each section are given at the end of the main text of this Chapter.

1. Functions

Given two sets A and B, f is said to be a *function* or *mapping* from A into B if for each $x \in A$ there exists a unique $y \in B$ such that $y = f(x)$. The set A is called the *domain* of f and the subset of B consisting of points y such that $y = f(x)$ for some $x \in A$ is called the *range* of f. We write $f\colon A \to B$ to mean f is a function from A to B.

The *inverse image of a point* $y \in B$ is $\{x \mid x \in A, y = f(x)\}$; the *inverse image of a set* B', $B' \subseteq B$ is $f^{-1}(B') = \{x \mid x \in A, f(x) \in B'\}$. If for every $y \in B$ the inverse image of y is at most a single point, f is said to be a *one-to-one* mapping. If the range of f is identical to B, f is said to be an *onto* mapping.

*We are grateful to Norman Clifford, Gerard Debreu, Michael Intriligator, Charles Kahn, Andreu Mas-Colell, R. Robert Russell, and Joel Sobel for useful suggestions.

Definition

If A and B are two sets, their *product*, $A \times B$ is the set $\{(x, y) | x \in A, y \in B\}$.

If $A = B = \boldsymbol{R}$, the set of real numbers, then their product, $\boldsymbol{R} \times \boldsymbol{R}$, denoted $\boldsymbol{R}^2$, is the set of all ordered pairs of real numbers.

If $A = B = S^1$, the circumference of a circle of unit radius, then their product is the *torus*. A point in the torus can be described by two numbers, θ_1 and θ_2, indicating the angles around the principal axis and around the cross-section of the torus that determine its location.

When the product of many sets is needed, it is sometimes convenient to enumerate them by an index i in a set I, which may be finite or infinite. If a class of sets $\{A_i\}$ is indexed by I, then the product of this class $\prod_{i \in I} A_i$ is the set of all functions, a, on I such that $a(i) \in A_i$.

If A is a set in a space X, then we denote by A^c, the *complement of* A, $A^c = \{x \in X | x \notin A\}$. If A and B are in X we denote $A \backslash B = \{x \in X | x \in A, x \notin B\}$.

Definition

A *projection* from $\prod_{i \in I} A_i$ into A_i is the function mapping a into $a(i)$.

Definition

The *graph* of a function $f: A \to B$ is the set $\{(x, y) | y = f(x), x \in A\}$.

Definition

If $f: A^1 \to A^2$ and $g: A^2 \to A^3$, the function $g \circ f: A^1 \to A^3$ is defined by $g \circ f(x) = g(f(x))$ for all $x \in A^1$. It is called the *composition* of the functions f and g.

2. Metric spaces

We now turn to the study of metric spaces. Here, the properties of the real numbers will be used extensively, and without proof. For a fundamental treatment of the structure of the space $\boldsymbol{R}$, the interested reader should consult Rudin (1964, ch. 1).

Definition

A *metric space* is a pair, (S, d), where S is a non-empty set and $d: S \times S \to \boldsymbol{R}$ satisfies

(i)	$d(x, y) \geqslant 0$,	for all $x, y \in S$,
(ii)	$d(x, y) = 0$,	if and only if $x = y$,
(iii)	$d(x, y) = d(y, x)$,	
(iv)	$d(x, z) \leqslant d(x, y) + d(y, z)$,	for all $x, y, z \in S$.

Examples

(1) $\boldsymbol{R}^1=(R^1, d)$, R^1 is the real line, $d=|x-y|$.

To see that this is a metric space we have to verify each of the conditions. The first three are obvious from the definition of absolute value, $|\cdot|$. To check (iv), observe that there are essentially two cases:

(1) $x \geqslant y \geqslant z$ and (2) $x \geqslant z \geqslant y$,

the others involving only permutations of the letters. If (1) holds, then $|x-y|+|y-z|=|x-z|$. If (2) holds, then $|x-z|+|z-y|=|x-y|$. Since $|z-y| \geqslant 0$, we have $|x-z| \leqslant |x-y| \leqslant |x-y|+|y-z|$.

(2) $\boldsymbol{R}^2_m=(R^2, d_m)$, R^2 is the real plane, $d_m(x, y)=\max\{|x_1-y_1|, |x_2-y_2|\}$.

Verify (iv) as follows:

$$\begin{aligned} d_m(x, y) &= \max\{|x_1-z_1+z_1-y_1|; |x_2-z_2+z_2-y_2|\} \\ &\leqslant \max\{|x_1-z_1|+|z_1-y_1|; |x_2-z_2|+|z_2-y_2|\} \\ &\leqslant \max\{|x_1-z_1|+|z_1-y_1|; |x_1-z_1|+|z_2-y_2|; \\ &\qquad |x_2-z_2|+|z_1-y_1|; |x_2-z_2|+|z_2-y_2|\} \\ &\leqslant d_m(x, z)+d_m(z, y). \end{aligned}$$

(3) $\boldsymbol{R}^2=(R^2, d_e)$, $d_e(x, y)=((x_1-y_1)^2+(x_2-y_2)^2)^{\frac{1}{2}}$.

The reader can verify that d_e defines a metric; it is called the *Euclidean metric* and corresponds to the ordinary notion of distance in the plane.

Definition

Two metric spaces $\langle S_1, d_1\rangle$ and $\langle S_2, d_2\rangle$ are *isometric* if there is a function f: $S_1 \to S_2$ which is one-to-one, onto and such that

$$d_2(f(x), f(y))=d_1(x, y), \quad \text{for all} \quad x, y \in S_1.$$

Metric spaces may have sets whose elements are actually functions, or have a more complex description than just "points" like those above. Nevertheless the elements are still referred to as "points".

Examples

(1) $\mathcal{C}[0,1]$.

The set is the set of all continuous functions on

$$[0,1]=\{x|x\in\boldsymbol{R},0\leqq x\leqq 1\} \text{ into } \boldsymbol{R}.$$

The metric, $d(f,g)$, is defined as

$$d_{\mathcal{C}}=\sup_{x\in[0,1]}|f(x)-g(x)|.$$

(2) $\mathcal{C}L[0,1]$.

Same set. Metric is

$$d_{\mathcal{C}L}(f,g)=\int_0^1|f(x)-g(x)|dx.$$

The proof that $\mathcal{C}$ and $\mathcal{C}L$ are metric spaces is left to the reader.

Definition

A metric space (S_1,d_1) is said to be a *subspace* of (S,d) if $S_1\subseteq S$ and $d_1(x,y)=d(x,y)$ for all $x,y\in S_1$.

Examples

(1) Let ℓ_∞ be the space of all bounded sequences in the real line, with the metric

$$d(x,y)=\sup_i|x_i-y_i|.$$

A subspace of ℓ_∞ is the metric space, c, of all convergent sequences in the real line. A subspace of c is the set of all sequences converging to zero.

(2) Consider S^2, the surface of a sphere of dimension 2, where distance is measured as the minimum length connecting the points, lying in the surface of the sphere (great circle distance). The space, S^1, which consists of a single great circle lying in the sphere is a subspace of S^2.

3. Topological structure of metric spaces

The concept of a metric is used to define a notion of distance on a space and carries with it a certain structure. The idea that points are close, measured by the

metric, can be used to make precise concepts such as "boundary", "interior" and "connected", which have a geometric interpretation in ordinary language. Moreover, certain properties of sets (for example, those that contain their own boundary) are useful in developing other mathematical concepts. Therefore, these inherit their structure indirectly from the metric.

Definition

If (S, d) is a metric space and $x \in S$, the *ε-sphere about x* is given by

$$S_\varepsilon(x) = \{y \mid y \in S, d(x, y) < \varepsilon\}.$$

Definition

A subset B of a metric space (S, d) is *open* (or, *open in S*) if for every $x \in B$ there exists $\varepsilon > 0$ such that $S_\varepsilon(x) \subseteq B$.

Clearly ε-spheres are open. The empty set is open in every metric space and the metric space is open in itself. However, it is possible to have a set that is a subset of a metric space and of some proper subspace that is open in the latter, but not in the former. For instance, $(0, 1]$ is open in $[0, 1]$ but not in $\boldsymbol{R}^1$.

Proposition 3.1

Let (S, d) be a metric space. $B \subseteq S$ is open if and only if B is the union of ε-spheres.

Proposition 3.2

The union of open sets is open. The intersection of finitely many open sets is open.

Proof

Follows directly from the definitions.

One can discuss open sets in a more general setting than that of metric spaces by introducing the notion of a topological space. This allows us to isolate those aspects of the structure of the space which are independent of the metric. The rest of this section is devoted to discussing the implications of the metric for the topological structure of the space.

Definition

A *topological space* is a pair $(S, \mathcal{T})$ where S is a set and $\mathcal{T}$ is a collection of subsets of S satisfying:

(i) $\phi, S \in \mathcal{T}$,
(ii) $B_\alpha \in \mathcal{T}$, for all $\alpha \in I$, implies $\cup_{\alpha \in I} B_\alpha \in \mathcal{T}$,
(iii) $B_1, \ldots, B_n \in \mathcal{T}$ implies $\cap_{\alpha=1}^{n} B_\alpha \in \mathcal{T}$.

In this definition the *open sets* are the elements of the collection $\mathfrak{T}$. They are primitives of the system in that they need not be derived from any other concept. The collection $\mathfrak{T}$ is said to be a *topology* on S.

Proposition 3.3

In any metric space, the family of all open sets is a topology for the space.

Proof

Follows from previous proposition.

It is to be noted, however, that not all topological spaces can be derived from an underlying metric on the space. For example, if

$$S=\{a, b\} \quad \text{and} \quad \mathfrak{T}=\{S, \{a\}, \phi\},$$

then $\mathfrak{T}$ is a topology on S, but $\mathfrak{T}$ is not derived from any metric. This is because if d were a metric, then $d(a, b)>0$ and hence $\{a\}$ and $\{b\}$ would have to be open, being ε-spheres centered at a and b, respectively, for $\varepsilon<d(a, b)$.

Given a set S, the metric d is said to *generate* the topology $\mathfrak{T}$ if $\mathfrak{T}$ is the class of open sets defined by d.

Definition

Two metrics d_1 and d_2 on S are said to be *topologically equivalent* if they generate the same topology $\mathfrak{T}$.

Proposition 3.4

If d_1 and d_2 are topologically equivalent then for each point $x \in S$, and positive numbers ε and δ, there exist positive numbers $\bar{\varepsilon}$ and $\bar{\delta}$ such that

$$S_{\bar{\varepsilon}}^{d_1}(x) \subseteq S_{\varepsilon}^{d_2}(x) \quad \text{and} \quad S_{\bar{\delta}}^{d_2}(x) \subseteq S_{\delta}^{d_1}(x),$$

where $S_{\varepsilon}^{d_i}$ are the ε-spheres using the metric d_i.

Definition

A point x in a subset, B, of a metric space (S, d) is said to be an *interior point* of B if there exists $\varepsilon>0$ such that $S_{\varepsilon}(x) \subseteq B$. The set of all interior points of B is called the *interior* of B (denoted $\mathring{B}$ or int B).

Proposition 3.5

(a) $A \subseteq B$, A open implies $A \subseteq \mathring{B}$.
(b) B is open if and only if $B=\mathring{B}$
(c) Let $\{B_\alpha\}$ be the collection of all open sets with $B_\alpha \subseteq B$. Then $\mathring{B}=\cup_\alpha B_\alpha$.
(d) $\mathring{A} \cap \mathring{B}=(A \mathring{\cap} B)$.
(e) $\mathring{A} \cup \mathring{B} \subseteq A \mathring{\cup} B)$.

Proof

Follows directly from the definitions.

To see that the set inclusion in part (e) may be strict, consider

$$A=[0,1], \qquad B=[1,2], \qquad S=\boldsymbol{R}^1.$$

Then,

$$\mathring{A}=(0,1), \qquad \mathring{B}=(1,2),$$

$$(A \mathring{\cup} B)=(0,2) \quad \text{but} \quad 1 \notin \mathring{A} \cup \mathring{B}.$$

Definition

Let B be a subset of a metric space. A point x is said to be a *limit point* of B if for any $\varepsilon>0$, $(S_\varepsilon(x)\backslash\{x\}) \cap B \neq \phi$.

The set of all limit points of B is written $\ell p(B)$. The *closure* of B is $B \cup \ell p(B)$, and is written $\overline{B}$.

A set B is said to be *closed* if $B=\overline{B}$.

Whether or not a set is closed may depend on the metric space it is considered to be a subset of. For example, $B=[\frac{1}{2}, 1)$ is closed in $S=[0,1]$ but not in $S=[0,1]$ or $S=\boldsymbol{R}^1$. Usually the space is understood, but it is sometimes important to be specific. In these cases one says "B is closed in S".

Proposition 3.6

(a) $\overline{\overline{B}}=\overline{B}$.
(b) B closed if and only if B^c open.
(c) Finite unions and arbitrary intersections of closed sets are closed.

Proof

Follows directly from the definitions.

Definition

A point $x \in S$ is said to be a *boundary point* of a subset B if $x \in \overline{B}$ and $x \in \overline{B^c}$. The set of all boundary points is denoted bdy$B$.

Examples

(a) Consider $S=\mathcal{C}[0,1]$ and let $B=\{f \mid |f(x)| \leqslant 1\}$. Then bdy$B=\{f \mid f(x)=+1$ or $f(x)=-1$ for some x and $|f(x)| \leqslant 1$ for all $x\}$.

Proof

Since $B \subseteq \overline{B}$ and any f in the indicated set is in B, it is also in $\overline{B}$. We must show that f is a limit point of B^c. If $f(x)=+1$ for some x consider $g_\varepsilon(x)=f(x)+\varepsilon$. For

any $\varepsilon>0$, $g_\varepsilon(x)\in B^c$. Given $\delta>0$, any $\varepsilon<\delta$ has the property that $d(g,f)=\sup_x|g_\varepsilon(x)-f(x)|=\varepsilon$. Therefore $g_\varepsilon\in S_\varepsilon(f)$, and hence $g_\varepsilon\in S_\varepsilon(f)\backslash\{f\}\cap B^c$. Thus $f\in\ell p(B^c)$.

(b) Let $S=\boldsymbol{R}^1$ and $B=\boldsymbol{Q}$, the set of all rational numbers. Bdy$\boldsymbol{Q}=\boldsymbol{R}^1$ since any ε-sphere in $\boldsymbol{R}^1$ contains both rationals and irrationals. This is an example where the boundary is strictly larger than the set itself.

Proof of Proposition 3.1

If B is open in S, then for each $x\in B$ there exists ε_x such that $S_{\varepsilon_x}(x)\subseteq B$. Thus

$$\bigcup_{x\in B} S_{\varepsilon_x}(x)\subseteq B,$$

and clearly,

$$\bigcup_{x\in B} S_{\varepsilon_x}(x)\supseteq B.$$

Therefore B is the union of ε-spheres. Conversely, let

$$B=\bigcup_{\alpha} S_{\varepsilon_\alpha}(x_\alpha) \quad \text{for} \quad \alpha\in I,$$

an indexing set. If $I=\phi$, then $B=\phi$ and is therefore open. If $I\neq\phi$, then each $x\in B$ is in $S_{\varepsilon_{\alpha_x}}(x_{\alpha_x})$ for some $\alpha_x\in I$. Let

$$r_x=\varepsilon_{\alpha_x}-d(x,x_{\alpha_x}).$$

Because d is a metric and hence satisfies the triangle inequality,

$$S_{r_x}(x)\subseteq B.$$

Therefore the existence of r_x as defined for each $x\in B$ suffices to show that B is open. ■

Proof of Proposition 3.4

Let $\mathfrak{T}$ be the topology generated by both d_1 and d_2. In particular, $S_\varepsilon^{d_2}(x)\in\mathfrak{T}$ for all ε and x.

But, $S_\varepsilon^{d_2}(x)$ is open in the topology generated by d_1. Thus there is an open sphere in the d_1 metric centered on x and of sufficiently small radius $\bar{\varepsilon}$ such that $S_{\bar{\varepsilon}}^{d_1}(x)\subseteq S_\varepsilon^{d_2}(x)$.

Reversing the roles of d_1 and d_2, the proof of the proposition is complete. ■

4. Sequences, complete spaces, and separable spaces

Let $\boldsymbol{Z}^+$ be the non-negative integers, in increasing order.

Definition

A *sequence* in (S,d) is a mapping

$$s: \boldsymbol{Z}^+ \to S.$$

Definition

A sequence s has *limit* $x \in S$, if for any $\varepsilon > 0$ there exists $N \in \boldsymbol{Z}^+$ such that $n \geqslant N$ implies $s(n) \in S_\varepsilon(x)$. If s has a limit x, it is said that s *converges to* x.

Definition

A *subsequence* of the sequence s is a sequence of the form

$$s \circ g,$$

where g is a strictly monotonic increasing function from $\boldsymbol{Z}^+$ to itself. Such a sequence is written $\{s_{g_k}\}$, meaning that

$$s \circ g(k) = s(g_k).$$

For example one can take a subsequence consisting of every other member of the sequence s, in which case $g_k = 2k$.

Sequences and subsequences are more often written with the index as a subscript than as the argument of a function — i.e., (s_k).

Proposition 4.1

Let s be a sequence and $x \in \ell p\{s_1, \ldots\}$. There exists a subsequence of s, s', converging to x.

Proof

Take a sequence of positive numbers $\varepsilon_k \to 0$. There exists an integer g_1 such that $s_{g_1} \in S_{\varepsilon_1}(x)$. Moreover there exists $g_2 > g_1$ such that $s_{g_2} \in S_{\varepsilon_2}(x)$, for if not, the sequence $s_{g_1}, s_{g_1+1}, \ldots$ would not have x as a limit point, and hence neither would s. Proceeding in this way we define the required increasing sequence $g = (g_k)$ such that $s \circ g$ converges to x.

Definition

A sequence s is said to be a *Cauchy sequence* if for any $\varepsilon > 0$ there exists an integer N such that $n, m \geqslant N$ implies $d(s_n, s_m) < \varepsilon$.

Definition

A metric space is *complete* if every Cauchy sequence in it converges to some point in the space.

One of the basic properties of $\boldsymbol{R}^n$ is that it is a complete metric space. This feature is "built into" the structure of the set $\boldsymbol{R}^n$ in a certain sense, rather than being a "derived" property. The set $\boldsymbol{R}$ is constructed from the natural numbers by first forming the rational numbers and then taking their "completion". Different axiomatizations perform this procedure in slightly different ways, but the result in each case is a complete space. Then $\boldsymbol{R}^n$ is constructed as the n-fold product of $\boldsymbol{R}$, and it is easy to see that it is complete.

Proposition 4.2

Let $(A, d) \subseteq (S, d)$ and let (S, d) be complete. Then (A, d) is complete if and only if A is closed as a subset S.

Let $\mathcal{B}[0,1]$ be the space of all bounded real-valued functions on $[0,1]$.

Proposition 4.3

$\mathcal{C}[0,1]$ is a closed subspace of $\mathcal{B}[0,1]$.

Corollary

$\mathcal{C}[0,1]$ is complete.

Examples

(1) $\mathcal{C}L[0,1]$ is *not* complete. (Take $f_n(x)=1$ for $x>2^{-n}$, $f_n(x)=x/2^{-n}$ for $0 \leqq x \leqq 2^{-n}$.)
(2) c_{00}, the space of sequences with only a finite number of non-zero terms is not complete. (Take $s_n=(1, 1/2, 1/3, \ldots, 1/n, 0, 0 \ldots)$.)
(3) Let c_0 be the space of sequences converging to zero; c_0 is complete, and $c_{00} \subseteq c_0$; therefore c_{00} is not closed.

Definition

Let B and A be subsets of (S, d). B is said to be *dense in* A if $A \subseteq \bar{B}$. If we take $A=S$, then B is said to be *dense* (or everywhere dense, or dense in S). B is said to be *nowhere dense* if $\operatorname{int} \bar{B}=\phi$.

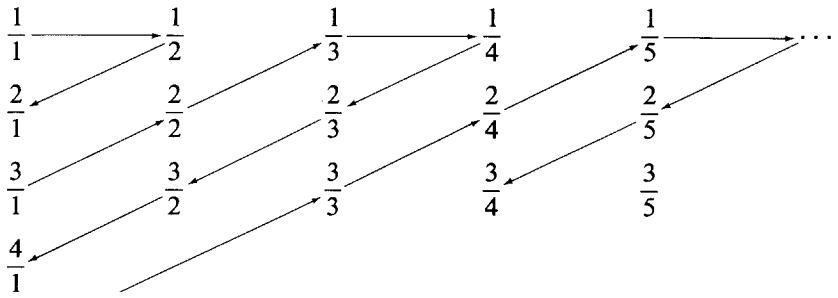
Definition

A set is said to be *countable* if it can be placed into a one-to-one relationship with the natural numbers.

Examples

(1) The space of rational numbers, $\boldsymbol{Q}$, is countable. This can be seen by arranging the rationals in a two-dimensional lattice according to the

numerator and denominator. Then, quotients that are redundant are deleted, and the rest can be enumerated by reading them off the diagonals of the array



(2) The real numbers are *not* countable. The usual proof of this proposition involves writing each real number between 0 and 1 in its binary expansion

$$0.a_1a_2a_3\ldots = \sum_{i=1}^{\infty} a_i 2^{-i},$$

where a_i is either 0 or 1 for each i. One must note that there can be at most two distinct binary expansions representing the same real number: when $a_i = 0$ for $i > k$ and $a_k = 1$, the same number is given by $a'_k = 0$, $a'_i = 1$ for $i > k$, and $a'_i = a_i$ for $i < k$. If the real numbers are countable, we can make a list of all alternative expressions for all real numbers in an array

$$0.a_1^1 a_2^1\ldots,$$
$$0.a_1^2 a_2^2\ldots,$$
$$0.a_1^3 a_2^3\ldots.$$

Now construct a new binary decimal $0.a_1a_2\ldots$ by setting

$$a_i = 0 \quad \text{if} \quad a_i^i = 1,$$
$$a_i = 1 \quad \text{if} \quad a_i^i = 0.$$

Clearly the real number represented by $0.a_1a_2\ldots$ is different from any of the members in the list. Hence the real numbers cannot be counted in this way.

Definition

The space (S, d) is said to be *separable* if it contains a countable dense subset. Separable spaces are very useful, particularly when approximation results are

desired. For this purpose one takes a subset of the countable dense set that "comes close to filling the space"...

Examples

(1) A is dense in A for any set $A \subseteq S$.
(2) $\boldsymbol{Q}$ is dense in $\boldsymbol{R}$.
(3) $\boldsymbol{Q}$ is dense in $(0, 1]$ as a subset of $\boldsymbol{R}$.
(4) $\boldsymbol{Z}^+$ is nowhere dense in $\boldsymbol{R}$.
(5) Any space with countably many points is separable.

Theorem

Let (S, d) be a metric space. Then there is a complete metric space, $(S, d)^*$, which is unique up to isometry, such that (S, d) is isometric to a dense subset of $(S, d)^*$.

For a proof, see Kolmogorov and Fomin (1970).

Example

If (S, d) is $\boldsymbol{Q}$, then $(S, d)^*$ can be taken to be $\boldsymbol{R}$.

Examples of separable spaces

(1) $\boldsymbol{R}^1$
(2) $\boldsymbol{Z}$, $\boldsymbol{Q}$
(3) $\boldsymbol{R}^n$
(4) $\mathcal{C}[0, 1]$ (Take polynomial functions with rational coefficients, or linear piecewise functions with graphs having kinks only at points in $\boldsymbol{Q} \times \boldsymbol{Q}$.)
(5) $\mathcal{B}[0, 1]$ and ℓ_∞ are *not* separable.

We say a sequence of sets $\{B_n\}_{n=1,\ldots}$ is *decreasing* if $B_n \supseteq B_{n+1}$ for all n. Also, define the *diameter* of a set $d(A) = \sup\{d(x, y) | x, y \in A\}$.

Theorem 4.4 (Cantor Intersection Theorem)

Let $\{B_n\}_{n=1,\ldots}$ be a decreasing sequence of non-empty, closed subsets of a complete metric space (S, d) such that $d(B_n)$ converges to zero. Then $B = \cap_{n=1}^{\infty} B_n$ contains exactly one point.

Proof of Proposition 4.2

Necessity. Assume (A, d) is complete. We will show that $x \in \ell p(A)$ implies $x \in A$. By virtue of Proposition 4.1, we construct a sequence $\{x_n\}$ in A such that x_n converges to $x \in S$ and such that $d(x, x_n) < 1/n$. Such a sequence is Cauchy. Since A is complete, $x \in A$.

Sufficiency. Assume A is closed and (x_n) is a Cauchy sequence in A. Since the metric in A and S is the same, (x_n) is also Cauchy in S; and since S is complete

it converges, say to $x \in S$. Suppose $x \notin A$, now $x \in \ell p(A)$, and A is closed. Therefore $x \in A$, and x_n converges to a point in A, proving that A is complete. ■

Proof of Proposition 4.3

Take $f \in \ell p \mathcal{C}[0,1]$. We will show $f \in \mathcal{C}[0,1]$. Let (f_n) be a sequence in $\mathcal{C}[0,1]$ converging to f. Since $\mathcal{B}[0,1]$ is complete, $f \in \mathcal{B}[0,1]$. We must show that f is continuous—that is, given $x \in [0,1]$ and $\varepsilon > 0$ there is $\delta > 0$ such that for $r \in S_\delta(x)$, $|f(x) - f(r)| < \varepsilon$. By the triangle inequality,

$$|f(x) - f(r)| \leqslant |f(x) - f_n(x)| + |f_n(x) - f_n(r)| + |f_n(r) - f(r)|.$$

Take $d(f, f_n) < \varepsilon/3$ for $n \geqslant N$. Thus the first and third terms are made smaller than $\varepsilon/3$. Since $f_n \in \mathcal{C}[0,1]$, there exists $\delta > 0$ such that

$$|f_n(x) - f_n(r)| < \varepsilon/3,$$

when $|x - r| < \delta$. Taking $|x - r| < \delta$ and $n \geqslant N$, $|f(x) - f(r)| < \varepsilon/3 + \varepsilon/3 + \varepsilon/3 = \varepsilon$. Therefore $f \in \mathcal{C}[0,1]$, and $\mathcal{C}[0,1]$ is therefore closed. ■

Proof of Theorem 4.4

For each n, select $x_n \in B_n$. $\{x_n\}$ is a Cauchy sequence since $d(B_n)$ converges to zero. Since (S, d) is complete, x_n converges to x. To show that $x \in B_n$ for all n: Suppose $x \notin B_k$ for some k. Then $x \in B_k^c$ and B_k^c is open in S as B_k is closed. But since $\{B_n\}$ is a decreasing sequence, $B_k^c \cap B_n = \phi$ for $n \geqslant k$.

This contradicts $\lim x_n = x$ since B_k^c is an open set around x which is disjoint from $x_n (\in B_n)$ for n sufficiently large. Thus $\cap B_n \neq \phi$. There cannot be two points in it, as $d(B_n)$ can be made arbitrarily small. ■

5. Continuity

One of the most important topological concepts is that of continuity of functions. Intuitively, we want to express the idea that a function varies only slightly when its argument varies slightly. But the idea of small variation is connected to the metric on the space. For large, complex spaces, this may be a subtle issue. It turns out that continuity is in a sense equivalent to the definition of topology. The basic structure of spaces is preserved under continuous deformation. Continuous functions also have other features of particular importance to economics — for example, they are needed to assert the existence of maxima and of fixed points (see below). Therefore, the concept of continuity is one of the central mathematical ideas to be studied.

Definition

A mapping f from a metric space (S, d) into a metric space (T, ρ) is *continuous at* $x \in S$ if for any $\varepsilon > 0$ there exists $\delta > 0$ such that $d(x, y) < \delta$ implies $\rho(f(x), f(y)) < \varepsilon$.

If f is continuous at x for all $x \in S$, then f is said to be *continuous*.

Theorem 5.1

Let f: $(S, d) \to (T, \rho)$. The following are equivalent:

(i) f is continuous.
(ii) For all $x \in S$ and $\varepsilon > 0$ there exists $\delta > 0$ such that $f(S_\delta(x)) \subseteq S_\varepsilon(f(x))$.
(iii) For all open sets B in (T, ρ), $f^{-1}(B)$ is open in (S, d).
(iv) If $\{x_n\}$ converges to x in S then $f(x_n)$ converges to $f(x)$ in T.

Proposition 5.2

Let f: $(R, \gamma) \to (S, d)$ and g: $(S, d) \to (T, \rho)$ be continuous functions. Then $g \circ f$: $(R, \gamma) \to (T, \rho)$ is continuous.

Proof

Immediate on applying (iii) in the previous proposition.

Definition

Let f: $(S, d) \to (T, \rho)$. f is said to be *uniformly continuous* if for all $\varepsilon > 0$ there exists $\delta > 0$ such that if $d(x, y) < \delta$, then $\rho(f(x), f(y)) < \varepsilon$. (Here δ depends on ε but not on x.)

Proposition 5.3

Let f be uniformly continuous, then $\{x_n\}$ Cauchy implies $\{f(x_n)\}$ Cauchy.

Proof

Immediate from definition.

Definition

A map f: $(S, d) \to (T, \rho)$ is a *homeomorphism* if it is one-to-one, onto, continuous and f^{-1} is continuous.

If there exists a homeomorphism between (S, d) and (T, ρ) they are said to be *homeomorphic*. Any property preserved under homeomorphisms is said to be a *topological property*.

In the special case where (S, d) is homeomorphic to (S, ρ), the metrics d and ρ are topologically equivalent. This can be seen from the definition in Section 3 and Theorem 5.1.

Proof of Theorem 5.1

(i) implies (ii). Follows directly upon noting that $S_\delta(x)=\{x'|d(x,x')<\delta\}$ and $S_\varepsilon(f(x))=\{y=f(x')|\rho(f(x'),f(x))<\varepsilon\}$.

(ii) implies (iii). Take B open in (T,ρ), and $a\in f^{-1}(B)$. Since B is open, there exists $\varepsilon_a>0$ such that $S_{\varepsilon_a}(f(a))\subseteq B$. By virtue of (ii) there exists $\delta_a>0$ such that $f(S_{\delta_a}(a))\subseteq S_{\varepsilon_a}(f(a))\subseteq B$. Applying f^{-1} to this set inclusion, $S_{\delta_a}(a)\subseteq f^{-1}(B)$. Thus $f^{-1}(B)$ contains an open sphere about each of its points and is therefore an open set.

(iii) implies (iv). Take $\{x_n\}$ converging to $x\in(S,d)$. Let $B\subseteq(T,\rho)$ be an arbitrarily small open set containing $f(x)$. By virtue of (iii) $f^{-1}(B)$ is open and contains x. Thus there exists N such that $n\geqslant N$ implies $x_n\in f^{-1}(B)$.

Let $\varepsilon>0$ be any number such that $S_\varepsilon(x)\subset f^{-1}(B)$. Since $x_n\to x$, we know that there exists N' such that $x_n\in S_\varepsilon(x)$ for all $n\geqslant N'$. But then $x_n\in f^{-1}(B)$ so $f(x_n)\in B$ for all $n\geqslant N'$. Since B was an arbitrarily small open set containing $f(x)$, $\{f(x_n)\}$ converges to $f(x)$.

(iv) implies (i). Assume that f is not continuous at $x\in S$. Then for some $\varepsilon>0$ and each $\delta>0$ there exists $y\in S$ with $d(x,y)<\delta$ but $\rho(f(x),f(y))\geqslant\varepsilon$. Take $\{x_n\}$ such that $d(x,x_n)<1/n$ and $\rho(f(x),f(x_n))\geqslant\varepsilon$. Then x_n converges to x but $\{f(x_n)\}$ does not converge to $f(x)$, violating (iv).

6. Compactness

The notion of compactness is a basic topological tool. Indeed, it is possible to obtain all of the results in topology by taking the compact sets as the basic primitive concept instead of the open sets. The usefulness of compactness derives from the special properties possessed by continuous functions defined on compact sets, by the maximizers of real-valued functions on compact sets, and by the usefulness of alternative characterizations of compact sets and the analytical ease with which they can be checked. In addition to compact sets in Euclidean spaces, we pay special attention to function spaces and to the characteristics of compact sets in that domain.

Definition

A collection $\{A_i\}_{i\in I}$ is a *cover* of a set B if $B\subseteq\cup_i A_i$. An *open cover* is a cover consisting only of open sets. A cover $\{A_i\}_{i\in I}$ of B is called a *subcover* of $\{A_j\}_{j\in J}$ if for each $i\in I$ there is a $j\in J$ such that $A_i=A_j$.

Definition

A set B in (S, d) is *compact* if every open cover of B contains a finite subcover. A *compact metric space* (S, d) is one in which S is a compact subset of itself.

Proposition 6.1

Let (S, d) be a compact metric space, then $A \subseteq S$ is compact if and only if A is closed.

Proposition 6.2

A compact subset in any metric space is bounded.

Definition

Let (S, d) be a metric space. Given $\varepsilon > 0$ an ε-*net* is a finite subset F of S such that $S = \cup_{x \in F} S_\varepsilon(x)$. The space (S, d) is said to be *totally bounded* if every $\varepsilon > 0$ has an ε-net.

Proposition 6.3

A compact subset of a metric space is totally bounded.

Proof

Take $\{S_\varepsilon(x)\}_{x \in A}$, A compact. Then the centers of the finite subcover suffice for an ε-net.

Definition

The metric space (S, d) is *sequentially compact* if every sequence in (S, d) contains a convergent subsequence.

Proposition 6.4

A sequentially compact metric space (S, d) is totally bounded.

Definition

A *Lebesgue number for the open cover* $\{U_i\}$ of the space (S, d) is a real number $L > 0$ such that if $A \subseteq S$ and $d(A) < L$ then $A \subseteq U_i$ for some i.

Example

If $s = [0, 1]$, $d = |x - y|$ is the usual metric, and $\{U_i\} = \{[0, \frac{3}{4}), (\frac{1}{4}, 1]\}$ is an open cover of S, then any $L < \frac{1}{2}$ is a Lebesgue number for $\{U_i\}$. Consider a set A containing points $x \leqslant \frac{1}{4}$ and $y \geqslant \frac{3}{4}$, so that $A \not\subset U_i \in \{U_i\}$, but then $d(A) \geqslant \frac{1}{2}$. Conversely, if $d(A) < \frac{1}{2}$ then it must be true that either $x < \frac{3}{4}$ or $x > \frac{1}{4}$ for all $x \in A$, and therefore either $A \subset [0, \frac{3}{4})$ or $A \subset (\frac{1}{4}, 1]$.

Notice that if the cover is not open, there may be no Lebesgue number. Consider as an example $\{U_i\} = \{[0, \frac{1}{2}], [\frac{1}{2}, 1]\}$. For any $\varepsilon > 0$, $[\frac{1}{2} - \varepsilon, \frac{1}{2} + \varepsilon]$ is not contained in either $U_i \in \{U_i\}$. Strict positivity of the Lebesgue number is the important feature.

Proposition 6.5

Every open cover of a sequentially compact metric space has a Lebesgue number.

Definition

The metric space (S, d) has the *Bolzano–Weierstrass property* if every infinite subset of S has a limit point. (This point does not necessarily lie in the subset.)

Theorem 6.6

Let (S, d) be a metric space. The following are equivalent:

(i) (S, d) is compact.
(ii) (S, d) has the Bolzano–Weierstrass property.
(iii) (S, d) is sequentially compact.
(iv) (S, d) is totally bounded and complete.

Proof

(i) implies (ii). Let A be a subset of S containing an infinite number of points but no limit point. Take $x \in A$.

There exists $\varepsilon_x > 0$ such that

$$\left(S_{\varepsilon_x}(x) \setminus \{x\}\right) \cap A = \phi.$$

Doing this for all $x \in A$, we have that $\{A^c, S_{\varepsilon_x}(x)\}_{x \in A}$ is an open cover of S with no finite subcover. Therefore A has limit points and the Bolzano–Weierstrass property is validated.

(ii) implies (iii). Assume (ii) holds and let $\{x_n\}$ be a sequence in S. If the set $\{x_1, \ldots\}$ has only finitely many distinct points, then there exists a constant, convergent, subsequence. If $\{x_1, \ldots\}$ is infinite, then it has a limit point by virtue of (ii). Thus we can find a subsequence with this point as its limit, and (S, d) is sequentially compact.

(iii) implies (iv). By Proposition 6.4, (S, d) is totally bounded. We next prove completeness. Any Cauchy sequence has a subsequence that converges to some point b, by virtue of (iii). For any $\varepsilon > 0$, there exists N such that $d(x_n, x_m) < \varepsilon/2$ for $n, m \geqslant N$. Also, $d(b, x_p) < \varepsilon/2$ for some $p > N$, by definition of b. Hence, $n > N$ implies $d(b, x_n) \leqslant d(b, x_p) + d(x_p, x_n) < \varepsilon$, as was to be shown.

We next show that condition (iii) implies (i). After that, we shall prove that (iv) implies (iii), thereby completing the proof of the theorem.

(iii) implies (i). Let (S, d) be sequentially compact and let $\{U_i\}_{i \in I}$ be an open cover of S. By Proposition 6.5, $\{U_i\}$ has a Lebesgue number $L > 0$. Then there exists for S an ε-net, $\{x_1, \ldots\}$, with $\varepsilon = L/3$, so that $\{S_\varepsilon(x_1), \ldots, S_\varepsilon(x_N)\}$ covers S. Observe that $d(S_\varepsilon(x_k)) \leqslant 2\varepsilon = \frac{2}{3} L < L$. Therefore for each $k = 1, \ldots, N$, there is $i_k \in I$ such that $S_\varepsilon(x_k) \subseteq U_{i_k}$. Hence $\{U_{i_k}\}_{k=1,\ldots,N}$ forms a finite open cover of S, and S is thus compact.

(iv) implies (iii). We shall show that every sequence has a Cauchy subsequence. Since S is complete, this will be sufficient. Consider any sequence of points in S, $\{x_i^1\}_{i=1}^\infty$. Since S is totally bounded, there is a finite set of open spheres, each with radius $\frac{1}{2}$, which cover S. There is therefore a subsequence of the original sequence, say $\{x_i^2\}_{i=1}^\infty$, which lies entirely in one of the open spheres of radius $\frac{1}{2}$. Similarly, we can argue that there is a subsequence of $\{x_i^2\}$ which lies entirely in a sphere of radius $\frac{1}{3}$, called $\{x_i^3\}$. This procedure may be continued indefinitely, with $\{x_i^n\}$ contained in a sphere of radius $1/n$. Now, consider the diagonal subsequence of $\{x_i^1\}_{i=1}^\infty : \{x_1^1, x_2^2, x_3^3, \ldots\}$. This is a Cauchy subsequence of the original sequence; for if $n > m$, then $d(x_m^m, x_n^n) < 1/m$. ■

Proposition 6.7

The continuous image of a compact set is compact.

The above proposition, when combined with Theorem 6.6, has a very useful corollary in many economic applications.

Corollary

Let (S, d) be a compact metric space and let f: $S \to \boldsymbol{R}$ be a continuous function. Then there exists $x \in S$ such that $f(x) \geqq f(y)$ for all $y \in S$.

In many economic contexts the relevant objects of choice are functions rather than numbers or vectors. Examples might be the rate of investment over time, tax rates as a function of income or probabilities (in a mixed strategy) as a function of an observable, continuously divisible, state of the system. It may be important to prove that the feasible set of such actions is compact. (See Sections 10 and 11 below.) We therefore study the structure of the space of continuous functions.

Definition

Let $\mathcal{F}$ be a collection of functions from a metric space A to a metric space B, with metric d. $\mathcal{F}$ is said to be *equicontinuous* at $x_0 \in A$ if for each $\varepsilon > 0$ there is a neighborhood U of x_0 such that

$$d(f(x), f(x_0)) < \varepsilon \quad \text{for any} \quad f \in \mathcal{F} \quad \text{and any} \quad x \in U.$$

If f *is* equicontinuous at all $x_0 \in A$, then $\mathcal{F}$ is *equicontinuous*.

Lemma

Let A be a compact metric space and let B be a compact metric space with metric d. A collection of continuous functions $\mathcal{F}$ from A to B is equicontinuous if and only if $\mathcal{F}$ is totally bounded under the metric ρ given by

$$\rho(f,g)=\sup_{x\in A} d(f(x)-g(x)).$$

Theorem 6.8 (Ascoli's Theorem)

Let A be a compact space. A subset $\mathcal{F}$ of the continuous functions from A into $\boldsymbol{R}^n$ is compact if and only if it is closed, bounded and equicontinuous.

Proof

Assume $\mathcal{F}$ is compact. By Propositions 6.1 and 6.2 it is closed and bounded. By Theorem 6.6 ((i) implies (iv)) it is totally bounded. By the Lemma, it is therefore also equicontinuous.

Conversely, assume $\mathcal{F}$ is closed, bounded and equicontinuous. We will now prove it is compact by demonstrating that it is sequentially compact and again applying Theorem 6.6 ((iii) implies (i)). Since each $f\in\mathcal{F}$ is bounded, Proposition 4.3 and its corollary imply that $\mathcal{F}$ is complete.

Take a sequence $\{f_i\}$ in $\mathcal{F}$. It will suffice to show that there is a Cauchy subsequence. The compactness of A implies that it is totally bounded. We can find a countable dense subset of A (for example the centers of the spheres defining ε-nets for a sequence of ε's converging to zero). Let this dense subset be arranged in some sequence $\{x_i\}$.

Consider the sequence of points in $\boldsymbol{R}^n$ defined by $\{f_i(x_1)\}$. Since $\mathcal{F}$ is bounded, there is a convergent subsequence. Call it $\{f_{i1}(x_1)\}$. Now consider $\{f_{i1}(x_2)\}$. It will contain a convergent subsequence $\{f_{i2}(x_2)\}$. Continue in this way and consider the subsequence composed of the "diagonal" elements $\{f_{ii}\}$ where f_{ii} is the ith function in the ith subsequence. We will show that $\{f_{ii}\}$ is Cauchy.

Fix $\varepsilon>0$. Since $\mathcal{F}$ is equicontinuous there exists $\delta>0$ such that if $d(x,x')<\delta$ then $\|f_{ii}(x)-f_{ii}(x')\|<\varepsilon/3$, for all f_{ii} in the diagonal subsequence.

Consider the δ-spheres centered on the x_i. They are an open cover of A and a finite subcover exists since A is compact. Let I be the largest index of the x_i generating this subcover. For any $x\in A$ $\inf_{i\leq I} d(x,x_i)<\delta$. We can find an N such that if $n,m\geqslant N$ then $\|f_{mm}(x_j)-f_{nn}(x_j)\|<\varepsilon/3$, as $\{f_{ii}(x_j)\}$ converges for each x_j (being a subsequence of a converging sequence). Therefore, for $m,n\geqq N$ and any $x\in A$, we can write

$$|f_{mm}(x)-f_{nn}(x)|\leqq|f_{mm}(x)-f_{mm}(x_i)|+|f_{mm}(x_i)-f_{nn}(x_i)|+|f_{nn}(x_i)-f_{nn}(x)|,$$

and if $d(x,x_i)<\delta'$, each of the terms on the right-hand side will be bounded by $\varepsilon/3$. Thus $\{f_{ii}\}$ is Cauchy. ■

Proof of Proposition 6.1

Sufficiency. Take $x \in A^c$. We will show that A^c is open by demonstrating the existence of an open set U containing x such that $U \cap A = \phi$. For each $y \in A$ choose $\varepsilon_y = \frac{1}{2} d(x, y)$. Then $\{S_{\varepsilon_y}(y)\}_{y \in A}$ is an open cover of A. Let $(S_1, \ldots, S_N)$ be a finite subcover of A. Such a subcover exists because A is assumed to be compact. We know that $x \notin \bar{S}_{\varepsilon_y}(y)$ and so $x \notin \bar{S}_k$ for any $k = 1, \ldots, N$. But then, $x \notin \cup \bar{S}_k$ and since this is a finite union of closed sets it is closed. Letting $U = (\cup \bar{S}_k)^c$ we have $x \in U$ and $U \cap A = \phi$.

Necessity. Let C be closed. Let $\{U_i\}_{i \in I}$ be any open cover of C. Since C^c is open, $\{U_i\}_{i \in I} \cup \{C^c\}$ is an open cover of S. Thus there exists a finite subcover of C, and hence it is compact. ■

Proof of Proposition 6.2

Let A be compact. Then $\{S_1(x)\}_{x \in A}$, a set of spheres with radius 1, is an open cover of A. By compactness there is a finite subcover, say with N spheres of radius 1 and centers, x_i, $i = 1, \ldots, N$. Then

$$d(A) \leqslant 2 + \max_{1 \leqslant i, j \leqslant N} d(x_i, x_j),$$

which is finite. ■

Proof of Proposition 6.4

Take $\varepsilon > 0$, and $x_1 \in S$. If $S = S_\varepsilon(x_1)$ then $\{x_1\}$ is an ε-net. If not, choose $x_2 \in (S_\varepsilon(x_1))^c$. If $S = S_\varepsilon(x_1) \cup S_\varepsilon(x_2)$ then $\{x_1, x_2\}$ is an ε-net. Continuing in this way, we note that the process must terminate since otherwise $\{x_1, \ldots\}$ would be a sequence with no convergent subsequence, contrary to hypothesis. Whenever it stops, the set $\{x_1, \ldots, x_N\}$ generated up to that point is an ε-net. Thus S is totally bounded. ■

Proof of Proposition 6.5

Let $\{U_i\}_{i \in I}$ be an open cover and suppose there is no Lebesgue number. Then there exists $\{A_n\}$ a sequence of subsets of (S, d) such that $d(A_n) = \delta_n$ and δ_n converges to zero, where no A_n is contained in any U_i. Choose a_n in A_n and define the sequence $\{a_n\}$. Since (S, d) compact there is a convergent subsequence $\{a'_n\}$. Let $p = \lim_i a'_i$. As $p \in S$ we know $p \in U_i$ for some i. Since U_i open, there exists $\varepsilon > 0$ such that $S_\varepsilon(p) \subseteq U_i$. Choose K such that $i \geqslant K$ implies

(1) $d(a_{n(i)}, p) < \varepsilon/2$, (2) $\delta_{n(i)} < \varepsilon/2$.

Now if $x \in A_N$,

$$d(x, p) \leqslant d(x, a_{n(K)}) + d(a_{n(K)}, p) \leqslant d(A_{n(K)}) + \varepsilon/2 = \varepsilon.$$

Therefore $A_N \subseteq S_\varepsilon(p) \subseteq U_i$, which is a contradiction. Thus, every open cover has a Lebesgue number, if (S, d) is sequentially compact. ■

Proof of Corollary

Let f: $(S, d) \to (T, \rho)$ be continuous and take A compact in S. Let $\{U_i\}_{i \in I}$ be an open cover of $f(A)$. Since the sets U_i are open, $f^{-1}(U_i)$ is open by the continuity of f. Thus $\{f^{-1}(U_i)\}_{i \in I}$ is an open cover of A. A finite subcover exists, say $\{f^{-1}(U_{i_1}), \ldots, f^{-1}(U_{i_N})\}$. Take $\{ff^{-1}(U_{i_1}), \ldots, ff^{-1}(U_{i_N})\} = \{U_{i_1}, \ldots, U_{i_N}\}$. This forms a finite subcover of $f(A)$, as required. ■

Proof of Proposition 6.7

Since S is compact, $f(S)$ is compact by Proposition 6.7. Let $a = \sup_{x \in S} f(x)$. For each $\varepsilon > 0$ there exists $x_\varepsilon \in S$ such that $a - f(x_\varepsilon) < \varepsilon$. Let (ε_i) be a sequence converging to zero and let (x_{ε_i}) be the sequence of points corresponding to ε_i as above. Since $f(S)$ is compact, by Theorem 6.6, $f(S)$ is sequentially compact. The sequence $(f(x_{\varepsilon_i}))$ has a subsequence, converging to $a \in f(S)$. Therefore, any element of $f^{-1}(a)$ suffices to maximize f over S. ■

Proof of Lemma

Given x_0 we first show that $\mathcal{F}$ is equicontinuous at x_0. Let $\varepsilon > 0$ be given, and take ε_1, ε_2 positive and such that $2\varepsilon_1 + \varepsilon_2 \leqslant \varepsilon$. Cover $\mathcal{F}$ by an ε_1-net,

$$S_{\varepsilon_1}(f_1), \ldots, S_{\varepsilon_1}(f_n).$$

Because $f_1, \ldots, f_n$ are continuous we can take an open neighborhood U of x_0 such that $x \in U$ implies

$$d(f_i(x), f_i(x_0)) < \varepsilon_2 \quad \text{for} \quad i = 1, \ldots, n.$$

Take $f \in \mathcal{F}$. Then $f \in S_{\varepsilon_1}(f_k)$ for some k. Hence

$$\begin{aligned} d(f(x), f(x_0)) &\geqq d(f(x), f_k(x)) + d(f_k(x), f_k(x_0)) + d(f_k(x_0), f(x_0)) \\ &\geqq \quad \varepsilon_1 \quad + \quad \varepsilon_2 \quad + \quad \varepsilon_1 \\ &\geqq \quad \varepsilon. \end{aligned}$$

Therefore $\mathcal{F}$ is equicontinuous at x_0.

Pick $\varepsilon_1 > 0$. Because $\mathcal{F}$ is equicontinuous, for any $x_0 \in A$ we can pick a neighborhood $U(x_0)$ such that $d(f(x), f(x_0)) < \varepsilon_1$ for all $f \in \mathcal{F}$, $x \in U(x_0)$. Let $\{U(x_0)\}_{x_0 \in A}$ be the set of these neighborhoods. This set is an open cover of A. By compactness, it has a finite subcover, say $U(x_1), \ldots, U(x_m)$. Within this subcover we still have, for any $x \in U(x_i)$,

$$d(f(x), f(x_i)) < \varepsilon_1 \quad \text{for any} \quad f \in \mathcal{F}.$$

Cover B by finitely many open sets $V_1, \ldots, V_n$ of diameter less than ε_2.

Let T be the set of all mappings τ of $\{1,\ldots,m\}$ into $\{1,\ldots,n\}$. For each τ we ask whether there is an $f\in\mathcal{F}$ such that $f(x_i)\in V_{\tau(i)}$ for all i. If so, denote one such function by f_τ. Thus there will be a finite collection of functions $\{f_\tau\}$ for those τ in T such that this property holds within $\mathcal{F}$.

Take $f\in\mathcal{F}$. We will show that $f\in S_\varepsilon(f_\tau)$ for some f_τ.

Let $x\in A$ and choose i so that $x\in U_i$. Then

$$d(f(x),f(x_i))<\varepsilon_1,$$

$$d(f(x_i),f_\tau(x_i))<\varepsilon_2,$$

$$d(f_\tau(x_i),f_\tau(x))<\varepsilon_1.$$

Since ε_1 and ε_2 are arbitrary, we can set them small enough that, as above, $d(f(x),f_\tau(x))<\varepsilon$. Because this holds for each $x\in X$,

$$\rho(f,f_\tau)=\sup\{d(f(x),f_\tau(x))\}<\varepsilon. \quad \blacksquare$$

7. Connected sets

A continuous curve in a topological space X is the image of the unit interval in $\boldsymbol{R}$ under a continuous function $f:[0,1]\to X$. A metric space X is *arc-connected* if there is a continuous curve linking any two points in X. Clearly, the set $S=(0,\frac{1}{2})\cup(\frac{1}{2},1)$ is not arc-connected in $\boldsymbol{R}^1$. A metric space X is *connected* if it is *not* the union of a pair of non-empty open sets which are disjoint. Equivalently, X is connected if and only if the only sets which are both open and closed in X are X itself and the empty set. It can be shown that arc-connectedness implies connectedness. There are examples of connected sets in $\boldsymbol{R}^2$ that are not arc-connected (see Dieudonné). One can show that a subset of the real line is connected if and only if it is an interval. An important property of a continuous function f is that the image of a set S under f, $f(S)$, is connected if S is connected. This result has a very useful implication: Let f be a continuous real-valued function on a connected subset A of $\boldsymbol{R}^m$ such that $f(a_1)<0<f(a_2)$ for some points $a_1, a_2\in A$. Then there exists a solution to the equation $f(x)=0$. To see this, not that $f(A)$ is a connected subset of $\boldsymbol{R}^1$, so $f(A)$ is an interval $[a,b]$. By assumption, $0\in[a,b]$, so there exists $x\in A$ such that $f(x)=0$. A useful corollary of this result is that a pair of real-valued continuous functions g, h on a connected subset A of $\boldsymbol{R}^m$ has a solution to the equation $g(x)=h(x)$ so long as there exists $a_1, a_2\in A$ such that $g(a_1)\geqslant h(a_1)$ and $g(a_2)\leqslant h(a_2)$.

8. Convex sets and cones in Euclidean spaces

Convexity is arguably the most important mathematical property in microeconomics (see Chapters 9–14). Without convexity of preferences, for example, demand and supply functions are not continuous, and so competitive markets

generally do not have equilibrium points. The extremely useful results of mathematical programming (see Chapter 2) and duality theory (see Chapter 12) do not hold in the absence of convexity. The economic interpretation of convex production sets is, of course, constant or decreasing returns to scale; the corresponding interpretation of convex indifference curves (or preference sets) in consumer theory is diminishing marginal rates of substitution. Relatively little is known about general economic equilibrium models that allow non-convex production sets (e.g., economics of scale) or non-convex preferences (e.g., the consumer prefers a glass of beer or a glass of champagne alone to any mixture of the two). The only exact results are for economies with an infinite number of participants.

All sets in this section are subsets of $\boldsymbol{R}^m$. A *convex set* $S \subset \boldsymbol{R}^m$ is one which contains a line joining any pair of points belonging to S. More formally, let $x, y \in \boldsymbol{R}^m$. The *line-segment joining* x *and* y, denoted $[x, y]$, is the set of points $z = \alpha x + (1-\alpha)y$ where $0 \leqslant \alpha \leqslant 1$; such a z is called a *convex combination* of x and y. A set S is *convex* if $[x, y] \subset S$ for any pair of points, $x, y \in S$. Any sphere $S_\varepsilon(x)$ is convex. The solution set of any system of linear equations or inequalities is convex.

Some simple properties of convex sets are: (1) $\bar{S}$ is convex if S is convex; (2) S and T convex sets implies $S \cap T$ convex; (3) $S \cup T$ is not necessarily convex even if S and T are convex; (4) $S + T \equiv \{s + t \mid s \in S \text{ and } t \in T\}$ is convex if S and T are convex (thus, the aggregate production possibility set of an economy is convex if each individual firm has a convex production set); (5) let $\lambda \geqslant 0$ be a scalar, then $\lambda S = \{\lambda s \mid s \in S\}$ is convex if S is convex. A very useful topological property of a convex set S is that if S has a non-empty interior then int S is *dense* in $\bar{S}$, i.e., $\bar{S} = \overline{\text{int } S}$. Heuristically, any sphere about $x \in S$ has a point in common with int S because of convexity. Thus x is the limit of points belonging to the interior of S.

Another very useful property of convex sets is that any disjoint pair of convex sets can be separated by a hyperplane. This fact is frequently used in economics to obtain a price system that decentralizes a Pareto-efficient allocation of resources, i.e., a price system that leads consumers and producers to choose this allocation. Decentralization is illustrated in Figure 8.1 for a two-good, one-producer, one-consumer economy. Here, Y is the production set, I is the highest indifference curve that shares a point with Y, z is the Pareto-efficient allocation, and H is the separating hyperplane: the set of pairs (x, y) such that $p_x x + p_y y = M$. The sets Y and $P(z)$ are separated by H, where $P(z)$ is the upper preference set bounded by I. When prices are (p_x, p_y) the producer is maximizing profits, and given those profits as income, M, the consumer is maximizing utility subject to his budget constraint.

More formally, a *hyperplane* $H(p, \alpha)$ in $\boldsymbol{R}^m$ is the set of all points $x \in \boldsymbol{R}^m$ satisfying $p \cdot x = \alpha$ for some non-zero vector $p \in \boldsymbol{R}^m$ and some scalar α, i.e., $H(p, \alpha) = \{x \in \boldsymbol{R}^m \mid p \cdot x = \alpha\}$. The vector p is called the *normal* to the hyperplane H (see Figure 8.2).

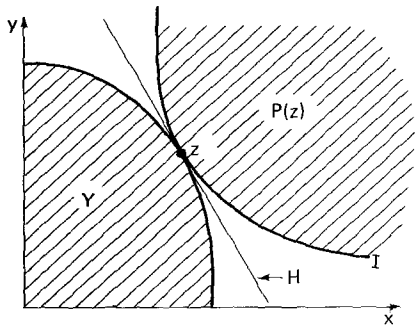


Figure 8.1

A *closed half-space* determined by the hyperplane $H(p, \alpha)$ is either the set of points 'below' or the set of points 'above' H, i.e., either the set $\{x \in \boldsymbol{R}^m \mid p \cdot x \leqslant \alpha\}$ or the set $\{x \in \boldsymbol{R}^m \mid p \cdot x \geqslant \alpha\}$. The set $\{x \in \boldsymbol{R}^m \mid p \cdot x \leqslant \alpha\}$ could be, for example, a consumer's budget set with prices p and income α. An *open half-space* is defined as the interior of a closed half-space. A *bounding hyperplane* $H(p, \alpha)$ to a set S is a hyperplane such that S is completely contained in one of the two closed half-spaces determined by $H(p, \alpha)$; see Figure 8.2. A *supporting hyperplane* $H(p, \alpha)$ for S is a bounding hyperplane that shares a point in common with the boundary of S (more precisely, $\inf\{p \cdot x \mid x \in S\} = \alpha$). A hyperplane $H(p, \alpha)$ *separates* two sets S and T if $p \cdot s \geqslant \alpha$ for all $s \in S$ and $p \cdot t \leqslant \alpha$ for all $t \in T$ (i.e., if S and T are completely contained in opposite closed half-spaces).

We have already given an example of a separating hyperplane above in Figure 8.1. Indeed, that same hyperplane supports the production set Y and $P(z)$, the upper preference set bounded by I, reflecting the fact that the producer is maximizing profits and the consumer is minimizing the cost of obtaining the utility level corresponding to I (and hence, in this case, maximizing utility subject to a budget constraint).

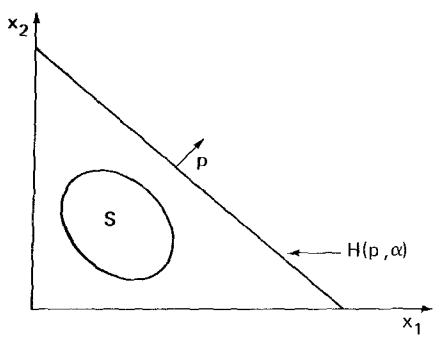


Figure 8.2

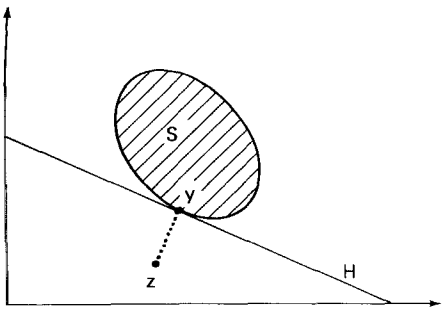


Figure 8.3

Lemma

If S is convex and if $z \notin \bar{S}$, then there exists $y \in \bar{S}$ and $p \neq 0$ such that $p \cdot y = \inf\{p \cdot x | x \in S\} > p \cdot z$ (i.e., there exists a supporting hyperplane H for S that separates S from the point z). In Figure 8.3, y is the point on the frontier of S that is closest to z.

Sketch of the Proof

Let $L = [z, y]$ denote the line joining z and y. Let $H(p, \alpha)$ be the hyperplane which is perpendicular to L and which passes through y. It is intuitive that H is tangent to S at y (i.e., $p \cdot s \geqslant p \cdot y$ for all $s \in S$ in some neighborhood of y). We claim that $p \cdot s \geqslant p \cdot y$ for all $s \in S$, not just in the neighborhood of y. Suppose not, i.e., the exists $x^\circ \in S$ such that $p \cdot x^\circ < p \cdot y$. The line segment $[x^\circ, y]$ belongs to S by convexity and $p \cdot x < p \cdot y$ for all $x \in [x^\circ, y]$ except when $x = y$ (since $p \cdot x = p \cdot (\theta x^\circ + (1-\theta)y) < \theta p \cdot y + (1-\theta) p \cdot y$). But this contradicts the tangency of H to S, since x can be chosen to be in an arbitrarily small neighborhood of y. ■

Supporting Hyperplane Theorem

If y is on the boundary of a convex set S, then there exists a supporting hyperplane for S that passes through y.

Minkowski Separating Hyperplane Theorem

If S and T are convex sets with disjoint interiors, then there exists a hyperplane that separates S and T.

Proof

This follows from the preceding Lemma. For simplicity, take the case where $S \cap T = \phi$. Let $S - T = \{s - t | s \in S, t \in T\}$. Then $0 \notin S - T$. If zero is on the boundary of $S - T$ then there exists a supporting hyperplane $H(p, \alpha)$ for $S - T$

that passes through zero. If not, $0 \notin \overline{S-T}$, so zero is separated from $S-T$. In either case, $0 = p \cdot 0 \leqslant p \cdot (s-t)$ for $s \in S$ and $t \in T$, so that $ps \leqslant pt$, as was asserted. ■

An immediate corollary of the Minkowski Theorem is that if Z is convex and disjoint from the non-negative orthant, then there is a separating hyperplane $H(p, \alpha)$ for Z and the non-negative orthant with $p > 0$ (i.e., $p \neq 0$ and $p_i \geqslant 0$ for all $i = 1, \ldots, m$) such that $p \cdot z \leqslant 0$ for all $z \in Z$. This result can be used to establish the existence of a "decentralizing" price system for Pareto efficient allocations in a multigood, multiperson economic system. See Figure 8.1 and the discussion following it for the basic idea.

A *cone* C in R^m is a set such that if $x \in C$ then $\lambda x \in C$ for any $\lambda \geqslant 0$. A *convex cone* is a cone that is also a convex set. It is easy to see that a convex cone C contains the sum of any two vectors in C and that $0 \in C$. An *extreme point* of a convex set S is any point $z \in S$ such that z cannot be expressed as a convex combination of points in S that are distinct from z. The vertices of a triangle in $\boldsymbol{R}^2$ are extreme points, for example. A convex cone is *pointed* if zero is an extreme point of C. A convex production set that contains 0 and satisfies constant returns and irreversibility is a pointed convex cone. It turns out that the sum of convex cones is a convex cone. The union of cones is also a cone (but the union of convex cones is not generally a convex set). The *dual* (or polar cone) of a convex cone C is the set $C^* = \{p \in R^m \mid p \cdot x \leqslant 0 \text{ for all } x \in C\}$. The dual is a convex cone.

Duality Theorem

If C is a closed convex cone, then the dual of the dual of C is C itself; i.e., $(C^*)^* = C$.

The *convex hull* of a set S, denoted con S, is the set of all convex combinations of points in S, i.e., $\operatorname{con} S = \{x \mid x = \sum_{i=1}^{n} \alpha_i s_i$ for some numbers α_i, $0 \leqslant \alpha_i \leqslant 1$, $\sum_{i=1}^{n} \alpha_i = 1$, and some set of points $s_i \in S\}$. Thus, the convex hull of a convex set S is S itself. The convex hull of any set S is a convex set. The convex hull of a pair of points $\{x, y\}$ is the line segment $[x, y]$.

Caratheodory's Theorem

Any point in con S where S is any subset of $\boldsymbol{R}^m$, is the convex combination of at most $m+1$ points of S.

A useful property of convex hulls is that $\operatorname{con}(\sum_{i=1}^{n} S_i) = \sum_{i=1}^{n} \operatorname{con}(S_i)$. Although the convex hull of a closed set is not necessarily closed, the convex hull of a compact set is itself a compact set.

Krein–Milman Theorem

Any compact convex set is equal to the convex hull of its extreme points.

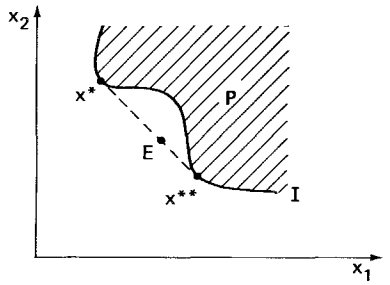


Figure 8.4

Although not much is known about exact solutions of descriptive general equilibrium models when convexity is absent, some important approximation results are known. The *Shapley–Folkman Theorem* forms the mathematical basis of these results; it discovered as a direct consequence of the posing of the question of non-convexities in economics.

Suppose there are a large number of consumers with identical non-convex preferences in a pure exchange economy. In Figure 8.4, I is a typical indifference curve and P is the corresponding upper preference set. Now, examine the fictitious economy generated by taking as preferences the convex hulls of upper preference sets such as P. Such an economy is convex and so has an equilibrium. Let the point E in the Figure 8.4 be the equilibrium to the representative consumer. Suppose E is, for example, $(302/713)x^* + (411/713)x^{**}$. Then, if there were exactly 713 consumers, one could obtain E on average by placing 302 consumers at x^* and 411 at x^{**}. But this would be a competitive equilibrium in the original economy if prices p^* corresponded to the normal of the line segment $[x^*, x^{**}]$. The argument generalizes when the number of consumers is sufficiently large.

Shapley–Folkman Theorem

Let S_i $(i=1,\ldots,n)$ be non-empty subsets of R^m, let $S=\sum_{i=1}^{n} S_i$, and let x be an element of the set $\operatorname{con} S$. Then there exist elements $x_i \in \operatorname{con} S_i$ such that $x=\sum_{i=1}^{n} x_i$ and $x_i \in S_i$ for all but (at most) m of the sets S_i. Thus, any element in $\operatorname{con} S$ can be expressed as the sum of elements of the S_i, with the exception of at most m elements, no matter how many such S_i there are.

Proof

Because a proof of the Shapley–Folkman Theorem is not readily available, we shall provide one here (due to J. P. Aubin and I. Ekeland). Let $x \in \operatorname{con}(S)$. Caratheodory's theorem guarantees that x can be expressed as the convex combination of a finite number of points, k, of S, $x=\sum_{j=1}^{k} \alpha_j y_j$ with $y_j \in S$, $\alpha_j > 0$ and $\sum_{j=1}^{k} \alpha_j = 1$.

Moreover, each y_j may be written as

$$y_j = \sum_{i=1}^{n} y_{ij} \quad \text{with} \quad y_{ij} \in S_i.$$

Let F_i be the set $\{y_{ij}\}_{j=1}^{k}$ consisting of k points. But then $x \in \text{con}(\sum_{i=1}^{n} F_i)$ since $y_j \in \sum_{i=1}^{n} F_i$. As mentioned earlier, the convex hull of a sum of sets is the sum of the convex hulls of those sets, so $\text{con}(\sum F_i) = \sum \text{con } F_i$.

We have replaced each (possibly non-compact) set S_i by a compact (indeed, finite) set F_i. By the Krein–Milman Theorem, any extreme point of con F_i is an element of F_i (this is not true of non-compact sets!). Consider the set $P = \{(x_i)_{i=1}^{n} | x_i \in \text{con } F_i \text{ and } \sum_{i=1}^{n} x_i = x\}$. This compact convex subset of $\boldsymbol{R}^{mn}$ is non-empty since $x \in \sum \text{con } F_i$. Let $(\bar{x}_i)_{i=1}^{n}$ be an extreme point of P. Again, by the Krein–Milman Theorem, $(\bar{x}_i)_{i=1}^{n}$ is an element of P, so $x = \sum_{i=1}^{n} \bar{x}_i$ with $\bar{x}_i \in \text{con } F_i$. We now claim that all but m (at most) of the $\bar{x}_i$ are extreme points of con F_i. Since extreme points of the convex hull of the F_i are members of F_i, we will have proved the theorem.

Suppose not, i.e., there exist $m+1$ of the $\bar{x}_i$ which are not extreme points of con F_i. Suppose these points are $\bar{x}_1, \ldots, \bar{x}_{m+1}$ after relabeling. For each non-extreme $\bar{x}_i$, there exists a non-zero vector $z_i \in \boldsymbol{R}^m$ and a number $\varepsilon_i > 0$ such that

(1) for all t, $|t| < \varepsilon_i$, we have $\bar{x}_i + tz_i \in \text{con } F_i$.

Define $\varepsilon = \min\{\varepsilon_i | 1 \leqslant i \leqslant m+1\}$. We have $m+1$ vectors z_i in m-dimensional space. Thus, they are linearly dependent, i.e., there exist numbers β_i $(i = 1, \ldots, m+1)$ not all zero, such that $\sum_{i=1}^{m+1} \beta_i z_i = 0$. We may assume $|\beta_i| \leqslant 1$.

Next, construct two points of $\boldsymbol{R}^{mn}$, $(x_i')_{i=1}^{n}$ and $(x_i'')_{i=1}^{n}$, as follows:

$$\begin{aligned} x_i' &= \bar{x}_i + \varepsilon\beta_i z_i \quad \text{for} \quad 1 \leqslant i \leqslant m+1, \\ x_i'' &= \bar{x}_i - \varepsilon\beta_i z_i \quad \text{for} \quad 1 \leqslant i \leqslant m+1, \\ x_i' &= x_i'' = x_i \quad \text{otherwise.} \end{aligned}$$

By (1), x_i' and x_i'' both belong to con F_i. Furthermore, $\sum_{i=1}^{n} x_i' = \sum_{i=1}^{n} \bar{x}_i + \varepsilon \sum_{i=1}^{n} \beta_i z_i = x$ and $\sum_{i=1}^{n} x_i'' = \sum_{i=1}^{n} \bar{x}_i - \varepsilon \sum_{i=1}^{n} \beta_i z_i = x$. Thus, $(x_i')_{i=1}^{n}$ and $(x_i'')_{i=1}^{n}$ both belong to P. But $(\bar{x}_i)_{i=1}^{n} = \frac{1}{2}(x_i')_{i=1}^{n} + \frac{1}{2}(x_i'')_{i=1}^{n}$, so $(\bar{x}_i)_{i=1}^{n}$ is not an extreme point of P, a contradiction. ■

We are interested in a corollary to the Shapley–Folkman Theorem which implies, in effect, that the distance between any element of con S and some element of S is bounded in a fashion that is independent of n, the number of sets in the sum. To this end, define $d(x, Y)$ as the distance between a point x and a set Y, $d(x, Y) = \inf_{y \in Y} \| x - y \|$, where $\| \cdot \|$ is Euclidean distance. Consider

the real-valued function $\rho(Y)=\sup_{x\in\operatorname{con} Y} d(x, Y)$. This number is the maximum of the distances between points in the convex hull from the original set, and so the function ρ is a measure of the non-convexity of any set. Also, $\rho(Y)$ is zero if and only if Y is convex. [See Heller (1972) for more detail.]

We use this measure to prove the following corollary, which is central to most of the approximate equilibrium results in the literature:

Corollary

Let $S_1, \ldots, S_n$ be subsets of $\boldsymbol{R}^m$ such that $\rho(S_i)\leqslant c$ for all i and some $c>0$. Let x be any element of $\operatorname{con}(\sum_{i=1}^n S_i)$. Then there are elements $a_i\in S_i$, $i=1,\ldots,n$, such that $\|x-\sum_{i=1}^n a_i\|\leqslant mc$.

Proof

As before, define $S=\sum_{i=1}^n S_i$. It is shown first that $\rho(S)\leqslant\sum_{i=1}^n\rho(S_i)$. For $n=1$, this result is trivial. We use induction on n. Let $x\in\operatorname{con} S$ and $S'=\sum_{i=1}^{n-1}S_i$. Define $x'\in\operatorname{con} S'$ and $x_n\in\operatorname{con} S_n$ so that $x=x'+x_n$. Let $b\in S$ and define $b'\in S'$ and $b_n\in S_n$ so that $b=b'+b_n$. Then, $\|x-b\|\leqslant\|x'-b'\|+\|x_n-b_n\|$. Hence,

$$\inf_{b\in S}\|x-b\|\leqslant\|x'-b'\|+\|x_n-b_n\|,$$

$$\inf_{b\in S}\|x-b\|\leqslant\inf_{b'\in S'}\|x'-b'\|+\inf_{b_n\in S_n}\|x_n-b_n\|,$$

$$\inf_{b\in S}\|x-b\|\leqslant\sup_{x'\in\operatorname{con} S'}\inf_{b'\in S'}\|x'-b'\|+\sup_{x_n\in\operatorname{con} S_n}\inf_{b_n\in S_n}\|x_n-b_n\|$$

$$=\rho(S')+\rho(S_n).$$

Therefore, $\rho(S)\leqslant\rho(S')+\rho(S_n)\leqslant\sum_{i=1}^n\rho(S_i)$, by the induction hypothesis.

Given $x\in\operatorname{con} S$, use the Shapley–Folkman Theorem to obtain a set of n points $\{x_i\}$ and a set of indices I (with at most m elements) such that for $i\in I$, $x_i\in\operatorname{con} S_i$, while for i not in I, $x_i\in S_i$ and $x=\sum_{i=1}^n x_i$.

We know that for each $i\in I$ there is an $a_i\in S_i$ $(i\in I)$, such that

$$(2)\quad \|\textstyle\sum_{i\in I}(x_i-a_i)\|\leqslant\rho\left(\sum_{i\in I}S_i\right)\leqslant\sum_{i\in I}\rho(S_i),$$

by the definition of $\rho(\cdot)$, and by the result of the preceding paragraph. For i not in I, define $a_i\equiv x_i$. Therefore, there exist points $a_i\in S_i$ such that

$$\left\|\sum_{i=1}^n x_i-\sum_{i=1}^n a_i\right\|=\left\|\sum_{i\in I}x_i-\sum_{i\in I}a_i\right\|\leqslant mc,$$

because $\rho(S_i)\leqslant c$ and because of inequalities (2), and the fact that I has no more than m elements. ■

The S_i can be taken to be non-convex production sets or preference sets or some combination of them. Thus, in the aggregate, the discrepancy between an allocation in the fictitious economy generated by con S and some allocation in the real economy is bounded in a way that is independent of the number of economic agents. Therefore, the average agent experiences a deviation from intended actions that vanishes in significance as the number of agents goes to infinity.

9. Concave functions, quasi-concave functions, and homothetic functions

A real-valued function on a convex subset S of $\boldsymbol{R}^m$ is *concave* if for any $x^\circ, x^1 \in S$, we have $f(\theta x^\circ + (1-\theta)x^1) \geqslant \theta f(x^\circ) + (1-\theta)f(x^1)$. Geometrically, the graph of the function lies above the line joining any two points on the graph (see Figure 9.1). The concavity of a function is equivalent to the convexity of the set of points lying below the graph of the function (i.e., the set G in Figure 9.1). A function f is *convex* if $(-f)$ is concave.

If a concave function f is differentiable, then one can show that the tangent plane to any point on the graph of f lies above the graph. Formally, if $\partial f(x^\circ)/\partial x$ is the vector of partial derivatives at some arbitrary x°, then $f(x^1) - f(x^\circ) \leqslant (x^1 - x^\circ)\cdot(\partial f(x^\circ)/\partial x)$, for any x^1 in the domain. If f is a function of a real variable, this fact can be seen as follows: suppose for simplicity $x^\circ = 0$ and $f(0) = 0$. Then we must show $f(x^1) \leqslant x^1 \cdot (\partial f(0)/\partial x)$. For this purpose, let $0 < y < x^1$. By concavity,

$$f(y) \geqslant \left[\frac{x^1 - y}{x^1}\right] f(0) + \left[\frac{y}{x^1}\right] f(x^1) = \frac{y}{x^1} f(x^1).$$

Thus $f(y)/y \geqslant f(x^1)/x^1$. Letting $y \to 0$, we find that $\partial f(0)/\partial x \geqslant f(x^1)/x^1$, as was to be shown.

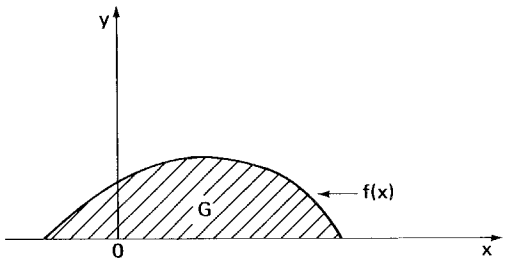


Figure 9.1

This preceding fact shows that the first-order conditions for a maximum, viz, $\partial f(x^\circ)/\partial x = 0$, are sufficient for a global maximum at x°, in the event that f is concave. For, then $f(x^1) - f(x^\circ) \leqslant 0 \cdot (x^1 - x^\circ) = 0$. [See Chapter 2 for more detail.]

A function f is *quasi-concave* if for all $\alpha \in R$, the set $\{x \mid f(x) \geqslant \alpha\}$ is convex. For example, if f is a utility function that has convex upper preference sets, then f is quasi-concave. A concave function is clearly quasi-concave. A point that satisfies the first-order Lagrangian conditions (see Chapter 2) is a global constrained maximum if the objective function f is quasi-concave and the constraint set is described by concave functions $g_i(x) \leqslant 0$ (so long as a weak technical condition is also satisfied).

A function f is *quasi-convex* if $(-f)$ is quasi-concave. The above results for quasi-concave functions can be recast, of course, for quasi-convex functions with the appropriate changes of inequalities and with the word "minimum" replacing "maximum".

We say that a function f: $X \to \boldsymbol{R}$, where $X \subset \boldsymbol{R}^m$ is *homogeneous of degree k* if, for any $x \in X$ and $\lambda > 0$, $f(\lambda x) = \lambda^k f(x)$. A production function that is homogeneous of degree 1 displays constant returns to scale. One can show that the expansion path for a homogeneous function is a ray through the origin: formally, that $\partial f(\lambda x)/\partial(\lambda x_i) = \lambda^{k-1}(\partial f(x)/\partial x_i)$. The expansion path in production theory is the set of input combinations that minimize cost (subject to a given output level) as the level of output expands (cf. Figure 9.2). Thus, the marginal rate of substitution $\mathrm{d}x_2/\mathrm{d}x_1|_{f=\text{const}}$ is unchanging along any ray from the origin. A *homothetic function h* is a monotone increasing transform of a homogeneous (of positive degree) function: $h(x) = g(f(x))$ where g is a monotone increasing function of a real variable and f is homogeneous of some degree $k > 0$. A homothetic function also has the property that the marginal rate of substitution is constant along any ray from the origin. In fact, this latter property is equivalent to homotheticity.

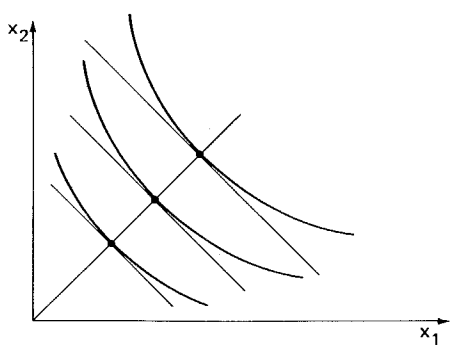


Figure 9.2

10. Hemi-continuity of correspondences and the maximum theorem

A correspondence is a point-to-set function. Formally speaking, let X and Y be any sets. A *correspondence* γ from X into Y is a relation which associates to every element $x \in X$ a unique non-empty subset $\gamma(x)$ of Y. Let 2^Y denote the set of all subsets of Y. Then, in functional notation, we can express the correspondence γ between X and Y as γ: $X \to 2^Y$ with $\gamma(x) \neq \phi$.

Correspondences are important in economics because, for example, there is frequently more than one solution point to a constrained maximum problem. Thus, in the absence of strict convexity of preferences, economists must deal with demand correspondences rather than demand functions.

The *graph* of the correspondence γ: $X \to 2^Y$ is the set $G(\gamma) = \{(x, y) | y \in \gamma(x)\}$. On the other hand, every set S in $X \times Y$ defines a relation ψ as follows: $\psi(x) = \{y | (x, y) \in S\}$; ψ will be a correspondence if $\psi(x) \neq \phi$ for all $x \in X$. The *sum of correspondences* γ_i: $X \to 2^Y$ $(i = 1, \ldots, n)$, where Y is a Euclidean space, is defined as $\sum_{i=1}^n \gamma_i(x) \equiv \{y \in Y |$ there exists $y_i \in \gamma_i(x)$, for all i, such that $y = \sum_{i=1}^n y_i\}$. The *cross-product of correspondences* γ_i: $X \to 2^Y$ $(i = 1, \ldots, n)$ is defined as $\gamma_1(x) \times \gamma_2(x) \times \cdots \times \gamma_n(x) \equiv \{(y_1, y_2, \ldots, y_n) | y_i \in \gamma_i(x)$ for each $i\}$. The *composition of two correspondences* γ: $Y \to 2^Z$ and ψ: $X \to 2^Y$ is defined as $\gamma \circ \psi(x) \equiv \{z |$ there exists $y \in \psi(x)$ such that $z \in \gamma(y)\}$.

For the rest of this section, suppose that X and Y are topological spaces. Hemi-continuity of correspondences is an essential property for obtaining the existence of fixed points. A correspondence γ is *upper hemi-continuous* (abbreviated *u.h.c.*) *at* $x^\circ \in X$ if for any open set V which contains all of $\gamma(x^\circ)$, there exists a neighborhood $U(x^\circ)$ such that $\gamma(x) \subset V$ for all $x \in U(x^\circ)$. The correspondence γ is *upper hemi-continuous* if it is u.h.c. at every $x \in X$. A correspondence which is *not* u.h.c. at x° "blows up" in any neighborhood of x° in the sense that part of $\gamma(x)$ lies outside some small open set containing $\gamma(x^\circ)$, as occurs in Figure 10.1.

It is easy to show that if γ is point valued (i.e., γ is a function), then γ is u.h.c. at x° if and only if it is a continuous function at x°. The term "upper semi-continuity" for correspondences is frequently used in the literature, rather than upper hemi-continuity. The latter term was recently adopted to avoid conceptual confusion with usage of upper semi-continuity for functions. [A function is upper semi-continuous at x° if for each $\varepsilon > 0$, there exists a neighborhood $N(x^\circ)$ such that $x \in N(x^\circ)$ implies $f(x) < f(x^\circ) + \varepsilon$.] There are numerous examples of functions that are upper semi-continuous but not continuous. We say $\gamma(x)$ is *compact-valued* if $\gamma(x)$ is a compact set for every $x \in X$. A correspondence γ is *closed at* x° if for every sequence $(x_n, y_n) \in G(\gamma)$ such that $(x_n, y_n) \to (x^\circ, y^\circ)$ it is true that $(x^\circ, y^\circ) \in G(\gamma)$, so that if $y^n \in \gamma(x^n)$ and $x^n \to x^\circ$ and $y^n \to y^\circ$, then $y^\circ \in \gamma(x^\circ)$. A correspondence is *closed* if it is closed at each x°. It is immediate that γ is closed if and only if $G(\gamma)$ is a closed set. Let the *image of a*

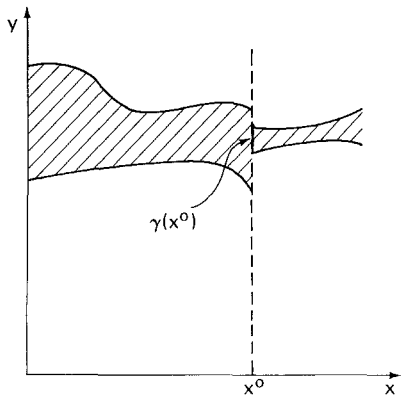


Figure 10.1

set $K \subset X$ *under* γ be defined as $\gamma(K) = \{y \mid y \in \gamma(x) \text{ for some } x \in K\}$. The *range* of γ is $\gamma(X)$. Suppose that the range of γ is compact and $\gamma(x)$ is a closed set for each x. It then turns out that γ is closed if and only if it is u.h.c. The latter result is frequently useful in establishing u.h.c. However, not all closed correspondences are u.h.c., even when compact-valued. Let γ: $\boldsymbol{R} \to 2^{\boldsymbol{R}}$ be such that

$$\begin{aligned} \gamma(x) &= \{0\}, \quad x = 0, \\ &= \{1/x\}, \quad x > 0. \end{aligned}$$

The following condition is sufficient for u.h.c. at x° when γ is compact-valued and X and Y are metric spaces: For every pair of sequences (x^n), (y^n) such that $x^n \to x^\circ$ and $y^n \in \gamma(x^n)$, there is a convergent subsequence of (y^n) whose limit belongs to $\gamma(x^\circ)$.

Let γ: $X \to 2^Y$, where X and Y are metric spaces; the following properties of u.h.c. correspondences are useful:

(1) $\overline{\gamma(x)}$ is u.h.c. if γ is u.h.c.
(2) $\cup_{i=1}^n \gamma_i$ is u.h.c. if the γ_i are u.h.c.
(3) If $(\gamma_i)_{i=1}^n$ are all u.h.c. and compact-valued, then $\sum_{i=1}^n \gamma_i$ is u.h.c. and compact-valued, where Y is a Euclidean space.
(4) The cross-product of u.h.c. and compact-valued correspondences is u.h.c. and compact-valued.
(5) If γ is u.h.c. and compact-valued, then the convex hull correspondence con γ (defined by $\text{con}\,\gamma(x) \equiv \text{con}[\gamma(x)]$) is also u.h.c. and compact-valued, when Y is a subset of Euclidean space.
(6) If K is a compact set and if γ is u.h.c. and compact-valued, the image of K under γ, $\gamma(K)$, is a compact set.

(7) The composition $\gamma \circ \psi$ of two u.h.c. correspondences is u.h.c.
(8) If γ_1 and γ_2 are u.h.c. and closed-valued, then $\gamma_1 \cap \gamma_2$ is u.h.c. if $\gamma_1 \cap \gamma_2(x) \neq \phi$ for all $x \in X$.

A correspondence is *lower hemi-continuous* (or *l.h.c.*) *at* x°, if for every open set V that meets $\gamma(x^\circ)$, i.e., $\gamma(x^\circ) \cap V \neq \phi$, there exists a neighborhood of x°, $U(x^\circ)$, such that $\gamma(x)$ also meets V for every $x \in U(x^\circ)$. We say that γ is *l.h.c.* if it is l.h.c. at every $x \in X$. Geometrically, the idea is that $\gamma(x)$ does not suddenly contract in size if we move slightly away from x°, as occurs in Figure 10.2 (where there is a "spike" at x°). Again, if $\gamma(x)$ is single-valued, then the *correspondence* γ is l.h.c. if and only if the *function* γ is continuous.

When X and Y are metric spaces the following sequence condition is necessary and sufficient for γ to be l.h.c. at x°: for every sequence (x^n) which converges to x° and for every $y^\circ \in \gamma(x^\circ)$, there exists a sequence (y^n) such that $y^n \in \gamma(x^n)$ (for all n) and $y^n \to y^\circ$.

The following is a list of useful properties of l.h.c. correspondences when X and Y are metric spaces:

(1) $\bar{\gamma}$ is l.h.c. if γ is l.h.c.
(2) $\cup_{i=1}^{n} \gamma_i$ is l.h.c. if the γ_i are l.h.c.
(3) The composition $\gamma_1 \circ \gamma_2$ is l.h.c. if each of the γ_i are l.h.c.
(4) The sum of l.h.c. correspondences is an l.h.c. correspondence, if Y is a Euclidean space.
(5) The cross-product of l.h.c. correspondences is l.h.c.
(6) Let Y be a subset of Euclidean space. Then $\operatorname{con}\gamma(x)$ is l.h.c., if γ is l.h.c.
(7) The intersection of two l.h.c. correspondences is *not* in general l.h.c.; however:
(8) Let X and Y be subsets of Euclidean space. If γ_i, $i = 1, 2$, are two l.h.c. convex-valued correspondences such that $\operatorname{int}\gamma_1(x^\circ) \cap \operatorname{int}\gamma_2(x^\circ) \neq \phi$, then $\gamma_1 \cap \gamma_2$ is l.h.c. at x°.

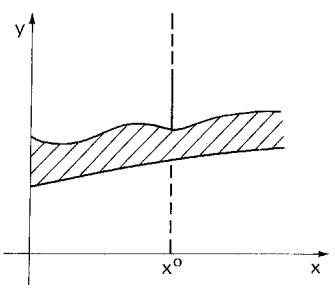


Figure 10.2

A correspondence is *continuous at* x° if it is both u.h.c. and l.h.c. at x°. A *continuous correspondence* is continuous at each $x \in X$. The budget correspondence $B(p, \omega) = \{x \in C \mid p \cdot (x - \omega) \leqslant 0\}$ is continuous if the endowment vector $\omega \in \operatorname{int} C$ and C is convex, where C is the set of possible consumption vectors and p is the price vector. This follows from property (8) for l.h.c. correspondences and from the facts that: (a) C and $\{x \mid p \cdot (x - \omega) \leqslant 0\}$ are convex sets, (b) $z = \omega - \varepsilon(1, 1, \ldots, 1) \in \operatorname{int} C$ for $\varepsilon > 0$ sufficiently small and $p \cdot (z - \omega) < 0$, since $z \in \operatorname{int} C \cap \{x \mid p \cdot (x - \omega) < 0\}$ in that case, and (c) $\{x \mid p \cdot (x - \omega) \leqslant 0\}$ is continuous in (p, ω) and $C(p, \omega) \equiv C$ is also continuous in (p, ω).

A *demand correspondence* $x(p, \omega)$ is the set of maximizers of $U(x)$ subject to $x \in B(p, \omega)$. Of great interest in economics are the continuity properties of $x(p, \omega)$.

Maximum Theorem

Let X be a topological space. If F is a continuous, real-valued function of X and B is a continuous compact-valued correspondence from Y to subsets of X, then the correspondence γ defined by $\gamma(y) = \{x \in B(y) \mid F(x) \geqslant F(x') \text{ for all } x' \in B(y)\}$ is u.h.c. and compact-valued, and the function f defined by $f(y) \equiv F(\gamma(y))$ is a continuous function.

Note that if either of the sets X or Y in the statement of the Maximum Theorem is a compact set, then γ is a closed correspondence; for if X is compact, then $\gamma(y) \subset X$ and $\gamma(y)$ compact-valued and u.h.c. implies γ is a closed correspondence. On the other hand, if Y is compact then $B(Y)$ [i.e., range of $B(y)$] is compact, since the image of a compact set under an u.h.c. correspondence is compact, so again $\gamma(y)$ is contained in the same compact set for all $y \in Y$.

A sketch of the proof of the first conclusion of the Maximum Theorem may be helpful in grasping its meaning. Suppose, for simplicity, that X and Y are metric spaces and that Y is compact. We shall show that γ is closed, i.e., $y^n \to y^\circ$, $x^n \in \gamma(y^n)$ (for all n) and $x^n \to x^\circ$ imply $x^\circ \in \gamma(y^\circ)$. Since $B(y)$ is u.h.c., $x^\circ \in B(y^\circ)$. Let z° be any element of $B(y^\circ)$. By the l.h.c. of $B(y)$, there exists a sequence $z^n \in B(y^n)$ such that $z^n \to z^\circ$. But $x^n \in \gamma(y^n)$ means that $F(x^n) \geqslant F(z^n)$. The function F is continuous, so $F(x^\circ) \geqslant F(z^\circ)$. But this means precisely that $x^\circ \in \gamma(y^\circ)$, as was to be shown.

11. Fixed point theorems

A *fixed point of a function* f: $X \to Y$ (where $X \cap Y \neq \phi$) is a point $\bar{x} \in X$ such that $\bar{x} = f(\bar{x})$. Fixed point theorems are useful for establishing the existence of solutions to a system of nonlinear equations. In economics, fixed point theorems are most often used to guarantee the existence of equilibrium in a wide variety

of models of the economy. For instance, let $f(p)$ be an m-dimensional excess demand function. A vector $\bar{p}$ such that $f(\bar{p})=0$ is a competitive equilibrium, since demand equals supply at $\bar{p}$. A fixed point $\bar{p}$ of the function $f(p)+p$ would therefore be a competitive equilibrium. (Unfortunately, establishing the existence of competitive equilibrium is more complex than the preceding sentence suggests.)

The importance of establishing the existence of solutions to economic models is sometimes underemphasized. It is not uncommon for the researcher to find that an economic model is vacuous in having no solutions. The well-known Cournot duopoly model is a case in point. Except in the unlikely circumstances of identical firms or concave demand functions, there may be no Cournot (pure-strategy) equilibrium [see Roberts and Sonnenschein (1977)].

Brouwer Fixed Point Theorem

If X is a compact convex subset of $\boldsymbol{R}^m$ and $f: X \to X$ is a continuous function, then there is a fixed point.

The Brouwer theorem is relatively easy to establish when f is a continuous function of a real variable. Thus, let $X=[0,1]$. Consider Figure 11.1: If f crosses the 45° line, then f has a fixed point. We might as well assume that $f(0)>0$, since otherwise zero is a fixed point. Suppose f lies above the 45° line everywhere. But this is impossible since $f(1) \leqslant 1$ by the assumption that $f: X \to X$.

A function f: $(S, d) \to (S, d)$ is said to be a *contraction* if there exists $r<1$ such that for all $x, y \in S$, $d(f(x), f(y)) < r\ d(x, y)$. Clearly, any contraction is a continuous function.

Contraction Mapping Theorem

A contraction on a complete metric space has a unique fixed point.

Although no assumptions about convexity or finite dimensionality of S are needed for this theorem, contractions are rather special functions.

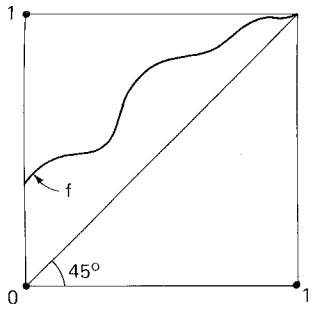


Figure 11.1

A *fixed point of a correspondence* $\gamma(x)$ (where γ: $X \to 2^X$) is a point $\bar{x}$ such that $\bar{x} \in \gamma(\bar{x})$.

Kakutani Fixed Point Theorem

Let X be a compact convex subset of R^m, and let γ be a closed correspondence from X into subsets of X. If $\gamma(x)$ is a convex set for every $x \in X$, then there is a fixed point.

An interesting exercise for the reader is to give counter-examples to this result in each circumstance in which any one of the assumptions is not satisfied. For instance, let $\gamma(x) = X \setminus B_\varepsilon(x)$, i.e., $\gamma(x)$ is the complement of an open ball centered on x. Then $\gamma(x)$ is connected, but not convex, γ is closed, and γ has no fixed point.

Reference notes

A good, rigorous survey of some economic applications of the methods discussed here is the textbook by Takayama. His book also includes proofs of many of the less advanced mathematical results. Detailed below are some good expositions of the proofs of the theorems cited in this chapter, as well as their extensions. Many of these sources also have excellent bibliographies (e.g., Arrow–Hahn, Hildenbrand, Klein, Rockafeller, Smart, and Takayama).

Sections 1–6. Good intermediate-level sources are Rudin (1964) and Simmons (1963).

Section 7. A good source is Dieudonné (1960, ch. 3).

Section 8. Arrow–Hahn (1971, app. B), Hildenbrand–Kirman (1976, app. II), Karlin (1959, app. B), Klein (1973, pp. 72–76, 323–341), and Nikaido (1968, pp. 15–44) contain proofs of most of the results of this section. Excellent advanced treatments of convex sets are in Berge (1963, ch. 7 and pp. 158–168) and Rockafeller (1970, pp. 3–22, 43–81, 95–101, 153–212). The first proof of the Shapley–Folkman Theorem appeared in Starr (1969). Artstein (1976) has an interesting general result that implies both the Carathéodory and the Shapley–Folkman Theorems, among others.

Section 9. Nikaido (1968, pp. 44–53) is a good source for proofs of most of the results in this section. Rockafeller is the classic treatise on convex functions.

Section 10. Hildenbrand–Kirman (1976, app. III) and Klein (1973, ch. 6) are good elementary sources. Berge (1963, ch. 6) and Nikaido (1968, pp. 70–73) are good intermediate-level sources. Hildenbrand (1974, pp. 21–35) and Heller (1978) contain useful results for economics about the continuity properties of correspondences.

Section 11. Good elementary proofs of Brouwer's Theorem are in Burger (1963, app.), Klein (1973, ch. 7) and Tompkins (1964). Kakutani's Theorem is also proved by Burger and by Klein. More advanced treatments of the two theorems are given in Berge (1963, pp. 168–176) and Nikaido (1968, pp. 53–70). Arrow–Hahn (1971, app. C) contains proofs of the Brouwer and Kakutani Theorems based on Scarf's algorithm. A good advanced reference for a wide variety of fixed point theorems is Smart.

References

Arrow, K. J. and F. Hahn (1971), General competitive analysis. San Francisco, CA: Holden-Day. Now distributed by North-Holland, Amsterdam.
Artstein, Z. (1976), "Look at extreme points", Manuscript. Rehovat: Weitzman Institute of Science.
Aubin, J. P. and I. Ekelund (1974), "A discrete approach to the bang-bang principle", Mimeo.
Berge, C. (1963), Topological spaces. New York: Macmillan.
Burger, E. (1963), Introduction to the theory of games. Englewood Cliffs, NJ: Prentice-Hall.
Debreu, G. (1959), Theory of value, Cowles Foundation Monograph, Vol. 17. New York: Wiley.
Dierker, E. (1974), Topological methods in Walrasian economics, Lecture notes in economics and mathematical systems, Vol. 92. Berlin: Springer-Verlag.
Dieudonné, J. (1960), Foundations of modern analysis. New York: Academic Press.
Dunford, N. and J. T. Schwartz (1964), Linear operators, Part 1: General theory. New York: Wiley.
Heller, W. P. (1972), "Transactions with set-up costs", Journal of Economic Theory, 4:465–478.
Heller, W. P. (1978), "Continuity in general nonconvex economies (with applications to the convex case)", in: G. Schwödiauer, ed., Equilibrium and disequilibrium in economic theory. Deventer: Reidel.
Hildenbrand, W. (1974), Core and equilibria of a large economy. Princeton, NJ: Princeton University Press.
Hildenbrand, W. and A. Kirman (1976), Introduction to equilibrium analysis. Amsterdam: North-Holland.
Kakutani, S. (1941), "A generalization of Brouwer's fixed point theorem". In 1968 reprinted in: P. Newman, ed., Readings in mathematical economics, Part I. Baltimore, MD: The Johns Hopkins Press.
Karlin, S. (1959), Mathematical methods and theory in games, Programming and economics, Vol. 1. Reading, MA: Addison-Wesley.
Kelley, J. L. (1955), General topology. New York: Van Nostrand.
Klein, E. (1973), Mathematical methods in theoretical economies. New York: Academic Press.
Kolmogorov, A. N. and S. V. Fomin (1970), Introductory real analysis. Englewood Cliffs, NJ: Prentice-Hall.
Newman, P., ed. (1968), Readings in mathematical economics, Part I. Baltimore, MD: The Johns Hopkins Press.
Nikaido, H. (1968), Convex structures and economic theory. New York: Academic Press.
Roberts, J. and H. Sonnenschein (1977), "On the foundations of monopolistic competition", Econometrica, 45:101–114.
Rockafeller, R. T. (1970), Convex analysis. Princeton, NJ: Princeton University Press.
Rudin, W. (1964), Principles of mathematical analysis, 2nd ed. New York: McGraw-Hill.
Scarf, H. (1973), The computation of economic equilibria. New Haven, CT: Yale University Press.
Simmons, G. F. (1963), Introduction to topology and modern analysis. New York: McGraw-Hill.
Smart, D. (1974), Fixed point theorems. Cambridge: Cambridge University Press.
Starr, R. M. (1969), "Quasi-equilibria in markets with non-convex preferences", Econometrica, 37:25–38.
Takayama, A. (1974), Mathematical economics. Hillsdale, NJ: Dryden Press.
Tompkins, C. B. (1964), "Sperner's lemma and some extensions". In 1968 reprinted in: P. Newman, ed., Readings in mathematical economics, Part I. Baltimore, MD: The John Hopkins Press.

Chapter 2

MATHEMATICAL PROGRAMMING WITH APPLICATIONS TO ECONOMICS

MICHAEL D. INTRILIGATOR*

University of California, Los Angeles

1. Introduction and overview

Mathematical programming refers to the basic mathematical problem of maximizing a function subject to constraints.[1] The nature of this problem and its various solution concepts are discussed in Section 2. Historically this problem has its roots in the development of the calculus.[2] Indeed, one of the first uses of the calculus was to treat the simplest problem of mathematical programming, that of *unconstrained maximization*, as discussed in Section 3. A basic motivation for the further development of the calculus was that of solving a more general type of mathematical programming problem. This problem, the *classical programming problem* of maximization of a given function subject to a set of equality constraints, is discussed in Section 4. Other problems of mathematical programming, some of which were influenced by the study of certain economic problems, were not treated until the twentieth century. One such problem is the *nonlinear programming problem* of maximization of a given function subject to a set of inequality constraints, as discussed in Section 5. A special case, important in itself and one which was extremely influential in the development of the theory of mathematical programming, is the *linear programming problem* of maximization of a given linear form subject to a set of linear inequality constraints, as discussed in Section 6.

Applications of the mathematical programming problem are legion. In economics the theory of mathematical programming has been applied to a wide variety of problems. It has been used to characterize the solution of fundamental problems in virtually all areas of economics. It has also led to the comparative

*The author would like to acknowledge, with appreciation, the helpful suggestions of Kenneth Arrow, Jeffrey Conner, Erwin Diewert, Richard Ernst, James Friedman, Arthur Geoffrion, Magnus Hestenes, James Quirk, John Riley, Knut Sydsaeter, Leigh Tesfatsion, and Daniel Vandermeulen.

[1]The problem will be stated here as one of maximization. A problem of minimization can be treated as one of maximization simply by changing the sign of the function to be minimized.

[2]For a discussion of the historical development of mathematical programming, see Dantzig (1963). The term "programming" is based on scheduling of activities, which led to the development of linear programming. Use of this term was then extended via the development of nonlinear programming.

statics analysis of these problems. The mathematical programming problem provides one of the main approaches to the study of *microeconomics*, as discussed in Section 7. Application of mathematical programming to two principal areas of study in microeconomics, the *neoclassical theory of the household* and the *neoclassical theory of the firm*, are discussed in Sections 8 and 9, respectively.

In addition to the basic mathematical theory, as discussed in Sections 2–6 and applications to economics, as discussed in Sections 7–8 mathematical programming also encompasses applications to other areas (e.g., engineering, physics) and computational techniques. While not treated here, these other applications and computational techniques are discussed in the references cited in the bibliography. Also omitted here are problems with discrete variables (integer programming), problems with random variables (stochastic programming), and problems with vector-valued objective functions (multi-criterion problems), which, again, are discussed in the references cited in the bibliography.

2. The mathematical programming problem and solution concepts[3]

The general form of the *mathematical programming problem* can be stated

$$\max_{\boldsymbol{x}} F(\boldsymbol{x}) \quad \text{subject to} \quad \boldsymbol{x} \in X. \tag{2.1}$$

Here $\boldsymbol{x}$ is a column vector of n choice variables,

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_n)' \tag{2.2}$$

(the prime denotes the transpose of the row vector), $F(\boldsymbol{x})$ is a given real-valued function of these variables,

$$F(\boldsymbol{x}) = F(x_1, x_2, \ldots, x_n), \tag{2.3}$$

and X is a given subset of Euclidean n-space (the space of all n-tuples of real numbers),[4]

$$X \subset E^n. \tag{2.4}$$

[3] Basic references on mathematical programming include Hadley (1964), Luenberger (1969, 1973), Intriligator (1971), Aoki (1971), Geoffrion (1972), and Hestenes (1975).

[4] The more general problem of mathematical optimization can be stated as in (2.1) where X can be a subset of any appropriately defined space. For example, if X is a subset of the finite dimensional space E^n then the problem is one of mathematical programming, while if X is a subset of the infinite dimensional space of piecewise continuous functions then the problem is one of mathematical control. For a discussion of control theory see Chapter 4 by Kendrick. [See also Intriligator (1971).] For a discussion of E^n (which is sometimes written R^n) and other spaces see Chapter 1 by Green and Heller.

It will generally be assumed that X is not empty, that is, that there exists a feasible vector $\boldsymbol{x}$, where $\boldsymbol{x}$ is *feasible* if and only if $\boldsymbol{x} \in X$.

In economics the vector $\boldsymbol{x}$ is frequently called the vector of *instruments*, the function $F(\boldsymbol{x})$ is frequently called the *objective function* (or *criterion function*), and the set X of feasible instrument vectors is frequently called the *opportunity set*. The basic economic problem of allocating scarce resources among competing ends can then be interpreted as one of mathematical programming, where a particular resource allocation is represented by the choice of a particular vector of instruments; the scarcity of the resources is represented by the opportunity set, reflecting constraints on the instruments; and the competing ends are represented by the objective function, which gives the value attached to each of the alternative allocations. Problem (2.1) can therefore be interpreted in the language of economics as that of choosing instruments within the opportunity set so as to maximize the objective function.[5]

There are various solution concepts for the basic problem (2.1). A *global maximum* (or *solution*) is a vector $\boldsymbol{x}^*$ for which

$$\boldsymbol{x}^* \in X \quad \text{and} \quad F(\boldsymbol{x}^*) \geqslant F(\boldsymbol{x}) \quad \text{for all} \quad \boldsymbol{x} \in X. \tag{2.5}$$

It is a solution in that the instrument vector yields a value for the objective function that is no less than its value at any feasible instrument vector. A *strict global maximum* is a vector $\boldsymbol{x}^*$ which satisfies

$$\boldsymbol{x}^* \in X \quad \text{and} \quad F(\boldsymbol{x}^*) > F(\boldsymbol{x}) \quad \text{for all} \quad \boldsymbol{x} \in X, \quad \boldsymbol{x} \neq \boldsymbol{x}^*. \tag{2.6}$$

2.1. *Weierstrass theorem*[6]

According to the Weierstrass Theorem if the function $F(\boldsymbol{x})$ is continuous and the set X is closed and bounded (hence compact) and non-empty then there exists a global maximum. The proof of this theorem is based upon the fact that the *image* of X under F, defined as

$$F(X) = \{F(\boldsymbol{x}) | \boldsymbol{x} \in X\}, \tag{2.7}$$

is a closed and bounded set on the real line and therefore must contain a

[5]For a further discussion of the intimate connections between the problem of mathematical programming and that of economic allocation, including the basic theory of mathematical programming and applications to economics, see Intriligator (1971, 1975, 1977). See also Lancaster (1968); El-Hodiri (1971); Takayama (1974); and Dixon, Bowles and Kendrick (1980).

[6]In general all theorems will be given names here. Some of these names such as the "Weierstrass Theorem" or the "Theorem on First-Order Conditions" are well known and appear in the literature. Other theorem names, such as the "Local–Global Theorem" or the "Demand Theorem" are not well-known, but they provide useful and descriptive names for these theorems.

maximal element, which is $F(x^*)$. It should be noted that the conditions of the theorem are sufficient but are not necessary for the existence of a maximum, i.e., a maximum may exist even if these conditions are not met. (For example, the problem of maximizing x^2 subject to $0<x\leqslant 2$ has a solution.) The Weierstrass Theorem can be strengthened by relaxing the assumption on $F(x)$ to that of $F(x)$ being upper semicontinuous.[7]

2.2. *Local–global theorem*

A *local maximum* is a vector $x^*\in X$ for which, for some $\varepsilon>0$,

$$F(x^*)\geqslant F(x) \quad \text{for all} \quad x\in X\cap N_\varepsilon(x^*). \tag{2.8}$$

Here $N_\varepsilon(x^*)$ is an ε neighborhood of x^*, i.e., the set of all points no more than ε distance from x^*.[8] The maximum is "local" in that the instrument vector yields a value for the objective function that is no less than its value at any point that is both feasible (i.e., in X) and sufficiently "close" (i.e., in $N_\varepsilon(x^*)$ for some $\varepsilon>0$). A *strict local maximum* is a vector $x^*\in X$ that satisfies, for some $\varepsilon>0$,

$$F(x^*)>F(x) \quad \text{for all} \quad x\in X\cap N_\varepsilon(x^*), \quad x\neq x^*. \tag{2.9}$$

Obviously, a global maximum is a local maximum but not vice-versa, a strict (global or local) maximum is also a (global or local) maximum but not vice-versa, and a strict global maximum is unique.

According to the Local–Global Theorem, if the objective function $F(x)$ is a concave function and the opportunity set X is a convex set then every local maximum is a global maximum, the set of all such solutions is convex, and the solution is unique if $F(x)$ is a strictly concave function.[9] Generalizing the last

[7]As discussed in Chapter 1, the function $F(x)$ is *upper semicontinuous at* x_0 if and only if given any $\varepsilon>0$ there exists a $\delta>0$ such that $|x-x_0|<\delta$ implies $F(x)-F(x_0)<\varepsilon$, and $F(x)$ is *upper semicontinuous* if and only if it is upper semicontinuous at all points in its domain. Here $|x-x_0|$ is the *Euclidean distance* between x and x_0, defined as

$$|x-x_0|=\sqrt{\sum_{j=1}^{n}\left(x_j-x_j^0\right)^2}\ .$$

[8]Using the Euclidean distance function of the last footnote, the ε *neighborhood* of the point x^* is defined as $N_\varepsilon(x^*)=\{x\,|\,|x-x^*|\leqslant\varepsilon\}$. See also Chapter 1.

[9]As discussed in Chapter 1, the set X is a *convex set* if and only if all convex combinations of points in X are also in X, i.e., given $x^1, x^2\in X$, $\alpha x^1+(1-\alpha)x^2\in X$ for all α, $0\leqslant\alpha\leqslant 1$. The function $F(x)$ is a *concave function* if and only if the value of the function at a convex combination of points is never less than the linearly interpolated value of the function, i.e., given $x^1, x^2\in X$, $F(\alpha x^1+(1-\alpha)x^2)\geqslant\alpha F(x^1)+(1-\alpha)F(x^2)$ for all α, $0\leqslant\alpha\leqslant 1$. The function $F(x)$ is a *strictly concave function* if and only if, given $x^1, x^2\in X$, $x^1\neq x^2$, the above inequality holds strictly for all α, $0<\alpha<1$.

part of this theorem, if $F(\boldsymbol{x})$ is strictly quasiconcave then a local maximum is the unique global maximum.[10]

The Local–Global Theorem is extremely important because virtually all methods for solving mathematical programming problems identify local rather than global maxima. With the Local-Global Theorem it is possible to infer, with the proper assumptions on concavity and convexity, that a local solution is also a global solution.

3. The unconstrained problem[11]

The *problem of unconstrained maximization* is that of choosing values of n variables so as to maximize a function of these variables

$$\max_{\boldsymbol{x}} F(\boldsymbol{x}). \tag{3.1}$$

In this case the opportunity set X in (2.1) is the entire space E^n (or an open subset of E^n).

3.1. Theorem on first-order conditions

According to the Theorem on First-Order Conditions, if $F(\boldsymbol{x})$ is a differentiable function, then a first-order necessary condition for $\boldsymbol{x}^*$ to be a local maximum of $F(\boldsymbol{x})$ is that $\boldsymbol{x}^*$ be a *stationary point* at which all first-order partial derivatives vanish,

$$\frac{\partial F}{\partial \boldsymbol{x}}(\boldsymbol{x}^*)=\left(\frac{\partial F}{\partial x_1}(\boldsymbol{x}^*), \frac{\partial F}{\partial x_2}(\boldsymbol{x}^*),\ldots, \frac{\partial F}{\partial x_n}(\boldsymbol{x}^*)\right)=\mathbf{0}. \tag{3.2}$$

Here $(\partial F/\partial \boldsymbol{x})(\boldsymbol{x}^*)$ is the *gradient vector*, the $1\times n$ row vector of all first-order partial derivatives of $F(\boldsymbol{x})$ and $\mathbf{0}$ is the $1\times n$ zero vector, all elements of which are 0.[12] Thus the theorem states that if $\boldsymbol{x}^*=(x_1^*, x_2^*,\ldots, x_n^*)$ is a local maximum, then

$$\frac{\partial F}{\partial x_j}(x_1^*, x_2^*,\ldots, x_n^*)=0, \qquad j=1,2,\ldots,n. \tag{3.3}$$

[10]The function $F(\boldsymbol{x})$ is a *quasiconcave function* if and only if, given $\boldsymbol{x}^1, \boldsymbol{x}^2\in X$, where $F(\boldsymbol{x}^1)\geqslant F(\boldsymbol{x}^2)$,

$$F(\alpha\boldsymbol{x}^1+(1-\alpha)\boldsymbol{x}^2)\geqslant F(\boldsymbol{x}^2) \quad \text{for all} \quad \alpha, \quad 0\leqslant\alpha\leqslant 1,$$

while it is *strictly quasiconcave* if and only if, given $\boldsymbol{x}^1, \boldsymbol{x}^2\in X$, $\boldsymbol{x}^1\neq\boldsymbol{x}^2$, where $F(\boldsymbol{x}^1)>F(\boldsymbol{x}^2)$, the above inequality holds strictly for all α, $0<\alpha<1$. Note that a concave function is quasiconcave but a quasiconcave function need not be concave. See also Chapter 1 and footnotes 26 and 34.

[11]See Cournot (1947), Apostol (1957), Fleming (1965), and Intriligator (1971).

[12]Use is made of the convention that the derivative of a scalar ($F(\boldsymbol{x})$) with respect to a column vector ($\boldsymbol{x}$) is a row vector. See Intriligator (1971, app. B).

This theorem can be proved by using a Taylor's series expansion for the value of the function around x^*.

3.2. *Theorem on second-order conditions*

According to the Theorem on Second-Order Conditions, if $F(x)$ is a twice differentiable function with continuous second-order partial derivatives, then second-order necessary conditions for x^* to be a local maximum of $F(x)$ are that the $n \times n$ Hessian matrix of second-order partial derivatives of $F(x)$,

$$\frac{\partial^2 F}{\partial x^2}(x) = \begin{bmatrix} \frac{\partial^2 F}{\partial x_1^2}(x) & \frac{\partial^2 F}{\partial x_1 \partial x_2}(x) & \cdots & \frac{\partial^2 F}{\partial x_1 \partial x_n}(x) \\ \vdots & & & \\ \frac{\partial^2 F}{\partial x_n \partial x_1}(x) & \frac{\partial^2 F}{\partial x_n \partial x_2}(x) & \cdots & \frac{\partial^2 F}{\partial x_n^2}(x) \end{bmatrix}, \tag{3.4}$$

be negative semidefinite at x^*.[13] Thus, according to the theorem, if x^* is a local maximum, then the Hessian matrix evaluated at x^* is negative semidefinite. This theorem can also be proved using Taylor's theorem.

3.3. *Theorem on sufficient conditions*

According to the Theorem on Sufficient Conditions, if the function $F(x)$ is twice differentiable with continuous second-order partial derivatives and the first-order conditions on the vanishing of the gradient vector (3.2) are met, then the *strengthened* second-order conditions, which state that the Hessian matrix is negative definite, imply that x^* is a (strict) local maximum for $F(x^*)$.[14] Once more Taylor's theorem can be used to prove this theorem.

The three conditions introduced here for the unconstrained problem have analogues in the constrained cases, as discussed in Sections 4 and 5.

[13] The matrix is *negative semidefinite* at x^* if the related quadratic form is non-positive, i.e.,

$$h' \frac{\partial^2 F}{\partial x^2}(x^*) h \leq 0 \quad \text{for all} \quad n \times 1 \text{ column vectors } h.$$

See Intriligator (1971, app. B).

[14] The Hessian matrix is *negative definite* at x^* if the related quadratic form is negative, i.e., the inequality of the previous footnote holds strictly for all $n \times 1$ column vectors $h \neq 0$. Note that the case of the function x^3 at $x = 0$ satisfies the second-order conditions but not the strengthened second-order conditions. The function x^3 is not at a maximum at $x = 0$, illustrating the fact that the Theorem on Second-Order Conditions provides necessary but not sufficient conditions for a local maximum.

3.4. An example: The quadratic objective function

An example of the unconstrained problem is that of maximizing the *quadratic objective function*,

$$\max_{x} F(x) = cx + \frac{1}{2}x'Qx = \sum_{j=1}^{n} c_j x_j + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} q_{ij}x_i x_j, \tag{3.5}$$

where c is a given $1 \times n$ row vector (c_j) and Q is a given $n \times n$ symmetric matrix (q_{ij}). The first part of the objective function is the linear form cx, and the second is the quadratic form $x'Qx$ (scaled by $\frac{1}{2}$ to simplify later expressions). The first-order necessary conditions for a local maximum state that the gradient vector vanishes,

$$\frac{\partial F}{\partial x}(x^*) = c + x^{*\prime}Q = 0, \tag{3.6}$$

and the second-order necessary conditions state that Q is negative semidefinite. By the sufficiency theorem, if Q is negative definite then conditions (3.6) imply that x^* is a strict local maximum. In fact, if Q is negative definite $F(x)$ is strictly concave, so x^* is also a global maximum. Furthermore Q is then non-singular, so solving for x^* yields

$$x^* = -Q^{-1}c'. \tag{3.7}$$

The maximized value of the objective function is then

$$F(x^*) = -cQ^{-1}c' + \frac{1}{2}(cQ^{-1})Q(Q^{-1}c') = -\frac{1}{2}cQ^{-1}c' > 0, \tag{3.8}$$

which is positive assuming Q (and Q^{-1}) are negative definite. Related examples to this one appear in Sections 4 and 5.

4. Classical programming: Lagrange multipliers[15]

The *problem of classical programming* is that of choosing values of n variables so as to maximize a function of these variables subject to equality constraints,

$$\max_{x} F(x) \quad \text{subject to} \quad g(x) = b. \tag{4.1}$$

[15]Basic references on classical programming include Courant (1947), Apostol (1957), Hadley (1964), Fleming (1965), Luenberger (1969, 1973), and Intriligator (1971). Unlike the terms "linear programming" and "nonlinear programming", the term "classical programming" is not in general use. This terminology is used because the problem is one of mathematical programming and because its origins are classical, extending back to the beginning of the calculus. It is, in fact, sometimes referred to as the "problem of classical constrained maximization".

Here the vector of *instruments* $\boldsymbol{x}$ and the *objective function* $F(\boldsymbol{x})$ are as in (2.1), where $F(\boldsymbol{x})$ is a real-valued function defined on E^n. The vector-valued function $\boldsymbol{g}(\boldsymbol{x})$ is a mapping from E^n into E^m, representing m *constraint functions*, and the column vector $\boldsymbol{b}$ is a $m \times 1$ vector of *constraint constants*,[16]

$$\boldsymbol{g}(\boldsymbol{x}) = \begin{bmatrix} g_1(x_1, x_2, \ldots, x_n) \\ g_2(x_1, x_2, \ldots, x_n) \\ \vdots \\ g_m(x_1, x_2, \ldots, x_n) \end{bmatrix}, \qquad \boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}. \tag{4.2}$$

In terms of the basic problem (2.1) the classical programming problem corresponds to the case in which the opportunity set can be written as

$$\begin{aligned} X &= \{\boldsymbol{x} \in E^n \mid \boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{b}\} \\ &= \{(x_1, x_2, \ldots, x_n)' \mid g_i(x_1, x_2, \ldots, x_n) = b_i, \quad i = 1, 2, \ldots, m\}. \end{aligned} \tag{4.3}$$

4.1. Theorem on Lagrange multipliers

A characterization of the solution to the problem of classical programming that is analogous to the Theorem on First-Order Conditions for unconstrained problems is provided by the Theorem on Lagrange Multipliers. For this theorem, introducing a row vector of m additional new variables called *Lagrange multipliers*,

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_m), \tag{4.4}$$

one for each constraint, the *Lagrangian function* is defined as the following real-valued function of the n original and the m added variables,

$$\begin{aligned} L(\boldsymbol{x}, \boldsymbol{y}) &= F(\boldsymbol{x}) + \boldsymbol{y}(\boldsymbol{b} - \boldsymbol{g}(\boldsymbol{x})) \\ &= F(x_1, x_2, \ldots, x_n) + \sum_{i=1}^{m} y_i(b_i - g_i(x_1, x_2, \ldots, x_n)), \end{aligned} \tag{4.5}$$

where the last term is the inner product of the row vector and the column vector of constraint constants less constraint functions.[17] Then, according to the Theorem on Lagrange Multipliers, assuming that $n > m$ (where $n - m$ is the

[16] There is no loss in generality in setting $\boldsymbol{b} = \boldsymbol{0}$. The constraint constants will be written $\boldsymbol{b}$, however, to facilitate analysis of the effect of changing these constraints, as in (4.11).

[17] The Lagrange multipliers are written $\boldsymbol{y}$ rather than the more common $\boldsymbol{\lambda}$ to ensure a consistent notation in all mathematical programming problems.

degrees of freedom of the problem), that $F(\boldsymbol{x})$ and $\boldsymbol{g}(\boldsymbol{x})$ are $m+1$ functions with continuous first-order partial derivatives, and that the constraints are linearly independent at the solution, i.e., if $\boldsymbol{x}^*$ is a local maximum of the problem,

$$\rho\left(\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}(\boldsymbol{x}^*)\right)=\rho\begin{bmatrix}\frac{\partial g_1}{\partial x_1}(\boldsymbol{x}^*) & \cdots & \frac{\partial g_1}{\partial x_n}(\boldsymbol{x}^*)\\ \vdots & & \\ \frac{\partial g_m}{\partial x_1}(\boldsymbol{x}^*) & \cdots & \frac{\partial g_m}{\partial x_n}(\boldsymbol{x}^*)\end{bmatrix}=m, \tag{4.6}$$

(that is, the $m\times n$ Jacobian matrix of all first-order partial derivatives of the constraint functions is of full row rank at the solution), the first-order necessary conditions are the $n+m$ conditions on the vanishing of all first-order partial derivatives of $L(\boldsymbol{x}, \boldsymbol{y})$,

$$\frac{\partial L}{\partial \boldsymbol{x}}(\boldsymbol{x}^*, \boldsymbol{y}^*)=\frac{\partial F}{\partial \boldsymbol{x}}(\boldsymbol{x}^*)-\boldsymbol{y}^*\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}(\boldsymbol{x}^*)=\boldsymbol{0} \quad (n \text{ conditions}), \tag{4.7}$$

$$\frac{\partial L}{\partial \boldsymbol{y}}(\boldsymbol{x}^*, \boldsymbol{y}^*)=\boldsymbol{b}-\boldsymbol{g}(\boldsymbol{x}^*)=\boldsymbol{0} \quad (m \text{ conditions}), \tag{4.8}$$

where the last m conditions simply require that the constraints be met at $\boldsymbol{x}^*$. Thus the theorem states that at a local maximum $\boldsymbol{x}^*$ there exists a vector of m Lagrange multipliers $\boldsymbol{y}^*$ such that, from (4.7), the gradient of $F(\boldsymbol{x})$ at $\boldsymbol{x}^*$ is a linear combination of the gradients of the $g_i(\boldsymbol{x})$ functions at this point, the Lagrange multipliers being the coefficients,[18]

$$\frac{\partial F}{\partial \boldsymbol{x}}(\boldsymbol{x}^*)=\boldsymbol{y}^*\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}(\boldsymbol{x}^*) \quad \text{i.e.} \quad \frac{\partial F}{\partial x_j}(\boldsymbol{x}^*)=\sum_{i=1}^{m} y_i^*\frac{\partial g_i}{\partial x_j}(\boldsymbol{x}^*), \quad j=1,2,\ldots,n. \tag{4.9}$$

These n conditions are analogous to the first-order conditions (3.2) on the vanishing of the gradient vector. In fact, this theorem reduces to the Theorem on First-Order Conditions if $m=0$, which is the unconstrained case. The theorem is usually proved using the Implicit Function Theorem.

A second part of the Theorem on Lagrange Multipliers gives an interpretation to these m additional variables. Consider not one problem of classical programming but a set of such problems characterized by the constraint constants $\boldsymbol{b}$. As any of these constants changes the maximized value of the objective function

[18]The Lagrange multipliers are unique since, by the rank condition, the gradients of the $g_i(\boldsymbol{x})$ functions at $\boldsymbol{x}^*$ are linearly independent, these gradients being the rows of the Jacobian matrix of (4.6).

will also change. This maximized value is given by

$$F^* = F(x^*) = L(x^*, y^*), \tag{4.10}$$

where the second equality follows from the fact that the constraints are satisfied at the solution (4.8). The Lagrange multipliers at their optimal values y^* measure the rate of increase of the maximized value F^* as the corresponding constraint constant is changed,[19]

$$y^* = \partial F^*/\partial b \quad \text{i.e.} \quad y_i^* = \partial F^*/\partial b_i, \quad i = 1,2,\ldots,m. \tag{4.11}$$

Thus each Lagrange multiplier measures the sensitivity of the maximized value of the objective function to changes in the corresponding constraint constants, all other parts of the problem remaining the same. In economic problems in which F has the dimensions of a value (price$\times$quantity) such as profit or revenue and b has the dimension of a quantity such as output or input the Lagrange multipliers y^* have the interpretation of a price, called a *shadow price* to distinguish it from a market price. They measure the increase in the value as the quantity constraint changes.

A geometric interpretation can be given for the classical programming problem and the characterization of its solution via Lagrange multipliers. The equality constraints define the opportunity set X in (4.3), which, by assumption (4.6), is of dimension $n-m$. The independence assumption in (4.6) implies that at the solution x^*, any direction $\mathrm{d}x$ satisfying

$$\frac{\partial g}{\partial x}(x^*)\mathrm{d}x = 0 \quad \text{i.e.} \quad \sum_{j=1}^{n} \frac{\partial g_i}{\partial x_j}(x^*)\mathrm{d}x_j = 0, \quad i = 1,2,\ldots,m, \tag{4.12}$$

lies in the tangent surface to X at x^*. The gradient vectors of the constraint functions, $(\partial g_i/\partial x_j)\ (x^*)$ are orthogonal to this tangent surface at x^*. The first-order conditions (4.9) mean, geometrically, that the gradient vector of the objective function $(\partial F/\partial x)\ (x^*)$, which points in the direction of maximum increase (steepest ascent) of $F(x)$ at x^*, is a weighted combination of the gradient vectors of the constraint functions, the weights being the Lagrange multipliers y^*. Thus $\partial F/\partial x\ (x^*)$ is also orthogonal to the tangent surface to X at x^* in that, given a direction $\mathrm{d}x$ in the tangent surface,

$$\frac{\partial F}{\partial x}(x^*)\mathrm{d}x = y^* \frac{\partial g}{\partial x}(x^*)\mathrm{d}x = 0. \tag{4.13}$$

[19] To ensure that (4.11) holds a second-order regularity condition of the form

$$\begin{vmatrix} 0 & \partial g/\partial x \\ \partial g/\partial x & \partial^2 L/\partial x^2 \end{vmatrix} \neq 0,$$

is assumed at (x^*, y^*). This condition ensures the existence of $\partial F^*/\partial b$. It is also assumed that $F(x)$ and $g(x)$ have continuous first- and second-order partial derivatives. See Intriligator (1971).

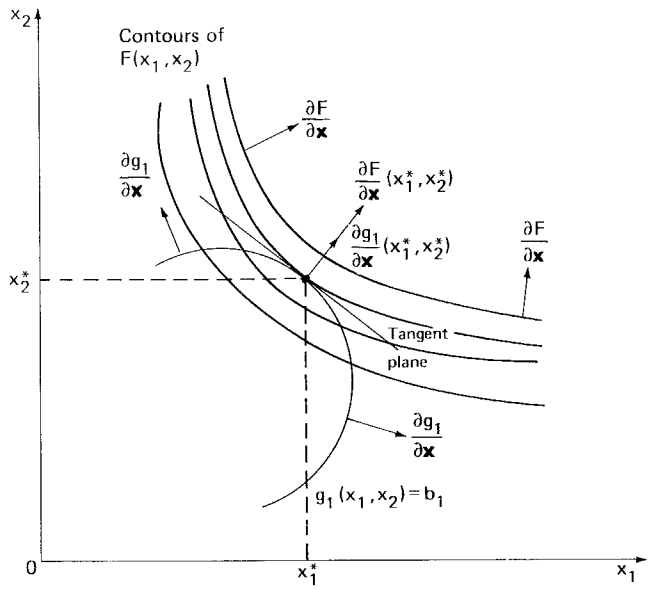


Figure 4.1. Solution to the classical programming problem: max $F(x_1, x_2)$ subject to $g_1(x_1)=b_1$.

The simplest problem, where $n=2$ and $m=1$, is illustrated in Figure 4.1, where $\partial F/\partial \boldsymbol{x}$ is orthogonal to the contours of $F(x_1, x_2)$ defined by $F(x_1, x_2)=$constant and where $\partial g_1/\partial \boldsymbol{x}$ is orthogonal to the tangent plane to the locus defined by $g_1(x_1, x_2)=b_1$. The solution is where $\partial F/\partial \boldsymbol{x}$ is pointing in the same direction as $\partial g_1/\partial \boldsymbol{x}$.

4.2. Theorem on the bordered Hessian

The analogue in this case of classical programming to the Theorem on Second-Order Conditions for unconstrained problems is provided by the Theorem on the Bordered Hessian. According to this theorem the Hessian matrix of second-order partial derivatives of the Lagrangian function with respect to the instruments,

$$\frac{\partial^2 L}{\partial \boldsymbol{x}^2}=\begin{bmatrix} \dfrac{\partial^2 L}{\partial x_1^2} & \dfrac{\partial^2 L}{\partial x_1 \partial x_2} & \cdots & \dfrac{\partial^2 L}{\partial x_1 \partial x_n} \\ \vdots & & & \\ \dfrac{\partial^2 L}{\partial x_n \partial x_1} & & \cdots\cdots & \dfrac{\partial^2 L}{\partial x_n^2} \end{bmatrix}, \tag{4.14}$$

must be negative semidefinite when evaluated at the local maximum point (x^*, y^*) subject to the n conditions

$$dg = \frac{\partial g}{\partial x}(x^*)dx = 0. \tag{4.15}$$

4.3. *Theorem on sufficient conditions for classical programming*

The final analogue is that to the Theorem on Sufficient Conditions. According to the Theorem on Sufficient Conditions for Classical Programming, if the $n+m$ first-order conditions (4.7) and (4.8) are satisfied at x^*, then the *strengthened* bordered Hessian conditions, which state that the Hessian matrix in (4.14) is negative definite when evaluated at the point (x^*, y^*) subject to the n conditions in (4.15), imply that x^* is a local maximum for $F(x)$ subject to the m constraints.

Equivalently, the condition requires that the *bordered Hessian*, defined as the Hessian of $L(x, y)$ with respect to all variables,

$$\left[\begin{array}{c|c} \mathbf{0} & \dfrac{\partial g}{\partial x} \\ \hline \dfrac{\partial g'}{\partial x} & \dfrac{\partial^2 L}{\partial x^2} \end{array}\right] = \left[\begin{array}{c|c} 0\cdots 0 & \dfrac{\partial g_1}{\partial x_1}\cdots\dfrac{\partial g_1}{\partial x_n} \\ \vdots\ \ \vdots & \vdots\ \ \vdots \\ 0\cdots 0 & \dfrac{\partial g_m}{\partial x_1}\cdots\dfrac{\partial g_m}{\partial x_n} \\ \hline \dfrac{\partial g_1}{\partial x_1}\cdots\dfrac{\partial g_m}{\partial x_1} & \dfrac{\partial^2 L}{\partial x_1^2}\cdots\dfrac{\partial^2 L}{\partial x_1 \partial x_n} \\ \vdots\ \ \vdots & \vdots\ \ \vdots \\ \dfrac{\partial g_1}{\partial x_n}\cdots\dfrac{\partial g_m}{\partial x_n} & \dfrac{\partial^2 L}{\partial x_n \partial x_1}\cdots\dfrac{\partial^2 L}{\partial x_n^2} \end{array}\right], \tag{4.16}$$

where $\partial g/\partial x$ is the Jacobian matrix of (4.6), satisfy the $n-m$ conditions that the last $n-m$ leading principal minors alternate in sign, the sign of the first being $(-1)^{m+1}$. Note that both this theorem and the previous one reduce to the corresponding theorems in the unconstrained case when $m=0$.

4.4. *An example: The quadratic-linear problem*

An example of the classical programming problem, which follows that of Section 3.4, is the *quadratic-linear problem*,

$$\max_x F(x) = cx + \frac{1}{2}x'Qx \quad \text{subject to} \quad Ax = b. \tag{4.17}$$

Here the objective function is the same as that in (3.5), and the constraints are the m linear equalities,

$$\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b} \quad \text{i.e.} \quad \sum_{j=1}^{n} a_{ij}x_j=b_i, \quad i=1,2,\ldots,m, \tag{4.18}$$

determined by the $m\times n$ matrix $\boldsymbol{A}$ and the $m\times 1$ column vector $\boldsymbol{b}$. The Lagrangian is

$$\boldsymbol{L}(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{c}\boldsymbol{x}+\frac{1}{2}\boldsymbol{x}'\boldsymbol{Q}\boldsymbol{x}+\boldsymbol{y}(\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}), \tag{4.19}$$

where $\boldsymbol{y}$ is the vector of Lagrange multipliers. Using the $n+m$ first-order conditions (4.7), (4.8),

$$\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{x}}=\boldsymbol{c}+\boldsymbol{x}^{*\prime}\boldsymbol{Q}-\boldsymbol{y}^{*}\boldsymbol{A}=\boldsymbol{0}, \tag{4.20}$$

$$\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{y}}=\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}^{*}=\boldsymbol{0}. \tag{4.21}$$

These $n+m$ conditions require that

$$\boldsymbol{x}^{*}=-\boldsymbol{Q}^{-1}(\boldsymbol{c}'-\boldsymbol{A}'\boldsymbol{y}^{*\prime}). \tag{4.22}$$

The Lagrange multiplier can be identified by multiplying by $\boldsymbol{A}$ and using the constraint

$$\boldsymbol{A}\boldsymbol{x}^{*}=-\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{c}'+(\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}')\boldsymbol{y}^{*\prime}=\boldsymbol{b}. \tag{4.23}$$

Thus, solving for the vector of Lagrange multipliers,

$$\boldsymbol{y}^{*}=(\boldsymbol{b}'+\boldsymbol{c}\boldsymbol{Q}^{-1}\boldsymbol{A}')(\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}')^{-1}, \tag{4.24}$$

and inserting this result in (4.22),

$$\boldsymbol{x}^{*}=-\boldsymbol{Q}^{-1}\left[\boldsymbol{c}'-\boldsymbol{A}'(\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}')^{-1}(\boldsymbol{b}+\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{c}')\right]. \tag{4.25}$$

Letting $\bar{\boldsymbol{x}}^{*}$ be the solution to the unconstrained problem in (3.1) as given by (3.7), the solution to the constrained problem can be written

$$\boldsymbol{x}^{*}=\bar{\boldsymbol{x}}^{*}+\boldsymbol{Q}^{-1}\boldsymbol{A}'(\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}')^{-1}(\boldsymbol{b}-\boldsymbol{A}\bar{\boldsymbol{x}}^{*}). \tag{4.26}$$

Thus if $\bar{\boldsymbol{x}}^{*}$ satisfies the constraints then it also solves the constrained problem. Furthermore the difference between the constrained and unconstrained solutions, $\boldsymbol{x}^{*}-\bar{\boldsymbol{x}}^{*}$ is a linear function of amounts by which the unconstrained solution fail to satisfy the constraints $\boldsymbol{b}-\boldsymbol{A}\bar{\boldsymbol{x}}^{*}$.

5. Nonlinear programming: Kuhn–Tucker conditions[20]

The *problem of nonlinear programming* is that of choosing non-negative values of n variables so as to maximize a function of those variables subject to m inequality constraints,

$$\max_{\boldsymbol{x}} F(\boldsymbol{x}) \quad \text{subject to} \quad \boldsymbol{g}(\boldsymbol{x}) \leqslant \boldsymbol{b}, \quad \boldsymbol{x} \geqslant \boldsymbol{0}. \tag{5.1}$$

Here the vector of *instruments* $\boldsymbol{x}$ and the *objective function* $F(\boldsymbol{x})$ are as in (2.1), where $F(\boldsymbol{x})$ is a real-valued continuously differentiable function defined on E^n. The vector-valued *constraint function* $\boldsymbol{g}(\boldsymbol{x})$ and *constraint vector* $\boldsymbol{b}$ are as in (3.1), where $\boldsymbol{g}(\boldsymbol{x})$ is a continuously differentiable mapping from E^n into E^m. In terms of the basic problem (2.1), the nonlinear programming problem corresponds to the case in which the opportunity set can be written[21]

$$\begin{aligned} X &= \{\boldsymbol{x} \in E^n \mid \boldsymbol{g}(\boldsymbol{x}) \leqslant \boldsymbol{b}, \boldsymbol{x} \geqslant \boldsymbol{0}\} \\ &= \{(x_1, x_2, \ldots, x_n)' \mid g_i(x_1, x_2, \ldots, x_n) \leqslant b_i, \quad i = 1,2,\ldots,m, \\ &\qquad x_j \geqslant 0, \quad j = 1,2,\ldots,n\}. \end{aligned} \tag{5.2}$$

This problem is a generalization of the classical programming problem (4.1) since equality constraints are a special case of inequality constraints.[22]

5.1. *Theorem on Kuhn–Tucker conditions*

A characterization of the solution to the problem of nonlinear programming that is analogous to both the Theorem on First-Order Conditions for unconstrained problems and the Theorem on Lagrange Multipliers for classical programming programs is provided by the Theorem on Kuhn–Tucker Conditions. As in the case of classical programming, by introducing a row vector of m additional new variables, called *Lagrange multipliers*,

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_m), \tag{5.3}$$

[20] Basic references on nonlinear programming include Fiacco and McCormick (1968), Hadley (1969), Mangasarian (1969), Luenberger (1969, 1973), Intriligator (1971), Hestenes (1975), and Avriel (1976). The classic paper on the subject is Kuhn and Tucker (1951), but the basic ideas had been developed much earlier in conjunction with the development of the calculus of variations. See Hestenes (1966, 1975). For a discussion of computational algorithms for numerical solutions to nonlinear programming problems, see Zangwill (1969), Polak (1971), Avriel (1976), and Bazaraa and Shetty (1979).

[21] The inequality constraints $\boldsymbol{g}(\boldsymbol{x}) \leqslant \boldsymbol{b}$ mean that each component of $\boldsymbol{g}(\boldsymbol{x})$ is no more than the corresponding component of $\boldsymbol{b}$. Similarly the non-negativity constraints $\boldsymbol{x} \geqslant \boldsymbol{0}$ mean that each component of $\boldsymbol{x}$ is non-negative.

[22] For example, the equality constraint $x_1 + 6x_2 = 5$ can be written as the two inequality constraints $x_1 + 6x_2 \leqslant 5$ and $-x_1 - 6x_2 \leqslant -5$.

one for each inequality constraint, the *Lagrangian function* can be defined as the following real-valued function of the n original and the m added variables:

$$\begin{aligned} L(\boldsymbol{x}, \boldsymbol{y}) &= F(\boldsymbol{x}) + \boldsymbol{y}(\boldsymbol{b} - \boldsymbol{g}(\boldsymbol{x})) \\ &= F(x_1, x_2, \ldots, x_n) + \sum_{i=1}^{m} y_i(b_i - g_i(x_1, x_2, \ldots, x_n)), \end{aligned} \tag{5.4}$$

as in (4.5). The *Kuhn–Tucker conditions* are then defined at the point $\boldsymbol{x}^*$, $\boldsymbol{y}^*$ as the $2n+2m$ inequalities and 2 equalities,

$$\begin{aligned} &\frac{\partial L}{\partial \boldsymbol{x}}(\boldsymbol{x}^*, \boldsymbol{y}^*) \leqslant 0, && \frac{\partial L}{\partial \boldsymbol{y}}(\boldsymbol{x}^*, \boldsymbol{y}^*) \geqslant \boldsymbol{0} && (n+m \text{ conditions}), \\ &\frac{\partial L}{\partial \boldsymbol{x}}(\boldsymbol{x}^*, \boldsymbol{y}^*)\boldsymbol{x}^* = 0, && \boldsymbol{y}^* \frac{\partial L}{\partial \boldsymbol{y}}(\boldsymbol{x}^*, \boldsymbol{y}^*) = 0 && (2 \text{ conditions}), \\ &\boldsymbol{x}^* \geqslant \boldsymbol{0}, && \boldsymbol{y}^* \geqslant \boldsymbol{0} && (n+m \text{ conditions}). \end{aligned} \tag{5.5}$$

Of the inequalities, $n+m$ represent the constraints of the original problem

$$\frac{\partial L}{\partial \boldsymbol{y}}(\boldsymbol{x}^*, \boldsymbol{y}^*) = \boldsymbol{b} - \boldsymbol{g}(\boldsymbol{x}^*) \geqslant \boldsymbol{0} \qquad (m \text{ conditions}), \tag{5.6}$$

$$\boldsymbol{x}^* \geqslant \boldsymbol{0} \qquad (n \text{ conditions}), \tag{5.7}$$

while the added $n+m$ inequalities require that

$$\frac{\partial L}{\partial \boldsymbol{x}}(\boldsymbol{x}^*, \boldsymbol{y}^*) = \frac{\partial F}{\partial \boldsymbol{x}}(\boldsymbol{x}^*) - \boldsymbol{y}^* \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}(\boldsymbol{x}^*) \leqslant \boldsymbol{0} \qquad (n \text{ conditions}), \tag{5.8}$$

$$\boldsymbol{y}^* \geqslant \boldsymbol{0} \qquad (m \text{ conditions}). \tag{5.9}$$

The n conditions in (5.8) are written as inequalities rather than equalities [as in (4.7)] because of the non-negativity restrictions on $\boldsymbol{x}$ (5.7), or, more generally, because boundary solutions are permitted. The m conditions in (5.9), requiring that the Lagrange multipliers be non-negative, stem from the fact that the constraints in (5.6) are written as inequalities rather than as equalities: If a constraint is an equality then the corresponding element of $\boldsymbol{y}^*$ is unrestricted, as in the classical programming case.

The two equality Kuhn–Tucker conditions

$$\frac{\partial L}{\partial \boldsymbol{x}}(\boldsymbol{x}^*, \boldsymbol{y}^*)\boldsymbol{x}^* = \sum_{j=1}^{n} \left(\frac{\partial F}{\partial x_j}(\boldsymbol{x}^*) - \boldsymbol{y}^* \frac{\partial \boldsymbol{g}}{\partial x_j}(\boldsymbol{x}^*) \right) x_j^* = 0, \tag{5.10}$$

$$\boldsymbol{y}^* \frac{\partial L}{\partial \boldsymbol{y}}(\boldsymbol{x}^*, \boldsymbol{y}^*) = \sum_{i=1}^{m} y_i^*(b_i - g_i(x^*)) = 0, \tag{5.11}$$

together with the other conditions, require that every term in both of these sums vanish. Thus when one of the inequality conditions is satisfied at the solution as a strict inequality then the corresponding (dual) variable vanishes,

$$\frac{\partial F}{\partial x_j}(\boldsymbol{x}^*) - \boldsymbol{y}^* \frac{\partial \boldsymbol{g}}{\partial x_j}(\boldsymbol{x}^*) < 0 \quad \text{implies} \quad x_j^* = 0, \quad j = 1,2,\ldots,n, \tag{5.12}$$

$$g_i(\boldsymbol{x}^*) < b_i \quad \text{implies} \quad y_i^* = 0, \quad i = 1,2,\ldots,m. \tag{5.13}$$

These conditions are known as the *complementary slackness conditions of nonlinear programming*. Condition (5.11) also implies that at the solution the value of the Lagrangian is the maximized value of the objective function

$$L(\boldsymbol{x}^*, \boldsymbol{y}^*) = F(\boldsymbol{x}^*) = F^*. \tag{5.14}$$

According to the Theorem on Kuhn–Tucker Conditions, if a suitably strong constraint qualification is satisfied, then the Kuhn–Tucker conditions (5.5) [or, equivalently, (5.6)–(5.11)] are necessary conditions for the nonlinear programming problem, so if $\boldsymbol{x}^*$ solves (5.1) then there exists a vector of Lagrange multipliers $\boldsymbol{y}^*$ satisfying (5.5).[23]

Just as in the case of classical programming, the solutions for the Lagrange multipliers have an interpretation as the sensitivities of the maximized value of the objective function to changes in the constraint constants,[24]

$$\boldsymbol{y}^* = \frac{\partial F^*}{\partial \boldsymbol{b}} \quad \text{i.e.} \quad y_i^* = \frac{\partial F^*}{\partial b_i}, \quad i = 1,2,\ldots,m, \tag{5.15}$$

[23]There are, in fact, many alternative forms of the constraint qualification condition. One is the same as that for classical programming, requiring that $m < n$ (which generally need not be the case in nonlinear programming) and that the (linearized) constraints be linearly independent, as in (4.6). A second, the *Slater constraint qualification* requires that there be a point $\boldsymbol{x}^0 \geq \boldsymbol{0}$ for which $\boldsymbol{g}(\boldsymbol{x}^0) < \boldsymbol{b}$, that is, a point at which all inequality constraints are satisfied as strict inequalities. For a discussion of these and other constraint qualification conditions, see Arrow, Hurwicz, and Uzawa (1958, 1961); Mangasarian (1969); Canon, Cullum, and Polak (1970); Bazaraa, Goode, and Shetty (1972); Peterson (1973); and Bazaraa and Shetty (1976). If the problem does not satisfy the constraint qualification condition then it is necessary to add another Lagrange multiplier y_0 to the objective function, so (5.4) becomes

$$L(\boldsymbol{x}, \boldsymbol{y}, y_0) = y_0 F(\boldsymbol{x}) + \boldsymbol{y}(\boldsymbol{b} - \boldsymbol{g}(\boldsymbol{x})).$$

In fact, one of the main advances of Kuhn and Tucker (1951) over earlier developments, such as John (1948), was the formulation of the constraint qualification condition, so that y_0 is guaranteed to be positive and hence can be taken to be 1 (by dividing all $m+1$ Lagrange multipliers by y_0). It is important that y_0 be non-zero for the solution not to be independent of the objective function $F(\boldsymbol{x})$.

[24]To be more precise, given concavity and differentiability of $F(\boldsymbol{x})$, convexity and differentiability of $\boldsymbol{g}(\boldsymbol{x})$, and Slater's constraint qualification condition, it follows that

$$\partial F^*/\partial b_i|_+ \leq y_i^* \leq \partial F^*/\partial b_i|_-,$$

that is, that the Lagrange multipliers are bounded by the left and right partial derivatives of the maximized objective function with respect to the constraint constants.

where F^* is defined as

$$F^* = F(x^*) = L(x^*, y^*). \tag{5.16}$$

In particular, from the complementarity slackness conditions (5.13) if a constraint is met as a strict inequality at the solution then the corresponding Lagrange multiplier is zero, so increasing the constraint constant by a suitably small amount will not change the maximized value of the objective function.

5.2. Kuhn–Tucker saddle point theorem

A theorem that is analogous to both the Theorem on Sufficient Conditions for unconstrained problems and the Theorem on Sufficient Conditions for classical programming problems is provided by the Kuhn–Tucker Saddle Point Theorem. Given the Lagrangian function defined in (5.4), the *saddle point problem* is defined as

$$\max_x \min_y L(x, y) \quad \text{subject to} \quad x \geqslant \mathbf{0}, \quad y \geqslant \mathbf{0}. \tag{5.17}$$

Thus x^*, y^* solves the saddle point problem if and only if, for all $x \geqslant \mathbf{0}$, $y \geqslant \mathbf{0}$,

$$L(x, y^*) \leqslant L(x^*, y^*) \leqslant L(x^*, y). \tag{5.18}$$

According to the Kuhn–Tucker Saddle Point Theorem, a sufficient condition for x^* to solve the nonlinear programming problem (5.1) is that there exist a y^* such that x^*, y^* solves the saddle point problem (5.17). Thus if x^*, y^* satisfy the saddle-point conditions in (5.18) then x^* solves the nonlinear programming problem. While this part of the theorem does not require any convexity or constraint qualification assumptions, the converse of the theorem does require such assumptions. According to this second part of the theorem, if x^* solves the nonlinear programming problem and it is assumed both that a suitable constraint qualification condition is met and that the problem is one of *concave programming* in which $F(x)$ is a concave function and each constraint function $g_i(x)$ is a convex function, then there exists a non-zero vector y^* such that x^*, y^* solves the saddle point problem.[25] Thus under these assumptions the two

[25] For the definition of a concave function, see footnote 9. A convex function is defined similarly except the inequality is reversed. Thus $g_i(x)$ is a *convex function* if and only if, given any $x^1, x^2 \in X$, $g_i(\alpha x^1 + (1-\alpha)x^2) \leqslant \alpha g_i(x^1) + (1-\alpha) g_i(x^2)$ for all α, $0 \leqslant \alpha \leqslant 1$. Equivalently, $g_i(x)$ is a convex function if and only if $-g_i(x)$ is a concave function. If all constraint functions $g(x)$ are convex then the opportunity set is convex, so the problem becomes one of maximizing a concave function over a convex set. Thus by the Local–Global Theorem of Section 2 all local solutions are global solutions and the set of solutions is convex.

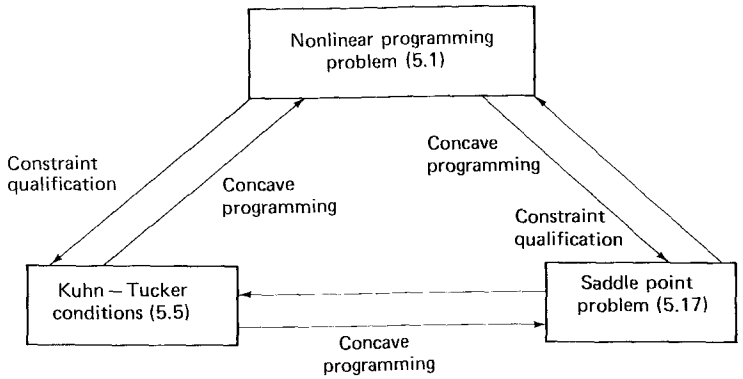


Figure 5.1. Relations between the nonlinear programming problem, the Kuhn–Tucker conditions, and the saddle point problem.

problems are equivalent. It should be noted that neither part of this saddle point characterization theorem requires the assumption of differentiability of $F(x)$ or $g(x)$. Given differentiability, however, if the problem is one of concave programming, then the Kuhn–Tucker conditions are sufficient conditions in that if x^*, y^* satisfies (5.5), then x^* solves (5.1).[26] Thus, for a concave programming problem in which a suitable constraint qualification condition is met the Kuhn–Tucker conditions are both necessary and sufficient for x^* to solve the nonlinear programming problem. The relations between the nonlinear programming problem, the saddle point problem, and the Kuhn–Tucker conditions are shown in Figure 5.1.[27] For example, if the problem is one of concave programming, then, assuming x^*, y^* satisfy the Kuhn–Tucker conditions, x^*, y^* also solve the saddle point problem and x^* solves the nonlinear programming problem. If, in addition, a suitable constraint qualification condition is met then all three problems are equivalent.

As in the case of classical programming, a geometric interpretation can be given for the nonlinear programming problem and its solution via the two

[26]More generally, the Kuhn–Tucker conditions are sufficient if $F(x)$ is concave and each of the $g_i(x)$ is quasiconvex. See Arrow and Enthoven (1961) and Mangasarian (1969). The function $g_i(x)$ is a *quasiconvex function* if and only if, given any $x^1, x^2 \in X$, where $g_i(x^1) \leq g_i(x^2)$,

$$g_i(\alpha x^1 + (1-\alpha)x^2) \leq g_i(x^2) \quad \text{for all } \alpha, \quad 0 < \alpha < 1.$$

Equivalently $g_i(x)$ is quasiconvex if and only if $-g_i(x)$ is quasiconcave, where a quasiconcave function is defined in footnote 10. This extension to quasiconvex constraint functions is a generalization since a function which is convex (concave) is also quasiconvex (quasiconcave), but not vice-versa. The Kuhn–Tucker conditions are also sufficient if $F(x)$ is twice differentiable and quasiconcave, each of the $g_i(x)$ is quasiconvex and, in addition, $(\partial F/\partial x)(x^*) \neq \mathbf{0}$.

[27]See Mangasarian (1969, p. 110) for a generalization of Figure 5.1.

Kuhn–Tucker theorems. From the Kuhn–Tucker conditions (5.8) and (5.9), at an interior solution, where all $\boldsymbol{x}^* > \mathbf{0}$ (or if non-negativity of the x's is not part of the problem), conditions (5.8) and (5.9), if all $\boldsymbol{x}^* > \mathbf{0}$ (or if the non-negativity of the x's is not part of the problem),

$$\frac{\partial F}{\partial \boldsymbol{x}}(\boldsymbol{x}^*) = \boldsymbol{y}^* \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}, \qquad \boldsymbol{y}^* \geqslant \mathbf{0}. \tag{5.19}$$

Thus the gradient of the objective function must, at the solution, be a non-negative weighted combination of the gradients of the constraint function. The gradient vector of the objective function must therefore lie within the cone spanned by the outward pointing normals to the opportunity set at $\boldsymbol{x}^*$. This solution is illustrated in Figure 5.2 for the problem in which $n=2$, $m=3$. The gradient vector $\partial F/\partial \boldsymbol{x}$ is orthogonal to the contours of $F(\boldsymbol{x})$, as in Figure 4.1, and the gradient vectors for the constraint functions are the outward pointing normals. At the solution shown the gradient of the objective function, $\partial F/\partial \boldsymbol{x}$, lies within the cone spanned by $\partial g_1/\partial \boldsymbol{x}$ and $\partial g_2/\partial \boldsymbol{x}$, the outward pointing normals for the constraints that are satisfied as equalities.

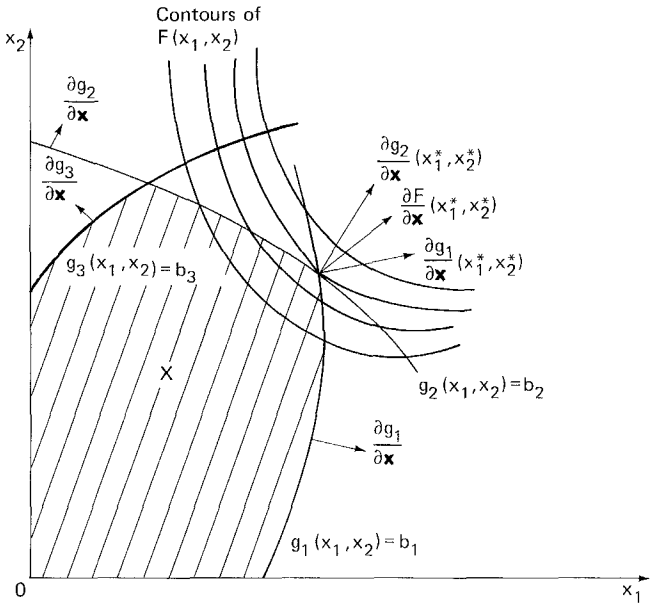


Figure 5.2. Solution to the nonlinear programming problem: max $F(x_1, x_2)$ subject to $g_1(x_1, x_2) \leqslant b_1$, $g_2(x_1, x_2) \leqslant b_2$, $g_3(x_1, x_2) \leqslant b_3$, $x_1 \geqslant 0$, $x_2 \geqslant 0$.

5.3. *An example: The quadratic programming problem*

An example of the nonlinear programming problem is the *quadratic programming problem* [as in (4.17), where the constraint is of the form of a set of inequalities],

$$\max_{x} F(x) = cx + \frac{1}{2}x'Qx \quad \text{subject to} \quad Ax \leqslant b, \quad x \geqslant 0. \tag{5.20}$$

Here c is a given $1 \times n$ row vector, Q is a given $n \times n$ negative semidefinite symmetric matrix, A is a given $m \times n$ matrix, and b is a given $m \times 1$ column vector. The Lagrangian is given in (4.19) and the Kuhn–Tucker conditions are

$$\begin{aligned} &\frac{\partial L}{\partial x} = c + x^{*\prime}Q - y^{*}A \leqslant 0, && \frac{\partial L}{\partial y} = b - Ax^{*} \geqslant 0, \\ &\frac{\partial L}{\partial x}x^{*} = (c + x^{*\prime}Q - y^{*}A)x^{*} = 0, && y^{*}\frac{\partial L}{\partial y} = y^{*}(b - Ax^{*}) = 0, \\ &x^{*} \geqslant 0, && y^{*} \geqslant 0. \end{aligned} \tag{5.21}$$

These conditions characterize the solution to the problem. Because Q is negative semidefinite, the objective function $F(x)$ is concave and the linear transformation Ax is convex. Furthermore the constraint qualification condition is met. The problem is therefore one of concave programming, in which the Kuhn–Tucker conditions (5.21) are both necessary and sufficient. The vector x^{*} thus solves the quadratic programming problem (5.20) if and only if there is a y^{*} such that x^{*}, y^{*} satisfy the Kuhn–Tucker conditions (5.21).

6. Linear programming[28]

The *problem of linear programming* is that of choosing non-negative values of n variables so as to maximize a linear form in these variables subject to m linear inequality constraints,

$$\max_{x} cx \quad \text{subject to} \quad Ax \leqslant b, \quad x \geqslant 0. \tag{6.1}$$

Here the vector of instruments x is as in (2.1), (3.1), and (4.1); A is a given $m \times n$ matrix (a_{ij}); b is a given column vector of m elements, as in (4.1) and (5.1); and c is a given row vector of n elements. In terms of the nonlinear programming

[28] Basic references on linear programming include Dorfman, Samuelson, and Solow (1958); Gale (1960); Hadley (1963); Dantzig (1963); Simmonard (1966); Intriligator (1971); Luenberger (1973); and Gass (1975).

problem (5.1) the linear problem corresponds to the case in which the objective function is the linear form

$$F(\boldsymbol{x})=\boldsymbol{c}\boldsymbol{x}=\sum_{j=1}^{n} c_j x_j, \tag{6.2}$$

and each of the constraint functions is also a linear form

$$\boldsymbol{g}(\boldsymbol{x})=\boldsymbol{A}\boldsymbol{x} \quad \text{i.e.} \quad g_i(x_1, x_2, \ldots, x_n)=\sum_{j=1}^{n} a_{ij}x_j, \quad i=1,2,\ldots,m. \tag{6.3}$$

The problem is then a special case of the nonlinear programming problem that is doubly linear in that it is linear both in the objective function and in the constraint functions.[29] Since a linear form is both a concave and a convex function, the problem, considered a special case of that of nonlinear programming, is equivalent to the saddle point problem

$$\max_{\boldsymbol{x}} \min_{\boldsymbol{y}} L(\boldsymbol{x}, \boldsymbol{y})=\boldsymbol{c}\boldsymbol{x}+\boldsymbol{y}(\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}) \quad \text{subject to} \quad \boldsymbol{x}\geqslant\boldsymbol{0}, \quad \boldsymbol{y}\geqslant\boldsymbol{0}. \tag{6.4}$$

Associated with every linear programming problem is a dual problem. If the primal problem is given as in (6.1) the *dual problem* is

$$\min_{\boldsymbol{y}} \boldsymbol{y}\boldsymbol{b} \quad \text{subject to} \quad \boldsymbol{y}\boldsymbol{A}\geqslant\boldsymbol{c}, \quad \boldsymbol{y}\geqslant\boldsymbol{0}. \tag{6.5}$$

This problem is also of finding an extremum of a linear form subject to a set of linear inequality constraints by choice of non-negative values of variables. The variables of the dual problem, $\boldsymbol{y}$, are the Lagrange multipliers of the original (primal) problem. The dual to the dual is the primal problem, since the dual to a minimization problem is one of maximization, in the dual problem the constraint constants become the coefficients of the objective function while the coefficients of the objective function become the constraint constants, and in the dual problem the coefficients postmultiply rather than premultiply both the coefficient vector of the objective function and the coefficient matrix of the constraint functions. The saddle point problem for the dual problem is

$$\min_{\boldsymbol{y}} \max_{\boldsymbol{x}} L(\boldsymbol{y}, \boldsymbol{x})=\boldsymbol{y}\boldsymbol{b}+(\boldsymbol{c}-\boldsymbol{y}\boldsymbol{A})\boldsymbol{x} \quad \text{subject to} \quad \boldsymbol{y}\geqslant\boldsymbol{0}, \boldsymbol{x}\geqslant\boldsymbol{0}, \tag{6.6}$$

so the Lagrangian function is the same for both primal and dual

$$L(\boldsymbol{x}, \boldsymbol{y})=L(\boldsymbol{y}, \boldsymbol{x})=\boldsymbol{c}\boldsymbol{x}+\boldsymbol{y}\boldsymbol{b}-\boldsymbol{y}\boldsymbol{A}\boldsymbol{x}. \tag{6.7}$$

[29] In terms of the quadratic programming problem (5.20), linear programming is the special case in which the matrix $\boldsymbol{Q}$ vanishes.

The Kuhn–Tucker conditions, which are the same for both primal and dual problems, are

$$\begin{aligned} &\frac{\partial L}{\partial x} = c - y^* A \leqslant 0, && \frac{\partial L}{\partial y} = b - Ax^* \geqslant 0, \\ &\frac{\partial L}{\partial x} x^* = (c - y^* A)x^* = 0, && y^* \frac{\partial L}{\partial y} = y^*(b - Ax^*) = 0, \\ &x^* \geqslant 0, && y^* \geqslant 0. \end{aligned} \tag{6.8}$$

The three major theorems of linear programming — the existence theorem, the duality theorem, and the complementary slackness theorem — can be proved on the basis of these conditions.

6.1. *Existence theorem*

According to the Existence Theorem, if feasible points exist for both primal and dual problems, then solutions exist for both. Thus if there exist x^0, y^0 such that

$$Ax^0 \leqslant b, \quad x^0 \geqslant 0, \qquad y^0 A \geqslant c, \quad y^0 \geqslant 0, \tag{6.9}$$

then there exist x^*, y^* solving both primal and dual problems.

6.2. *Duality theorem*

According to the Duality Theorem, for any feasible vectors for both primal and dual problems x^0, y^0 it follows that

$$cx^0 \leqslant y^0 b. \tag{6.10}$$

Furthermore, feasible vectors that satisfy this inequality as an equality provide solutions x^*, y^* to the dual problems where

$$cx^* = y^* b. \tag{6.11}$$

6.3. *Complementary slackness theorem*

According to the Complementary Slackness Theorem, x^* and y^*, which are feasible vectors for the dual problems, solve these problems if and only if they satisfy the two equality conditions of the Kuhn–Tucker conditions (6.8), given as

$$(c - y^* A)x^* = 0, \qquad y^*(b - Ax^*) = 0. \tag{6.12}$$

From these conditions the optimized values of the dual objective functions are equal to one another and also to the values of both Lagrangian functions at the solution

$$\boldsymbol{cx}^* = \boldsymbol{y}^*\boldsymbol{Ax}^* = \boldsymbol{y}^*\boldsymbol{b} = L(\boldsymbol{x}^*, \boldsymbol{y}^*) = L(\boldsymbol{y}^*, \boldsymbol{x}^*). \tag{6.13}$$

Together with the other Kuhn–Tucker conditions, the conditions in (6.12) imply that when any one of the inequality constraints is satisfied in the solution as a strict inequality, then the corresponding dual variable vanishes, i.e.,

$$\begin{aligned} &(c_j - \Sigma y_i^* a_{ij}) < 0 \quad \text{implies} \quad x_j^* = 0, \quad j = 1, 2, \ldots, n, \\ &(b_i - \Sigma a_{ij} x_j^*) > 0 \quad \text{implies} \quad y_i^* = 0, \quad i = 1, 2, \ldots, m. \end{aligned} \tag{6.14}$$

These conditions are known as the *complementary slackness conditions of linear programming.*

As in the last two sections, a geometric interpretation can be given for the linear programming problem and its solution. The opportunity set is a polyhedral closed convex set since it is the intersection of $m+n$ half spaces defined by the m inequality and n non-negativity constraints. The contours of the

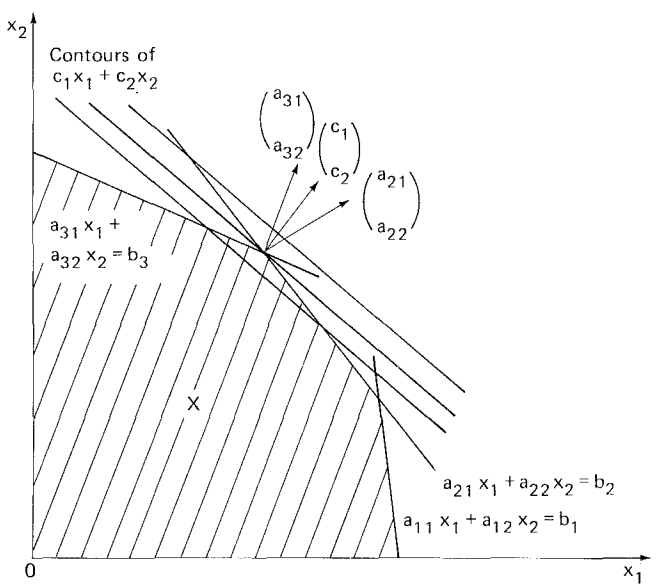


Figure 6.1. Solution to the linear programming problem: max $c_1x_1+c_2x_2$ subject to $a_{11}x_1+a_{12}x_2 \leqslant b_1$, $a_{21}x_1+a_{22}x_2 \leqslant b_2$, $a_{31}x_1+a_{32}x_2 \leqslant b_3 (x_1 \geqslant 0, x_2 \geqslant 0)$.

objective function are hyperplanes, and the problem is solved on the highest hyperplane within the polyhedral set. This solution cannot be at an interior point. A solution must occur at a vertex (in which case it is unique) or along a bounding face (in which case it is non-unique). A vertex solution is illustrated in Figure 6.1 for the problem in which $n=2$ and $m=3$. As in the nonlinear programming problem the solution occurs at a point where the gradient vector of the objective function [here the constant vector $(c_1, c_2)'$] lies in the cone spanned by the outward pointing normals to the opportunity set [here $(a_{21}, a_{22})'$ and $(a_{31}, a_{32})'$].

7. Microeconomics: Mathematical programming and comparative statics

Microeconomic problems are typically formulated as those of economic agents (e.g. households, firms) attempting to maximize an objective function subject to certain constraints. They are therefore typically formulated as problems of mathematical programming. The theory of mathematical programming is then used to analyze these problems — specifically to characterize the equilibrium solution and to determine how the solution varies as the parameters of the problem change. The latter determination of how changes in the parameters influence the solution is called *comparative statics* since it compares two equilibrium situations — an initial equilibrium and an equilibrium after one or more of the parameters change.[30] The characterization of the solution is generally based on the first-order conditions of the mathematical programming problem, and the comparative static analysis of how the solution varies as the parameters change is based on differentiation of the first-order conditions. The resulting qualitative or quantitative determination of how parameters influence the solution yields certain restrictions on the solution. These restrictions make the theory operationally meaningful in that the restrictions could be refuted empirically.

7.1. Comparative statics theorem

Suppose the problem of a certain economic agent can be characterized as the choice of certain variables $\boldsymbol{x}$ as in the problem of classical programming (4.1) with a single constraint. The objective function and the constraint may each depend on a q-dimensional column vector of parameters $\boldsymbol{a}$, so the problem can

[30] For a discussion of comparative static analysis, see Samuelson (1947) and Kalman and Intriligator (1973).

be stated

$$\max_{\boldsymbol{x}} F(\boldsymbol{x}, \boldsymbol{a}) \quad \text{subject to} \quad g(\boldsymbol{x}, \boldsymbol{a}) = b. \tag{7.1}$$

The solution to this problem is characterized by the first-order conditions (4.7) and (4.8) which here are

$$b - g(\boldsymbol{x}, \boldsymbol{a}) = 0, \tag{7.2}$$

$$\frac{\partial F}{\partial \boldsymbol{x}}(\boldsymbol{x}, \boldsymbol{a}) - y \frac{\partial g}{\partial x}(\boldsymbol{x}, \boldsymbol{a}) = \mathbf{0}, \tag{7.3}$$

where y is the single Lagrange multiplier, corresponding to the single constraint. The solutions $\boldsymbol{x}^*$, y^* will generally depend on the $q+1$ parameters of the problem $(\boldsymbol{a}, b)$

$$\boldsymbol{x}^* = \boldsymbol{x}^*(\boldsymbol{a}, b), \tag{7.4}$$

$$y^* = y^*(\boldsymbol{a}, b). \tag{7.5}$$

Inserting these solutions in the first-order conditions yields the $n+1$ identities

$$b - g(\boldsymbol{x}(\boldsymbol{a}, b), \boldsymbol{a}) \equiv 0, \tag{7.6}$$

$$\frac{\partial F}{\partial \boldsymbol{x}}(\boldsymbol{x}(\boldsymbol{a}, b), \boldsymbol{a}) - y(\boldsymbol{a}, b) \frac{\partial g}{\partial \boldsymbol{x}}(\boldsymbol{x}(\boldsymbol{a}, b), \boldsymbol{a}) \equiv \mathbf{0}. \tag{7.7}$$

Assuming the functions $F(\boldsymbol{x})$ and $g(\boldsymbol{x})$ are continuously differentiable, these identities can be differentiated to obtain[31]

$$\mathrm{d}b - \frac{\partial g}{\partial \boldsymbol{x}} \mathrm{d}\boldsymbol{x} - \frac{\partial g}{\partial \boldsymbol{a}} \mathrm{d}\boldsymbol{a} = 0, \tag{7.8}$$

$$\frac{\partial^2 F}{\partial \boldsymbol{x}^2} \mathrm{d}\boldsymbol{x} + \frac{\partial^2 F}{\partial \boldsymbol{x} \partial \boldsymbol{a}} \mathrm{d}\boldsymbol{a} - \left(\frac{\partial g}{\partial \boldsymbol{x}} \right)' \mathrm{d}y - y \frac{\partial^2 g}{\partial \boldsymbol{x}^2} \mathrm{d}\boldsymbol{x} - y \frac{\partial^2 g}{\partial \boldsymbol{x} \partial \boldsymbol{a}} \mathrm{d}\boldsymbol{a} = \mathbf{0}, \tag{7.9}$$

[31] As in footnote 12, the derivative of a scalar with respect to a column vector is a row vector. Thus $\partial g / \partial \boldsymbol{x}$ and $\partial g / \partial \boldsymbol{a}$ are both row vectors. It should be noted that the principal mathematical theorem underlying comparative statics analysis is the Implicit Function Theorem, under which the rank condition on the Jacobian matrix guarantees that the solutions to a set of equations involving certain parameters are differentiable functions of these parameters. The existence of solutions to (7.6) and (7.7) is, however, not ensured by the Implicit Function Theorem.

where

$$\frac{\partial g}{\partial \boldsymbol{a}} = \left(\frac{\partial g}{\partial a_1}, \frac{\partial g}{\partial a_2}, \ldots, \frac{\partial g}{\partial a_q} \right), \tag{7.10}$$

$$\mathrm{d}\boldsymbol{x} = (\mathrm{d}x_1, \mathrm{d}x_2, \ldots, \mathrm{d}x_n)', \tag{7.11}$$

$$\mathrm{d}\boldsymbol{a} = (\mathrm{d}a_1, \mathrm{d}a_2, \ldots, \mathrm{d}a_q)'. \tag{7.12}$$

Solving for the changes in y and $\boldsymbol{x}$ yields, in matrix notation,

$$\begin{pmatrix} \mathrm{d}y \\ \mathrm{d}\boldsymbol{x} \end{pmatrix} = \begin{bmatrix} 0 & -\left(\frac{\partial g}{\partial \boldsymbol{x}}\right) \\ -\left(\frac{\partial g}{\partial \boldsymbol{x}}\right)' & \frac{\partial^2 L}{\partial \boldsymbol{x}^2} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial g}{\partial \boldsymbol{a}} \mathrm{d}\boldsymbol{a} - \mathrm{d}\boldsymbol{b} \\ -\left(\frac{\partial^2 L}{\partial \boldsymbol{x} \partial \boldsymbol{a}}\right) \mathrm{d}\boldsymbol{a} \end{bmatrix}, \tag{7.13}$$

where it is assumed that the bordered Hessian matrix to be inverted is non-singular.

Using this result, the Comparative Statics Theorem states that, assuming $F(\boldsymbol{x})$ and $g(\boldsymbol{x})$ are continuously differentiable, there is a feasible point, and the bordered Hessian matrix is non-singular there exists almost everywhere a *generalized Slutsky equation* of the form[32]

$$\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{a}} = \left(\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{a}} \right)_{\text{comp}} + \frac{1}{y} \left(\frac{\partial \boldsymbol{x}}{\partial b} \right) \left(\frac{\partial L}{\partial \boldsymbol{a}} \right). \tag{7.14}$$

Here "comp" refers to a compensated change in $\boldsymbol{a}$ under which b is adjusted so

[32] See Kalman and Intriligator (1973). This result follows from (7.13) using the result on inverting a partitioned matrix. In the special case of unconstrained maximization (7.13) implies that

$$\partial \boldsymbol{x}/\partial \boldsymbol{a} = -(\partial^2 F/\partial \boldsymbol{x}^2)^{-1} (\partial^2 F/\partial \boldsymbol{x} \partial \boldsymbol{a}).$$

If the equilibrium is characterized by the n equations,

$$\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{a}) = \boldsymbol{0},$$

then Samuelson (1947) proved that

$$\partial \boldsymbol{x}/\partial \boldsymbol{a} = -(\partial \boldsymbol{f}/\partial \boldsymbol{x})^{-1} (\partial \boldsymbol{f}/\partial \boldsymbol{a}).$$

In the case of maximizing without a constraint, however, the first-order conditions (3.2) are $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{a}) = (\partial F/\partial \boldsymbol{x})(\boldsymbol{x}, \boldsymbol{a}) = \boldsymbol{0}$, implying the same result as above for the unconstrained case.

that F is held constant. This generalized Slutsky equation can also be written

$$\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{a}} + \frac{\partial \boldsymbol{x}}{\partial b}\frac{\partial g}{\partial \boldsymbol{a}} = \left(\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{a}}\right)_{\text{comp}} + \frac{1}{y}\frac{\partial \boldsymbol{x}}{\partial b}\frac{\partial F}{\partial \boldsymbol{a}} = S(\boldsymbol{a}, b). \tag{7.15}$$

Here the terms on the left are the "observables", the changes in the choice variables with respect to the $q+1$ parameters, the change with respect to b weighted by the change in g with respect to $\boldsymbol{a}$. The terms on the right are the "non-observables", the first being the matrix of compensated changes and the second being non-observable if the objective function is unique only up to a monotonic transformation. The $n \times q$ matrix on the right, $S(\boldsymbol{a}, b)$, is the *generalized matrix of substitution effects*. A second part of the theorem states that if $q=n$, so $S(\boldsymbol{a}, b)$ is square, then it is symmetric if and only if both the objective function $F(\boldsymbol{x}, \boldsymbol{a})$ and the constraint function $g(\boldsymbol{x}, \boldsymbol{a})$ can be written

$$F(\boldsymbol{x}, \boldsymbol{a}) = A_F \boldsymbol{a}'\boldsymbol{x} + \beta_F(\boldsymbol{x}) + \gamma_F(\boldsymbol{x}), \tag{7.16}$$

$$g(\boldsymbol{x}, \boldsymbol{a}) = A_g \boldsymbol{a}'\boldsymbol{x} + \beta_g(\boldsymbol{x}) + \gamma_g(\boldsymbol{x}), \tag{7.17}$$

where A_F and A_g are constants. Finally the quadratic form of $S(\boldsymbol{a}, b)$ is negative semidefinite if

$$A_F - yA_g \geqslant 0. \tag{7.18}$$

8. Neoclassical theory of the household[33]

The household and the firm are the two most important microeconomic agents. As an economic agent, the household is assumed to behave so as to maximize utility subject to a budget constraint. Assuming there are n goods (and services) available, let $\boldsymbol{x}$ be the column vector of the goods purchased and consumed by the household,

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_n)'; \tag{8.1}$$

$U(\boldsymbol{x})$ be the utility function of the household,

$$U(\boldsymbol{x}) = U(x_1, x_2, \ldots, x_n), \tag{8.2}$$

giving utility as a function of consumption levels; $\boldsymbol{p}$ be the row vector of

[33] Basic references on the neoclassical theory of the household include Hicks (1946), Samuelson (1947), Wold and Jureen (1953), Intriligator (1971), and Phlips (1974). See also Chapter 9 by Barton and Böhm. For a presentation of the duality approach to the theory of the household, see Chapter 12 by Diewert.

(positive) given prices of the goods,

$$\boldsymbol{p}=(p_1, p_2, \ldots, p_n); \tag{8.3}$$

and I be the (positive) given income available to the household. The problem of the household is then

$$\max_{\boldsymbol{x}} U(\boldsymbol{x}) \quad \text{subject to} \quad \boldsymbol{px} \leqslant I, \quad \boldsymbol{x} \geqslant \boldsymbol{0}. \tag{8.4}$$

Thus the household chooses nonnegative amounts of goods $\boldsymbol{x}$ so as to maximize the utility function $U(\boldsymbol{x})$ subject to the budget constraint

$$\boldsymbol{px} = \sum_{j=1}^{n} p_j x_j \leqslant I, \tag{8.5}$$

which states that expenditure on all n goods cannot exceed income. This problem is one of nonlinear programming, so, following the approach of Section 5, introduce the Lagrange multiplier y and define the Lagrangian as

$$L(\boldsymbol{x}, y) = U(\boldsymbol{x}) + y(I - \boldsymbol{px}). \tag{8.6}$$

The Kuhn–Tucker conditions state that at the solution $\boldsymbol{x}^*$, y^*,

$$\begin{aligned} &\frac{\partial L}{\partial \boldsymbol{x}} = \frac{\partial U}{\partial \boldsymbol{x}} - y\boldsymbol{p} \leqslant \boldsymbol{0}, && \frac{\partial L}{\partial y} = I - \boldsymbol{px} \geqslant 0, \\ &\frac{\partial L}{\partial \boldsymbol{x}}\boldsymbol{x} = \left(\frac{\partial U}{\partial \boldsymbol{x}} - y\boldsymbol{p}\right)\boldsymbol{x} = 0, && \frac{\partial L}{\partial y} = y(I - \boldsymbol{px}) = 0, \end{aligned} \qquad \boldsymbol{x} \geqslant \boldsymbol{0}, \quad y \geqslant 0. \tag{8.7}$$

Furthermore y^* has the interpretation of the marginal utility of money (or marginal utility of income), MU_m,

$$y^* = \partial U^* / \partial I = MU_m, \tag{8.8}$$

where U^* is the maximized level of utility,

$$U^* = U(x^*). \tag{8.9}$$

Given the positive prices and income, if utility is monotonically increasing in all consumption levels,

$$\partial U / \partial x_j = MU_j > 0, \tag{8.10}$$

where MU_j is the (positive) marginal utility of good j, it follows that added

income enables the household to buy more goods and hence increase utility. Thus y^*, the marginal utility of added income, is positive and, from the complementary slackness condition,

$$\boldsymbol{p}\boldsymbol{x}^* = I, \tag{8.11}$$

that is, all income is spent.

From the Kuhn–Tucker conditions it follows that the product of the marginal utility of income and the price of a good sets an upper limit to the marginal utility of each good,

$$MU_j \leqslant y^* p_j, \qquad j = 1, 2, \ldots, n. \tag{8.10}$$

From the complementary slackness condition it follows that if a good is purchased ($x_j^* > 0$) condition (8.10) holds as an equality. Thus if good j is purchased,

$$MU_j / p_j = y^* = MU_m, \tag{8.11}$$

so the ratio of marginal utility to price is the same for all goods that are actually purchased, the common ratio being the marginal utility of money. If (8.10) holds as a strict inequality then by the complementary slackness condition the good is not purchased ($x_j^* = 0$).

8.1. Demand theorem

According to the Demand Theorem, there exist solutions for the purchases of goods $\boldsymbol{x}^*$ and the marginal utility of money y^*, which can be considered functions of $n+1$ parameters, namely the n prices and income $\boldsymbol{p}$, I,

$$\boldsymbol{x}^* = \boldsymbol{x}^*(\boldsymbol{p}, I), \tag{8.12}$$

$$y^* = y^*(\boldsymbol{p}, I), \tag{8.13}$$

assuming $\boldsymbol{x}^* > \boldsymbol{0}$, $U(\boldsymbol{x})$ is twice continuously differentiable in a neighborhood of $\boldsymbol{x}^*$, $\boldsymbol{p}\boldsymbol{x}^* = I$ (non-satiation), and the Hessian matrix

$$\boldsymbol{H} = \frac{\partial^2 U}{\partial \boldsymbol{x}^2} = \frac{\partial}{\partial \boldsymbol{x}}\left(\frac{\partial U}{\partial \boldsymbol{x}}\right) \tag{8.14}$$

is non-singular. The functions in (8.12) are the *demand functions* for the n goods, the existence of which is guaranteed by the Implicit Function Theorem. Restricting attention to goods that are actually purchased, the first-order conditions,

using the solutions, can be written as the $n+1$ identities

$$\frac{\partial U}{\partial \boldsymbol{x}}(\boldsymbol{x}^*(\boldsymbol{p}, I)) \equiv y^*(\boldsymbol{p}, I)\boldsymbol{p}, \tag{8.15}$$

$$\boldsymbol{p}\boldsymbol{x}^*(\boldsymbol{p}, I) \equiv I. \tag{8.16}$$

(Restricting attention to goods that are actually purchased excludes the situation in which, by changing a parameter a good that had not been purchased could be purchased.) According to the theorem, these conditions characterize the equilibrium of the household. If the utility function $U(\boldsymbol{x})$ is strictly concave they are both necessary and sufficient conditions for an equilibrium.[34] Furthermore by the theorem, the n demand functions in (8.12) are positive homogeneous of degree zero in all prices and income,

$$\boldsymbol{x}^*(\lambda \boldsymbol{p}, \lambda I) = \boldsymbol{x}^*(\boldsymbol{p}, I), \quad \text{for all } \lambda, \quad \lambda > 0, \tag{8.17}$$

since changing $\boldsymbol{p}, I$ to $\lambda \boldsymbol{p}, \lambda I$ does not change the problem if $\lambda > 0$. (Only the constraint is affected, and $\lambda \boldsymbol{p}\boldsymbol{x} \leqslant \lambda I$ is equivalent to $\boldsymbol{p}\boldsymbol{x} \leqslant I$ if $\lambda > 0$.) Choosing λ to be $1/I$ the demand functions can be written

$$\boldsymbol{x}^* = \boldsymbol{x}^*\left(\frac{1}{I}\boldsymbol{p}\right) = \boldsymbol{x}^*(\boldsymbol{p}^*), \tag{8.18}$$

where $\boldsymbol{p}^*$ is the vector of prices relative to income,

$$\boldsymbol{p}^* = (p_1/I, p_2/I, \ldots, p_n/I) \tag{8.19}$$

Thus demand depends only on prices relative to income.[35] The demand theorem therefore characterizes the demand functions, states their homogeneity, and indicates their dependence on relative prices.

8.2. Slutsky theorem

The Slutsky Theorem summarizes the comparative statics of the household, as obtained by differentiating the first-order conditions (8.15) and (8.16) with respect to both prices and income. Following the approach of Section 7 yields

[34] The utility function is usually assumed to be strictly quasiconcave, as defined in footnote 10. Assuming $U(\boldsymbol{x})$ is strictly quasiconcave means that the set of all points on or above any indifference curve is convex and the indifference curves have no linear segments, where an *indifference curve* is a locus of points $\boldsymbol{x}$ for which $U(\boldsymbol{x})$ is constant. Any monotonic transformation of a utility function is also a utility function, however, and under a certain regularity condition any quasiconcave function can be transformed, by such a monotonic transformation, into a concave function. Thus, assuming the regularity condition is met, if $U(\boldsymbol{x})$ is strictly quasiconcave then there is at least one member of the class of allowable utility functions that is strictly concave.

[35] Alternatively, choosing one price, say p_n as *numeraire*, setting $\lambda = 1/p_n$ the demand functions can be written as functions of the price ratios,

$$\boldsymbol{x}^* = \boldsymbol{x}^*(p_1/p_n, p_2/p_n, \ldots, p_{n-1}/p_n, I/p_n).$$

the *fundamental matrix equation of the theory of the household*,

$$\begin{bmatrix} \dfrac{\partial y^*}{\partial I} & \dfrac{\partial y^*}{\partial \boldsymbol{p}} & \left(\dfrac{\partial y^*}{\partial \boldsymbol{p}}\right)_{\text{comp}} \\ \dfrac{\partial \boldsymbol{x}^*}{\partial I} & \dfrac{\partial \boldsymbol{x}^*}{\partial \boldsymbol{p}} & \left(\dfrac{\partial \boldsymbol{x}^*}{\partial \boldsymbol{p}}\right)_{\text{comp}} \end{bmatrix} = \begin{pmatrix} 0 & -\boldsymbol{p} \\ -\boldsymbol{p}' & \boldsymbol{H} \end{pmatrix}^{-1} \begin{pmatrix} -1 & \boldsymbol{x}^{*\prime} & \boldsymbol{0} \\ \boldsymbol{0} & y^*\boldsymbol{I}_n & y^*\boldsymbol{I}_n \end{pmatrix}, \quad (8.20)$$

where the comparative static results are summarized by the changes in the solutions y^*, $\boldsymbol{x}^*$ as the parameters I and $\boldsymbol{p}$ change,

$$\begin{aligned} \frac{\partial y^*}{\partial I} &= \frac{\partial^2 U^*}{\partial I^2}, \\ \frac{\partial \boldsymbol{x}^*}{\partial I} &= \left(\frac{\partial x_1^*}{\partial I} \; \frac{\partial x_2^*}{\partial I} \cdots \frac{\partial x_n^*}{\partial I}\right), \\ \frac{\partial y^*}{\partial \boldsymbol{p}} &= \left(\frac{\partial y^*}{\partial p_1} \; \frac{\partial y^*}{\partial p_2} \cdots \frac{\partial y^*}{\partial p_n}\right), \\ \frac{\partial \boldsymbol{x}^*}{\partial \boldsymbol{p}} &= \begin{bmatrix} \dfrac{\partial x_1^*}{\partial p_1} & \dfrac{\partial x_1^*}{\partial p_2} & \cdots & \dfrac{\partial x_1^*}{\partial p_n} \\ \vdots & & & \\ \dfrac{\partial x_n^*}{\partial p_1} & \dfrac{\partial x_n^*}{\partial p_2} & \cdots & \dfrac{\partial x_n^*}{\partial p_n} \end{bmatrix}, \end{aligned} \quad (8.21)$$

and all variables and derivatives are computed at the solution values $\boldsymbol{x}^*$, y^*.[36] Here "comp" refers to a compensated change in price, where income is compensated so as to keep utility constant; $\boldsymbol{H}$ is the Hessian matrix of (8.14), which is assumed negative definite so the matrix to be inverted, the bordered Hessian, is non-singular; and $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. Solving the fundamental equation, by inverting the partitioned matrix, leads to the *Slutsky equation*,

$$\begin{aligned} \frac{\partial \boldsymbol{x}^*}{\partial \boldsymbol{p}} &= \left(\frac{\partial \boldsymbol{x}^*}{\partial \boldsymbol{p}}\right)_{\text{comp}} - \left(\frac{\partial \boldsymbol{x}^*}{\partial I}\right)\boldsymbol{x}^* \quad \text{i.e.} \\ \left(\frac{\partial x_j^*}{\partial p_k}\right) &= \left(\frac{\partial x_j^*}{\partial p_k}\right)_{\text{comp}} - \left(\frac{\partial x_j^*}{\partial I}\right)x_k^*, \quad \text{all } j, k, \end{aligned} \quad (8.22)$$

stating that the *total effect* of a change in price on demand is the sum of the *substitution effect* of a compensated change in price on demand and the *income effect* of a change in income on demand, where the income effect involves the

[36] See Barten (1964), Intriligator (1971), and Theil (1975).

weighting by $-x^*$. This equation is the first part of the Slutsky Theorem. The second part of the theorem states that the matrix of substitution effect is symmetric and negative semidefinite,[37]

$$\left(\frac{\partial x^*}{\partial p}\right)_{\text{comp}} \text{ is symmetric} \quad \text{i.e.} \quad \frac{\partial x_j^*}{\partial p_k}+\frac{\partial x_j^*}{\partial I}x_k^*=\frac{\partial x_k^*}{\partial p_j}+\frac{\partial x_k^*}{\partial I}x_j^*, \quad \text{all } j,k, \tag{8.23}$$

$$z\left(\frac{\partial x^*}{\partial p}\right)_{\text{comp}} z' \leqslant 0 \quad \text{and} \quad =0 \quad \text{if} \quad z=\alpha p. \tag{8.24}$$

The final parts of the theorem are the *Engel aggregation condition*,

$$p\left(\frac{\partial x^*}{\partial I}\right)=1, \quad \text{i.e.} \quad \sum_{j=1}^{n} p_j\frac{\partial x_j^*}{\partial I}=1; \tag{8.25}$$

the *Cournot aggregation condition*,

$$p\left(\frac{\partial x^*}{\partial p}\right)+x^{*\prime}=\mathbf{0} \quad \text{i.e.} \quad \sum_{j=1}^{n} p_j\left(\frac{\partial x_j^*}{\partial p_l}\right)+x_l^*=0, \quad \text{all } l; \tag{8.26}$$

and the *homogeneity condition*,[38]

$$\frac{\partial x^*}{\partial \tilde{p}}p'+\frac{\partial x^*}{\partial I}I=\mathbf{0} \quad \text{i.e.} \quad \sum_{k=1}^{n}\frac{\partial x_j^*}{\partial p_k}p_k+\frac{\partial x_j^*}{\partial I}I=0, \quad \text{all } j. \tag{8.27}$$

9. Neoclassical theory of the firm[39]

The firm, as an economic agent, is assumed to behave so as to maximize profit subject to the technological constraint of the production function. Assuming the firm uses n inputs to produce a single output, let x be the column vector of inputs,

$$x=(x_1, x_2, \ldots, x_n)'; \tag{9.1}$$

[37]Kalman and Intriligator (1973) treat the case when prices enter the utility function. It is shown that there exists a generalized Slutsky equation in which the generalized matrix of substitution effects is still symmetric and negative semidefinite.

[38]This condition can be derived from the homogeneity condition (8.17) using Euler's theorem on homogeneous functions.

[39]Basic references on the neoclassical theory of the firm include Hicks (1946), Samuelson (1947), and Intriligator (1971). See also Chapter 10 by Nadiri. For a presentation of the duality approach to the theory of the firm, see Chapter 12 by Diewert.

q be the output; $f(\boldsymbol{x})$ be the production function of the firm,

$$q=f(\boldsymbol{x})=f(x_1, x_2, \ldots, x_n), \tag{9.2}$$

giving output as a function of the inputs; $\boldsymbol{w}$ be a row vector of (positive) given wages of the inputs,

$$\boldsymbol{w}=(w_1, w_2, \ldots, w_n); \tag{9.3}$$

and p be the (positive) given price of output. The problem of the (competitive) firm is then

$$\max_{q,\boldsymbol{x}} \pi=pq-\boldsymbol{w}\boldsymbol{x} \quad \text{subject to} \quad q=f(\boldsymbol{x}), \quad \boldsymbol{x}\geqslant \boldsymbol{0}. \tag{9.4}$$

Thus the firm chooses levels of inputs and output so as to maximize profits π, given in (9.4) as the difference between revenue, pq, and cost, given as the total expenditure on all inputs,

$$\boldsymbol{w}\boldsymbol{x}=\sum_{j=1}^{n} w_j x_j. \tag{9.5}$$

The production function can be incorporated into the objective function directly, so the problem can be stated

$$\max_{\boldsymbol{x}} \pi(\boldsymbol{x})=pf(\boldsymbol{x})-\boldsymbol{w}\boldsymbol{x} \quad \text{subject to} \quad \boldsymbol{x}\geqslant \boldsymbol{0}. \tag{9.6}$$

The Kuhn–Tucker conditions then state that at the solution $\boldsymbol{x}^*$,

$$\begin{aligned} &\frac{\partial \pi}{\partial \boldsymbol{x}}=p\frac{\partial f}{\partial \boldsymbol{x}}-\boldsymbol{w}\leqslant \boldsymbol{0}, \\ &\frac{\partial \pi}{\partial \boldsymbol{x}}\boldsymbol{x}=\left(p\frac{\partial f}{\partial \boldsymbol{x}}-\boldsymbol{w}\right)\boldsymbol{x}=0, \\ &\boldsymbol{x}\geqslant \boldsymbol{0}. \end{aligned} \tag{9.7}$$

Thus the ratio of input price to output price sets an upper limit to the marginal productivity of each input,

$$MP_j\equiv \partial f/\partial x_j \leqslant w_j/p, \qquad j=1,2,\ldots,n. \tag{9.8}$$

From the complementary slackness condition it follows that if an input is purchased ($x_j^*>0$) condition (9.8) holds as equality, thus if input j is purchased,

$$MP_j=w_j/p, \tag{9.9}$$

so the ratio of marginal product to wage is the same for all inputs actually purchased, the common ratio being the reciprocal of the output price.

9.1. *Supply theorem*

According to the Supply Theorem, there exist solutions for the purchase of inputs x^*, which can be considered functions of $n+1$ parameters, namely the n wages and output price w, p,

$$x^* = x^*(w, p), \tag{9.10}$$

assuming $x^* > \mathbf{0}$, $f(x)$ is twice continuously differentiable in a neighborhood of x^*, and the Hessian matrix

$$H = \frac{\partial^2 f}{\partial x^2} = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right) \tag{9.11}$$

is non-singular. The functions in (9.10) are the *input demand functions*, the existence of which is guaranteed by the Implicit Function Theorem. The output supply function is then

$$q^* = q^*(w, p) = f(x^*). \tag{9.12}$$

Restricting attention to inputs that are actually purchased, the first-order conditions, using the solutions, are the identities

$$p\frac{\partial f}{\partial x}(x^*(w, p)) \equiv w, \tag{9.13}$$

$$q^*(w, p) \equiv f(x^*(w, p)). \tag{9.14}$$

(As in the case of the household, restricting attention to inputs that are actually purchased excludes the situation in which, by changing a parameter, an input that had not been purchased could be purchased). According to the theorem these conditions characterize the equilibrium of the firm. If the production function $f(x)$ is strictly concave they are both necessary and sufficient conditions for an equilibrium. Furthermore by the theorem the n input demand functions in (9.10) and the output supply function in (9.12) are positive homogeneous of degree zero in all wages and output prices,

$$\begin{aligned} x^*(\lambda w, \lambda p) &= x(w, p), \\ q^*(\lambda w, \lambda p) &= q^*(w, p), \end{aligned} \quad \text{all} \quad \lambda > 0, \tag{9.15}$$

since changing w, p to $\lambda w, \lambda p$ only changes π in (9.4) to $\lambda\pi$, and maximizing $\lambda\pi$ yields the same solution as maximizing π if $\lambda > 0$. Choosing λ to be $1/p$ the input

demand functions and output supply function can be written

$$\begin{aligned} x^* &= x^*\left(\frac{1}{p}w\right) = x^*(w^*), \\ q^* &= q^*\left(\frac{1}{p}w\right) = q^*(w^*), \end{aligned} \tag{9.16}$$

where w^* is the vector of real wages, that is, wages relative to output price,

$$w^* = (w_1/p, w_2/p, \ldots, w_n/p). \tag{9.17}$$

Thus input demand depends only on the n real wages. The supply theorem therefore characterizes both the input demand and output supply functions, states their homogeneity, and indicates their dependence on real wages.

9.2. *Theorem on comparative statics for the firm*

The Theorem on Comparative Statics for the Firm is obtained by differentiating the first-order conditions (9.13) and (9.14) with respect to both input prices w and output price p. Following the approach of Section 7 yields the *fundamental matrix equation of the theory of the firm*,

$$\begin{bmatrix} \dfrac{\partial q^*}{\partial p} & \left(\dfrac{\partial q^*}{\partial w}\right)' \\ \dfrac{\partial x^*}{\partial p} & \dfrac{\partial x^*}{\partial w} \end{bmatrix} = \begin{bmatrix} -1 & \dfrac{\partial f}{\partial x} \\ \mathbf{0} & pH \end{bmatrix}^{-1} \begin{bmatrix} 0 & \mathbf{0} \\ -\left(\dfrac{\partial f}{\partial x}\right)' & I_n \end{bmatrix}, \tag{9.18}$$

where the comparative statics results are summarized by the change in the solutions q^*, x^* as the parameters p and w change,

$$\begin{aligned} &\frac{\partial q^*}{\partial p} \\ &\frac{\partial x^*}{\partial p} = \left(\frac{\partial x_1^*}{\partial p}\ \frac{\partial x_2^*}{\partial p} \cdots \frac{\partial x_n^*}{\partial p}\right)', \\ &\frac{\partial q^*}{\partial w} = \left(\frac{\partial q^*}{\partial w_1}\ \frac{\partial q^*}{\partial w_2} \cdots \frac{\partial q^*}{\partial w_n}\right)', \\ &\frac{\partial x^*}{\partial w} = \begin{bmatrix} \dfrac{\partial x_1^*}{\partial w_1} & \dfrac{\partial x_1^*}{\partial w_2} & \cdots & \dfrac{\partial x_1^*}{\partial w_n} \\ \vdots & & & \\ \dfrac{\partial x_n^*}{\partial w_1} & \dfrac{\partial x_n^*}{\partial w_2} & \cdots & \dfrac{\partial x_n^*}{\partial w_n} \end{bmatrix}, \end{aligned} \tag{9.19}$$

and all variables and derivatives are computed at the solution values $q^*, \boldsymbol{x}^*$. Here $\partial f/\partial \boldsymbol{x}$ is the vector of marginal products; $\boldsymbol{H}$ is the Hessian matrix of (9.11), which is assumed negative definite; and $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. Solving the fundamental equation leads to the equality[40]

$$\partial q^*/\partial \boldsymbol{w} = -\partial \boldsymbol{x}^*/\partial p \quad \text{i.e.} \quad \partial q^*/\partial w_j = -\partial x_j^*/\partial p, \quad \text{all } j, \tag{9.20}$$

stating that the effect of any wage on output is identical but opposite in sign to the effect of output price on the same input. This equation is the first part of the theorem. The second part of the theorem states that the matrix of effects of wages on input demands is symmetric and negative definite,

$$\partial \boldsymbol{x}^*/\partial \boldsymbol{w} \text{ is symmetric} \quad \text{i.e.} \quad \partial x_j^*/\partial w_k = \partial x_k^*/\partial w_j, \quad \text{all } j, k, \tag{9.21}$$

$$z(\partial \boldsymbol{x}^*/\partial \boldsymbol{w})z' \leqslant 0 \quad \text{and} \quad = 0 \quad \text{if} \quad z = \alpha \boldsymbol{w}. \tag{9.22}$$

The final part of the theorem states that an increase in output price will increase the supply of output,[41]

$$\partial q^*/\partial p > 0. \tag{9.23}$$

The theory of linear programming can be applied to a firm that produces using an activity analysis technology. In such a case the firm produces n outputs $x_1, x_2, \ldots, x_n$ using m inputs $b_1, b_2, \ldots, b_m$.[42] To produce one unit of output x_j requires a_{ij} units of input i. Suppose that in the short run all inputs are fixed so the only choice for the firm is that of deciding what mix of outputs to produce given these inputs. The problem is then the standard linear programming one,

$$\max_{\boldsymbol{x}} \boldsymbol{cx} \quad \text{subject to} \quad \boldsymbol{Ax} \leqslant \boldsymbol{b}, \quad \boldsymbol{x} \geqslant \boldsymbol{0}, \tag{9.24}$$

as in (6.1). The objective function to be maximized is total revenue, given by

$$\boldsymbol{cx} = c_1 x_1 + c_2 x_2 + \cdots + c_n x_n, \tag{9.25}$$

where c_j is the given price and x_j is the chosen level of output j. The m

[40] See Intriligator (1971). Inputs j for which $\partial x_j^*/\partial p < 0$ are called *inferior inputs*. Not all inputs can be inferior. See Bear (1965).

[41] For a comparison of this classical case of profit maximization to the case of a firm maximizing sales subject to a profit constraint, as discussed in Baumol (1967), see Kalman and Intriligator (1973). It is shown that some inferior inputs are possible for a classical firm but that such inputs are not possible for a Baumol firm.

[42] Note that here outputs rather than inputs are the x's and that inputs are given by the b's. This switch in notation facilitates use of the standard notation for linear programming problems, as presented in Section 6.

constraints are of the form

$$a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n \leqslant b_i, \qquad i = 1,2,\ldots,m, \tag{9.26}$$

stating that the total amount of input i used to produce the output vector $\boldsymbol{x}$ cannot exceed the level of input i available, b_i. The problem is thus one of choosing nonnegative outputs so as to maximize profit, given the technology and the available inputs.

The dual problem is

$$\min_{\boldsymbol{y}} \boldsymbol{yb} \quad \text{subject to} \quad \boldsymbol{yA} \geqslant \boldsymbol{c}, \quad \boldsymbol{y} \geqslant \boldsymbol{0}, \tag{9.27}$$

as in (6.5). This problem can be interpreted as one of choosing non-negative values (shadow prices) for the inputs $y_1, y_2, \ldots, y_m$ so as to minimize the cost of the inputs,

$$\boldsymbol{yb} = y_1 b_1 + y_2 b_2 + \cdots + y_m b_m, \tag{9.28}$$

where y_i is the chosen value and b_i is the given level of input i. The n constraints are of the form

$$y_1 a_{1j} + y_2 a_{2j} + \cdots + y_m a_{mj} \geqslant c_j, \qquad j = 1,2,\ldots,n, \tag{9.29}$$

stating that the unit cost of good j, obtained by summing the cost of producing one unit over all inputs, is no less than the price of this good. The dual to a problem of allocation, the primal problem, (9.24), is therefore one of valuation, the dual problem (9.27). According to the complementary slackness conditions (6.14), if for any output j inequality (9.29) holds as a strict inequality, so unit cost exceeds price (the output is produced at a loss), then this output is not produced ($x_j^* = 0$). Similarly if for any input i inequality (9.26) holds as a strict inequality, so not all of the input is used (it is in excess supply), then this input is a free good ($y_i^* = 0$). Furthermore, from (6.13),

$$\boldsymbol{cx}^* = \boldsymbol{y}^*\boldsymbol{b}, \tag{9.30}$$

so at the solution to the dual problems total revenue from the output equals the total cost of the inputs, i.e., the firm produces at a zero profit level.

10. Conclusions

Two conclusions naturally emerge from this survey of mathematical programming with applications to economics. *First*, the various mathematical programming problems treated here — the unconstrained problem, classical programming, nonlinear programming, and linear programming — are all

closely interrelated, with analogous theorems in all cases. *Second*, the same mathematical programming problems have important applications to economics, particularly to the microeconomic theory of the household and the firm. The results of mathematical programming lead to both a characterization of the equilibrium of each of these agents and an analysis of their comparative statics responses to changes in parameters, such as prices and income.

References

Aoki, M. (1971), Introduction to optimization techniques. New York: Macmillan.

Apostol, T. (1957), Mathematical analysis. Reading, MA: Addison-Wesley.

Arrow, K. J. and A. C. Enthoven (1961), "Quasiconcave programming", Econometrica, 29:779–800.

Arrow, K. J., L. Hurwicz and H. Uzawa (1958), "Constraint qualifications in maximization problems", in: K. J. Arrow, L. Hurwicz and H. Uzawa, eds., Studies in linear and nonlinear programming. Stanford, CA: Stanford University Press.

Arrow, K. J., L. Hurwicz and H. Uzawa (1961), "Constraint qualifications in maximization problems", Naval Research Logistics Quarterly, 8:175–191.

Avriel, M. (1976), Nonlinear programming. Englewood Cliffs, NJ: Prentice-Hall.

Barten, A. P. (1964), "Consumer demand functions under conditions of almost additive preferences", Econometrica, 32:1–38.

Baumol, W. J. (1967), Business behavior, value, and growth, Rev. ed. New York: Harcourt, Brace, and World.

Bazaraa, M. S. (1979), Nonlinear programming: Theory and algorithms. New York: Wiley.

Bazaraa, M. S. and C. M. Shetty (1976), Foundations of optimization. Berlin: Springer-Verlag.

Bazaraa, M. S., J. J. Goode and C. M. Shetty (1972), "Constraint qualifications revisited", Management Science, 18:567–573.

Bear, D. V. T. (1965), "Inferior inputs and the theory of the firm", Journal of Political Economy, 73:287–289.

Canon, M. D., C. D. Cullum, Jr. and E. Polak (1970), Theory of optimal control and mathematical programming. New York: McGraw-Hill.

Courant, R. (1947), Differential and integral calculus, 2nd ed. New York: Interscience Publishers.

Dantzig, G. (1963), Linear programming and extensions. Princeton, NJ: Princeton University Press.

Dixon, P. B., S. Bowles and D. Kendrick (1980), Notes and problems in microeconomic theory. Amsterdam: North-Holland.

Dorfman, R., P. A. Samuelson and R. M. Solow (1958), Linear programming and economic analysis. New York: McGraw-Hill.

El-Hodiri, M. A. (1971), Constrained extrema: Introduction to the differentiable case with economic applications. Berlin: Springer-Verlag.

Fiacco, A. V. and G. P. McCormick (1968), Nonlinear programming: Sequential unconstrained minimization techniques. New York: Wiley.

Fleming, W. H. (1965), Functions of several variables. New York: McGraw-Hill.

Gale, D. (1960), The theory of linear economic models. New York: McGraw-Hill.

Gass, S. I. (1975), Linear programming: Methods and applications, 4th ed. New York: McGraw-Hill.

Geoffrion, A. M., ed. (1972), Perspective on optimization. Reading, MA: Addison-Wesley.

Hadley, G. (1963), Linear programming. Reading, MA: Addison-Wesley.

Hadley, G. (1964), Nonlinear and dynamic programming. Reading, MA: Addison-Wesley.

Hestenes, M. R. (1966), Calculus of variations and optimal control theory. New York: Wiley.

Hestenes, M. R. (1975), Optimization theory: The finite dimensional case. New York: Wiley.

Hicks, J. R. (1946), Value and capital, 2nd ed. New York: Oxford University Press.

Intriligator, M. D. (1971), Mathematical optimization and economic theory. Englewood Cliffs, NJ: Prentice-Hall.

Intriligator, M. D. (1975), "Applications of optimal control theory in economics", Synthese, 31:271–288.

Intriligator, M. D. (1977), "Economic systems", in: C. T. Leondes, ed., Control and dynamic systems, Advances in theory and applications, Vol. 13. New York: Academic Press.
John, F. (1948), "Extremum problems with inequalities as side conditions", in: K. O. Friedrichs, O. W. Neugebauer and J. J. Stoker, eds., Studies and essays: Courant anniversary volume. New York: Interscience Publishers.
Kalman, P. J. and M. D. Intriligator (1973), "Generalized comparative statics, with applications to consumer and producer theory", International Economic Review, 14:473–486.
Kuhn, H. W. and A. W. Tucker (1951), "Nonlinear programming", in: J. Neyman, ed., Proceedings of the second Berkeley symposium on mathematical statistics and probability. Berkeley, CA: University of California Press.
Lancaster, K. (1968), Mathematical economics. New York: Macmillan.
Luenberger, D. G. (1969), Optimization by vector space methods. New York: Wiley.
Luenberger, D. G. (1973), Introduction to linear and nonlinear programming. Reading, MA: Addison-Wesley.
Mangasarian, O. L. (1969), Nonlinear programming. New York: McGraw-Hill.
Peterson, D. W. (1973), "A review of constraint qualifications in finite-dimensional spaces", SIAM Review, 15:
Phlips, L. (1974), Applied consumption analysis. Amsterdam: North-Holland.
Polak, E. (1971), Computational methods in optimization: A unified approach. New York: Academic Press.
Samuelson, P. A. (1947), Foundations of economic analysis. Cambridge, MA: Harvard University Press.
Simmonard, M. (1966), Linear programming. Englewood Cliffs, NJ: Prentice-Hall.
Takayama, A. (1974), Mathematical economics. Hillsdale, NJ: Dryden Press.
Theil, H. (1975), The theory and measurement of consumer demand, Vol. 1. Amsterdam: North-Holland.
Wold, H. and L. Jureen (1953), Demand analysis. New York: Wiley.
Zangwill, W. I. (1969), Nonlinear programming: A unified approach. Englewood Cliffs, NJ: Prentice-Hall.

Chapter 3

DYNAMICAL SYSTEMS WITH APPLICATIONS TO ECONOMICS*

HAL R. VARIAN

University of Michigan

This chapter provides a survey of some basic mathematical results concerning dynamical systems which have proved useful in economics. There is some degree of overlap in subject matter with some other chapters in this book, especially the chapters on control theory (Chapter 4), global analysis (Chapter 8), and the stability of competitive equilibrium (Chapter 16). For this reason I have not attempted to provide an extensive survey of economic applications in these areas, but have instead concentrated on describing the basic forms of the main mathematical tools that can be used to answer questions common to many economic applications.

1. Basic concepts

1.1. Dynamical systems in R^n

The *state* of a system consists of a description of everything one needs to know in order to describe how the system will change. In most economic applications the state of a system can be described by some n-tuple of real numbers. The *state space* of a system consists of all feasible or relevant states. In almost all economic applications the state space can be regarded as a subset of R^n. In many economic applications, the state space can be regarded as being topologically equivalent to the unit disk,

$$D^n = \{x \text{ in } R^n\colon \|x\| \leqq 1\}.$$

Example 1

Consider a standard general equilibrium model where the k-vector of excess demands, $z(p)$, is a homogeneous function of k non-negative prices. Then we can take the state space of the economy to be the set of all non-negative prices, R^k_+. A more convenient choice of a state space can be found by noticing that

*The writing of this paper was financed in part by a grant from the National Science Foundation.

prices can be normalized by the requirement that $\sum p_i^2 = 1$. Thus the state space will just be the positive orthant of the unit sphere,

$$S_+^{k-1} = \left\{ x \text{ in } D^k \colon \|x\| = 1, \quad x \geqq 0 \right\}.$$

Note that S_+^{k-1} is topologically equivalent to the unit disk of dimension $k-1$.

Let X denote the state space of some system under consideration. A *state transition function*, T, is a function from $X \times R$ to X. The real line is interpreted as time, and $T(x, t)$ gives us the state of the system at time t if the system was in state x at time 0. In most applications the state transition function is not given explicitly, but rather is given implicitly by some *system of differential equations*,

$$\begin{aligned} &\dot{x}_i(t) = \mathrm{d}x_i(t)/\mathrm{d}t = f_i(x_1(t), \ldots, x_n(t)), \\ &x_i(0) = x_{0i}, \end{aligned} \qquad i = 1, \ldots, n.$$

In vector notation, we write this system as

$$\begin{aligned} &\dot{x}(t) = f(x(t)), \\ &x(0) = x_0. \end{aligned}$$

Let x: $R \to X$ be a solution to this system of differential equations with initial condition $x(0) = x_0$. Then $x(t)$ defines a state transition function as follows:

$$T(x_0, t) \equiv x(t).$$

Sometimes we want to emphasize the dependence of the position of the state at time t on the initial position x. In this case we can define the *flow* of the differential equation $\phi_t(x)$ as

$$\phi_t(x) \equiv T(x, t).$$

A *dynamical system* is just a state space along with a state transition function.

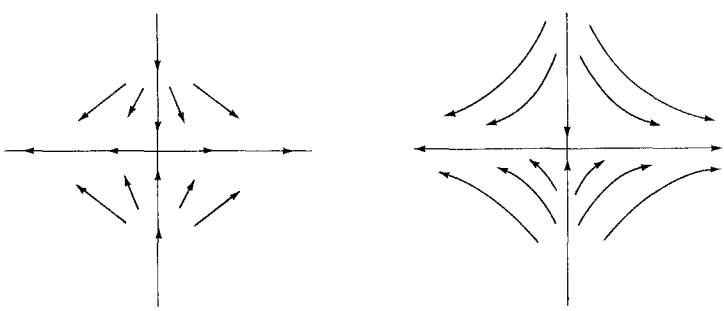


Figure 1.1. Vector field and solution curves.

A nice way to visualize these concepts is through the use of a *vector field*. Here we think of attaching a vector $f(x)$ to each point x in the state space. The *solution curves* (trajectories, orbits, etc.) for the differential equation system $\dot{x}=f(x)$ will just be the images of the function $\phi_t(x)$ as t ranges over all of R and x ranges over S. It is easy to see that if x is a point on some solution curve $\phi_t(\cdot)$, then $f(x)$ is a tangent vector to the curve at x. See Figure 1.1 for an illustration.

1.2. Dynamical systems on manifolds

For some applications in economics we want a state space that is more general then R^n or D^n. The appropriate concept is that of a *manifold*. In this section we will sketch a brief introduction to the theory of dynamical systems on manifolds.

First we define the closed halfspace $H^m=\{(x_1,\ldots,x_m)$ in R^m: $x_m\geqq 0\}$. Next we define the concept of a *diffeomorphism*: f: $X\to Y$ is a diffeomorphism if f is a homeomorphism and both f and f^{-1} are differentiable. Finally we can define the concept of a *manifold*.

Definition

A subset X of R^k is a smooth m-manifold if each x in X has a neighborhood $U\cap X$ diffeomorphic to an open subset $V\cap H^m$ of H^m.

(The reader should be warned that this is usually called a *manifold with boundary*. Since most of the manifolds we will discuss will have boundaries, it seems more economical to use this terminology.)

Let x be a point in an m-manifold X. Let g be a particular diffeomorphism between $U\cap X$ and $V\cap H^m$. Then g is called a *parameterization* of $U\cap X$.

Since g is a map between R^k and R^m, its derivative can be represented by a $k\times m$ matrix $Df(x)$. The *tangent space* of X at x is just the image of R^m under the linear map $Dg^{-1}(y)$ where $y=f(x)$.

Geometrically speaking, a manifold is just a generalization of the idea of an m-dimensional surface, and a tangent space is a generalization of the idea of a tangent hyperplane.

A *vector field* on a manifold X is a map f from X to R^m such that $f(x)$ is in the tangent space of X at x. We can think of f as defining an ordinary system of differential equations on a subset of R^k. By the usual existence and uniqueness theorems, we can find a solution x: $R\to R^k$ to this system of differential equations. Since the tangent vector to $x(t)$ at x is always tangent to the surface of X, the solution curves to the differential equation system must lie in the manifold X. Thus the vector field f defines in a natural way a dynamical system on X.

2. Basic tools[1]

Given a system of differential equations and a state space X there are many questions that naturally arise. For example:

(1) *Existence of solutions.* Given $\dot{x}=f(x)$, $x(0)=x_0$, is there necessarily a solution $x(t)$? What properties does $x(t)$ have?
(2) *Existence of equilibria.* Are there points x^* in X such that $f(x^*)=0$?
(3) *Number of equilibria.* How many equilibria are there?
(4) *Local stability of equilibria.* If we are perturbed slightly from an equilibrium will the system return to it?
(5) *Global stability of equilibrium.* If we start at an arbitrary state x, will we be led to an equilibrium?
(6) *Existence of cycles.* If we start at a state x will we eventually return to x?

In the following sections we will describe some of the mathematical tools used to answer such questions and give some examples of how these questions arise in economic applications.

2.1. *Existence, uniqueness, and continuity of solutions*

Let f: $X \to R^n$ and let $\dot{x}=f(x)$ define a system of differential equations with initial conditions $x(0)=x_0$. A *solution* to this system is just a differentiable function x: $I \to X$, where I is some interval in R, such that (1) $\mathrm{d}x(t)/\mathrm{d}t=f(x(t))$ and (2) $x(0)=x_0$.

The basic result on existence and uniqueness of solutions is:

Theorem

Let X be an open subset of R^n and let x_0 be an element of X. Let f: $X \to R^n$ be a continuously differentiable function. Then there is some $a>0$ and a unique solution x: $(-a, a) \to X$ of the differential equation $\dot{x}=f(x)$ which satisfies the initial condition $x(0)=x_0$.

Proof

See Hirsch and Smale (1974, p. 163).

It turns out that if we are just interested in *existence* of a solution it is enough to assume that f is continuous. However, the uniqueness result is very useful since it implies the important topological restriction that *solution curves can not*

[1]Standard references on differential equations and dynamical systems are Hirsch and Smale (1974), Hartman (1964), and Coddington and Levinson (1955).

cross. This kind of regularity is well worth the additional restriction of continuous differentiability. (Even weaker conditions will suffice.)

It is often of considerable interest to know how solution curves will behave as we vary the initial conditions. It turns out that they vary continuously; that is, if x_0 and y_0 are sufficiently close, then $\phi_t(x_0)$ and $\phi_t(y_0)$ will be close.

Theorem

Let f be as above and let y: $[t_0; t_1] \to X$ be a solution with $y(t_0) = y_0$. Then there is a neighborhood U of y_0 such that for any x_0 in U, there is a solution x: $[t_0; t_1] \to X$ with $x(t_0) = x_0$ and some constant K such that

$$|y(t) - x(t)| \leqq K\,|y_0 - x_0|\exp|(K(t - t_0))|, \quad \textit{for all } t \textit{ in } [t_0; t_1].$$

Proof

See Hirsch and Smale (1974, p. 173).

This theorem says that the flow of the differential equation ϕ_t: $X \to X$ is continuous as a function of x.

2.2. *Existence of equilibria*[2]

An *equilibrium* of a dynamical system $\dot{x} = f(x)$ is a point x^* in X such that $f(x^*) = 0$. If a dynamical system is in an equilibrium state it will remain there forever. The question arises: When does a dynamical system possess equilibrium states?

Theorem

Let f: $D^n \to R^n$ be a continuous vector field on the unit disk that points in on the boundary of D^n; that is, $x \cdot f(x) < 0$ for all x such that $\|x\| = 1$. Then there exists an x^* in D^n such that $f(x^*) = 0$.

Proof

See Spanier (1966, p. 197).

Of course the theorem is also true for any state space homeomorphic to a disk.

[2]An excellent historical and bibliographical survey of the problem of the existence of a Walrasian equilibrium is given in Arrow and Hahn (1971, ch. 1). The boundary condition trick was suggested by Varian (1977a) in a different context.

Example 2

Consider the Walrasian model described in Example 1. We think of $z(p)$ as being a function on S_+^{k-1}. We make three assumptions about z:

(1) *Continuity*: z: $S_+^{k-1} \to R^k$ is continuous.
(2) *Walras' Law*: $p \cdot z(p) = 0$ for p in S_+^{k-1}.
(3) *Desirability*: $z_i(p) > 0$ if $p_i = 0$, $i = 1, \ldots, k$.

Then there exists a p^* in S_+^{k-1} such that $z(p^*) = 0$. To see this, note that Walras' Law implies that $z(p)$ must lie in the tangent space of S_+^{k-1}, and that desirability implies $z(p)$ points in on the boundary of S_+^{k-1}. The result now follows from the above theorem.

The assumptions of the theorem can be weakened in several ways. For example, the following assumption can replace Walras' Law:

(4) *No inflation*: For all p in S_+^{k-1}, there is no $t \neq 0$ such that $z(p) = tp$.

Simply note that we can project $z(p)$ onto the tangent space of S_+^{k-1} without introducing any new equilibria.

Similarly, the boundary condition in the existence theorem may sometimes be too restrictive. A weaker replacement is the assumption that f never points directly out along the boundary of D^n:

(5) *Never points out*: For all x such that $\|x\| = 1$, $f(x) \neq tx$ for $t > 0$.

To reduce this case to the original case simply note that we can enclose D^n in a ball of radius 2. Along the boundary of this ball we define the vector field $\dot{x} = -x/\|x\|$ which clearly points in. Now we continuously extend this vector field to the one on D^n by taking a convex combination of $f(x/\|x\|)$ and $-x/\|x\|$. It is easy to see that this construction introduces no new zeros so the existence theorem applies directly.

2.3. *Uniqueness of equilibria*[3]

Suppose now we have a smooth dynamical system on the disk that points in on the boundary of the disk. By the last section we know that there will be at least one equilibrium x^*. Under what conditions will there be exactly one equilibrium?

[3] For a survey of results on uniqueness, see Arrow and Hahn (1971, ch. 9). The importance of the index of a fixed point was first realized by Dierker (1972, 1974). For some economic interpretations of the uniqueness condition, see Varian (1975). For an application of the index theorem to a *non-uniqueness* question, see Varian (1977b).

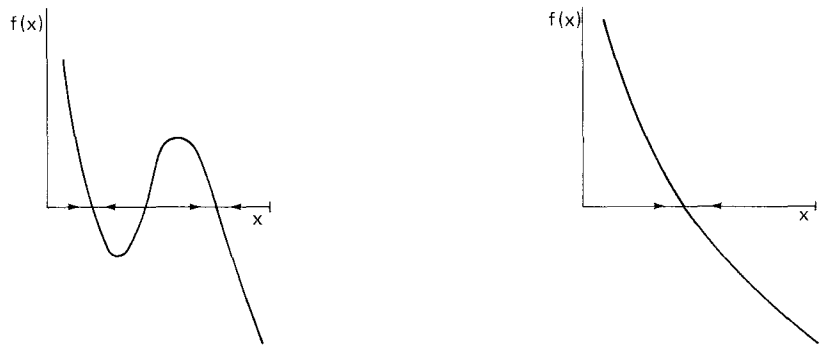


Figure 2.1. Uniqueness of equilibria.

The basic tool to answer this question comes from differential topology and is known as the *Poincaré index of a vector field*. An excellent discussion of this important topic can be found in Guillemin and Pollack (1974) and Milnor (1965).

Let us first consider the one-dimensional case to get some intuition. Let $\dot{x}=f(x)$ define a smooth vector field on the unit interval that points in on the boundary; i.e., such that $f(0)>0$, $f(1)<0$. Then several observations present themselves:

(1) Except in "degenerate" cases there are a finite number of equilibria.
(2) In general this number is odd.
(3) If $f'(x^*)$ is of one sign at all equilibria then there can be only one equilibrium. (See Figure 2.1.)

It turns out that all of these remarks generalize to higher-dimensional cases. In this case, we let f: $D^n \to R^n$ be a smooth vector field on the disk, D^n, that points in on the boundary of D^n. Let x^* be an equilibrium. The *index* of x^*, $I(x^*)$, is defined to be

$$+1 \quad \text{if} \quad \det(-Df(x^*))>0,$$

$$-1 \quad \text{if} \quad \det(-Df(x^*))<0,$$

an integer depending on topological considerations
if $\det(-Df(x^*))=0$.

We now have a fundamental theorem of differential topology:

Theorem (Poincaré–Hopf)

Suppose f: $D^n \to R^n$ *has a finite number of isolated equilibria* (x_i), $i=1,\ldots,k$, *and that* f *points in on the boundary of* D^n. *Then*

$$\sum_{i=1}^{k} I(x_i) = +1.$$

Example 3

Let us apply this theorem to the problem of uniqueness of Walrasian equilibrium. Here we have a vector field given by z: $S_+^{k-1} \to R^k$. In order to compute the index of each equilibrium we need to choose a local parametrization g: $S_+^{k-1} \to R^{k-1}$. It is geometrically clear that projection onto R^{k-1} can serve as an appropriate parametrization. Algebraically this just means we write down the k by k Jacobian matrix $Dz(p^*)$ and drop (say) the last row and column. The index of the equilibrium p^* is the determinant of the negative of this $(k-1)$ by $(k-1)$ matrix. We can now apply the argument of Milnor (1964, p. 8) to conclude that if $\det(-Dz(p^*)) \neq 0$ at all equilibrium values of p^* there can only be a finite number of equilibria.

The uniqueness result then follows easily: if $\det(-Dz(p^*)) > 0$ at all equilibria, there can be only one. If there is only one equilibrium then $\det(-Dz(p^*)) \geqq 0$.

2.4. *Local stability of equilibria*[4]

Let x^* be an equilibrium of a dynamical system f: $X \to R^n$. Roughly speaking, this equilibrium is locally stable if the system returns to x^* when perturbed to nearby states. If an equilibrium is to be economically relevant in that the system remains equilibrated for any length of time it seems that it would have to be locally stable. We formulate a precise notion of this concept and investigate a criterion for stability below:

Definition

An equilibrium is locally asymptotically stable if there is some $e > 0$ *such that* $|x_0 - x^*| < e$ *implies that* $\phi_t(x_0)$ *converges to* x^* *as* t *goes to infinity.*

Theorem

Let x^* *be an equilibrium of* f: $X \to R^n$ *and let* $Df(x^*)$ *have all negative eigenvalues. Then* x^* *is locally asymptotically stable.*

[4] For a discussion of local stability results, see Arrow and Hahn (1971, ch. 11, sec. 6, and ch. 12, sec. 5). For a derivation of Slutsky's equation, see for example Varian (1978, ch. 3).

Proof

See Hirsch and Smale (1974).

Example 4

Let us consider the Walrasian equilibrium model described earlier. Here it is convenient to choose a slightly different normalization for prices. Let us set the kth price equal to 1 and measure all other prices relative to it. By abuse of notation, we will let z denote the mapping which sends these $k-1$ normalized prices to the $k-1$ excess demands. By Walras' Law, if $p^* \gg 0$ and $z_1(p^*), \ldots, z_{k-1}(p^*)$ are zero, then $z_k(p^*)$ is zero; thus the equilibria of the system $\dot{p} = z(p)$ are precisely the Walrasian equilibria p^* which will be locally stable when $Dz(p^*)$ has all negative eigenvalues. What is an economic interpretation of this condition?

By Slutsky's equation, we can write $Dz(p^*)$ as

$$Dz(p^*) = \sum_{i=1}^{n} S_i(p^*) + \sum_{i=1}^{n} Y_i(p^*) = S(p^*) + Y(p^*),$$

where $S_i(p^*)$ is the substitution matrix of the ith consumer (which is known to be negative definite) and $Y_i(p^*)$ is the "income effect" for the ith consumer. The matrix $S(p^*)$ is negative definite and therefore has all negative eigenvalues; thus if the "aggregate income effects" $Y(p^*)$ are not too big, the system $\dot{p} = z(p)$ will be locally stable at p^*.

2.5. *Global stability of equilibria*[5]

Let x^* be an equilibrium of a dynamical system. Then x^* is *globally stable* if $x(t)$ approaches x^* as t goes to infinity, for any initial condition x_0. That is, x^* is globally stable if $\lim_{t \to \infty} \phi_t(x) = x^*$ for all x.

Clearly global stability implies local stability; however, global stability is a much stronger condition. How do we tell when a dynamical system is globally stable? The main tool is the concept of a *Liaponov function*.

Definition

Let $\dot{x} = f(x)$ be a dynamical system on X with equilibrium x^*. Suppose we can find a differentiable function V: $X \to R$ such that

$$V(x^*) = 0, \quad V(x) > 0 \quad \text{for} \quad x \neq x^*,$$

$$dV(x(t))/dt < 0 \quad \text{for} \quad x \neq x^*.$$

[5]For a survey of global stability theorems in economics, see Arrow and Hahn (1971, chs. 11–13).

Then V is called a Liaponov function.

The basic result is:

Theorem

Let f: $X \to R^n$ be a dynamical system with X compact and with equilibrium x^*. Suppose that we can find a Liaponov function for this system. Then x^* is a globally stable equilibrium.

Proof

See Hirsch and Smale (1974, p. 193).

Unfortunately there is no simple way to find Liaponov functions in general. However, in most economic applications Liaponov functions turn out to be fairly natural.

Example 5

Let p^* be an equilibrium of the Walrasian system $\dot{p} = z(p)$. Suppose that $z(p)$ obeys the "weak axiom of revealed preference" so that $p^* \cdot z(p) > 0$ for all $p \neq p^*$. Then p^* is a globally stable equilibrium. In order to prove this we need to show that the state space can be chosen to be compact and that the system admits a Liaponov functions. We will omit the first part of the proof and simply show that $V(p)$ can be chosen to be $V(p) = \| p - p^* \|^2 = \sum_{i=1}^{k} (p_i - p_i^*)^2$.

To see this we just differentiate $V(p(t))$,

$$\frac{dV(p(t))}{dt} = 2 \sum_{i=1}^{k} (p_i(t) - p_i^*) \dot{p}_i(t).$$

Now use the fact that $\dot{p}_i(t) = z_i(p(t))$,

$$\frac{dV(p(t))}{dt} = 2 \left[\sum_{i=1}^{k} p_i(t) z_i(p_i(t)) - \sum_{i=1}^{k} p_i^*(t) z_i(p(t)) \right] = -2p^* \cdot z(p(t)) < 0,$$

where the last step follows from Walras' Law and the weak axiom of revealed preference hypothesis.

2.6. *Existence of cycles*[6]

Let f: $X \to R^n$, $\dot{x} = f(x)$ be a smooth dynamical system. A point x is in a *closed orbit* if x is not an equilibrium but $\phi_t(x) = x$ for some $t \neq 0$. That is, a state is in a

[6] Very little work has been done on the existence of closed orbits in Walrasian economics since there are no good mathematical criteria for existence when the dimension of the state space is greater than 2. There has been some study of the existence of cycles in the macroeconomic literature. See for example Ichimura (1954), Chang and Smyth (1971), and Varian (1979).

closed orbit if the system eventually returns to that state. Closed orbits are commonly referred to as *cycles*. A useful criterion for the existence of closed orbits is the Poincaré–Bendixson Theorem. In order to state the theorem we need some definitions.

A point y in X is an *ω-limit point* of x if there is a sequence $t_n \to \infty$ such that $\lim_{n\to\infty} \phi_{t_n}(x) = y$. An *$\omega$-limit set* of y, $L_\omega(y)$ is the set of all ω-limit points of y.

If x^* is an equilibrium point then $L_\omega(x^*)$ consists only of x^*. If x^* is a globally stable equilibrium, then $L_\omega(x) = \{x^*\}$ for any x in X. If x is on some closed orbit, C, then $L_\omega(x) = C$. In high dimensions, limit sets can have very complicated structures. However, in two-dimensional systems their structure is rather simple:

Theorem (Poincaré–Bendixson)

A non-empty compact limit set of a continuously differentiable system in R^2, which contains no equilibrium point, is a closed orbit.

Proof

See Hirsch and Smale (1974, p. 248).

Example 6

Let us consider a Walrasian system with three goods so that $\dot{p} = z(p)$ defines a dynamical system on S^2_+. We assume that this system points in on the boundary of S^2_+, and think of this system as a dynamical system on D^2. We know that there will exist at least one equilibrium p^* where $z(p^*) = 0$. Suppose that all equilibria are *totally unstable* in the sense that the eigenvalues of $Dz(p^*)$ are strictly positive. Then there must exist a closed orbit — a "business cycle" if you will.

The proof is a direct application of the Poincarè–Bendixson Theorem. First we note that by an index argument there can be only one equilibrium p^*. Choose any other p in D^2 and consider its limit set $L_\omega(p)$. This is a non-empty, closed and thus compact subset of D^2. Furthermore it does not contain an equilibrium since p^* is the unique equilibrium and it is unstable. Hence $L_\omega(p)$ must be a closed orbit.

3. Some special kinds of dynamical systems

Up until now we have been concerned with general dynamical systems. In this section we consider two special kinds of dynamical systems that are often encountered in economics.

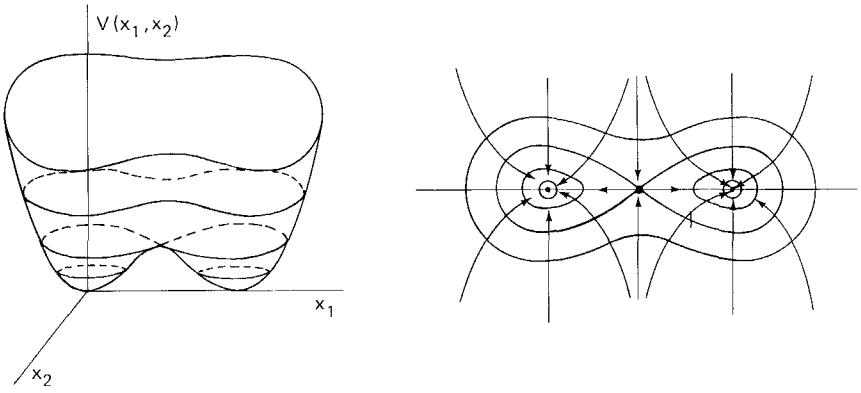


Figure 3.1. Gradient systems.

3.1. *Gradient systems*[7]

A dynamical system on X, $\dot{x}=f(x)$ is a *gradient system* if there is some function $V: X\to R$ such that $f(x)\equiv -\mathrm{D}V(x)$. The function $V(x)$ is often referred to as the *potential function* of the system; $f(x)$ is called the gradient of V at x.

There is an important geometrical interpretation of gradient systems. In Figure 3.1 we have drawn a graph of a potential function V: $R^2\to R$ as well as some level sets of this function in R^2.

The *directional derivative* of $V(x)$ in the direction $h=(h_1,\ldots,h_n)$, $\|h\|=1$, is defined to be $\mathrm{D}V(x)\cdot h$. The directional derivative measures how fast the function V is increasing in the direction h; as the above formula indicates, it is just the projection of $\mathrm{D}V(x)$ on the vector h. It is therefore clear that this projection will be maximized when $\mathrm{D}V(x)$ itself points in the direction h. Thus we have a nice geometrical interpretation of the gradient: *it points in the direction where V increases most rapidly*.

Furthermore, it is not hard to see that $\mathrm{D}V(x)$ *must be orthogonal to the level set of V at x*. For the level set of V at x is just that set of points where the value of V remains constant. Hence the directional derivative of V in a direction tangent to the level set of V at x must be zero. But this says that $\mathrm{D}V(x)$ is orthogonal to any such tangent vector, and hence is orthogonal to the level set itself.

These observations make it quite easy to construct the trajectories of $\dot{x}=-\mathrm{D}V(x)$ once the function V is known. A typical case is given in Figure 3.1. Some other special properties of gradient systems are given in:

[7]Gradient systems arise naturally in economics whenever one considers algorithms for maximizing or minimizing some function. See for example, Arrow–Hurwicz–Uzawa (1958).

Theorem

Let f: $X \to R^n$ *be given by* $\dot{x} = f(x) = -\mathrm{D}V(x)$ *where* V: $X \to R$ *is some smooth function. Then*:

(1) *if* x^* *is an isolated minimum of* V, x^* *is an asymptotically stable equilibrium of* $\dot{x} = -\mathrm{D}V(x)$;
(2) *any* ω*-limit point of a trajectory is an equilibrium*;
(3) *the eigenvalues of* $\mathrm{D}f(x)$ *are real at all* x.

Proof

See Hirsch and Smale (1974, pp. 199–209).

Item (3) of the theorem follows because $\mathrm{D}f(x)$ is just $\mathrm{D}^2V(x)$ and thus must be a real symmetric matrix. It is often useful to know that the converse is true. If we have a dynamical system on X, $\dot{x} = f(x)$ such that $\mathrm{D}f(x)$ is everywhere a real symmetric matrix, then there exists some potential function V: $X \to R$ such that $f(x) = -\mathrm{D}V(x)$. [For a precise statement of this result, the Frobenius Theorem, see Hartman (1969, ch. 6).]

Example 7

We consider a rather stylized Walrasian model where all consumers have utility functions linear in money. The utility maximization problem for consumer i is just

$$\max u_i(x_i) + m_i \quad \text{subject to} \quad p \cdot x_i + m_i = w_i,$$

where

$x_i = i$'s demand for goods $(x_i^1, \ldots, x_i^k)$,
$m_i = i$'s demand for money,
$w_i = i$'s initial holdings of money,
p = vector of prices $p_1 \cdots p_k$.

Agent i's demand function $x_i(p)$ must satisfy the first-order conditions,

$$\partial u_i(x_i(p))/\partial x_i^j = p_j, \qquad j = 1, \ldots, k,$$

or, in vector notation,

$$\mathrm{D}u_i(x_i(p)) = p.$$

Differentiating this identity with respect to p we get

$$D^2u_i(x_i(p))\cdot Dx_i(p)=I,$$

or

$$Dx_i(p)=\left[D^2u_i(x_i(p))\right]^{-1}.$$

Thus the Jacobian of each agent's demand function is just the inverse of the Hessian of the utility function.

Now we let ω be some aggregate supply of the k goods and define the aggregate excess demand function $z(p)=\sum_{i=1}^{n}x_i(p)-\omega$. Consider the dynamical system $\dot{p}=z(p)$. By the above calculations $Dz(p)$ is a real symmetric matrix so we have a gradient system. It is not too difficult to discover the potential function for this system. We let $v_i(p)=u_i(x_i(p))$ be the *indirect utility function* of agent i. Then a potential function for the system $\dot{p}=z(p)$ is just given by

$$V(p)=\sum_{i=1}^{n}v_i(p)+p\cdot\omega.$$

Further properties follow rather quickly. If we assume $u_i(x_i)$ is a strictly concave function, $D^2u_i(x)$ will be a negative definite matrix. It thus has all negative eigenvalues. Applying some previous results we see that the system has a unique globally stable equilibrium p^* which in fact minimizes the sum of the indirect utility functions.

3.2. *Hamiltonian systems*[8]

Let $\dot{x}=f(x,y)$, $\dot{y}=g(x,y)$ be a dynamical system for x and y on $X\times Y$ contained in $R^n\times R^n$. This system is called a *Hamiltonian system* if there is some function H: $X\times Y\to R$, the Hamiltonian function, such that

$$\dot{x}=f(x,y)=D_yH(x,y),$$
$$\dot{y}=g(x,y)=-D_xH(x,y).$$

Hamiltonian systems arise quite naturally in classical mechanics and serve to unify the study of many phenomena in this area. Economists have recently become aware of their many natural applications in economics.

[8]For many applications of Hamiltonian systems to problems in economic growth, see Cass and Shell (1976) and the cited works therein.

The primary feature of Hamiltonian systems for economic applications is that they have certain desirable stability properties. In the classical theory of Hamiltonian mechanics, H was quadratic so that the Hamiltonian system was a linear system of differential equations. In this case a classical theorem of Poincaré shows that if λ is an eigenvalue of the linear system at (x^*, y^*) then $-\lambda$ is also an eigenvalue. Thus the equilibrium of the Hamiltonian system are symmetric saddle points. In the general case, where the Hamiltonian is nonlinear, the same kind of saddle point property occurs when the function is concave in x and convex in y.

4. Some newer techniques

In this section we will survey two newer areas of the study of dynamical systems and discuss their potential applications in economics.

4.1. Structural stability[9]

Let f: $X \to R^n$ define a vector field on some state space X. Then, roughly speaking, this system is *structurally stable* if small perturbations in the function f do not change the topological structure of the vector field $\dot{x} = f(x)$. Consider for example the case where $X = R^2$ and $f(x) = Ax$, where A is a 2×2 non-singular matrix. Then we know that the origin is the unique equilibrium of the system and the topological nature of the flow around the origin is determined by the nature of the eigenvalues of the matrix A.

For "most" choices of A, the system given by $\dot{x} = Ax$ will be structurally stable since small perturbations in A will not change the signs of the eigenvalues. The one exception is when both eigenvalues have real part zero. In this case, the flow of the system consists of closed orbits surrounding the origin. However, any small perturbation of A that gives the eigenvalues a nonzero real part will exhibit a flow with no closed orbits at all. The topological structure of the system exhibits a drastic change — we have a case of structural instability.

Let us now return to the general setting of a vector field $\dot{x} = f(x)$. We will take the state space of this system to be D^n. We let $\mathcal{V}$ be the space of all continuously differentiable functions from D^n to R^n, and we endow $\mathcal{V}$ with the standard C^1 norm; i.e., two functions are close if their values are close and their derivatives are close. We can then think of a *perturbation* of f as being a choice of any function in some ε-ball around f.

[9] A good reference for the mathematical results concerning structural stability is the book of Nitecki (1971). A brief survey is given in Hirsch and Smale (1974, ch. 16).

We want the topological structure of $\dot{x}=f(x)$ to be invariant with respect to small perturbations of f. What does this mean? How do we describe the notion that two vector fields have the same qualitative features?

The relevant concept is that of *topological equivalence*. Roughly speaking the flow of two dynamical systems on D^n are topologically equivalent if there is a homeomorphism h: $D^n \to D^n$ that carries the orbits of one flow onto the orbits of the other. We can think of this homeomorphism as being some continuous change of co-ordinates, so that topological equivalence of two flows just means that we can find a continuous change of coordinates so that one flow looks like the other.

Finally we define the concept of *structural stability*. A dynamical system $\dot{x}=f(x)$ on D^n is structurally stable if there is some neighborhood of f such that for every function g in that neighborhood, the flow induced by $\dot{x}=g(x)$ is topologically equivalent to the flow of f. Loosely speaking, a dynamical system is structurally stable if small perturbations in the underlying function f do not change the qualitative nature of the flow.

4.2. *Catastrophe theory*[10]

Let us consider some dynamical system given by f: $X \times A \to R^n$, $\dot{x}=f(x, a)$. Here the system is thought of as *parameterized* by some parameters $a=(a_1, \ldots, a_r)$. Now suppose we think of the parameters a as changing slowly over time. Most of the time small changes in a will not result in radical changes in the qualitative nature of the dynamical system. However, sometimes we will get real structural change.

For example, consider the system on R^1 given by

$$\dot{x}=x^2+a.$$

If a is positive, there are no equilibria of this system. If a is zero there is one equilibrium, $x^*=0$; and if a is negative, there are two equilibria at $x_1^*=-a^{1/2}$, $x_2^*=+a^{1/2}$.

The topological nature of the system undergoes a radical change as a passes through zero. We say zero is a *catastrophe point* for the system $\dot{x}=x^2+a$.

The goal of catastrophe theory is to classify all the ways in which a system can undergo structural change. Unfortunately, this goal is a long way off. The current state of theory is well developed only for studying *local catastrophes of gradient systems*.

[10] Basic expositions of catastrophe theory can be found in Golubitsky (1978) and Thom (1975). A nice application of the theory is given in Zeeman (1972). Economic applications are presented in Varian (1979) and Zeeman (1974).

Let V: $R^n \times R^r \to R$ be a potential function for a gradient system. Here R^n is interpreted as the state space of the system and R^r is interpreted as a parameter space. Then the equilibria of the system,

$$\dot{x} = D_x V(x, a),$$

are precisely the singularities of the function $V(x, a)$; i.e., x^* is an equilibrium if and only if $D_x V(x, a)$ vanishes. Thus the study of how the nature of the system $\dot{x} = D_x(x, a)$ changes as a changes can be reduced to the study of the singularities of $V(x, a)$.

The example given earlier of $\dot{x} = x^2 + a$ fits into this framework since it is a gradient system with $V(x, a) = x^3/3 + ax$.

Now the remarkable thing is that for $r \leqq 4$, there are only seven distinct kinds of "stable" singularities. These are the seven elementary catastrophes of Thom's Classification Theorem. Roughly speaking, any "non-degenerate" singularity of $V(x, a)$ can be classified as one of these seven elementary types. The example given earlier where $V(x, a) = x^3/3 + ax$ is an example of the *fold catastrophe*, the simplest of the elementary catastrophes.

References

Arrow, K. and F. Hahn (1971), General competitive analysis. San Francisco, CA: Holden Day. Now distributed by North-Holland, Amsterdam.

Arrow, K., L. Hurwicz and H. Uzawa (1958), Studies in linear and nonlinear programming. Stanford, CA: Stanford University Press.

Cass, D. and K. Shell (1976), "Introduction to Hamiltonian dynamics in economics", Journal of Economic Theory, 12:1–10.

Chang, W. and D. Smyth (1971), "The existence and persistence of cycles in a nonlinear model: Kaldors' 1940 model re-examined", Review of Economic Studies, 38:37–45.

Coddington, E. and N. Levinson (1955), Theory of ordinary differential equations. New York: McGraw-Hill.

Dierker, E. (1972), "Two remarks on the number of equilibria of an economy", Econometrica, 40:951–953.

Dierker, E. (1974), Topological methods in Walrasian economics, Lecture notes in economics and mathematical systems, Vol. 92. Berlin: Springer-Verlag.

Golubitsky, M. (1978), "An introduction to catastrophe theory and its applications", SIAM Review, 20:352–387.

Guillemin, V. and A. Pollack (1974), Differential topology. Englewood Cliffs, NJ: Prentice-Hall.

Hartman, P. (1964), Ordinary differential equations. New York: Wiley.

Hirsch, M. and S. Smale (1974), Differential equations, dynamical systems, and linear algebra. New York: Academic Press.

Ichimura, S. (1954), "Towards a general nonlinear macrodynamic theory of economic fluctuations", in: K. Kurihara, ed., Post Keynesian economics. New Brunswick, NJ: Rutgers University Press.

Luenberger, D. (1979), Introduction to dynamical systems. New York: Wiley.

Milnor, J. (1965), Topology from the differentiable viewpoint. Charlottesville, VA: University of Virginia Press.

Nitecki, Z. (1971), Differentiable dynamics, Cambridge, MA: M.I.T. Press.

Spanier, E. (1966), Algebraic topology. New York: McGraw-Hill.

Thom, R. (1975), Structural stability and morphogenesis. Reading, MA: W. A. Benjamin.
Varian, H. (1975), "A third remark on the number of equilibria of an economy", Econometrica, 43: 985–986.
Varian, H. (1977a), "A remark on boundary restrictions in the global Newton method", Journal of Mathematical Economics, 4:127–130.
Varian, H. (1977b), "Nonwalrasian equilibria", Econometrica, 45:573–590.
Varian, H. (1978), Microeconomic analysis. New York: W. W. Norton.
Varian, H. (1979), "Catastrophe theory and the business cycle", Economic Inquiry, 17:14–28.
Zeeman, E. (1972), "Differential equations for the heartbeat and nerve impulses", in: C. H. Waddington, ed., Towards a theoretical biology, Vol. 4. Edinburgh: Edinburgh University Press. Also in: M. M. Peixoto, ed., Dynamical systems. New York: Academic Press.
Zeeman, E. (1974), "On the unstable behavior of stock market exchanges", Journal of Mathematical Economics, 1:39–50.

Chapter 4

CONTROL THEORY WITH APPLICATIONS TO ECONOMICS

DAVID KENDRICK*

University of Texas

1. Introduction

Control theory methods are used to find the optimal set of policies over time for a deterministic or stochastic system. Since a large number of economic problems are naturally described as dynamic systems which can be influenced by policies in an attempt to improve their performance, control theory has gained widespread application by economists. Also, stochastic elements are common in economics in equations errors, unknown parameters, and measurement errors, so the methods of stochastic and adaptive control are finding substantial numbers of applications in economics.[1]

This paper describes deterministic, stochastic, and adaptive control theory methods. In deterministic methods there are no uncertain elements, in stochastic approaches there are random elements but there is no purposeful effort to learn about (i.e., improve estimates of) these elements, and in adaptive (or dual) control there is an attempt to actively learn the value of uncertain elements. As a prelude to this material the reader who has not studied the calculus of variations, dynamic programming, and control theory might do well to read the chapters in Intriligator (1971) on these subjects.

The approach to control theory taken in this paper is partly mathematical and partly algorithmic, i.e., not only is the derivation of the optimality conditions given by also there is a discussion of the numerical methods employed to obtain the solution to the problem. This approach arises out of the author's conviction that many economic problems of interest cannot be solved analytically. Also, the focus is on the path of dynamic systems from the present status to an improved state rather than on the steady state solution to dynamic problems.

*The author wishes to express appreciation to the following individuals: Rick Ashley, Yaakov Bar-Shalom, Roger Craine, Ray Fair, Ken Garbade, Leif Johansen, Bo Hyun Kang, David Livesey, Peggy Mills, Homa Motamen, Fred Norman, Bob Pindyck, Jorge Rizo-Patron, Edison Tse, Stephen Turnovsky, and John Westcott. This research was supported by NSF SOC 76-11187.

[1]Surveys of applications of control theory to economics have been written by Arrow (1968), Dobell (1969), Aoki (1974b), Intriligator (1975), Athans and Kendrick (1974), and Kendrick (1976). Some of the principal books in the field of economics and control theory are Aoki (1976), Chow (1975), Pitchford and Turnovsky (1977), and Murphy (1965).

Thus the concern here is not so much on where the system will ultimately be as on the study of the paths from present circumstance to future position and on the paths of the policy variables during the period. The reader whose primary interest is in analytical solutions and steady state solutions is referred to Shell (1967).

2. Deterministic control

Consider the problem of finding $[u_k]_{k=0}^{N-1}=(u_0, u_1, \ldots, u_{N-1})$ to minimize the criterion function

$$J=L_N(x_N)+\sum_{k=0}^{N-1} L_k(x_k, u_k), \tag{2.1}$$

subject to the system equations

$$x_{k+1}=f_k(x_k, u_k), \quad k=0,1,\ldots,N-1, \tag{2.2}$$

$$x_0 \quad \text{given}, \tag{2.3}$$

where $x=n$-element state vector, $u=m$-element control vector, and $f=$vector-valued function specifying the n systems equations.

Problem (2.1)–(2.3) can be solved by a variety of methods among which two of the most common are the successive approximations[2] approach used by Garbade (1975a),[3] and gradient methods, viz the conjugate gradient method employed by Kendrick and Taylor (1970).

The successive approximation approach is a good beginning point not only because it has been used in solving economic models but also because the approximation employed is like the quadratic-linear tracking problem which has also been widely used in formulating economic problems as control theory models.

The approximation involves a second-order expansion of the criterion function about the path $[x_{ok}, u_{ok}]_{k=0}^{N}$,

$$\begin{aligned} J \approx & L'_{xN}(x_N-x_{oN})+\tfrac{1}{2}(x_N-x_{oN})'L_{xx,N}(x_N-x_{oN}) \\ & +\sum_{k=0}^{N-1}\left\{(L'_{xk}, L'_{uk})\begin{bmatrix} x_k-x_{ok} \\ u_k-u_{ok}\end{bmatrix}\right. \\ & \left.+\frac{1}{2}[(x_k-x_{ok})', (u_k-u_{ok})']\left[\begin{array}{c|c} L_{xx} & L_{xu} \\ \hline L_{ux} & L_{uu}\end{array}\right]_k\begin{bmatrix} x_k-x_{ok} \\ u_k-u_{ok}\end{bmatrix}\right\}, \end{aligned} \tag{2.4}$$

[2]Also sometimes called linearized linear-quadratic.
[3]In Garbade's model the systems equations (2.2) are in implicit rather than in explicit form.

and a first-order expansion of the systems equations,

$$x_{k+1} \approx f_k(x_{ok}, u_{ok}) + f_{xk}(x_k - x_{ok}) + f_{uk}(u_k - u_{ok}), \quad k = 0, 1, \ldots, N-1, \quad (2.5)$$

where[4]

$$L_{xk} = \begin{bmatrix} \dfrac{\partial L_k}{\partial x_{1k}} \\ \vdots \\ \dfrac{\partial L_k}{\partial x_{nk}} \end{bmatrix}, \qquad L_{xx,k} = \begin{bmatrix} \dfrac{\partial^2 L_k}{\partial x_{1k}\partial x_{1k}} & \cdots & \dfrac{\partial^2 L_k}{\partial x_{1k}\partial x_{nk}} \\ \vdots & & \vdots \\ \dfrac{\partial^2 L_k}{\partial x_{nk}\partial x_{1k}} & \cdots & \dfrac{\partial^2 L_k}{\partial x_{nk}\partial x_{nk}} \end{bmatrix},$$

$$L_{xu,k} = \begin{bmatrix} \dfrac{\partial^2 L_k}{\partial x_{1k}\partial u_{1k}} & \cdots & \dfrac{\partial^2 L_k}{\partial x_{1k}\partial u_{mk}} \\ \vdots & & \vdots \\ \dfrac{\partial^2 L_k}{\partial x_{nk}\partial u_{1k}} & \cdots & \dfrac{\partial^2 L_k}{\partial x_{nk}\partial u_{mk}} \end{bmatrix}, \qquad L_{ux,k} = L'_{xu,k},$$

$$L_{uu,k} = \begin{bmatrix} \dfrac{\partial^2 L_k}{\partial u_{1k}\partial u_{1k}} & \cdots & \dfrac{\partial^2 L_k}{\partial u_{1k}\partial u_{mk}} \\ \vdots & & \vdots \\ \dfrac{\partial^2 L_k}{\partial u_{mk}\partial u_{1k}} & \cdots & \dfrac{\partial^2 L_k}{\partial u_{mk}\partial u_{mk}} \end{bmatrix},$$

$$f_{xk} = \begin{bmatrix} f_{xk}^{1'} \\ \vdots \\ f_{xk}^{n'} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial f_k^1}{x_{1k}} & \cdots & \dfrac{\partial f_k^1}{x_{nk}} \\ \vdots & & \vdots \\ \dfrac{\partial f_k^n}{\partial x_{1k}} & \cdots & \dfrac{\partial f_k^n}{\partial x_{nk}} \end{bmatrix}, \qquad f_{uk} = \begin{bmatrix} f_{uk}^{1'} \\ \vdots \\ f_{uk}^{n'} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial f_k^1}{\partial u_{1k}} & \cdots & \dfrac{\partial f_k^1}{\partial u_{mk}} \\ \vdots & & \vdots \\ \dfrac{\partial f_k^n}{\partial u_{1k}} & \cdots & \dfrac{\partial f_k^n}{\partial u_{mk}} \end{bmatrix},$$

and all derivatives are evaluated on the path $[x_{ok}, u_{ok}]_{k=0}^{N}$.[5]

[4] This notation differs from the usual procedure of treating the gradient vector ∇f of a single function with respect to a number of arguments as a row vector. This is done so that all vectors are treated as column vectors unless they are explicitly transposed. Note here that subscript o is used to denote the nominal path, and that this should be distinguished from the subscript 0 which is used to denote time period zero.

[5] f^i refers to the ith function in the vector of functions in the set of systems equations (2.2).

An initial path for $[x_{ok}, u_{ok}]_{k=0}^{N}$ is chosen in the neighborhood of the expected optimal solution and the problem (2.4)–(2.5) is solved for the $[u_k^*]_{k=0}^{N-1}$ which minimizes (2.4). Then a new path $[u_{ok}^{\text{new}}]_{k=0}^{N-1}$ is chosen with

$$u_{ok}^{\text{new}} = \alpha[u_k^* - u_{ok}] + u_{ok}, \qquad k = 0, 1, \ldots, N-1, \tag{2.6}$$

where α = step size; and the new state path $[x_{ok}^{\text{new}}]$ is calculated from

$$x_{o,k+1}^{\text{new}} = f(x_{ok}^{\text{new}}, u_{ok}^{\text{new}}), \qquad k = 0, 1, \ldots, N-1. \tag{2.7}$$

The old path $[x_{ok}, u_{ok}]_{k=0}^{N}$ is then replaced with the new path $[x_{ok}^{\text{new}}, u_{ok}^{\text{new}}]_{k=0}^{N}$, and the problem (2.4)–(2.5) is solved again. This procedure is repeated until convergence is obtained.[6]

2.1. *Quadratic-linear problems*

In many applications of control theory to economics, viz Pindyck (1972, 1973a) and Chow (1975), the quadratic-linear tracking model is used.[7] Also, approximation methods are used to solve many nonlinear deterministic and stochastic problems, and one element of these approximation procedures is frequently the solution of quadratic-linear problems. In the quadratic-linear tracking problem one seeks to find the optimal control $[u_k^*]_{k=0}^{N-1}$ to guide the economic systems as closely as possible to a desired path $[\tilde{x}_k]_{k=1}^{N}$ without deviating too far from a desired control path $[\tilde{u}_k]_{k=0}^{N-1}$. The criterion function employed in this approach is normally of the form

$$J = \tfrac{1}{2}(x_N - \tilde{x}_N)' W_N (x_N - \tilde{x}_N) + \tfrac{1}{2} \sum_{k=0}^{N-1} \left[(x_k - \tilde{x}_k)' W_k (x_k - \tilde{x}_k) + (u_k - \tilde{u}_k)' \Lambda_k (u_k - \tilde{u}_k) \right], \tag{2.8}$$

and the systems equations are[8]

[6] See Garbade (1975a, ch. 2, and 1975b).

[7] Other examples of the applications of deterministic quadratic-linear control methods in economic problems are Tustin (1953), Bogaard and Theil (1959), van Eijk and Sandee (1959), Holt (1962), Theil (1964), (1965), Erickson, Leondes and Norton (1970), Sandblom (1970), Thalberg (1971a, b), Paryani (1972), Friedman (1972), Erickson and Norton (1973), Shupp (1976a, 1977), Tinsley, Craine and Havenner (1974), You (1975), Kaul and Rao (1975), Fischer and Uebe (1975), and Oudet (1976).

[8] The difference equations in the economic model frequently have lags longer than the single period shown here. The common practice in control theory is to convert these systems of nth order difference equations to a set of n first-order difference equations by augmenting the state variable. Norman and Jung (1977) show that from a computational point of view, it may be better not to augment the system.

$$x_{k+1}=A_k x_k+B_k u_k+c_k, \qquad k=0,1,\ldots,N-1, \tag{2.9}$$

with

$$x_0 \quad \text{given}, \tag{2.10}$$

W, Λ are weighting matrices $(n\times n)$ and $(m\times m)$, and A, B, c are parameter matrices $(n\times n)$ and $(n\times m)$ and constant term vector $(n\times 1)$, respectively.

In order to discuss the solution of the two models (2.4)–(2.5) and (2.8)–(2.10), it is useful to relate them to a common problem. This problem is written as: find $[u_k]_{k=0}^{N-1}$ to minimize

$$\begin{aligned} J&=L_N(x_N)+\sum_{k=0}^{N-1} L_k(x_k,u_k)\\ &=\tfrac{1}{2}x'_N W_N x_N+w'_N x_N\\ &\quad+\sum_{k=0}^{N-1}\left\{\tfrac{1}{2}x'_k W_k x_k+w'_k x_k+x'_k F_k u_k+\tfrac{1}{2}u'_k \Lambda_k u_k+\lambda'_k u_k\right\}, \end{aligned} \tag{2.11}$$

subject to

$$x_{k+1}=A_k x_k+B_k u_k+c_k, \qquad k=0,1,\ldots,N-1, \tag{2.12}$$

$$x_0 \quad \text{given}. \tag{2.13}$$

Table 2.1 provides the equivalence between the notations of problem (2.11)–(2.13) and problems (2.4)–(2.5) and (2.8)–(2.10). The constant terms have been dropped from the criterion function since they do not affect the choice of the optimal control.

Table 2.1
Notational equivalence for quadratic-linear problems.

Problem (2.11)–(2.13)	Problem (2.4)–(2.5)	Problem (2.8)–(2.10)
W_N	$L_{xx,N}$	W_N
w_N	$L_{x,N}-L'_{xx,N}x_{on}$	$-W_N\tilde{x}_N$
w_k	$L_{xk}-L'_{xx,k}x_{ok}-L_{xu,k}u_{ok}$	$-W_k\tilde{x}_k$
F_k	$L_{xu,k}$	0
Λ_k	$L_{uu,k}$	Λ_k
λ_k	$L_{uk}-L'_{uu,k}u_{ok}-L'_{xu}x_{ok}$	$-\Lambda_k\tilde{u}_k$
A_k	f_{xk}	A_k
B_k	f_{uk}	B_k
c_k	$-(f_{xk}x_{ok}+f_{uk}u_{ok})$	c_k

Now assume that the optimal cost-to-go[9] can be written as the quadratic form[10]

$$J^*(k)=\tfrac{1}{2}x_k'K_kx_k+p_k'x_k. \tag{2.14}$$

Then, the optimal cost-to-go at time N is

$$J^*(N)=\tfrac{1}{2}x_N'K_Nx_N+p_N'x_N. \tag{2.15}$$

Then by inspection of (2.11) and (2.15), we have

$$K_N=W_N, \tag{2.16}$$

$$p_N=w_N. \tag{2.17}$$

Next apply the principle of optimality from dynamic programming to compute the optimal control for $N-1$. From (2.11) and (2.15),

$$J^*(N-1)=\min_{u_{N-1}}\{J^*(N)+L_{N-1}(x_{N-1},u_{N-1})\}, \tag{2.18}$$

and

$$\begin{aligned}J^*(N-1)=\min_{u_{N-1}}\{&\tfrac{1}{2}x_N'K_Nx_N+p_N'x_N+\tfrac{1}{2}x_{N-1}'W_{N-1}x_{N-1}+w_{N-1}'x_{N-1}\\&+x_{N-1}'F_{N-1}u_{N-1}+\tfrac{1}{2}u_{N-1}'\Lambda_{N-1}u_{N-1}+\lambda_{N-1}'u_{N-1}\}.\end{aligned} \tag{2.19}$$

Substitute (2.12) into (2.19),

$$\begin{aligned}J^*(N-1)=\min_{u_{N-1}}\{&\tfrac{1}{2}(A_{N-1}x_{N-1}+B_{N-1}u_{N-1}+c_{N-1})'\\&\cdot K_N(A_{N-1}x_{N-1}+B_{N-1}u_{N-1}+c_{N-1})\\&+p_N'(A_{N-1}x_{N-1}+B_{N-1}u_{N-1}+c_{N-1})+\tfrac{1}{2}x_{N-1}'W_{N-1}x_{N-1}\\&+w_{N-1}'x_{N-1}+\tfrac{1}{2}u_{N-1}'\Lambda_{N-1}u_{N-1}+\lambda_{N-1}'u_{N-1}\\&+x_{N-1}'F_{N-1}u_{N-1}\},\end{aligned} \tag{2.20}$$

and

$$\begin{aligned}J^*(N-1)=\min_{u_{N-1}}\{&\tfrac{1}{2}x_{N-1}'\Phi_{N-1}x_{N-1}+\tfrac{1}{2}u_{N-1}'\Theta_{N-1}u_{N-1}+x_{N-1}'\Psi_{N-1}u_{N-1}\\&+\phi_{N-1}'x_{N-1}+\theta_{N-1}'u_{N-1}+\eta_{N-1}\},\end{aligned} \tag{2.21}$$

[9]The cost-to-go is frequently written with a constant term in addition to the K and p terms. However, the constant term does not affect the solution and so it is dropped here.

[10]For a discussion of the dynamic programming methods that are used to solve this class of problems, see Intriligator (1971, ch. 13) or Kendrick (1981, ch. 2).

where

$$\begin{aligned}
\Phi_{N-1} &= A'_{N-1}K_N A_{N-1} + W_{N-1},\\
\Theta_{N-1} &= B'_{N-1}K_N B_{N-1} + \Lambda_{N-1},\\
\Psi_{N-1} &= A'_{N-1}K_N B_{N-1} + F_{N-1},\\
\phi_{N-1} &= A'_{N-1}(K'_N c_{N-1} + p_N) + w_{N-1},\\
\theta_{N-1} &= B'_{N-1}(K'_N c_{N-1} + p_N) + \lambda_{N-1},\\
\eta_{N-1} &= c'_{N-1}K_N c_{N-1} + p'_N c_{N-1}.
\end{aligned} \tag{2.22}$$

Then minimizing over u_{N-1} in (2.21) yields

$$u'_{N-1}\Theta_{N-1} + x'_{N-1}\Psi_{N-1} + \theta'_{N-1} = 0. \tag{2.23}$$

Thus the optimal feedback rule is obtained from (2.23) as[11]

$$u_{N-1} = G_{N-1}x_{N-1} + g_{N-1}, \tag{2.24}$$

where

$$G_{N-1} = -\Theta_{N-1}^{-1}\Psi'_{N-1}, \qquad g_{N-1} = -\Theta_{N-1}^{-1}\theta_{N-1}. \tag{2.25}$$

Next substitute (2.24) back into (2.21) to obtain

$$\begin{aligned}
J^*(N-1) = {} & \tfrac{1}{2}x'_{N-1}(\Phi_{N-1} + G'_{N-1}\Theta_{N-1}G_{N-1} + 2\Psi_{N-1}G_{N-1})x_{N-1}\\
& + (g'_{N-1}[\Theta_{N-1}G_{N-1} + \Psi'_{N-1}] + \phi'_{N-1} + \theta'_{N-1}G_{N-1})x_{N-1}\\
& + \tfrac{1}{2}g'_{N-1}\Theta_{N-1}g_{N-1} + \theta'_{N-1}g_{N-1} + \eta_{N-1}.
\end{aligned} \tag{2.26}$$

Thus, from (2.14) and (2.26), we have

$$J^*(N-1) = \tfrac{1}{2}x'_{N-1}K_{N-1}x_{N-1} + p'_{N-1}x_{N-1}, \tag{2.27}$$

with

$$K_{N-1} = \Phi_{N-1} + G'_{N-1}\Theta_{N-1}G_{N-1} + 2\Psi_{N-1}G_{N-1}, \tag{2.28}$$

$$p_{N-1} = (\Psi_{N-1} + G'_{N-1}\Theta'_{N-1})g_{N-1} + G'_{N-1}\theta_{N-1} + \phi_{N-1}. \tag{2.29}$$

The steps above can be repeated for J^*_{N-2} with the result that

$$u_{N-2} = G_{N-2}x_{N-2} + g_{N-2}, \tag{2.30}$$

[11]The assumption that W and Λ are symmetric is used here. There is no loss of generality since they appear only in a quadratic form.

where

$$G_{N-2} = -\Theta_{N-2}^{-1}\Psi_{N-2}', \qquad g_{N-2} = -\Theta_{N-2}^{-1}\theta_{N-2}, \tag{2.31}$$

with

$$\begin{aligned}
\Phi_{N-2} &= A_{N-2}' K_{N-1} A_{N-2} + W_{N-2},\\
\Theta_{N-2} &= B_{N-2}' K_{N-1} B_{N-2} + \Lambda_{N-2},\\
\Psi_{N-2} &= A_{N-2}' K_{N-1} B_{N-2} + F_{N-2},\\
\phi_{N-2} &= A_{N-2}' K_{N-1} c_{N-2} + A_{N-2}' p_{N-1} + w_{N-2},\\
\theta_{N-2} &= B_{N-2}' K_{N-1} c_{N-2} + B_{N-2}' p_{N-1} + \lambda_{N-2},\\
\eta_{N-2} &= c_{N-2}' K_{N-1} c_{N-2} + p_{N-1}' c_{N-2}.
\end{aligned} \tag{2.32}$$

Then substitution of (2.30) back into the expression for $J^*(N-2)$ as in (2.26) for period $N-1$ yields

$$J^*(N-2) = \tfrac{1}{2} x_{N-2}' K_{N-2} x_{N-2} + p_{N-2}' x_{N-2}, \tag{2.33}$$

where

$$K_{N-2} = \Phi_{N-2} + G_{N-2}' \Theta_{N-2} G_{N-2} + 2\Psi_{N-2} G_{N-2}, \tag{2.34}$$

$$p_{N-2} = (\Psi_{N-2} + G_{N-2}' \Theta_{N-2}') g_{N-2} + G_{N-2}' \theta_{N-2} + \phi_{N-2}. \tag{2.35}$$

Comparing (2.28)–(2.29) and (2.34)–(2.35), we have

$$K_j = \Phi_j + G_j' \Theta_j G_j + 2\Psi_j G_j, \tag{2.36}$$

$$p_j = (\Psi_j + G_j' \Theta_j') g_j + G_j' \theta_j + \phi_j, \tag{2.37}$$

and from (2.22) and (2.32),

$$\begin{aligned}
\Phi_j &= A_j' K_{j+1} A_j + W_j,\\
\Theta_j &= B_j' K_{j+1} B_j + \Lambda_j,\\
\Psi_j &= A_j' K_{j+1} B_j + F_j,\\
\phi_j &= A_j' (K_{j+1} c_j + p_{j+1}) + w_j,\\
\theta_j &= B_j' (K_{j+1} c_j + p_{j+1}) + \lambda_j,\\
\eta_j &= c_j' K_{j+1} c_j + p_{j+1}' c_j.
\end{aligned} \tag{2.38}$$

So, in summary from (2.24)–(2.25) and (2.30)–(2.31) the optimal control is given by

$$u_j = G_j x_j + g_j, \tag{2.39}$$

where[12]

$$G_j = -\Theta_j^{-1}\Psi_j' = -\left[B_j' K_{j+1} B_j + \Lambda_j\right]^{-1}\left[F_j' + B_j' K_{j+1} A_j\right], \tag{2.40}$$

$$g_j = -\Theta_j^{-1}\theta_j = -\left[B_j' K_{j+1} B_j + \Lambda_j\right]^{-1}\left[B_j'(K_{j+1} c_j + p_{j+1}) + \lambda_j\right]. \tag{2.41}$$

Now (2.36)–(2.37) can be simplified somewhat by substituting (2.40), (2.41) and then (2.38) into them to obtain

$$\begin{aligned} K_j &= \Phi_j + \Psi_j \Theta_j^{-1}\Psi_j' - 2\Psi_j\Theta_j^{-1}\Psi_j' = \Phi_j - \Psi_j\theta_j^{-1}\Psi_j \\ &= A_j' K_{j+1} A_j - (A_j' K_{j+1} B_j + F_j)(B_j' K_{j+1} B_j + \Lambda_j)^{-1}(F_j' + B_j' K_{j+1} A_j) + W_j, \end{aligned} \tag{2.42}$$

with [from (2.16)] $K_N = W_N$. Also

$$\begin{aligned} p_j &= \left(\Psi_j - \Psi_j\Theta_j^{-1}\Theta_j\right) g_j - \Psi_j\Theta_j^{-1}\theta_j + \phi_j = -\Psi_j\theta_j^{-1}\theta_j + \phi_j \\ &= -\left[A_j' K_{j+1} B_j + F_j\right]\left[B_j' K_{j+1} B_j + \Lambda_j\right]^{-1}\left[B_j'(K_{j+1} c_j + p_{j+1}) + \lambda_j\right] \\ &\quad + A_j'(K_{j+1} c_j + p_{j+1}) + w_j, \end{aligned} \tag{2.43}$$

with [from (2.17)] $p_N = w_N$.

Thus the solution to both the quadratic-linear approximation problem (2.4)–(2.5) and the quadratic-linear tracking problem is given by the feedback rule (2.39) and the relationships (2.40)–(2.43) along with the notational equivalence given in Table 2.1.

The problem is solved numerically by integrating the Riccati equations for K and p [(2.42) and (2.43)] backward in time from the terminal conditions. The matrices of the feedback rule (2.39) can be obtained from (2.40) and (2.41). Then (2.12) and (2.39) are integrated forward in time from the initial state x_0 to obtain the control path $[u_k]_{k=0}^{N-1}$ and the state path $[x_k]_{k=1}^{N}$.

[12] For a discussion of the condition under which Θ is non-singular and what should be done when singularity occurs, see Garbade (1976).

2.2. *General nonlinear problems*[13]

A variety of methods have been employed to solve general nonlinear control problems.[14] Among the most widely used are methods of successive approximation and gradient methods. The successive approximation method was discussed in the introduction to this section, therefore, only gradient methods remain.

Consider again the problem (2.1)–(2.3) and define a scalar sequence H_k (which is analogous to the Hamiltonian in the continuous formulation) as

$$H_k = L_k(x_k, u_k) + \lambda'_{k+1} f_k(x_k, u_k). \tag{2.44}$$

Then the first-order conditions for the problem (2.1)–(2.3) may be written as[15]

optimality conditions

$$\partial H_k/\partial u_k = \partial L_k/\partial u_k + \lambda'_{k+1}(\partial f_k/\partial u_k) = 0, \tag{2.45}$$

costate equations

$$\lambda'_k = \partial L_k/\partial x_k + \lambda'_{k+1}(\partial f_k/\partial x_k), \qquad k = 1, \ldots, N-1, \tag{2.46}$$

with *terminal condition*

$$\lambda'_N = \partial L_N/\partial x_N, \tag{2.47}$$

and *system equations*

$$x_{k+1} = f_k(x_k, u_k), \qquad k = 0, \ldots, N-1, \tag{2.48}$$

and with *initial condition*

$$x_0 \quad \text{given.} \tag{2.49}$$

The gradient algorithm may then be described as:

(a) select a nominal control path $[u_k]_{k=0}^{N-1}$;

[13] Example of the application of nonlinear control theory to economic problems include Livesey (1971, 1978), Cheng and Wan (1972), Shupp (1972), Norman and Norman (1973), Fitzgerald, Johnston and Bayes (1973), Holbrook (1973, 1974, 1975), Woodside (1973), Friedman and Howrey (1973), Healey and Summers (1974), Sandblom (1974), Fair (1974, 1975, 1978a, b), Rouzier (1974), Healey and Medina (1975), Gupta, Meyer, Raines and Tarn (1975), Craine, Havenner and Tinsley (1976), Ando, Norman and Palash (1978), Athans, Kuh, et al. (1977), Palash (1977), and Klein (1979).

[14] For example, see Drud (1976).

[15] See Kendrick and Taylor (1971) and Bryson and Ho (1969, ch. 7, secs. 7 and 8).

(b) integrate the systems equations (2.48) forward in time using the nominal control path;
(c) at terminal time evaluate λ_N with (2.47);
(d) integrate the costate equations (2.46) backward in time using the nominal control variables and state variables from step (b);
(e) calculate the Hamiltonian (2.44) and its gradient with respect to u (2.45);[16]
(f) make a one-dimensional search in the gradient direction until the Hamiltonian is minimized, i.e., choose α to minimize

$$g(\alpha)=H[u-\alpha\nabla H'(u)], \tag{2.50}$$

where u is the current control and ∇H is the gradient of the Hamiltonian;
(g) calculate the new control path from the relationship

$$u_{\text{new}}=u-\alpha\nabla H(u) \tag{2.51}$$

and return to step (b).

Steps (b) through (g) are repeated until satisfactory convergence is obtained.

A variety of gradient methods can be employed in step (e). For example, the Murtagh and Saunders (1977) code MINOS includes a quasi-Newton method and five conjugate gradient methods:

(i) Fletcher and Reeves (1964),
(ii) Polak and Ribiere (1969),
(iii) Perry (1976),
(iv) Memoryless DFP [Davidon (1959), Fletcher and Powell (1963)],
(v) Memoryless Complementary DFP.

Also a variety of line search methods may be used in step (f). For example, Murtagh and Saunders use subroutines LNSRCH and NEWPTC from the NPL Algorithm Library,[17] and Mantell and Lasdon (1977) employ a method due to Shanno (1977).

Two characteristics of the deterministic control problem which have motivated much of the recent work in this field are:

(i) the large size of the models,
(ii) problems of numerical accuracy.

The accuracy problem arise from the fact that the systems equations in macroeconometric models are frequently in implicit function form, i.e., instead of

$$x_{k+1}=f_k(x_k, u_k), \tag{2.52}$$

[16]In practice it is usually preferable to use a modified gradient direction such as the conjugate gradient direction used by Kendrick and Taylor (1970). For a description of this procedure, see Lasdon, Mitter and Warren (1967) and Fletcher and Reeves (1964).
[17]See Gill, Murray, Pichen, Barber and Wright (1976).

they are in the form

$$g(x_{k+1}, x_k, u_k) = 0, \tag{2.53}$$

and it is necessary to solve a large system of simultaneous nonlinear equations in order to obtain the form (2.52) from (2.53). This leads to round-off problems in computing derivatives.[18]

The large size of the problems has resulted in the use of sparse matrix techniques such as those embodied in Drud (1977) and in the Mantell and Lasdon (1977) code cited above. Both of these codes employ the reduced gradient techniques and permit the inclusion of inequality as well as equality constraints in the problem.

3. Stochastic control: Passive learning[19]

Stochastic control methods can be divided into two groups—passive and active learning. In passive learning algorithms the effect of the choice of control on future learning is not considered; in active learning algorithms it is considered.

Bar-Shalom, Tse and Larson (1974) and Bar-Shalom and Tse (1976b) make a distinction between open-loop feedback and closed-loop policies. This distinction is helpful in understanding the difference between passive and active stochastic control. Open-loop is like deterministic control, feedback is like passive learning stochastic control, and closed-loop is like active learning stochastic control. In defining these concepts it is convenient to develop some additional notation.

Consider a problem with systems equations

$$x_{k+1} = f_k(x_k, u_k, \xi_k), \qquad k = 0, \ldots, N-1,$$

where x = states, u = controls, ξ = process noise, and x_0 = initial state, and where ξ_k for all k and x_0 are random variables. The system is not measured exactly but rather through a noisy measurement relationship,

$$y_k = h_k(x_k, w_k), \qquad k = 0, \ldots, N,$$

where y_k = measurement vector and w_k = measurement noise.

The criterion function is set to minimize the expected cost of a function C defined over the state and policy trajectories,

$$J = E\left\{C\left(X_0^N, U_0^{N-1}\right)\right\},$$

[18] See Norman, Norman and Palash (1974), and Ando, Norman and Palash (1975).

[19] In preparing this and the following section, I have benefited from reading Rausser (1978). In particular, I have used his way of dividing methods into passive and active.

where E = expectation operator, C = real-valued function, and

$$X_0^N \equiv \{x_j\}_{j=0}^N, \qquad U_0^{N-1} \equiv \{u_j\}_{j=0}^{N-1}.$$

Also the set of observations Y^k is defined as

$$Y^k = \{y_j\}_{j=1}^k.$$

Knowledge about the measurement is defined as

$$M_k = \{h_j(x_j, w_j)\}_{j=1}^k,$$

and the joint probability distribution of the random variables is given by

$$S^k = P(x_0, \xi_0, \ldots, \xi_{N-1}, w_1, \ldots, w_k), \qquad k = 0, \ldots, N-1,$$

and

$$S^0 = P(x_0, \xi_0, \ldots, \xi_{N-1}).$$

With this notation, Bar-Shalom and Tse (1976b) define:

open-loop policy

$$u_k^{\mathrm{OL}} = g_k(S^0), \qquad k = 0, \ldots, N-1,$$

feedback policy

$$u_k^{\mathrm{F}} = g_k(Y^k, U^{k-1}; M^k, S^k), \qquad k = 0, \ldots, N-1,$$

closed-loop policy

$$u_k^{\mathrm{CL}} = g_k(Y^k, U^{k-1}; M^{N-1}, S^{N-1}), \qquad k = 0, \ldots, N-1.$$

The essential difference is that the open-loop policy does not use any measurement information and that the feedback policy uses real time data which is fed back to determine the control. Moreover, the feedback policy does not anticipate future measurements while the closed-loop policy does make use of information about those future measurements. That is, the closed-loop policy anticipates that the choice of present control may make a difference in future uncertainty about states and parameters. For this reason the closed-loop policy is said to do active learning.

Another way to characterize the difference between passive and active learning stochastic control algorithms is to note that in computing the optimal control u_k^* for period k with passive schemes, it is assumed that the covariance of the parameters will remain the same in all future time periods. In contrast, in active learning schemes the effect of the current control on the future covariance of the parameters (and/or the states when measurement error is present) is considered while choosing the optimal control. In summary, it is useful to distinguish between deterministic control (open-loop) and two kinds of stochastic control, i.e., stochastic control with passive learning (feedback) and stochastic control with active learning (closed-loop).

Uncertainty is a pervasive characteristic of most dynamic economic problems. A variety of types of uncertainty can be analyzed in the framework of stochastic control. Among these types are:

(i) additive errors (or noise terms) in the systems equations,
(ii) parameters which are constant but unknown and which must therefore be treated as uncertain,
(iii) parameters which are random,
(iv) measurement error terms.

Type (i) uncertainty above is the usual error term of econometric models, and it is the simplest kind of uncertainty to analyze with control theory methods. Types (ii) and (iii) are closely related. In type (ii) uncertainty it is assumed (as is commonly assumed in econometric models) that the parameters of the model are constant but unknown. The estimation process then yields means and covariances of the estimates. So the problem is the choice of optimal economic policies (controls) in the face of the fact that one is uncertain of the true values of the parameters but has available means and covariances of the parameter estimates. Type (iii) uncertainty is like type (ii) except that the true values of the parameters are themselves assumed to be random, viz crop yields in simple agricultural models. Type (iv) is discussed somewhat in the "errors-in-variables" literature in econometrics. It is appropriate for economic problems in which one cannot measure the state of the system perfectly but rather only through imperfect (or noisy) measurement processes.

This section is devoted to stochastic control with passive learning. Here problems with additive error terms [type (i)] and unknown but constant parameters [type (ii)][20] are considered. The following section on stochastic control with active learning will consider all four types of uncertainty.

[20] Measurement uncertainty [type (iv)] can also be considered in the stochastic control context, viz Athans (1972, sec. 4.32, p. 469).

3.1. Additive uncertainty

Consider the general nonlinear deterministic problem of Section 2 with the addition of an additive error term to the systems equations and a change of criterion to minimize the expected value, i.e., choose $[u_k]_{k=0}^{N-1}$ to minimize

$$J = E\{C_N\}, \tag{3.1}$$

where

$$C_N = L_N(x_N) + \sum_{k=0}^{N-1} L_k(x_k, u_k),$$

subject to the systems equations

$$x_{k+1} = f_x(x_k, u_k) + \xi_k, \tag{3.2}$$

$$x_0 \quad \text{given.} \tag{3.3}$$

Assume that ξ is normally distributed with mean zero and covariance Q and is serially uncorrelated,[21] i.e.,

$$E\{\xi_k\} = 0, \qquad E\{\xi_k \xi_k'\} = Q, \qquad E\{\xi_k \xi_\theta'\} = 0. \tag{3.4}$$
$$\theta \neq k$$

Special methods can be used to solve problem (3.1)–(3.4) when restrictions are placed on the form of L and f. For example, when L is quadratic and f is linear, the certainty equivalence results of Simon (1956) and Theil (1957) apply, so the expected value of the uncertain elements can be taken and the problem solved as a deterministic model. Alternatively, when f is linear and ξ is not necessarily normally distributed the postponed linear approximation method of Ashley (1976) is applicable.

For the general case of problem (3.1)–(3.4) the mean disturbance approach discussed by Athans (1972), and applied by Garbade (1975a) to a macroeconomic stabilization problem and by Kim, Goreux and Kendrick (1975) to a cocoa market stabilization problem may be employed.[22] This method involves

[21] For a discussion of control with correlated error terms, see Pagan (1975). Also two readers of a draft of this article have pointed out that for certainty equivalence to hold, it is not necessary that ξ be normally distributed.

[22] Some other applications of stochastic control to models with additive error terms include:
(a) monetary sector of the U.S. economy — Pindyck and Roberts (1974),
(b) macro models of the British economy — Bray (1974, 1975) and Wall and Westcott (1974, 1975),
(c) macro models of the U.S. economy — Chow (1972), Brito and Hester (1974), and Gordon (1974),
(d) theoretical macroeconomic models — Kareken, Muench and Wallace (1973), Phelps and Taylor (1977), and Sargent and Wallace (1975),
(e) water pollution control — Kendrick, Rao and Wells (1970).

several steps:

(1) Set all random elements to their mean values and solve for the deterministic path $[x_{o,k+1}, u_{ok}]_{k=0}^{N-1}$. For this step Garbade used the successive approximations method, and Kim, Goreux and Kendrick employed a differential dynamic programming procedure from Jacobson and Mayne (1970); however, any of the deterministic nonlinear optimizing methods described in Section 2 could be used.
(2) Make a second-order Taylor expansion of the criterion function and a first-order expansion of the systems equations as in (2.4)–(2.5) about the deterministic optimum[23] path $[x_{o,k+1}, u_{o,k}]_{k=0}^{N-1}$.
(3) Solve the resulting quadratic-linear problem for the feedback rule of the form (2.39), i.e., $\mathrm{d}u_k = G_k \mathrm{d}x_k + g_k$ where $\mathrm{d}u_k = u_k - u_{ok}$ and $\mathrm{d}x_k = x_k - x_{ok}$, i.e.,

$$u_k = u_{ok} + G_k(x_k - x_{ok}) + g_k. \tag{3.5}$$

The feedback rule (3.5) is then used as the control rule for the problem. For economic stabilization problems of either the macroeconomic or microeconomic variety this procedure results in stabilization about the deterministic optimal path. This has the desirable effect that the quality of the approximation used is improved since one is stabilizing about the expansion path, but the undesirable effect that it may not be desirable to stabilize about the deterministic optimum path.

3.2. *Multiplicative uncertainty*[24, 25]

If the uncertainty in an econometric model is in the coefficients rather than in an additive term the certainty equivalence principle cannot be used. In fact,

[23] It can be argued that the expansion should not be about the deterministic optimum path but rather about a path which is chosen with an eye toward the effects of the uncertainty on the optimal path, see for example Denham (1964). For a discussion of the problem which can arise from using the deterministic path as the nominal path, see Kendrick and Majors (1974).

[24] The basic method of the derivation used here comes from Farison, Graham, and Shelton (1967) and from Aoki (1967, pp. 44–47). For related algorithms, see Bar-Shalom and Sivan (1969), Curry (1969), Tse and Athans (1972), and Ku and Athans (1973). In preparing the specific derivation shown here, I have drawn on private communication with Yaakov Bar-Shalom and a few elements from Tse, Bar-Shalom and Meier (1973), and Bar-Shalom, Tse and Larson (1974). For a similar derivation, see Chow (1975, ch. 10). For an alternative treatment of multiplicative uncertainty, see Turnovsky (1975, 1977).

[25] Examples of the application of stochastic control methods to economic problems with multiplicative random variables are Fisher (1962), Zellner and Geisel (1968), Burger, Kalish and Babb (1971), Henderson and Turnovsky (1972), Bowman and Laporte (1972), Chow (1973), Turnovsky (1973, 1974, 1975, 1977), Kendrick (1973), Aoki (1974a, 1975a), Cooper and Fischer (1975), Shupp (1976b, c), and Walsh and Cruz (1975).

control problems with stochastic coefficients cannot generally be solved for analytical solutions.[26] Therefore, most of the stochastic control algorithms are approximations of one kind or another and involve the use of numerical methods. For example, consider the problem (2.11)–(2.13) and recall that it was derived both from a quadratic-linear approximation to a general nonlinear control problem and from a quadratic-linear tracking problem.

Assume now that the coefficients in A, B, and c in

$$x_{k+1}=A_k x_k+B_k u_k+c_k+\xi_k \tag{3.6}$$

are unknown, but that estimates of their means and covariances are available so that the control rule can be based on such estimates. In this case the criterion function (2.11) may be rewritten to minimize the expected value, i.e., find $[u_k]_{k=0}^{N-1}$ to minimize

$$J=E(C_N), \tag{3.7}$$

where

$$\begin{aligned} C_N &= L_N(x_N)+\sum_{k=0}^{N-1} L_k(x_k,u_k) \\ &= \tfrac{1}{2}x_N W_N x_N + w_N' x_N \\ &\quad + \sum_{k=0}^{N-1}\left[\tfrac{1}{2}x_k' W_k x_k + w_k' x_k + x_k' F_k u_k + \tfrac{1}{2}u_k' \Lambda_k u_k + \lambda_k' u_k\right]. \end{aligned} \tag{3.8}$$

The cost-to-go for the last $N-i$ steps is

$$C_{N-i}=L_N(x_N)+\sum_{k=i}^{N-1} L_k(x_k,u_k). \tag{3.8a}$$

The optimal cost-to-go j periods from the terminal time N is defined as[27]

$$J_{N-j}^{*}=\min_{u_{N-j}} E\left\{\ldots \min_{u_{N-2}} E\left\{\min_{u_{N-1}} E\left\{C_j \mid \mathcal{P}^{N-1}\right\}\ldots \mathcal{P}^{N-2}\right\}\ldots \mid \mathcal{P}^{N-j}\right\}, \tag{3.9}$$

where

$\mathcal{P}^j=(\mu,\Sigma)=$ mean and covariance of the unknown elements.

[26] Viz. Aoki (1967, p. 113).

[27] Actually this is an approximate cost-to-go which depends only on the first two moments.

Then, by the principle of optimality,

$$J^*_{N-j} = \min_{u_{N-j}} E\{L_{N-j}(x_{N-j}, u_{N-j}) + J^*_{N-j-1} | \mathcal{P}^{N-j}\}. \tag{3.10}$$

As in the deterministic problem, it is assumed that the optimal cost-to-go can be written as a quadratic form in the state variables, i.e.,

$$J^*_0 = E\{\nu_N + p'_N x_N + \tfrac{1}{2} x'_N K_N x_N | \mathcal{P}^N\}. \tag{3.10a}$$

Also from (3.9),

$$\begin{aligned} J^*_0 &= E\{C_0 | \mathcal{P}^{N-1}\} \\ &= E\{L_N(x_N)\} = E\{\tfrac{1}{2} x'_N W_N x_N + w'_N x_N\} \\ &= \tfrac{1}{2} x'_N E\{W_N\} x_N + E\{w_N\}' x_N. \end{aligned} \tag{3.10b}$$

Comparison of (3.10a) and (3.10b) yields

$$K_N = E\{W_N\} = W_N, \qquad p_N = E\{w_N\} = w_N, \qquad \nu_N = 0. \tag{3.10c}$$

The general form of (3.10a) is

$$J^*_{N-j} = E\{\nu_{N-j} + p'_{N-j} x_{N-j} + \tfrac{1}{2} x'_{N-j} K_{N-j} x_{N-j} | \mathcal{P}^{N-j}\}. \tag{3.11}$$

Then, substitution of (3.11) into (3.10) yields

$$\begin{aligned} J^*_{N-j} = \min_{u_{N-j}} & E\{L_{N-j}(x_{N-j}, u_{N-j}) \\ & + E\{\nu_{N-j-1} + p'_{N-j-1} x_{N-j-1} \\ & + \tfrac{1}{2} x'_{N-j-1} K_{N-j-1} x_{N-j-1} | \mathcal{P}^{N-j-1}\} | \mathcal{P}^{N-j}\}. \end{aligned} \tag{3.12}$$

In order to simplify the notation somewhat, let period k be the jth period from the terminal time N, then

$$k = N-j, \qquad k+1 = N-(j+1) = N-j-1. \tag{3.13}$$

Then using (3.13) and the implied definition of L_k from (3.8) in (3.12) yields

$$\begin{aligned} J^*_k = \min_{u_k} & E\{\tfrac{1}{2} x'_k W_k x_k + w'_k x_k + x'_k F_k u_k + \tfrac{1}{2} u'_k \Lambda_k u_k + \lambda'_k u_k \\ & + E\{\nu_{k+1} + p'_{k+1} x_{k+1} + \tfrac{1}{2} x'_{k+1} K_{k+1} x_{k+1} | \mathcal{P}^{k+1}\} | \mathcal{P}^k\}. \end{aligned} \tag{3.14}$$

Substitution of the systems equation (3.6) into (3.14) and dropping the inside expectation since all random $k+1$ terms have now been substituted out, leaves

$$\begin{aligned} J_k^* = \min_{u_k} E\Big\{ & \tfrac{1}{2}x_k' W_k x_k + w_k' x_k + x_k' F_k u_k \\ & + \tfrac{1}{2}u_k' \Lambda_k u_k + \lambda_k' u_k + \nu_{k+1} + p_{k+1}'(A_k x_k + B_k u_k + c_k + \xi_k) \\ & + \tfrac{1}{2}(\xi_k' + c_k' + u_k' B_k' + x_k' A_k') K_{k+1}(A_k x_k + B_k u_k + c_k + \xi_k) | \mathcal{P}^k \Big\}, \end{aligned} \tag{3.15}$$

$$\begin{aligned} J_k^* = \min_{u_k} E\Big\{ & \tfrac{1}{2}x_k' \Phi_k x_k + \phi_k' x_k + x_k' \Psi_k u_k + \tfrac{1}{2}u_k' \Theta_k u_k + \theta_k' u_k \\ & + \tfrac{1}{2}\xi_k' \Omega_k \xi_k + \omega_k' \xi_k + \eta_k | \mathcal{P}^k \Big\}, \end{aligned} \tag{3.16}$$

where

$$\begin{aligned} \Phi_k &= W_k + A_k' K_{k+1} A_k, \\ \phi_k &= A_k'(K_{k+1} c_k + p_{k+1}) + w_k, \\ \Psi_k &= F_k + A_k' K_{k+1} B_k, \\ \Theta_k &= \Lambda_k + B_k' K_{k+1} B_k, \\ \theta_k &= B_k'(K_{k+1} c_k + p_{k+1}) + \lambda_k, \\ \omega_k &= K_{k+1}(A_k x_k + B_k u_k + c_k) + p_{k+1}, \\ \eta_k &= \nu_{k+1} + p_{k+1}' c_k + \tfrac{1}{2} c_k' K_{k+1} c_k. \end{aligned} \tag{3.17}$$

Then taking the expectation in (3.16) yields

$$\begin{aligned} J_k^* = \min_{u_k} \Big[& \tfrac{1}{2}x_k'(E\Phi_k)x_k + (E\phi_k)' x_k + x_k'(E\Psi_k)u_k \\ & + \tfrac{1}{2}u_k'(E\Theta_k)u_k + (E\theta_k)' u_k + \tfrac{1}{2}E\{\xi_k' \Omega_k \xi_k\} + E\eta_k \Big], \end{aligned} \tag{3.18}$$

since x_k is assumed to be perfectly observed and therefore known with certainty and the control u_k is deterministic. Also $E(\xi_k) = 0$.

The expectations in (3.18) can also be written using (3.17) as

$$\begin{aligned} E\Phi_k &= W_k + E\{A_k' K_{k+1} A_k\}, \\ E\phi_k &= E\{A_k' K_{k+1} c_k\} + E\{A_k\}' p_{k+1} + w_k, \\ E\Psi_k &= F_k + E\{A_k' K_{k+1} B_k\}, \\ E\Theta_k &= \Lambda_k + E\{B_k' K_{k+1} B_k\}, \\ E\theta_k &= E\{B_k' K_{k+1} c_k\} + E\{B_k\}' p_{k+1} + \lambda_k, \\ E\{\xi_k' \Omega_k \xi_k\} &= \operatorname{tr}[\Omega_k Q_k] \quad \text{where} \quad Q_k = E\{\xi_k \xi_k'\}, \\ E\eta_k &= \nu_{k+1} + E\{c_k\}' p_{k+1} + \tfrac{1}{2}E\{c_k K_{k+1} c_k\}. \end{aligned} \tag{3.19}$$

The expectations above can be calculated using the result that[28]

$$E\{x'Ax\} = \hat{x}'A\hat{x} + \mathrm{tr}[A\Sigma], \tag{3.20}$$

where x = random vector of dimension n, $A = n \times n$ constant matrix, $\hat{x} = E\{x\}$, tr = trace operator, and Σ = covariance of $x = E\{(x-\hat{x})(x-\hat{x})'\}$.

Similarly, let

$$D = AKB, \tag{3.21}$$

then

$$d_{ij} = a_i K b_j, \tag{3.22}$$

where $d_{ij} = ij$th element of the matrix D, $a_i = i$th row of the matrix A' (the ith column of A), and $b_j = j$th column of the matrix B.

Then, using (3.20), one obtains from (3.22)

$$E(d_{ij}) = E\{a_i' K b_j\} = \hat{a}_i' K \hat{b}_j + \mathrm{tr}\left[K \Sigma_{b_j a_i}\right], \tag{3.23}$$

where

$$\hat{a}_i = E\{a_i\}, \qquad \hat{b}_i = E\{b_i\}, \qquad \Sigma_{b_j a_i} = E\left\{\left(b_j - \hat{b}_j\right)(a_i - \hat{a}_i)\right\}.$$

Then, taking the minimum of (3.18) yields the first-order condition

$$x_k'(E\Psi_k) + u_k'(E\Theta_k) + (E\theta_k)' = 0, \tag{3.24}$$

or

$$(E\Theta_k)'u_k = -(E\Psi_k)'x_k - E\theta_k, \tag{3.25}$$

so that the feedback rule is

$$u_k = G_k^{\dagger} x_k + g_k^{\dagger}, \tag{3.26}$$

where [assuming Θ is symmetric]

$$\begin{aligned} G_k^{\dagger} &= -(E\Theta_k)^{-1}(E\Psi_k)', \\ g_k^{\dagger} &= -(E\Theta_k)^{-1}(E\theta_k). \end{aligned} \tag{3.27}$$

[28] See Goldberger (1963, p. 166).

Then (3.27) can be rewritten with (3.19) to

$$\begin{aligned} G_k^\dagger &= -\left[\Lambda_k + E\{B_k' K_{k+1} B_k\}\right]^{-1}\left[F_k' + E\{B_k' K_{k+1} A_k\}\right], \\ g_k^\dagger &= -\left[\Lambda_k + E\{B_k' K_{k+1} B_k\}\right]^{-1}\left[E\{B_k' K_{k+1} c_k\} + E\{B_k\}' p_{k+1} + \lambda_k\right]. \end{aligned} \tag{3.28}$$

Note the similarity between the feedback rules for the deterministic problem (2.39) and for the stochastic problem with unknown parameters (3.26). They differ only in the expectations operators.

Substitution of the feedback rule (3.26) into (3.18) then enables one to obtain the optimal cost-to-go entirely in terms of the current state variable and thereby to compute the K and p recursions which are required to evaluate the expectations in (3.19),

$$\begin{aligned} J_k^* = {} & \tfrac{1}{2} x_k'(E\Phi_k) x_k + (E\phi_k)' x_k \\ & - x_k'(E\Psi_k)(E\Theta_k)^{-1}(E\Psi_k)' x_k - x_k'(E\Psi_k)(E\Theta_k)^{-1}(E\theta_k) \\ & + \tfrac{1}{2}\left[(E\theta_k)' + x_k'(E\Psi_k)\right](E\Theta_k)^{-1}(E\Theta_k)(E\Theta_k)^{-1}\left[(E\Psi_k)' x_k + (E\theta_k)\right] \\ & - (E\theta_k)'(E\Theta_k)^{-1}\left[(E\Psi_k)' x_k + (E\theta_k)\right] + \tfrac{1}{2} E\{\xi_k' \Omega_k \xi_k\} + E\eta_k, \end{aligned} \tag{3.29}$$

or

$$\begin{aligned} J_k^* = {} & \tfrac{1}{2} x_k'\left[E\Phi_k - (E\Psi_k)(E\Theta_k)^{-1}(E\Psi_k)'\right] x_k \\ & + \left[(E\phi_k)' - (E\theta_k)'(E\Theta_k)^{-1}(E\Psi_k)'\right] x_k \\ & - \tfrac{1}{2}(E\theta_k)'(E\Theta_k)^{-1}(E\theta_k) + \tfrac{1}{2} E\{\xi_k' \Omega_k \xi_k\} + E\eta_k, \end{aligned} \tag{3.30}$$

or

$$J_k^* = \tfrac{1}{2} x_k' K_k x_k + p_k' x_k + v_k, \tag{3.31}$$

where

$$\begin{aligned} K_k &= E\Phi_k - (E\Psi_k)(E\Theta_k)^{-1}(E\Psi_k)', \\ p_k &= E\phi_k - (E\Psi_k)(E\Theta_k)^{-1}(E\theta_k), \\ v_k &= -\tfrac{1}{2}(E\theta_k)'(E\Theta_k)^{-1}(E\theta_k) + \tfrac{1}{2} E\{\xi_k' \Omega_k \xi_k\} + E\eta_k. \end{aligned} \tag{3.32}$$

Then (3.32) can be rewritten in terms of the original parameters of the problem by substituting in (3.19) to obtain

$$
\begin{aligned}
K_k &= W_k + E\{A_k' K_{k+1} A_k\} \\
&\quad - [F_k + E\{A_k' K_{k+1} B_k\}][\Lambda_k + E\{B_k' K_{k+1} B_k\}]^{-1} \\
&\quad \times [E\{B_k' K_{k+1} A_k\} + F_k'], \\
p_k &= E\{A_k' K_{k+1} c_k\} + E\{A_k\}' p_{k+1} + w_k \\
&\quad - [F_k + E\{A_k' K_{k+1} B_k\}][\Lambda_k + E\{B_k' K_{k+1} B_k\}]^{-1} \\
&\quad \times [E\{B_k' K_{k+1} c_k\} + E\{B_k\}' p_{k+1} + \lambda_k], \qquad (3.33) \\
v_k &= -\tfrac{1}{2}[\lambda_k' + p_{k+1}' E\{B_k\} + E\{c_k' K_{k+1} B_k\}][\Lambda_k + E\{B_k' K_{k+1} B_k\}]^{-1} \\
&\quad \times [E\{B_k' K_{k+1} c_k\} + E\{B_k\}' p_{k+1} + \lambda_k] \\
&\quad + \tfrac{1}{2}\operatorname{tr}[\Omega_k Q_k] + v_{k+1} + E\{c_k\}' p_{k+1} + \tfrac{1}{2}\{c_k' K_{k+1} c_k\}.
\end{aligned}
$$

So, in summary, the solution to the stochastic control problem with unknown parameters is embodied in the feedback rule (3.26),

$$u_k = G_k^{\dagger} x_k + g_k^{\dagger},$$

with (3.28),

$$
\begin{aligned}
G_k^{\dagger} &= -[\Lambda_k + E\{B_k' K_{k+1} B_k\}]^{-1}[F_k' + E\{B_k' K_{k+1} A_k\}], \\
g_k^{\dagger} &= -[\Lambda_k + E\{B_k' K_{k+1} B_k\}]^{-1}[E\{B_k' K_{k+1} c_k\} + E\{B_k\}' p_{k+1} + \lambda_k],
\end{aligned}
$$

and where the K and p recursions are defined in (3.33).

3.3. *Updating in stochastic control with passive learning*

Many stochastic control algorithms combine the control selection procedures described above with methods for updating parameter estimates (usually means and covariances). Two methods are as follows:

Updated certainty equivalence[29]

The control is selected by assuming that the mean values of each parameter is its true value. However, the estimates for the parameters are updated after each period.

[29] Like Norman's (1976) heuristic certainty equivalence. See Rausser (1977).

Open-loop feedback[30]

The control is computed as in Section 3.2 above on multiplicative uncertainty. The parameters are updated after each period.

4. Stochastic control: Active learning

The stochastic control problem with active learning [sometimes called adaptive control or dual control] may be viewed as the logical extension of the stochastic control problem with unknown (and either constant or time varying) parameters. The extension is the notion that the parameters can be learned over time (i.e., that parameter estimates can be improved) and that the choice of control variables in each time period should be made not only with a view to its effect on the economic system but also to its effect on parameter estimation. It is this dual effect of the control, i.e., the target achievement effect and the learning effect, which gives rise to the name of dual control. It should be emphasized that there is no term in the criterion function that directly rewards improved estimation; rather the gain from better estimation at a point in time comes only in so far as that gain enables one to control the system better in future time periods.

As discussed above, a distinction is made in stochastic control between active and passive learning. Passive learning comes about via the fact that both state variables and parameters may be re-estimated each time period and their estimates may thereby improve. Active learning means that one takes into account the potential learning effect of the choice of control at the time of its determination. If an economic system is bounced around enough by random shocks and/or if the parameter estimates have very narrow confidence intervals at the initial time, then passive learning will be sufficient, and there will be little gain with active learning methods over passive learning control methods. If, on the other hand, confidence bands are very wide at the initial time and if there are only small natural shocks to the system there may be large gain from using active learning strategies.

A number of different adaptive control algorithms have been applied to economic problems. These include studies by MacRae (1972, 1975), Abel (1975) using the Chow (1975) algorithm, Upadhyay (1975) using the Desphande, Upadhyay and Lainiotis (1973) algorithm, and Kendrick (1979) using the Tse,

[30]Like Norman's (1976) open-loop mean variance, Rausser (1977) makes a distinction between open-loop feedback and sequential stochastic control. He includes the studies of Aoki (1967), Bar-Shalom and Sivan (1969), Curry (1969), Ku and Athans (1973), and Tse and Athans (1972) in the first category, and those of Rausser and Freebairn (1974), Zellner (1971), Chow (1975, ch. 10), and Prescott (1971) in the second category. He characterizes sequential stochastic control as a case where the derivation of the control rule is based on the assumption that future observations will be made but that they will not be used to adapt the probability distribution of the parameters.

Bar-Shalom and Meier (1973) algorithm.[31] No clear ranking among these algorithms has yet emerged but rather their relative performance seems to be somewhat problem specific, viz Norman (1976) and Bar-Shalom and Tse (1976a).

For purposes of the exposition here a single algorithm will be discussed in detail, namely, that of Tse, Bar-Shalom and Meier (1973). This algorithm is relatively general since it is designed to handle both linear and nonlinear problems and problems with and without measurement errors. A short discussion of the MacRae and Chow algorithms is then given.

4.1. *Problem statement*

An adaptive control problem may be written as that of selecting $[u_k]_{k=0}^{N-1}$ to minimize the cost functional[32]

$$J_N = E\{C_N\}, \tag{4.1}$$

where

$$C_N = L_N(x_N) + \sum_{k=0}^{N-1} \left[L_k(x_k) + \phi_k(u_k) \right],^{33} \tag{4.2}$$

subject to

$$x_{k+1} = f_k(x_k, u_k) + \xi_k, \tag{4.3}$$

and the measurement equations

$$y_k = h_k(x_k) + \zeta_k, \tag{4.4}$$

where y is an m element observation vector.

It is assumed that x_0 and $\{\xi_k, \zeta_{k+1}\}_{k=0}^{N-1}$ are independent Gaussian vectors with statistics

$$\begin{aligned} E\{x\} &= \hat{x}_{0|0}, & \operatorname{cov}(x_0) &= \Sigma_{0|0}, \\ E\{\xi_k\} &= 0, & \operatorname{cov}(\xi_k) &= Q_k, \\ E\{\zeta_k\} &= 0, & \operatorname{cov}(\zeta_k) &= R_k. \end{aligned} \tag{4.5}$$

[31] See also Taylor (1973, 1974).

[32] The discussion in this and the following sections draws on chapters by Bo Hyun Kang and the author in Kendrick (1981, chs. 9, 10). They provide a detailed discussion and derivation of results in Tse, Bar-Shalom and Meier (1973), Bar-Shalom, Tse and Larson (1974), and Tse and Bar-Shalom (1973).

[33] The criterion function is separated here for convenience into terms in x and terms in u; however, the method discussed can be applied to problems with cross terms in x and u with appropriate modifications.

The new element here is the measurement equation (4.4) which represents the type (iv) uncertainty described in the previous section, i.e., measurement error. The expression (4.4) permits x and y to be of different dimensions so that there may not be a one-to-one relation between states and observation variables but rather some states may not be observed and others may affect several of the observation variables y. Parameter uncertainties of type (ii) and (iii) will be discussed later in this section in an application of the method outlined here.

The notation of the form $\hat{x}_{k|j}$ introduced above indicates the mean vector as estimated at time k with data available through period j. Similarly $\Sigma_{k|j}$ means the covariance of a vector as estimated at time k with data available through period j. Thus $\Sigma_{0|0}$ above means the covariance of the state variable estimates at time 0 using data through period 0, i.e., it is assumed that data on the economic system is available for periods -1, -2, -3, etc. prior to the solution of the control problem. Contrary to the rather implausible assumption of perfect measurement made in many economic analyses, the measurement error ζ_k in (4.4), implies that the state variable x is measured imperfectly.

4.2. *Description of the algorithm*

Figure 4.1 provides a flow chart and thus an overview of the Tse, Bar-Shalom and Meier algorithm. One way to think of the algorithm is as three nested do-loops. The index for the outside loop is the counter for the Monte Carlo run, the index for the middle loop runs through the time periods within the planning horizon, and the inside loop index is an iteration counter for the search or gradient algorithm to find the optimal control u_k^* for a single time period. The discussion here begins with the outside loop.

4.2.1. *The Monte Carlo procedure*

There are three sets of random elements for the problem:

(i) the initial state estimate with mean $\hat{x}_{0|0}$ and covariance $\Sigma_{0|0}$,
(ii) the system noise ξ_k with mean zero and covariance Q_k,
(iii) the measurement noise ζ_k with mean zero and covariance R_k.

Only a single set of random values for the initial state must be generated, but values of ξ_k and ζ_k for all time periods are required.

4.2.2. *The N-period adaptive control problem*

The time period counter, k, is initialized at zero. A value for the period k control, u_k, must then be chosen in order to initiate the search for optimal

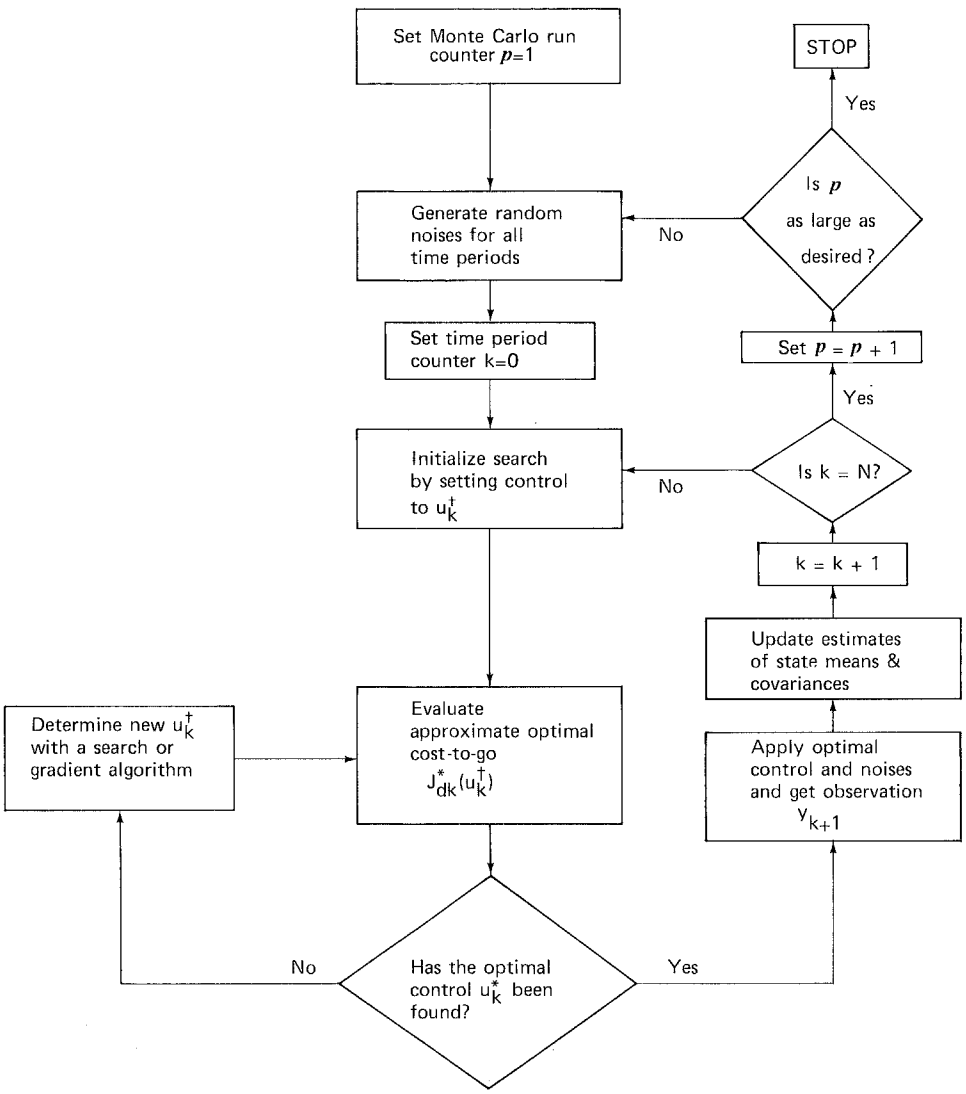


Figure 4.1. Flowchart of an adaptive control algorithm.

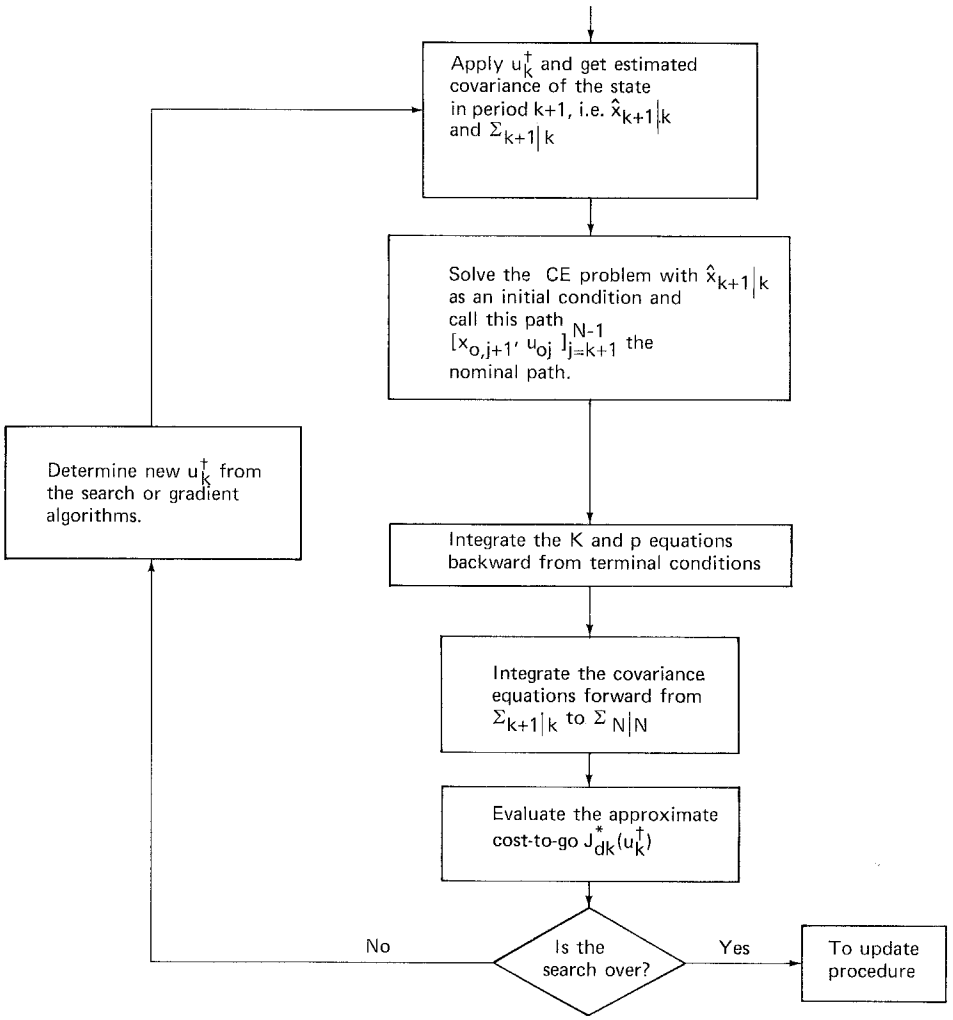


Figure 4.2. Flowchart of the search for u_k^*.

control u_k^*. In this algorithm this value is obtained by solving the certainty equivalence problem for all time periods (i.e., setting all random values to their mean values and solving the resulting deterministic problem) to obtain the optimal C.E. path $[x_{o,j+1}, u_{oj}]_{j=k}^{N-1}$ and setting the initial search value $u_k^\dagger$ to u_{ok}.[34]

Next the approximate optimal cost to go $J_{d,k}^*(u_k^\dagger)$ is computed for this value of the control. At this point it is convenient for the exposition to shift to the third do-loop, namely, the search for the optimal control u_k^* for period k. At the completion of that description the discussion of the middle do-loop will be continued.

4.2.3. *The search for the optimal control in period k*

Figure 4.2 provides a more detailed flowchart of this portion of the algorithm than was shown in Figure 4.1.

A dynamic programming procedure is employed to find the optimal control u_k^* which minimizes the expression

$$J_k^* = \min_{u_k} E\{L_k(x_k) + \phi_k(u_k) + J_{k+1}^* | Y^k, U^{k-1}\}, \tag{4.6}$$

where

$$J_{N-j}^* = J_k^*$$

(optimal cost-to-go j periods from the terminal time N equals optimal cost-to-go at period k), and

$$Y^k = \{y_0, y_1, \ldots, y_k\}, \qquad U^{k-1} = \{u_0, u_1, \ldots, u_{k-1}\}.$$

Hence, (4.6) indicates that the optimal cost-to-go at the present period (k) is the sum of the cost incurred in this period (L_k and ϕ_k) and the optimal cost-to-go at the next period ($k+1$), i.e., $J_{N-(k+1)}^* = J_{N-k-1}^*$. It is not possible to evaluate J_{k+1}^* exactly; therefore it is necessary to approximate this function by a second-order expansion about a nominal path.

So the procedure is to search the space u_k in order to find u_k^* which minimizes a second-order approximation to (4.6). This search is initiated from the certainty equivalence control and then varied according to the dictates of the particular search algorithm employed. For each choice of u_k the approximate cost-to-go must be computed.

[34] Note here again that the subscript o is used to denote the nominal path, and that this should be distinguished from the subscript 0 which is used to denote time period zero.

Before initiating the search it is necessary to choose a nominal path about which to do the second-order expansion of the cost-to-go function. This is done by using u_k in the system equation (4.3) in order to generate $\hat{x}_{k+1|k}$ which is an estimate of the value of the state at time $k+1$ (using data through time k) if the control u_k were employed. Then $\hat{x}_{k+1|k}$ serves as the starting point for the solution of a certainty equivalence problem in which all future random quantities are set to their mean values. This yields a nominal path of the state and control and the associated optimal cost $[x_{o,j+1}, u_{oj}, J^*_{oj}]_{j=k+1}^{N-1}$ which is conditional on the choice of control u_k in the search. A second-order Taylor expansion of the optimal cost-to-go J^*_{k+1} is then made about this nominal path.

This expansion, along with a second-order expansion of the system equations (4.3) about the same path, yields a stochastic control problem with a quadratic performance function and quadratic system equations in the perturbation about the nominal path. This is called the perturbation problem. It is assumed that the optimal cost-to-go at time k is a quadratic function of the perturbation of the current state of the form

$$\Delta J^*_{k+1} = E\left[\delta x'_{k+1} K_{k+1} \delta x_{k+1} + p'_{k+1} \delta x_{k+1} + v_{k+1}\right], \tag{4.7}$$

where

$$\delta x_{k+1} = x_{k+1} - x^{\mathrm{CE}}_{k+1},$$

and K, p and v are parameters to be derived later. It is then proven by induction that this function is quadratic by showing that the cost-to-go for the perturbation problem at time N is quadratic in the state and then showing that it is also quadratic at time $N-1$, etc.

The optimal cost-to-go for the perturbation problem ΔJ^*_{k+1} is then added to the nominal cost-to-go J^*_o to obtain the approximate cost-to-go at $k+1$, i.e.,

$$J^*_{k+1} = J^*_{o,k+1} + \Delta J^*_{k+1}. \tag{4.8}$$

Then, (4.8) is used in the dynamic programming expression (4.6).

Finally, the terms in the cost-to-go which do not depend on u_k are dropped since they do not appear in the first-order conditions and the cost-to-go for the remaining terms is defined as $J^*_{d,k}$, i.e., the approximate optimal cost-to-go at time k for terms which depend on u_k. This may then be written as

$$J^*_{d,k} = \min_{u_k} \left\{ \phi_k(u_k) + C_{o,k+1} + \gamma_{k+1} + \tfrac{1}{2} \operatorname{tr}\left[K_{k+1} \Sigma_{k+1|k}\right] + \tfrac{1}{2} \sum_{j=k+1}^{N-1} \operatorname{tr}\left[K_{j+1} Q_j + \mathcal{Q}_{xx,j} \Sigma_{j|j}\right] \right\}, \tag{4.9}$$

where

$$C_{o,k+1} = L_N(x_{o,N}) + \sum_{j=k+1}^{N-1} L_j(x_{oj}) + \phi_j(u_{oj}), \tag{4.10}$$

$$\gamma_k = \gamma_{k+1} - \tfrac{1}{2} H'_{uk} \mathcal{H}^{-1}_{uu,k} H_{uk}, \qquad \gamma_N = 0, \tag{4.11}$$

$$H_k \equiv L_k(x_k) + \phi_k(u_k) + p'_{k+1} f_k, \tag{4.12}$$

$$H_{uk} = \phi_{uk} + p'_{j+1} f_{uk}, \tag{4.13}$$

$$\mathcal{H}_{uu,k} = H_{uu,k} + f'_{uk} K_{k+1} f_{uk}, \tag{4.13a}$$

$$H_{uu,k} = \phi_{uu,k} + \sum_{i=1}^{n} e^i p'_{k+1} f^i_{uu,k}, \tag{4.14}$$

$$\begin{aligned}
&f^i_k = i\text{th system equation at period } k,\\
&e^i = \text{vector with one in the } i\text{th position and zeros elsewhere},\\
&p = \text{Lagrangian or costate type of variable},\\
&K = \text{Riccati matrix},\\
&\text{tr} = \text{trace operator},\\
&\Sigma_{k+1|k} = \text{covariance of the states at time } k+1\\
&\qquad\quad \text{as estimated with data through period } k,
\end{aligned} \tag{4.15}$$

$$Q_j = \text{cov}(\xi_j) \quad \text{where} \quad \xi_j = \text{system noise term}, \tag{4.16}$$

$$\mathcal{C}_{xx,k} = \mathcal{H}'_{ux,k} \mathcal{H}^{-1}_{uu,k} \mathcal{H}_{ux,k}, \tag{4.17}$$

$$\mathcal{H}_{ux,k} = H_{ux,k} + f'_{uk} K_{k+1} f_{xk}, \tag{4.18}$$

$$H_{ux,k} = H'_{xu,k} = \sum_{i=1}^{n} e^i p'_{k+1} f_{xu,k}. \tag{4.19}$$

The meaning of (4.9) can be more easily deciphered by separating it into three components: (i) deterministic, (ii) cautionary, and (iii) probing[35] as

$$J^*_{d,k} = \min_{u_k} \{J_{D,k} + J_{C,k} + J_{P,k}\}, \tag{4.20}$$

[35]See Bar-Shalom and Tse (1976a).

where

$$J_{D,k} = \phi_k(u_k) + C_{o,k+1} + \gamma_{k+1}, \tag{4.21}$$

$$J_{C,k} = \tfrac{1}{2}\operatorname{tr}\left[K_{k+1|k}\Sigma_{k+1|k}\right] + \tfrac{1}{2}\sum_{j=k+1}^{N-1}\operatorname{tr}\left[K_{j+1}Q_j\right], \tag{4.22}$$

$$J_{P,k} = \tfrac{1}{2}\sum_{j=k+1}^{N-1}\operatorname{tr}\left[\mathcal{Q}_{xx,j}\Sigma_{j|j}\right]. \tag{4.23}$$

The deterministic term (4.21) contains no stochastic terms. It is the sum of the cost of control in the current period (ϕ), the future cost of the nominal states and controls (C), and a term which depends on the second-order derivative of the Hamiltonian with respect to the control vector (γ).

The cautionary term (4.22) is a function of $\Sigma_{k+1|k}$ which can be interpreted as follows: one should be cautious in choosing a control u_k when the state of the system is unknown because a large control might drive one further from the desired path rather than bringing one closer to it. This term does not have a summation in front of it and depends only on the value of the covariance in period $k+1$. At the end of period $k+1$ it will be possible to observe y again and to choose a new control. Another way to say this is that caution applies only to period $k+1$ since for later periods it will be possible to observe the system again and to choose a new control to offset any past errors.

The cautionary term also depends on Q_j and the interpretation of this is less clear. In the trace term KQ, only K can be affected by the choice of u, so it is possible to interpret this as: the degree of cautionary necessary depends on the covariance of the random shocks which hit the system.

The probing term (4.23) depends on $\Sigma_{j|j}$ for all periods from $k+1$ to the terminal period N. This matrix is the post-observation covariance for the states. In some problems this term may encourage one to choose more active controls in early periods in order to improve state estimates in later periods and thereby to decrease the size of terms in $\Sigma_{j|j}$. Thus this term embodies the notion of active learning.

In order to find the optimal control u_k^* for period k it is necessary to evaluate the approximate optimal cost-to-go $J_{d,k}^*$ for different values of u_k using (4.9). In order to obtain the deterministic and cautionary terms one needs the Riccati matrix K and the vector p for each time period from $k+1$ to N. These are obtained by integrating the following equations backwards from terminal conditions:

for K

$$K_k = \mathcal{H}_{xx,k} - \mathcal{Q}_{xx,k}, \qquad K_N = L_{xx,N}, \tag{4.24}$$

where

$$\mathcal{H}_{xx,k} = H_{xx,k} + f'_{xk} K_{k+1} f_{xk}, \tag{4.25}$$

$$H_{xx,k} = L_{xx,k} + \sum_{i=1}^{n} e^i p'_{k+1} f^i_{xx,k}, \tag{4.26}$$

and $\mathcal{Q}_{xx,k}$ is defined in (4.17) above;

for p

$$p_k = H_{xk} - \mathcal{H}'_{ux,k} \mathcal{H}^{-1}_{uu,k} H_{uk}, \qquad p_N = L_{x,N}, \tag{4.27}$$

where

$$H_{xk} = L_{xk} + p'_{k+1} f_{xk}, \tag{4.28}$$

and $\mathcal{H}_{ux,k}$, $\mathcal{H}_{uu,k}$ and H_{uk} are defined in (4.18), (4.13a) and (4.13), respectively.

The procedure for obtaining these expressions for K and p is the same as the method used for obtaining the equivalent expression for the deterministic and stochastic control problem in Sections 2 and 3, respectively.[36]

In order to evaluate the probing term (4.23) one needs not only the K and p terms but also the post-observation covariance matrix $\Sigma_{j|j}$ for periods $k+1$ to N. These matrices are obtained from

$$\Sigma_{k+1|k} = f_{xk} \Sigma_{k|k} f'_{xk} + Q_k + \tfrac{1}{2} \sum_i \sum_j e^i e^{j'} \operatorname{tr}\left[f^i_{xx} \Sigma_{k|k} f^j_{xx} \Sigma_{k|k} \right], \tag{4.29}$$

and

$$\Sigma_{k+1|k+1} = (I - V_{k+1} h'_{x,k+1}) \Sigma_{k+1|k}, \tag{4.30}$$

where

$$V_{k+1} = \Sigma_{k+1|k} h'_{x,k+1} \Big(h'_{x,k+1} \Sigma_{k+1|k} h_{x,k+1} + R_{k+1} + \tfrac{1}{2} \sum_i \sum_j e^i e^{j'} \operatorname{tr}\left[h^i_{xx} \Sigma_{k+1|k} h^j_{xx} \Sigma_{k+1|k} \right] \Big)^{-1}. \tag{4.30a}$$

[36] For the complete derivation, see a chapter by the author and Kang in Kendrick (1981, ch. 9) or Bar-Shalom, Tse and Larson (1974, esp. (A.13)–(A.16)).

The expression (4.29) uses $\Sigma_{k|k}$ to obtain $\Sigma_{k+1|k}$ and depends in part on the covariance of the system equation noise term Q, i.e., the greater Q the more the Σ at $k+1|k$ increased over its value at $k|k$ (in the interval between the application of the control and the taking of a new measurement on the system).

The expression (4.30) in turn shows that $\Sigma_{k+1|k+1}$ will be the same as $\Sigma_{k+1|k}$ but decreased by the amount $Vh'\Sigma_{k+1|k}$. The size of V in turn depends on the measurement error covariance, i.e., in general the larger the measurement error variances, the less the covariance will be reduced by the measurement (going from $\Sigma_{k+1|k}$ to $\Sigma_{k+1|k+1}$).

In summary, it is necessary to integrate the K and p equations backward and the Σ equations forward each time the approximate cost-to-go J_{dk}^* is evaluated at a different value of u_k with (4.9). It is of interest to note that the Σ terms (except for $\Sigma_{k+1|k}$) are not required to evaluate the deterministic and cautionary components of the cost-to-go, so that the computation time with this algorithm could be almost cut in half if it were not necessary to evaluate the probing term.

In research on small macroeconomic models of the U.S. economy, Kendrick (1979) has found that the probing term is small and changes very little with changes in u_k so that it has only very slight influence on the approximate cost-to-go. If this result should hold up after more extensive testing and for a larger class of problems, it could lead to substantial computational savings.[37]

This completes the algorithm step of evaluating $J_{dk}^*(u_k^\dagger)$. From Figure 4.2 the next step is to determine whether or not the search is over. The usual procedure in answering such a question is to check the convergence of either or both J_{dk}^* and $u_k^\dagger$. However, a word of warning is in order. The function J_{dk}^* may not be convex. Thus gradient procedures may reach local rather than global optimal points. For this reason it may prove worthwhile to start the gradient algorithm at a variety of points or to conduct a grid search across the feasible space for the control.

If convergence has not been obtained a new $u_k^\dagger$ is chosen and the evaluation procedure is repeated. A variety of methods may be used to choose the new u_k. Tse and Bar-Shalom (1973) used a quadratic fit to three evaluations of J_{dk}^* with the new $u_k^\dagger$ being the minimum point on this quadratic. Kendrick (1979) employed both grid search and a quasi-Newton gradient techniques.

If convergence has been obtained, then the algorithm branches to the update procedure which is a part of the middle do-loop. That loop was called the N-period control problem above.

[37]This result is for models in which the parameters are assumed to be time invariant. If the parameters are assumed to be time varying, then the initial covariance of the parameters would be expected to include wider confidence intervals. In this case active learning may be more important.

4.2.4. The N-period adaptive control problem continued

Once the optimal control u_k^* for period k is chosen, it can be used in the systems equations (4.3) along with the system noise ξ_k to obtain the state x_{k+1}. This state is used in turn in the measurement equation (4.4) along with the measurement noise ζ_k to obtain the observation variable y_{k+1}.

An update procedure is then used to obtain estimates of the mean and covariance of the state at time $k+1$, i.e., $\hat{x}_{k+1|k+1}$ and $\Sigma_{k+1|k+1}$. These estimates are obtained from Kalman filter methods which yield expressions for the mean and covariance as follows:[38]

$$\hat{x}_{k+1|k+1} = \hat{x}_{k+1|k} + V_{k+1}\left(y_{k+1} - h_{x,k+1}\hat{x}_{k+1|k}\right), \tag{4.30b}$$

with

$$\hat{x}_{k+1|k} = f_x\left(\hat{x}_{k|k}, u_k^*\right) + \tfrac{1}{2}\sum_i e^i \operatorname{tr}\left[f_{xx,k}^i \Sigma_{k|k}\right], \tag{4.30c}$$

with V defined in (4.30a).

Expression (4.30c) is used to project the mean from period k to period $k+1$ using the system equations f and a measure of the uncertainty Σ. Then (4.30b) corrects $\hat{x}_{k+1|k}$ to $\hat{x}_{k+1|k+1}$ depending on the separation between the actual measurement y and the expected measurement $h_x\hat{x}$. The weight given to the new observation depends on V which varies inversely with the measurement noise covariance R, i.e., the greater the uncertainty of the measurement the less reliance placed upon it.

The update expressions for the covariance are (4.29) and (4.30). They were used before in projecting the covariance to obtain $\Sigma_{j|j}$ for all time periods while calculating the approximate cost-to-go.

Once the update is completed the time period index is advanced by one period and the process is repeated until all N periods of the problem have been solved. Similarly, the entire problem is solved repeatedly with different random elements generated by the Monte Carlo procedure until enough runs have been completed to tell the investigator on average how well the adaptive control algorithm would do in controlling the economic system under study.

One class of economic problems that has received some analysis with adaptive control methods is linear models with unknown parameters.

[38]See Kendrick (1981, ch. 9 and app. D).

4.3. Application to a linear model with unknown parameters[39]

4.3.1. Problem statement

Select $[u_k]_{k=0}^{N-1}$ to minimize the cost functional

$$J = E\left\{ \tfrac{1}{2}(x_N - \tilde{x}_N)'W_N(x_N - \tilde{x}_N) + \tfrac{1}{2}\sum_{k=0}^{N-1}(x_k - \tilde{x}_k)'W_k(x_k - \tilde{x}_k) + (u_k - \tilde{u}_k)'\Lambda_k(u_k - \tilde{u}_k) \right\}, \tag{4.31}$$

subject to

$$x_{k+1} = A_k(\theta_k)x_k + B_k(\theta_k)u_k + \xi_k, \qquad k = 0, 1, \ldots, N-1, \tag{4.32}$$

$$\theta_{k+1} = D_k\theta_k + \eta_k, \tag{4.33}$$

$$y_k = H_k x_k + \zeta_k, \tag{4.34}$$

where $\tilde{x}$ and $\tilde{u}$ are the desired paths for state and controls, θ_k is a vector of unknown parameters, D_k is a known matrix, and η_k is a random vector.

Some or all of the elements of A and B may be functions of θ. For example, when A is 2×2 and B is 2×1 and all parameters are treated as unknown

$$\theta = \begin{bmatrix} a_{11} \\ a_{12} \\ a_{21} \\ a_{22} \\ b_1 \\ b_2 \end{bmatrix},$$

i.e., θ is simply a vector which stacks the unknown parameters in A and B.

The vectors ξ, ζ, η, x_o, and θ_o are assumed to be mutually independent normally distributed random vectors with known mean and covariance,

$$\begin{aligned} &x_0 \sim N(x_0, \Sigma^{xx}_{0|0}), && \theta_0 \sim N\big(\hat{\theta}_{0|0}, \Sigma^{\theta\theta}_{0|0}\big), \\ &\xi_k \sim N(0, Q_k), && \zeta_k \sim N(0, R_k), \\ &\eta_k \sim N(0, G_k), \end{aligned} \tag{4.35}$$

[39]This section follows the approach developed by Tse and Bar-Shalom (1973), and is developed in detail by Kang and the author in Kendrick (1981, ch. 10).

with $\Sigma_{0|0}^{xx}, \Sigma_{0|0}^{\theta\theta}, Q_k, R_k, G_k \geqslant 0$. Also, it is assumed that the unknown parameters enter linearly in A and B. Finally, the constant terms c in the systems equations may be incorporated into the matrix A by augmenting the state vector x with an additional variable which is always one.

The problem (4.31)–(4.35) can be converted to the form of the nonlinear adaptive control problem of the previous section by defining a new state vector z which augments the initial state vector x with the parameter vector θ, i.e.,[40]

$$z = \begin{bmatrix} x \\ \theta \end{bmatrix}. \tag{4.36}$$

Then the systems equations with the new state variable z become

$$z_{k+1} = f_k(z_k, u_k) = \left[\begin{array}{c} f^x(z_k, u_k) \\ \hline f^\theta(z_k, u_k) \end{array}\right] = \left[\begin{array}{c} A_k(\theta_k)x_k + B_k(\theta_k)u_k \\ \hline D_k\theta_k \end{array}\right], \tag{4.37}$$

and the measurement equations become

$$y_k = h_k(z_k) = \left[\begin{array}{c} H_k x_k \\ \hline 0 \end{array}\right]. \tag{4.38}$$

Also the criterion function is similarly modified to include the augmented state z.[41]

The algorithm for the nonlinear adaptive control problem of the previous section can be applied to this problem which has nonlinear state equations (4.37). It follows the procedures outlined in the previous section and shown in Figures 4.1 and 4.2.

4.3.2. *Algorithm*

(A) *Initialization of the search*

(1) Initialize with $k=0$.
(2) Generate $\theta_{o,j}$ with (4.33) and obtain $A_j(\theta_{o,j})$, $B_j(\theta_{o,j})$, and $C_j(\theta_{o,j})$ by using $\theta_{o,j}$ for $j=k$.
(3) Compute $\tilde{K}_j$ and $\tilde{p}_j$, $j=k+1,\ldots,N$, by solving (4.31)–(4.35) as a certainty equivalence problem without augmentation and using (2.42) and (2.43).

[40] Norman (1979) has developed a first-order version of the Tse, Bar-Shalom and Meier algorithm without measurement error. In one version he uses state augmentation of this type and in the other he does not augment the state. Computational storage requirements are the primary reason to avoid augmentation.

[41] See Kendrick (1981, ch. 10).

(4) Set $u_k^\dagger = u_{o,k}$ as given by (2.39), i.e., set the search value to the nominal solution.

(B) *Evaluation of* $J_d(u_k^\dagger)$

(1) Apply $u_k^\dagger$ to obtain the predicted state $\hat{z}_{k+1|k}$ and its covariance $\Sigma_{k+1|k}$ by using (4.30c) and (4.29) which, with the notation

$$\Sigma_{k|k} = \left[\begin{array}{c|c} \Sigma^{xx} & \Sigma^{x\theta} \\ \hline \Sigma^{\theta x} & \Sigma^{\theta\theta} \end{array}\right]_{k|k}, \qquad \tilde{Q}_k = \left[\begin{array}{c|c} Q & 0 \\ \hline 0 & G \end{array}\right]_k, \qquad \hat{z}_{k+1|k} = \begin{bmatrix} \hat{x}_{k+1|k} \\ \hat{\theta}_{k+1|k} \end{bmatrix},$$

specializes to the problem at hand to provide

$$\hat{x}_{k+1|k} = A_k(\theta_{o,k})\hat{x}_{k|k} + B_k(\theta_{o,k})u_k^\dagger + \sum_{i\in X} e^i \operatorname{tr}\left[a_\theta^i \Sigma_{k|k}^{\theta x}\right], \tag{4.39}$$

$$\hat{\theta}_{k+1|k} = D\hat{\theta}_{ok}, \tag{4.40}$$

$$\begin{aligned} \Sigma_{k+1|k}^{xx} = {} & A_k \Sigma_{k|k}^{xx} A_k' + A_k \Sigma_{k|k}^{x\theta} f_{\theta k}^{x'} \\ & + f_{\theta k}^{x} \Sigma_{k|k}^{\theta x} A_k' + f_{\theta k}^{x} \Sigma_{k|k}^{\theta\theta} f_{\theta k}^{x'} + Q_k \\ & + \sum_{i\in X}\sum_{j\in X} e^i e^{j'} \operatorname{tr}\left\{a_\theta^i \Sigma_{k|k}^{\theta x} a_\theta^j \Sigma_{k|k}^{\theta x} + a_\theta^i \Sigma_{k|k}^{\theta\theta} a_\theta^j \Sigma_{k|k}^{xx}\right\}, \end{aligned} \tag{4.41}$$

$$\Sigma_{k+1|k}^{\theta x} = D_k \Sigma_{k|k}^{\theta x} A_k' + D_k \Sigma_{k|k}^{\theta\theta} f_{\theta k}^{x'}, \tag{4.42}$$

$$\Sigma_{k+1|k}^{\theta\theta} = D_k \Sigma_{k|k}^{\theta\theta} D_k + G_k, \tag{4.43}$$

where

$$f_{\theta k}^{x} = \sum_i e_i \hat{x}_{k|k}' a_\theta^i + \sum_i e_i u_k^{\dagger\prime} b_\theta^i. \tag{4.44}$$

(2) Use $\hat{x}_{k+1|k}$ as the initial state and solve the certainty equivalence problem from period $k+1$ to period N by computing $\{x_{o,j}\}_{j=k+1}^N$ and $\{u_{o,j}\}_{j=k+1}^N$ using (2.39) and (2.12). This provides the nominal path corresponding to the choice of u_k used in the search, namely $u_k^\dagger = u_{o,k}$.

(3) Evaluate all derivatives along the nominal path.

(4) Compute $K_j^{\theta x}$ and $K_j^{\theta\theta}$ for $j = k+1, \ldots, N$, by specializing (4.24) to the

problem at hand to obtain

$$K_j^{xx} = \tilde{K}_j \tag{4.45}$$

$$K_j^{\theta x} = \left(f_\theta^{x'} K_{j+1}^{xx} + D' K_{j+1}^{\theta x}\right) A - \left\{\left(f_\theta^{x'} K_{j+1}^{xx} + D' K_{j+1}^{\theta x}\right) B + \sum e_i' p_{j+1}^x b_\theta^i\right)'\right\} \mu_j \left\{B' K_{j+1}^{xx} A\right\} + \sum e_i' p_{j+1}^x a_\theta^i, \qquad K_N^{\theta x} = 0, \tag{4.46}$$

$$K_j^{\theta\theta} = f_\theta^{x'}\left(K_{j+1}^{xx} f_\theta^x + K_{j+1}^{x\theta} D\right) + D'\left(K_{j+1}^{\theta x} f_\theta^x + K_{j+1}^{\theta\theta} D\right) - \left\{\left(f_\theta^x K_{j+1}^{xx} + D' K_{j+1}^{\theta x}\right) B + \left(\sum e_i' p_{j+1}^x b_\theta^i\right)'\right\} \mu \left\{B'\left(K_{j+1}^{xx} f_\theta^x + K_{j+1}^{x\theta} D\right) + \sum e_i' p_{j+1}^x b_\theta^i\right\}, \qquad K_N^{\theta\theta} = 0, \tag{4.47}$$

where

$$\mu = \left[\Lambda + B' K^{xx} B\right]^{-1},$$

and

$$p_j^x = \tilde{K}_j x_{oj} + \tilde{p}_j. \tag{4.48}$$

(5) Project the future covariance matrices $\Sigma_{j+1|j}$ for $j = k+1, \ldots, N-1$, and $\Sigma_{j+1|j+1}$ for $j = k, \ldots, N-1$. The first of these is obtained from (4.29) as outlined above. The second is obtained by specializing (4.30) and (4.30a) to the problem at hand to obtain

$$\Sigma_{k+1|k+1}^{xx} = \left(I - \Sigma_{k+1|k}^{xx} H_{k+1}' S_{k+1}^{-1} H_{k+1}\right) \Sigma_{k+1|k}^{xx}, \tag{4.49}$$

$$\Sigma_{k+1|k+1}^{\theta x} = \Sigma_{k+1|k+1}^{x\theta'} = \Sigma_{k+1|k}^{\theta x}\left(I - H_{k+1}' S_{k+1}^{-1} H_{k+1} \Sigma_{k+1|k}^{xx}\right), \tag{4.50}$$

$$\Sigma_{k+1|k+1}^{\theta\theta} = \Sigma_{k+1|k}^{\theta\theta} - \Sigma_{k+1|k}^{\theta x} H_{k+1}' S_{k+1}^{-1} H_{k+1} \Sigma_{k+1|k}^{x\theta}, \tag{4.51}$$

where

$$S_{k+1} = H_{k+1} \Sigma_{k+1|k}^{xx} H_{k+1}' + R_{k+1}. \tag{4.52}$$

(6) Evaluate the dual cost-to-go by using (4.20). One may wish to calculate the deterministic, cautionary, and probing components of the cost-to-go by using (4.21), (4.22) and (4.23).

(C) *The search*

The procedure stated in (A) and (B) is applied for a choice of u_k at $k=0$, until optimal u_k^* at $k=0$ is obtained to minimize the dual cost $J_d(u_k)$. This requires a search technique over the control space.[42] Once the search is over, the parameter and state estimates are updated from the new observation and the whole procedure is repeated until $k=N$.

(D) *Updating*

After the optimal control u_k^* is determined in the search, this control value is applied to the system equations (4.37) plus the additive random terms to obtain z_{k+1}, i.e.,

$$z_{k+1}=\begin{bmatrix} x_{k+1} \\ \theta_{k+1} \end{bmatrix}=\begin{bmatrix} A_k\big(\hat{\theta}_{k|k}\big)x_{k|k}+B_k\big(\hat{\theta}_{k|k}\big)u_k^*+\xi_k \\ D_k\hat{\theta}_{k|k} \qquad\qquad\qquad\qquad +\eta_k \end{bmatrix}. \tag{4.53}$$

The random variables ξ_k and γ_k are generated by a random number generator using the covariance matrices Q_k and G_k, respectively.

These values x_{k+1} and θ_{k+1} are then used in the measurement equation (4.38) with the error term to obtain y_{k+1}. The random variables ζ_{k+1} are obtained from a random number generator with covariance R_{k+1}.

The new estimate of the mean values of x and θ are then obtained by using the extended Kalman filter (4.30b), i.e.,

$$\hat{z}_{k+1|k+1}=\hat{z}_{k+1|k}+V_{k+1}\big(y_{k+1}-h_{z,k+1}\hat{z}_{k+1|k}\big), \tag{4.54}$$

where $\hat{z}_{k+1|k}$ is determined as given in (4.39)–(4.40). This expression (4.54) is specialized to the linear problem and becomes

$$\hat{x}_{k+1|k+1}=\hat{x}_{k+1|k}+\Sigma_{k+1|k}^{xx}H_{k+1}'S_{k+1}^{-1}\big(y_{k+1}-H_{k+1}\hat{x}_{k+1|k}\big), \tag{4.55}$$

and

$$\hat{\theta}_{k+1|k+1}=\hat{\theta}_{k+1|k}+\Sigma_{k+1|k}^{\theta x}H_{k+1}'S_{k+1}^{-1}\big(y_{k+1}-H_{k+1}\hat{x}_{k+1|k}\big), \tag{4.56}$$

where

$$S_{k+1}=H_{k+1}\Sigma_{k+1|k}^{xx}H_{k+1}'+R_{k+1}. \tag{4.57}$$

[42] Care must be taken in performing this search since $J_d(u_k)$ may be non-convex, viz Kendrick (1978).

4.3.3. *Applications*

There are a number of applications of adaptive control methods to macroeconomic problems. Abel (1975) applied Chow (1975) to a small macroeconomic model. Upadhyay (1975) applied the method of Desphande, Upadhyay and Lainiotis (1973) to Pindyck's (1972) model of the U.S. economy. Also Kendrick (1979) has applied the method outlined above to a three state variable macroeconomic model which included measurement errors.

4.4. *Other active learning algorithms*

Space does not permit a detailed description of each of the active learning stochastic control algorithms that have been developed but it is useful to attempt to relate them to the Tse, Bar-Shalom and Meier (TBM) algorithm with which the author is most familiar.

One distinction is that most of the other algorithms do not consider measurement error. This means (in the notation of the previous sections) that $\Sigma^{xx}_{k|k}=0$ for all k. Substantially simplified versions of the TBM algorithm can be produced with this assumption. One example of this is the algorithm of Norman (1976).

4.4.1. *Norman's algorithm*

These algorithms are developed in large part by simplifying the TBM algorithm in two ways: (1) assuming that there is no measurement error, and (2) using a first-order rather than a second-order expansion of the cost-to-go function (thus, the name "first-order dual control").

Norman exploits the simplification of lack of measurement noise by developing a computational method which does not require the augmentation of the initial state vector x with the parameter vector θ.[43] This procedure reduces substantially the storage requirements for computation. He then compares this algorithm to two passive learning strategies (heuristic certainty equivalence and open loop mean variance) and finds that the ordering across the methods is problem specific.

4.4.2. *MacRae's algorithm*

MacRae (1972, 1975) also uses the assumption of no measurement noise. Thus the covariance matrix Γ which she uses is not like the full Σ matrix used in TBM but rather like the $\Sigma^{\theta\theta}$ component of that matrix.

[43] See Norman (1979) also.

In MacRae (1975) she derives an update relationship for the inverse of this parameter covariance matrix, i.e., a relationship of the form[44]

$$\Gamma_k^{-1} = f\left(\Gamma_{k-1}^{-1}\right). \tag{4.58}$$

This type of relationship may be obtained in TBM by assuming $D=I$, $H=I$, $\Sigma^{xx}=0$, and $R=0$, and then substituting (4.43) into (4.51).

The relationship (4.58) embodies the central notion of active learning, namely, that the choice of control may affect the evolution of the parameter uncertainty (and state uncertainty when measurement error is present).

MacRae then appends the covariance update relationship (4.58) to the expected value of the criterion function with Lagrangian variables and minimizes the resulting function subject to the system equation by using dynamic programming methods.

4.4.3. Chow's algorithm

Chow (1975, ch. 10) develops an active learning algorithm which is more general than TBM's in that it includes in the criterion function cross terms across time periods but less general in that it does not include measurement error.

Both algorithms use a second order approximation of the cost-to-go function. However, in TBM the path about which this approximation is made is changed at each step in the search process while in Chow the path is determined outside of the algorithm. Also, TBM do the second-order approximation first and then take the expectation, and Chow takes the expectation first and then makes the second-order approximation.

5. Decentralization and game theory

There has so far been only limited use of decentralization and game theory results from control theory in economics. Some of the exceptions are McFadden (1969) and Aoki (1975c) using methods in decentralized control, and Kydland (1973, 1976), Myoken (1975a), Pau (1973), and Pindyck (1976, 1977) using game theory.[45] The present author has not worked enough with this class of problems to give the kind of clear review they deserve. However, the essential notions they embody of competing economic interest and of decentralized information are so pervasive in economic problems this area of research seems likely to grow rapidly.

[44] See MacRae (1975, (2.4) and (2.10)).

[45] For a survey of the control theory literature in this field, see Sandell, Varaiya, Athans and Safonov (1977).

6. Conclusion

The methodology of control theory embodies a variety of notions which make it a particularly attractive means of analyzing many economic problems. First is the focus on dynamics and thus on the evolution of an economic systems over time. Second is the orientation toward reaching certain targets or goals and/or of improving the performance of an economic system. Third is the treatment of uncertainty not only in additive equation error terms but also in uncertain initial states, uncertain parameter estimates, and measurement errors.

References

Abel, Andrew B. (1975), "A comparison of three control algorithms as applied to the monetarist-fiscalist debate", Annals of Economic and Social Measurement, 4:239–252.

Ando, Albert, Alfred Norman and Carl Palash (1978), "On the application of optimal control to a large scale econometric model", in: A. Bensoussan, T. Kleindorfer and S. H. S. Tapiero, eds., Applied optimal control, Studies in management sciences, Vol. 9. Amsterdam: North-Holland.

Aoki, Masanao (1967), Optimization of stochastic systems. New York: Academic Press.

Aoki, Masanao (1974a), "Non-interacting control of macroeconomic variables: Implications on policy mix considerations", Journal of Econometrics, 2:261–281.

Aoki, Masanao (1974b), "Stochastic control theory in economics: Applications and new problems", IFAC Symposium on Stochastic Control, Budapest.

Aoki, Masanao (1975a), "Control of linear-discrete-time dynamic systems with multiplicative stochastic disturbances in gain", IEEE Transactions on Automatic Control, AC-20:388–391.

Aoki, Masanao (1975b), "On a generalization of Tinbergen's condition in the theory of policy to dynamic models", Review of Economic Studies, 42:293–296.

Aoki, Masanao (1976), Dynamic economic theory and control in economics. New York: American Elsevier.

Arrow, Kenneth J. (1968), "Application of control theory to economic growth", in: Lectures in applied mathematics, Mathematics of the decision sciences, Part 2, Vol. 12. Providence, RI: American Mathematical Society.

Ashley, Richard Arthur (1976)," Postponed linear approximation in stochastic multiperiod problems", Ph.D. dissertation. San Diego, CA: Department of Economics, University of California.

Athans, Michael (1972), "The discrete time linear-quadratic-Gaussian stochastic control problem", Annals of Economic and Social Measurement, 1:449–492.

Athans, Michael and D. Kendrick (1974), "Control theory and economics: A survey, forecast, and speculations", IEEE Transactions on Automatic Control, AC-19:518–523.

Athans, Michael, Richard Ku and Stanley B. Gershwin (1977), "The uncertainty threshold principle: Some fundamental limitations of optimal decisions making under uncertainty", IEEE Transactions on Automatic Control, AC-22:491–495.

Athans, Michael, Edwin Kuh, Turgay Ozkan, Lucas Papademos, Robert Pindyck and Kent Wall (1977), "Sequential open loop optimal control of a nonlinear macroeconomic model", in: M. D. Intriligator, ed., Frontiers of quantitative economics, Vol. IIIA. Amsterdam: North-Holland.

Bar-Shalom, Yaakov and R. Sivan (1969), "On the optimal control of discrete-time linear systems with random parameters", IEEE Transactions on Automatic Control, AC-14:3–8.

Bar-Shalom, Yaakov and Edison Tse (1976a), "Caution, probing and the value of information in the control of uncertain systems", Annals of Economic and Social Measurement, 5:323–338.

Bar-Shalom, Yaakov and Edison Tse (1976b), "Concepts and methods in stochastic control", in: C. T. Leondes, ed., Control and dynamic systems, Advances in theory and application, Vol. 12. New York: Academic Press.

Bar-Shalom, Yaakov, Edison Tse, and R. E. Larson (1974), "Some recent advances in the development of closed-loop stochastic control and resource allocation algorithms", in: Proceedings of the IFAC symposium on adaptive control. Budapest.

Bogaard, P. J. M. van den and H. Theil (1959), "Macrodynamic policy making: An application of strategy and certainty equivalence concepts to the economy of the United States, 1933–36", Metroeconomica, 11:149–167.

Bowman, H. Woods and Anne Marie Laporte (1972), "Stochastic optimization in recursive equation systems and random parameters", Annals of Economic and Social Measurement, 1:419–436.

Bray, Jeremy (1974), "Predictive control of a stochastic model of the UK economy simulating present policy making practice by the UK government", Annals of Economic and Social Measurement, 3:239–256.

Bray, Jeremy (1975), "Optimal control of a noisy economy with the UK as an example", Journal of the Royal Statistical Society, A 138:339–366.

Brito, D. L. and D. D. Hester (1974), "Stability and control of the money supply", Quarterly Journal of Economics, 88:278–300.

Bryson, Arthur E. and Yu-Chi Ho (1969), Applied optimal control. Waltham, MA: Blaisdell.

Burger, Albert E., Lionel Kalish III and Christopher T. Babb (1971), "Money stock control and its implications for monetary policy", Federal Reserve Bank of St. Louis Review, 53:6–22.

Cheng, David C. and San Wan (1972), "Time optimal control of inflation". Atlanta, GA: College of Industrial Management, Georgia Institute of Technology.

Chow, Gregory C. (1972), "How much could be gained by optimal stochastic control policies?", Annals of Economic and Social Measurement, 1:391–406.

Chow, Gregory C. (1973), "Effect of uncertainty on optimal control policies", International Economic Review, 14:632–645.

Chow, Gregory C. (1975), Analysis and control of dynamic systems. New York: Wiley.

Cooper, J. Phillip and Stanley Fischer (1972), "Stochastic simulation of monetary rules in two macroeconomic models", Journal of the American Statistical Association, 67:750–760.

Cooper, J. Phillip and Stanley Fischer (1975), "A method for stochastic control of nonlinear econometric models", Econometrica, 43:147–162.

Craine, R., A. Havenner and P. Tinsley (1976), "Optimal macroeconomic control policies", Annals of Economic and Social Measurement, 5:191–203.

Curry, R. E. (1969), "A new algorithm for suboptimal stochastic control", IEEE Transactions on Automatic Control, AC-14:533–536.

Davidon, W. C. (1959), "Variable metric method for minimization", AEC Research and Development Report no. ANL-5990.

Denham, W. (1964), "Choosing the nominal path for a dynamic system with random forcing function to optimize statistical performance", Report no. TR449. Cambridge, MA: Division of Engineering and Applied Physics, Harvard University.

Deshpande, J. G., T. N. Upadhyay and D. G. Lainiotis (1973), "Adaptive control of linear stochastic systems", Automatica, 9:107–115.

Dobell, A. Rod (1969), "Some characteristic features of optimal problems in economic theory", IEEE Transactions on Automatic Control, AC-14:4–14.

Drud, Arne (1976),"Methods for control of complex dynamic systems", Paper no. 27. Lyngby: Institute of Mathematical Statistics and Operations Research.

Drud, Arne (1977), "An optimization code for nonlinear econometric models based on sparse matrix techniques and reduced gradients—Part I, Theory". Lyngby: Department of Mathematics, Technical University of Denmark.

Eijk, C. J. van and J. Sandee (1959), "Quantitative determination of an optimal economic policy", Econometrica, 27:1–13.

Erickson, D. L. and F. E. Norton (1973), "Application of sensitivity constrained optimal control to national economic policy", in: C. T. Leondes, ed., Control and dynamic systems, Advances in theory and application, Vol. 9. New York: Academic Press.

Erickson, D. L., C. T. Leondes and F. E. Norton (1970), "Optimal decision and control policies in the national economy", in: Proceedings of the 9th IEEE symposium on adaptive processes, decision and control. Austin, TX: University of Texas.

Fair, Ray C. (1974), "On the solution of optimal control problems as maximization problems", Annals of Economic and Social Measurement, 3:135–154.
Fair, Ray C. (1975), A model of macroeconomic activity, Vol. II: The empirical model. Cambridge, MA: Ballinger.
Fair, Ray C. (1978a), "The effects of economic events on votes for president", Review of Economics and Statistics, 60:159–173.
Fair, Ray C. (1978b), "The use of optimal control techniques to measure economic performance", International Economic Review, 19:289–309.
Farison, J. B., R. E. Graham and R. C. Shelton (1967), "Identification and control of linear discrete systems", IEEE Transactions on Automatic Control, AC-12:438–442.
Fischer, Joachim and Götz Uebe (1975), "Stability and optimal control of a large linearized econometric model of Germany". Munich: Institut für Statistik and Unternehmensforschung, Techische Universität München.
Fisher, W. D. (1962), "Estimation in the linear decision model", International Economic Review, 3:1–29.
FitzGerald, V. W., H. N. Johnston and A. J. Bayes (1973), "An interactive computing algorithm for optimal policy selection with nonlinear econometric models". Canberra: Commonwealth Bureau of Census and Statistics.
Fletcher, R. and M. J. D. Powell (1963), "A rapidly convergent descent method of minimization", The Computer Journal, 6:163–168.
Fletcher, R. and C. M. Reeves (1964), "Function minimization for conjugate gradients", British Computer Journal, 7:149–154.
Friedman, Benjamin M. (1972), "Optimal economic stabilization policy: An extended framework", Journal of Political Economy, 80:1002–1022.
Friedman, Benjamin M. and E. Philip Howrey (1973), "Nonlinear models and linearly optimal policies: An evaluation", Discussion Paper no. 316. Cambridge, MA: Harvard Institute for Economic Research.
Garbade, Kenneth D. (1975a), Discretionary control of aggregate economic activity. Lexington, MA: Lexington Books.
Garbade, Kenneth D. (1975b), "Discretion in the choice of macroeconomic policies", Annals of Economic and Social Measurement, 4:215–238.
Garbade, Kenneth D. (1976), "On the existence and uniqueness of solutions to multiperiod linear quadratic control problems", International Economic Review, 17:719–732.
Gill, P. E., W. Murray, S. M. Picken, H. M. Barber and H. M. Wright (1976), "Subroutine LNSRCH and NEWPTC", Ref. no. E4/16/0 Fortran/02/76, NPL Algorithm Library, DNAC. Teddington: National Physical Laboratory.
Goldberger, Arthur S. (1963), Econometric theory. New York: Wiley.
Gordon, Roger H. (1974), "The investment tax credit as a supplementary discretionary stabilization tool". Cambridge, MA: Department of Economics, Harvard University.
Gupta, Surender K., Laurence H. Meyer, Fredric Q. Raines and Tzyh-Jong Tarn (1975), "Optimal coordination of aggregate stabilization policies: Some simulation results", Annals of Economic and Social Measurement, 4:253–270.
Hay, George A. and Charles Holt (1975), "A general solution for linear decision rules: An optimal dynamic strategy applied under uncertainty", Econometrica, 43:231–260.
Healey, A. J. and F. Medina (1975), "Economic stabilization from the monetaristic viewpoint using the dynamic Phillips curve concept". Austin, TX: Department of Mechanical Engineering, University of Texas.
Healey, A. J. and S. Summers (1974), "A suboptimal method for feedback control of the St. Louis econometric model", Transactions of the A.S.M.E., Journal of Dynamic Systems, Measurement and Control, 96:446–454.
Henderson, D. W. and S. J. Turnovsky (1972), "Optimal macroeconomic policy adjustment under conditions of risk", Journal of Economic Theory, 4:58–71.
Holbrook, Robert S. (1972), "Optimal economic policy and the problem of instrument instability", American Economic Review, 62:57–65.
Holbrook, Robert S. (1973), "An approach to the choice of optimal policy using large econometric models", Bank of Canada Staff Research Studies no. 8. Ottawa.

Holbrook, Robert S. (1974), "A practical method for controlling a large nonlinear, stochastic system", Annals of Economic and Social Measurement, 3:155–176.
Holbrook, Robert S. (1975), "Optimal policy choice under a nonlinear constraint: An iterative application of linear techniques", Journal of Money, Credit and Banking, 7:33–49.
Holly, S., B. Rustem and M. B. Zarrop, eds. (1979), Optimal control for econometric models: An approach to economic policy formulation. London: Macmillan.
Holt, C. C. (1962), "Linear decision rules for economic stabilization and growth", Quarterly Journal of Economics, 76:20–45.
Intriligator, Michael D. (1971), Mathematical optimization and economic theory. Englewood Cliffs, NJ: Prentice-Hall.
Intriligator, Michael D. (1975), "Applications of optimal control theory in economics", Synthese, 31:271–288.
Jacobson, D. H. and D. Q. Mayne (1970), Differential dynamic programming. New York: American Elsevier.
Kareken, J. H., T. Muench and N. Wallace (1973), "Optimal open market strategy: The use of information variables", American Economic Review, 63:156–172.
Kaul, T. K. and K. S. Rao (1975), "Digital simulation and optimal control of international short-term capital movements". Pilani Rajasthan: Birla Institute of Technology and Science.
Kendrick, David A. (1973), "Stochastic control in macroeconomic models", IEE conference publication no. 101. London.
Kendrick, David A. (1976), "Applications of control theory to macroeconomics", Annals of Economic and Social Measurement, 5:171–190.
Kendrick, David A. (1978), "Non-convexities from probing in an adaptive control problem", Economics Letters, 1:347–351.
Kendrick, David A. (1979), "Adaptive control of macroeconomic models with measurement error", in: S. Holly, B. Rustem and M. B. Zarrop, eds., Optimal control for econometric models, Chapter 9. London: Macmillan.
Kendrick, David A. (1981), Stochastic control for economic models. New York: McGraw-Hill.
Kendrick, David A. and J. Majors (1974), "Stochastic control with uncertain macroeconomic parameters", Automatica, 10:587–594.
Kendrick, David A. and Lance Taylor (1970), "Numerical solution of nonlinear planning models", Econometrica, 38:453–467.
Kendrick, David and Lance Taylor (1971), "Numerical methods and nonlinear optimizing models for economic planning", in: Hollis B. Chenery, ed., Studies in development planning, Chapter 1. Cambridge, MA: Harvard University Press.
Kendrick, D. A., H. Rao and C. Wells (1970), "Optimal operation of a system of waste water treatment facilities", in: Proceedings 9th IEEE symposium on adaptive processes, decision and control. Austin, TX: University of Texas.
Kim, Han H., Louis M. Goreaux and David A. Kendrick (1975), "Feedback control rule for cocoa market stabilization", in: Walter C. Labys, ed., Quantitative models of commodity markets, Chapter 9. Cambridge, MA: Ballinger.
Klein, L. R. (1979), "Managing the modern economy: Econometric specification", in: S. Holly, B. Rustem and M. B. Zarrop, eds., Optimal control for econometric models: An approach to economic policy formulation, Chapter 11. London: Macmillan.
Ku, Richard T. and Michael Athans (1973), "On the adaptive control of linear systems using the open loop feedback optimal approach", IEEE Transactions on Automatic Control, AC-18:489–493.
Ku, Richard T. and Michael Athans (1977), "Further results on the uncertainty threshold principle", IEEE Transactions on Automatic Control, AC-22:866–868.
Kydland, Finn (1973), "Decentralized macroeconomic planning", Ph.D. dissertation. Pittsburgh, PA: Carnegie-Mellon University.
Kydland, Finn (1975), "Decentralized stabilization policies: Optimization and the assignment problem", Presented at the 4th NBER stochastic control conference. Bergen: Norwegian School of Economics and Business Administration.
Lasdon, L. S., S. K. Mitter and A. D. Warren (1967), "The conjugate gradient method for optimal control problems", IEEE Transactions on Automatic Control, AC-12:132–138.

Livesey, D. A. (1971), "Optimising short-term economic policy", Economic Journal, 81:525–546.
Livesey, David A. (1978), "Feasible directions in economic policy", Journal of Optimization Theory and Applications, 25:383–406.
McFadden D. (1969), "On the controllability of decentralized macroeconomic systems: The assignment problem", in: H. W. Kuhn and G. P. Szegö, eds., Mathematical systems theory and economics I. New York: Springer-Verlag.
MacRae, Elizabeth Chase (1972), "Linear decision with experimentation", Annals of Economic and Social Measurement, 1:437–447.
MacRae, Elizabeth Chase (1975), "An adaptive learning rule for multiperiod decision problems", Econometrica, 3:893–906.
Mantell, J. B. and L. S. Lasdon (1977), "Algorithms and software for large econometric control problems", Presented at the 6th NBER conference on economics and control. New Haven, CT.
Murphy, Roy E. (1965), Adaptive processes in economic systems. New York: Academic Press.
Murtagh, Bruce A. and Michael A. Saunders (1977), "MINOS, a large-scale nonlinear programming system", Technical Report no. SOL 77-9. Stanford, CA: Department of Operations Research, Stanford University.
Myoken, Hajime (1975a), "Non-zero-sum differential games for the balance-of-payments adjustment in an open economy", International Journal of System Sciences, 6:501–511.
Norman, Alfred L. (1976), "First order dual control", Annals of Economic and Social Measurement, 5:311–322.
Norman, Alfred L. (1979), "Dual control of perfect observations", in: J. N. L. Janssen, L. M. Pau and A. Straszak, eds., Models and decision making in national economics. Amsterdam: North-Holland.
Norman, Alfred L. and Woo Sik Jung (1977), "Linear quadratic control theory for models with long lags", Econometrica, 45:905–918.
Norman, Alfred L. and M. R. Norman (1973), "Behavioral consistency test of econometric models", IEEE Transactions on Automatic Control, AC-18:465–472.
Norman, Alfred L., M. R. Norman and Carl Palash (1974), "On the computation of deterministic optimal macroeconomic policy", Paper no. 7505. New York: Federal Reserve Bank.
Oudet, B. A. (1976), "Use of the linear quadratic approach as a tool for analyzing the dynamic behavior of a model of the French economy", Annals of Economic and Social Measurement, 5:205–210.
Pagan, Adrian (1975), "Optimal control of econometric models with autocorrelated disturbance terms", International Economic Review, 16:258–263.
Palash, Carl J. (1977), "On the specification of unemployment and inflation in the objective function", Annals of Economic and Social Measurement, 6:275–300.
Paryani, K. (1972), "Optimal control of linear discrete macroeconomic systems", Ph.D. Thesis. East Lansing, MI: Department of Electrical Engineering, Michigan State University.
Pau, L. F. (1973), "Differential game among sectors in a macroeconomy", in: IFAC/IFORS international conference on dynamic modelling and control of national economies. Warwick: The Institution of Electrical Engineers, Warwick University.
Perry, A. (1976), "An improved conjugate gradient algorithm," Technical Note. Evanston, IL: Department of Decision Sciences, Graduate School of Management, Northwestern University.
Phelps, Edmund S. and John B. Taylor (1977), "Stabilization properties of monetary policy under rational price expectations", Journal of Political Economy, 85:163–190.
Phillips, A. W. (1954), "Stabilization policy in a closed economy", Economic Journal, 64:290–323.
Phillips, A. W. (1957), "Stabilization policy and the time form of the lagged responses", Economic Journal, 67:265–277.
Pindyck, Robert S. (1972), "An application of the linear quadratic tracking problem to economic stabilization policy", IEEE Transactions on Automatic Control, AC-17:287–300.
Pindyck, Robert S. (1973a), Optimal planning for economic stabilization. Amsterdam: North-Holland.
Pindyck, Robert S. (1973b), "Optimal policies for economic stabilization", Econometrica, 41:529–560.

Pindyck, Robert S. (1975), "The cost of conflicting objectives in policy formation", Presented at the 4th NBER stochastic control conference. Cambridge, MA: Sloan School, Massachusetts Institute of Technology.

Pindyck, Robert S. (1977), "Optimal economic stabilization policies under decentralized control and conflicting objectives", IEEE Transactions on Automatic Control, AC-22:517–529.

Pindyck, Robert S. and Steven M. Roberts (1974), "Optimal policies for monetary control", Annals of Economic and Social Measurement, 3:207–238.

Pitchford, John and Steve Turnovsky, eds., (1977), Application of control theory to economic analysis. Amsterdam: North-Holland.

Polack, E. and G. Ribiere (1969), "Note sur la convergence de methodes de directions conjugees" Revue Francaise Inf. Rech. Oper., 16RI:35–43.

Prescott, E. C. (1967), "Adaptive decision rules for macroeconomic planning", Doctoral Dissertation. Pittsburg, PA: Graduate School of Industrial Administration, Carnegie-Mellon University.

Prescott, E. C. (1971), "Adaptive decision rules for macroeconomic planning", Western Economic Journal, 9:369–378.

Prescott, E. C. (1972), "The multi-period control problem under uncertainty", Econometrica, 40:1043–1058.

Rausser, Gordon (1978), "Active learning, control theory, and agricultural policy", American Journal of Agricultural Economics, 60:476–490.

Rausser, Gordon and J. Freebairn (1974), "Comparison of approximate adaptive control solution to the U.S. beef trade policy", Annals of Economic and Social Measurement, 3:177–204.

Rouzier, P. (1974), "The evaluation of optimal monetary and fiscal policy with a macroeconomic model for Belgium". Louvain: Catholic University of Louvain.

Sandblom, C. L. (1970), "On control theory and economic stabilization", Ph.D. Dissertation. Lund: National Economy Institution, Lund University.

Sandblom, C. L. (1974), "Stabilization of a fluctuating simple macroeconomic model", National Economic Planning Research Paper no. 79. Birmingham: University of Birmingham.

Sandell, Nils R., Pravin Varaiya, Michael Athans and Michael G. Safonov (1977), "Survey of decentralized control methods for large scale systems", Paper no. ESL-P-777. Cambridge, MA: Electronic Systems Laboratory, Massachusetts Institute of Technology.

Sargent, T. J. and N. Wallace (1975), "'Rational' expectations, the optimal monetary instrument and the optimal money supply rule", Journal of Political Economy, 83:241–254.

Shanno, D. F. (1977), "Conjugate gradient methods with inexact searches", Working paper. Tucson, AZ: Management Information Systems, College of Business and Public Administration, University of Arizona.

Shell, Karl, ed. (1967). Essays on the theory of optimal economic growth. Cambridge, MA: M.I.T. Press.

Shupp, Franklin R. (1972) "Uncertainty and stabilization policies for a nonlinear macroeconomic model", Quarterly Journal of Economics, 80:94–110.

Shupp, Franklin R. (1976a), "Optimal policy rules for a temporary incomes policy", Review of Economic Studies, 43:249–259.

Shupp, Franklin R. (1976b), "Uncertainty and optimal policy intensity in fiscal and incomes policies", Annals of Economic and Social Measurement, 5:225–238.

Shupp, Franklin R. (1976c), "Uncertainty and optimal stabilization policies", Journal of Public Finance, 6:243–253.

Shupp, Franklin R. (1977), "On optimal and ad hoc stabilization policy rules", Economic Inquiry, 15:183–198.

Simon, H. A. (1956), "Dynamic programming under uncertainty with a quadratic criterion function", Econometrica, 24:74–81.

Taylor, John B. (1973), "A criterion for multiperiod control in economic models with unknown parameters", Discussion paper no. 73-7406. New York: Columbia University.

Taylor, John B. (1974), "Asymptotic properties of multiperiod control rules in the linear regression model", International Economic Review, 15:472–484.

Thalberg, Bjorn (1971a), "Stabilization policy and the nonlinear theory of the trade cycle", The Swedish Journal of Economics, 73:294–310.

Thalberg, Bjorn (1971b), "A note on Phillips' elementary conclusions on the problems of stabilization policy", The Swedish Journal of Economics, 73:1385–1408.
Theil, H. (1957), "A note on certainty equivalence in dynamic planning", Econometrica, 25:346–349.
Theil, H. (1964), Optimal decision rules for government and industry. Amsterdam: North-Holland.
Theil, H. (1965), "Linear decision rules for macro-dynamic policy problems, in: B. Hickman, ed., Quantitative planning of economic policy. Washington, DC: The Brookings Institute.
Tinsley, P., R. Craine and A. Havenner (1974), "On NEREF solutions of macroeconomic tracking problems", Presented at the 3rd NBER stochastic control conference. Washington, DC: Federal Reserve Bank.
Tinsley, P., R. Craine and A. Havenner (1975), "Optimal control of large nonlinear stochastic econometric models", in: Proceedings of the summer computer simulation conference. San Francisco, CA.
Tse, Edison and Michael Athans (1972), "Adaptive stochastic control for a class of linear systems", IEEE Transactions on Automatic Control, AC-17:38–52.
Tse, Edison and Y. Bar-Shalom (1973), "An actively adaptive control for linear systems with random parameters", IEEE Transactions on Automatic Control, AC-18: 109–117.
Tse, Edison, Y. Bar-Shalom and L. Meier (1973), "Wide sense adaptive dual control for nonlinear stochastic systems", IEEE Transactions on Automatic Control, AC-18:98–108.
Turnovsky, Stephen J. (1973), "Optimal stabilization policies for deterministic and stochastic linear systems", Review of Economic Studies, 40:79–96.
Turnovsky, Stephen J. (1974), "Stability properties of optimal economic policies", American Economic Review, 64:136–147.
Turnovsky, Stephen J. (1975), "Optimal choice of monetary instruments in a linear economic model with stochastic coefficients", Journal of Money, Credit, and Banking, 7:51–80.
Turnovsky, Stephen J. (1977), "Optimal control of linear systems with stochastic coefficients and additive disturbances", in: J. Pitchford and S. J. Turnovsky, Application of control theory to economic analysis, Chapter 11. Amsterdam: North-Holland.
Tustin, A. (1953), The mechanism of economic systems. London: Heinemann; Cambridge, MA: Harvard University Press.
Upadhyay, Treveni (1975), "Application of adaptive control to economic stabilization policy", International Journal of System Science, 6:641–650.
Wall, K. D. and J. H. Westcott (1974), "Macroeconomic modeling for control", IEEE Transactions on Automatic Control, AC-19:862–873.
Wall, K. D. and J. H. Westcott (1975), "Policy optimization studies with a simple control model of the U.K. economy", Proceedings of the IFAC/75 congress. Boston–Cambridge, MA.
Walsh, Peter and J. B. Cruz (1975), "Neighboring stochastic control of an econometric model", Presented at the 4th NBER stochastic control conference. Urbana, IL: Coordinated Science Lab., University of Illinois.
Woodside, M. (1973), "Uncertainty in policy optimization—experiments on a large econometric model", in: IFAC/IFORS international conference on dynamic modelling and control of national economies, I.E.E. conference publication no. 101. Warwick: The Institution of Electrical Engineers, Warwick University.
You, Jong Keun (1975), "A sensitivity analysis of optimal stochastic control policies", Presented at the 4th NBER stochastic control conference. New Brunswick, NJ: Rutgers University.
Zellner, Arnold (1966), "On controlling and learning about a normal regression model". Chicago, IL: School of Business, University of Chicago.
Zellner, Arnold (1971), An introduction to Bayesian inference in econometrics. New York: Wiley.
Zellner, Arnold and M. S. Geisel (1968), "Sensitivity of control to uncertainty and form of the criterion function", in: D. G. Watts, ed., The future of statistics. New York: Academic Press.

Chapter 5

MEASURE THEORY WITH APPLICATIONS TO ECONOMICS

ALAN P. KIRMAN

Université d'Aix-Marseille

This chapter will first present problems arising from economic theory, the modelling of which has required, in an essential way, measure theory. Having explained why measure theory is needed, we will give, for reference, some basic measure theoretical concepts and results, and this will be followed by a development and discussion of some particularly useful and less accessible results.[1]

1. The use of measure theory in economics

1.1. Perfect competition: Large economies

The idea of "perfect" or "pure" competition is a very old one in economics.[2] Any economist will have an intuitive idea as to what is meant by it, though the definitions may vary. The underlying principle may be captured by saying that a situation in which a group of individuals together are involved in economic activity, exchange for example, and in which no individual can, and therefore no individual will try to, affect the outcome is one of perfect competition. The first and most obvious requirement for such a situation is that there should be "many" individuals. This is clearly not enough; we also need that none of these individuals should be "important". The conceptual difficulty arises if we really insist that each individual shall have no influence whatsoever. As an example, think of the productive sector of an economy, and then what we require for it to be "perfectly competitive" would be, for example, that if a producer stopped production completely, the price of the commodity he produces should not

[1] The presentation is aimed at the "informed reader", that is, someone acquainted with the basic ideas of measure theory as presented in a course on probability theory. The better informed mathematician will see from the first and third sections where notions familiar to him are used in economics. The reader who is not sure where he stands should read quickly through Section 2. If it has a familiar ring, he should find the chapter profitable. If not, he would be well advised to first consult any standard text on measure theory, the classic reference being Halmos (1961).

[2] Adam Smith is clearly aware in the *Wealth of Nations* (Book 1, ch. 7, for example) that the existence of large numbers of agents, that is of a situation approaching perfect competition, diminishes the power of an individual agent to influence prices in a market process.

change. If prices can vary, then for this to be strictly true for every producer, there must be an infinite number of producers, or, in other words, his "influence as a producer must be genuinely negligible". As the term "perfect competition" suggests, this is an idealisation, not a description of reality; but the examination of such an ideal case, as in other sciences, provides us with useful insights into the working of economics.[3]

The idea that individuals should have no weight but that collectively they have positive weight is a familiar one in mathematics, and it is the basis of measure theory. If we wished merely to describe a perfectly competitive economy, it would be enough to consider any infinite set as the set of agents, or agents' names, and to specify the "characteristics" of each agent. Thus, in an exchange economy, agents are characterized by two things: preferences and an initial bundle of goods. Without entering into any details, consider the set of possible preferences as $\mathcal{P}$ and the set of bundles of goods on which these preferences are defined as R^{ℓ}_{+}, that is, there are ℓ goods. Then an exchange economy $\mathcal{E}$ is given by

$$\mathcal{E}: A \rightarrow \mathcal{P} \times R^{\ell}_{+},$$

where A is some arbitrary set of agents. Now, if we require that there are an infinite number of agents in A and that no agent has too much of any goods, e.g. that we restrict attention to some bounded set of R^{ℓ}_{+} for initial endowments, then we have a description of a perfectly competitive exchange economy. Provided that all we require is a definition, this would indeed suffice. However, if we wish to work with this model, we will need more than this. Suppose that we are concerned with the problem of competitive or Walrasian equilibrium. We need now some way of expressing the idea that for some *allocation of goods* f "supply equals demand".[4] In a finite economy, we simply add the demands of all the individuals and check that this is equal to the sum of all the initial resources. Thus we require that, referring to the initial bundle of an individual as $e(a)$,[5] that

$$\sum_{a \in A} f(a) = \sum_{a \in A} e(a), \tag{1.1}$$

and that $f(a) \in \varphi(p, a)$ where φ, the demand of an individual a at prices p, is defined in the normal way. Now, in our infinite economy, we can no longer add

[3] A discussion of the notion of perfect competition and its relation to recent theoretical developments is to be found in Mas-Colell (1979).

[4] An allocation of goods is here $f: A \rightarrow R^{\ell}_{+}$.

[5] $e(.)$ is the projection of the mapping $\mathcal{E}$ onto R^{ℓ}_{+}.

supply and demand. Instead we can substitute the idea that *mean supply is equal to mean demand*. In a finite economy, we could write

$$\frac{1}{\#A}\sum_{a\in A} f(a)=\frac{1}{\#A}\sum_{a\in A} e(a), \tag{1.2}$$

which is clearly identical to (1.1). However, in the infinite case, we can resort to the equivalent idea, one which will be familiar to all those who know a little probability theory, and write

$$\int_A f(a)\,\mathrm{d}\nu=\int_A e(a)\,\mathrm{d}\nu. \tag{1.3}$$

In writing (1.3), although its intuitive meaning is clear, we have introduced a number of technical complications. We have *integrated* but for this to be well defined, we have to integrate "with respect to some measure", that is, we must define a function ν which attributes a certain weight to each set of individuals. Intuitively, we can think of this as a "counting measure", i.e., one which says what proportion of individuals are in each set. The only important thing for us, for the moment, is that for an infinite economy such a measure should give zero weight to individuals, and for convenience, that it should give weight one to the whole set. Such a function is an *atomless probability measure*.[6] The machinery of measure theory provides a convenient way of resolving many economic problems in the context of such ideal economies. To use the standard tools of this theory imposes some technical restrictions, which will be specified in the next two sections, but suffice it to say that to construct an idealised or *perfectly competitive economy*, we take the set of agents A to be represented by an *atomless measure space* $(A, \mathcal{A}, \nu)$, the three components being the set A, the collection $\mathcal{A}$ of subsets on which the measure ν is defined,[7] and the measure ν itself.

The notion of an ideal economy, in the context discussed, was introduced by Aumann (1964), but a continuum of agents had already been used in economics by Allen and Bowley (1935), and in game theory by Shapley (1953), and in a number of other papers in the early 1900's. The idea of perfect competition in the sense that individuals believe that prices are given and beyond their control has a long history, accounts of which can be found in Schumpeter (1954) and Blaug (1968) for example, but it is only with the introduction of the "continuum theory" that such a behaviour is strictly justified. Indeed in the work of Torrens, Cournot, and Edgeworth is to be found a discussion as to whether it is rational for individuals to behave in this way. As Viner remarked, the fact that it is not has remained a "skeleton in the cupboard of free trade". The use of a measure

[6] We will in Section 2 come back to the precise definition of "atomless", and the reader will hopefully pardon a slight looseness in the statements above.

[7] Unfortunately, this is not $P(A)$, the set of all subsets of A, but more of this later.

space of agents thus enables us to formalise the notion of perfect competition,[8] but the next question is obviously: "Does it enable us to develop stronger results?" A first result showing how the assumption that an economy is large, in the sense described, leads us to drop assumptions necessary in the finite case, concerns the existence of equilibrium.[9] In a finite economy, we typically need to make an assumption about the convexity of the preferences of individuals to prove the existence of equilibrium. If we make a strong assumption that preferences are strictly convex, then the bundle demanded by an individual a at prices p, denoted $\varphi(a, p)$, will be unique, that is, φ will be a function. If we weaken the assumption to making preferences convex, then $\varphi(a, p)$ will be a set but a convex one, and we will have

$$\sum_{a \in A} \varphi(a, p) \quad \text{is convex for all } p.$$

Provided that total demand, or equivalently mean demand, is a convex set, we can prove that equilibrium exists.[10] If however individuals have nonconvex preferences, their demand may not be a convex set for some prices and the proof of existence no longer goes through. To look at this question in the continuum case, we must first be able to define the integral of individuals' demands, which are set-valued functions or correspondences. The *integration of correspondences* is discussed in Section 3 of this chapter. The important result is that even if we do not assume individual's preferences to be convex, nevertheless $\int_A \varphi(a, p) \mathrm{d}\nu$ is convex.

Thus, what might be thought of as irregular behaviour in individuals becomes "well-behaved" in large economics. This fact enables one to prove the existence of equilibria in large economies under weaker assumptions than in the finite case. See Aumann (1966) and Hildenbrand (1970).

1.2. Different solutions for the market problem

A further important result proved by Aumann (1964) was the equivalence between two different solution concepts in a continuum economy. One based on

[8] The approach adopted here is not by any means the only possible one. We use σ additive measures, and it may be possible to work with only finitely additive measures, but this slight conceptual weakening of assumptions leads to other complications in the definition of "atomless" for example. Another different approach is to consider agents as infinitesimally small but not null. To do this involves using non-standard analysis developed by Robinson (1965) and used in economics by Brown–Robinson (1974) and Khan (1973). The disadvantage of this approach is that the mathematical apparatus employed is familiar to a very limited audience.

[9] The essential result in this connection is Liapunov's theorem which will be given later.

[10] The standard discussion of this problem is given in Debreu's *Theory of Value* (1959), and a complete survey of the work in this area is given in Chapter 15 of this Handbook.

the price mechanism gives us the set of allocations which are equilibria denoted $W(\mathcal{E})$ and the other, the *core*, is the set of allocations upon which no coalition S of individuals can improve. "Improve upon" in this sense means that a coalition S of agents could reallocate its initial resources to make its members better off. Thus in a continuum economy $\mathcal{E}$, for example, where the set of agents A is the closed unit interval of the real line, an allocation f would be improved upon by S if the members can find g with

$$g(a) \succ_a f(a) \quad \text{for all members of } S,^{11}$$

and (1.4)

$$\int_S g(a)\,\mathrm{d}\nu = \int_S e(a)\,\mathrm{d}\nu.^{12}$$

Allocations which can be improved by no coalition form the core of the economy $\mathcal{E}$ denoted $C(\mathcal{E})$.

Aumann's result is that for "continuum economies",

$$W(\mathcal{E}) = C(\mathcal{E}).$$

This exact equality for an ideal economy confirmed in a more general setting an old asymptotic result of Edgeworth (1881) and a later result of Debreu and Scarf (1963) and gave rise in turn to a whole series of very general asymptotic results which are treated in detail in Chapter 18 of this Handbook on the core, and to which we will return shortly.

In discussing perfect competition, we have given an idea as to why atomless measure spaces provide a useful formalisation of the idea of a large economy in which each agent is insignificant. If this were indeed the only value of such tools, then it would be difficult to persuade economic theorists of the virtue of acquiring them. In fact, measure theory provides extremely useful insights at a conceptual level.

1.3. Distributions of characteristics

In a large economy, listing the characteristics of all the individual agents would be both a tedious and an elaborate task. Indeed, economists often make the simplifying step of describing an economy by the *distribution of its agents' characteristics*. The idea of the income distribution and describing it by some

[11] $x \succ_a y$ denotes that agent a strictly prefers x to y.

[12] The informed reader will note that (1.4) is not, for technical reasons, defined for all coalitions S; details will be forthcoming in Section 2.

such function as the Pareto distribution is well established in economics. The use of such functions relies implicitly on the idea that a large economy may be represented as a continuum, and the measure space of agents approach leads naturally to the development of the distribution as a fundamental concept.

If we consider a mapping from a probability space into the space of characteristics, then it is clear that a natural probability measure is induced on the latter. If we take a subset B of the characteristics space then consider the set C in the original space whose image lies in that subset, that is $C=\mathfrak{S}^{-1}(B)$. Now, let the measure of B be the measure of C; this gives us a measure on the characteristics space itself. Thus, instead of asking which agent has which characteristic in an economy, we might ask what proportion of agents have certain characteristics? Instead of thinking of an economy as a detailed listing of all the characteristics of the agents in that economy, we can think of it as a distribution of characteristics. Indeed as we have said, economists are in the habit of viewing economies as characterized by their income distribution, for example. We might, indeed, reasonably say that two economies for which the distributions of agents characteristics are the same are effectively the same economy. For this to be acceptable, we would have to show that the equilibria of these economies are the same. A full treatment of this sort of problem may be found in Hildenbrand (1975).

A little more formally, consider $(A, \mathfrak{A}, \mu)$ a probability space, M the space of characteristics, and f a mapping of A into M. The distribution ν of f denoted by $\mu \circ f^{-1}$ is defined by

$$\nu(B)=\mu\{a\in A \mid f(a)\in B\} \quad \text{for every subset } B \text{ of } M.$$

The reader will already be familiar with this idea from probability theory and will recognise f as a random element and, in particular if M were the real line, would recognise f as a random variable.

Now, since for many purposes we take some arbitrary basic measure space as a starting point, it is frequently the distribution that conveys the real information with which we are concerned. For example, in studying "large economies", the choice of the unit interval $[0,1]$ where it is used as the space of agents is purely for convenience and has no particular significance itself. In fact, given a suitable distribution on the space of characteristics of agents, we could always construct an associated economy with the unit interval as the space of agents.

1.4. Limit theorems

The idea of using distributions as the description of the essential features of an economy proved extremely useful in translating results from ideal or continuum economies to large but finite economies. For results on ideal economies to be of

any interest, they must also hold, at least approximately, for large enough finite economies.

Thus, rather than make a statement that such and such a result is true for a continuum economy say $\mathcal{E}_\infty$, we would like to construct an increasing sequence of economies $\mathcal{E}_n$ converging in some sense to $\mathcal{E}_\infty$ and then make the assertion that our result is approximately true for large enough n. The problem is that if we think of our economies as being listings of all the characteristics of the agents, the dimension of this description changes as more agents are added and as the economies of the sequence increase in size. How then can we construct an increasing sequence of economies and in what sense can that sequence be said to converge to the limit, atomless, economy?

An important key to solving this problem is that we can construct a sequence of parallel "equivalent" economies each with a continuum of agents and establish our results via this "equivalent" sequence. However, we will need to establish the meaning of the "equivalence" between the original sequence of finite economies and the sequence of artificially constructed economies. To handle these problems, we will need a number of mathematical tools, in particular, we will need to study the *convergence of measures* or more exactly *weak convergence of measures*. We will need subsequently to develop the idea of *convergence in distribution* so that we can give precision to the requirement that for a given sequence of economies "the distribution of agents' characteristics" should be "close", for n large, to that of the limit economy.

1.5. Many but different agents

As must by now be evident, much of the value of measure theoretical tools is to handle situations in which there are "many" agents. We have discussed the weakening of assumptions possible in "ideal" economies to achieve standard results. Thus the assumption of large numbers may be seen to be a substitute for restrictive hypotheses at the individual level. Sometimes however we need more than simply "many" agents. We will need that the agents are, in some sense, different, thus not only numbers but variety will be important. If we think of the distribution of agents characteristics, then we could require for example that the *support* of that measure should not be "too small", the support of a measure being the smallest set that has full measure. Thus we would require that peoples' characteristics in an economy are not too similar.

For what sort of economic problem is this of interest? A well-known difficulty in economics is that associated with the assumption of strict convexity of preferences. Although every elementary text in micro-economics has diagrams of preferences which inevitably satisfy this hypothesis, only the most hardened economic theorist feels completely at ease with it. Plausible counter-examples are so easy to find that one would be happy to dispense with it. However, the

formal difficulties that arise when it is removed are far from trivial to overcome. In particular, as we have already remarked, since at given prices p the bundle of goods demanded by agent a, that is $\varphi(a, p)$, is not necessarily unique, one can no longer work with demand functions. However, intuitively it is clear that if there is a large number of agents and the number of these who have more than one element in their demand set, is "negligible", then we have essentially what we require. For this idea to make perfect sense, we must have an infinite number of agents. Now, if we have an infinite number of agents, what we need is that "mean demand" should be unique. For this we will have to integrate over our agents,[13] and hence what we must show is that the "bad" set of agents have *measure zero*. For this we obviously must require that the preferences are sufficiently "dispersed". Results in this direction using assumptions of differentiability have been obtained by Sondermann (1975), Dierker, Dierker and Trockel (1978) and Araujo and Mas Colell (1978). Hildenbrand (1979) has shown with a suitable assumption about dispersion of preferences that the almost sure uniqueness of maximisers and hence the continuity of mean demand functions can be obtained without any differentiability assumptions. Again the usefulness of measure-theoretic tools in making precise an intuitive idea should be emphasised. For the use of continuous demand functions to be strictly justified in a context of non-strictly convex preferences, an infinite number of agents is essential, and to use the natural notion of the mean demand function, the measure theoretical approach is necessary.

Before leaving this topic, an important observation should be made. How are the above results obtained? They depend on showing that a certain phenomenon is "exceptional" or "unusual". The significance of this is that for a long time, unless we made extremely restrictive assumptions in economics, we were unable to rule out intractable situations even though it seemed unlikely that they might occur. One approach to this is topological. Thus rather than make strong assumptions to rule out certain phenomena one can show that the set of economies that exhibits these phenomena is "negligible", that is, that the set of well-behaved economies is *open and dense* in the set of all the economies under consideration.[14]

Thus, one can in a certain sense ignore such phenomena. One might also like to say that, in a probabilistic sense, certain things are unlikely, or more precisely, that the set of objects, economies, for example, exhibiting certain phenomena has measure *zero*, or that such a phenomena is "almost sure" not to occur.[15] In

[13]Agents are, in fact, identified by preferences, and A is an open subset of R^n. Thus preferences or utility functions can be classified by n parameters.

[14]The pitfalls of too facile a use of this approach are alluded to in Grandmont, Kirman and Neuefeind (1974), and the same strictures of course apply to the measure-theoretic approach.

[15]A fundamental paper which shows that economies with an infinite number of equilibria are unlikely in both the topological and probabilistic sense is that of Debreu (1970).

the papers mentioned above on the uniqueness of maximising elements, it is precisely this notion that allows the passage from individual demand correspondences to mean demand functions.

1.6. Price forecasting, tight measures, and compactness

In many situations we are led to introduce restrictions of the opposite sort of those mentioned earlier. When, for example, we want to establish existence of an equilibrium, we will need certain "compactness" properties. In particular, if we define measures on a space which is not itself "compact", we will need to restrict ourselves to families of measures which are, in a technical sense, concentrated essentially on a compact set. This technical requirement arises naturally in work on temporary equilibria.[16]

Consider traders who base their forecasts of future prices on today's prices. Thus any price vector today generates a measure on the space of tomorrow's prices. In a model of this sort, to ensure the existence of an equilibrium, one is led to assume that tomorrow's prices do not depend "too strongly" on today's prices. In other words, if some prices today become very high, then individuals attach a low probability to their being exceeded tomorrow. This rules out, for example, the simple-minded forecast that tomorrow's prices will, with probability one, be equal to today's prices.

The underlying stabilising assumption is clear; what we want is that if prices become very high today, for example, traders will attach a high probability to their diminishing tomorrow, and it is this that prevents prices exploding. The formal requirement is that the family of measures or forecasts should be *tight*. This requirement also plays an important role in work on large economies.[17]

1.7. Social choice with many agents

Arrow (1963) proved a theorem which is widely regarded as the most important in the field of social choice.[18] What he showed was that there is no rule for aggregating individual preferences, which respects certain apparently reasonable conditions. It was later shown by Fishburn (1970) that Arrow's theorem is not true if there is an infinite number of individuals in the society in question. This has been interpreted as meaning that in large societies Arrow's result loses its significance and its importance is thus diminished. However with many agents, we may obtain a measure theoretic equivalent of Arrow's theorem; see Kirman and Sondermann (1972).

[16]See Chapter 19 by Grandmont.
[17]See Chapter 18 by Hildenbrand.
[18]See Chapter 22 by Sen.

To sketch the problem briefly, consider A the set of individuals, X a set of alternatives, and P the set of preorders (preferences) on X. What we are looking for is a rule that will associate with a given distribution of preferences among the individuals (we will call this a "situation"), preferences for the society.

Let $f: A \to P$ be a situation, then $\mathcal{F}$ is the set of all possible situations. Then a social preference rule is $\sigma: \mathcal{F} \to P$. What Arrow shows is that given certain reasonable restrictions on σ the only rule that exists is the following "dictatorial" one:

Choose one individual a and, no matter what the preferences of the other individuals, if a prefers x to y, then x is socially preferred to y. Written with the obvious notation

$$xf(a)y \quad \text{implies} \quad x\sigma(f)y.$$

Since Arrow rules out such a dictatorial function, no social welfare function σ is possible. The mathematical structure of this problem is now well-known. The Arrovian axioms impose a very specific structure on the sets of individuals who are "socially decisive". That is the set B is decisive if, when all the members of B prefer x to y, then x is socially preferred to y. In the case where the set A of all individuals is finite, these socially decisive sets consist of all the sets that contain a given individual a and, in particular, the set $\{a\}$ consisting of just a himself. Now, suppose that the set A is infinite, for example, the interval $[0, 1]$, then we could, from Arrow's axioms, define a measure which could give weight 1 to the decisive sets and 0 to the others. If Arrow's theorem translated directly to this case, the measure μ would necessarily have the form

$$\mu(C) = 1 \quad \text{if and only if} \quad a \in C.$$

Thus a would be the dictator. In particular, note that such a measure is *not atomless* and that, for this reason, unlike the other measures with which we shall work, it is defined on every subset of A.

However, we know that Arrow's result does not hold in this case, but we also know that to discuss single individuals in such a case does not make much sense. What we can show is a different sort of result. If A is $[0, 1]$ then, given Arrow's axioms, any social rule σ has the following property:

Given any ε, *there is a socially decisive set* C *with* $\mu(C) < \varepsilon$,

where μ is the natural Lesbegue measure, i.e., the "length" of the set C. Thus, though no single individual determines society's preferences, arbitrarily small coalitions do so. Thus, the measure theoretic approach enables us to show that Arrow's result remains essentially true even in the infinite case.

1.8. How to cut a cake fairly

An old problem which has intrigued mathematicians is that of how to divide up some object "fairly" in some sense, among n individuals. The object to be interesting, of course, must be differently appreciated by different individuals. One could think of a block of ice cream with different flavours. Thus one could think of each individual i assigning a "measure" μ_i to the parts of the ice cream, each attributing 1 to the whole for example. Thus, what we would like is to find a way of dividing the ice cream U, i.e. a partition of U, $\{U_1,\ldots,U_n\}$, such that

$$\mu_i(U_i) \geqq 1/n \quad \text{for} \quad i=1,\ldots,n.$$

This would be fair in the sense that each individual receives in his own eyes at least $1/n$ of the value of the ice cream. In the case of two individuals, all those who have children will know that the method of "divide and choose" solves the problem. However, much better results in the n person case have been proved by Steinhaus, Banach and Knaster, and references are given and very general theorems proved in an elegant paper by Dubins and Spanier (1961). A very striking result shows that one can partition the ice cream or cake in question in such a way that each individual n believes that *all* the pieces of the cake are worth $1/n$. That is, one can find a partition $\{U_1,\ldots,U_n\}$ such that

$$\mu_i(U_j)=1/n \quad \text{for} \quad i=1,\ldots,n \quad \text{and} \quad j=1,\ldots,n.$$

This rules out an individual getting $1/n$ of the cake but being jealous of another individual. This is in fact equivalent to the old problem of the agricultural land of an Egyptian village which is flooded by the Nile to different heights each year. How should the land be divided so that each of the n landowners always has $1/n$ of the land remaining above water?

In addition, Dubins and Spanier show that there are "optimal" partitions in different senses. For example, there are partitions $\{U_1,\ldots,U_n\}$ which maximise

$$\sum_{i=1}^{n} \mu_i(U_i),$$

thus which are optimal in a utilitarian sense. Connections with other mathematical results are shown in their paper and the central role played by *Liapunov's theorem* mentioned above is clear.

Here again, the measure-theoretical approach has solved a number of interesting problems arising in an economic context. Having given a number of examples to motivate the use of measure theory in economics, we now turn to the mathematical tools themselves.

2. Some basic measure theory

The area covered by measure theory may be thought of as that concerned with attributing numbers to the parts of an object or set in such a way that these numbers correspond intuitively to the "size" or "measure" of those parts. Physically one might think of the weight of some object and its component parts, or if one takes an interval of the real line, one might be interested in the "length" of some subset of that interval. Again, from the point of view of intuition, it is important that if the numbers assigned are to be meaningful, they should have certain properties of additivity. Thus, if one takes two disjoint parts of an object, one would naturally require that the weights of these two parts taken together should equal the sum of their separate weights. Indeed, we would require that this be true not just for any two sets, but for arbitrary collections of subsets. The passage from the simple idea of adding the weights of a finite collection of subsets to find the weight of their union to the problem of adding the weights of an arbitrary collection of subsets is not even in general possible, and we will have to restrict ourselves to a less ambitious task. The specification of the functions that designate the measure of each subset of some set, the collection of subsets on which they are defined, and the properties of those functions will be the concern of the second part of this chapter.

2.1. Classes of subsets and algebras

Before developing the theory of set functions and measures in particular, we must first study the classes of subsets on which they are defined. If we consider any set E then we will denote $\mathcal{P}(E)$ the set of all subsets of E.

Definition 1

An *algebra* or *Boolean algebra* of sets $\mathcal{A}$ is a non-empty class of subsets of E such that if

$$A \in \mathcal{A} \quad \text{and} \quad B \in \mathcal{A},$$

then

$$A \cup B \in \mathcal{A} \quad \text{and} \quad A \setminus B \in \mathcal{A},$$

and

$$E \in \mathcal{A}.$$

It follows obviously that if $\mathcal{A}$ is an algebra, then

$$A \in \mathcal{A} \quad \text{implies} \quad A^c \in \mathcal{A}.$$

Examples

It is clear that for any set E, $\mathcal{P}(E)$ is a Boolean algebra.

The set of all intervals on the real line does not form an algebra since it is closed neither under the operation of difference nor that of union. However, the reader will be able to show that the set of all finite unions of intervals is an algebra.

We will need to consider classes of sets where there are members which cannot only be formed by the finite union of other members but also by countable unions. That is, if we consider some set E then we will need to be able to talk of the "weight", "size", or "measure" of some set which can be made up of a countable number of "pieces" of E. Thus we have:

Definition 2

A σ *algebra* is an algebra with the property that if

$$A_i \in \mathcal{A} \quad \text{then} \quad \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}, \qquad i = 1,2,\ldots$$

It is clear that the countable intersection of sets in a σ algebra belongs to that σ algebra.

2.2. *Generated algebras and σ algebras*

If we could always work with the set of all subsets of some set E, that is with $\mathcal{P}(E)$, and could define a measure on such subsets, things would be very simple. However, this is not possible and we have to restrict ourselves to subclasses of $\mathcal{P}(E)$. It is for this reason that we have introduced notions of algebras and σ algebras. In particular, it will often be useful to start with some simple class of subsets and to construct from it a larger class. To this end, we give the following:

Theorem 1

If $\mathcal{A}$ is any class of sets then there exists a unique algebra (resp. σ algebra) such that $R \supset \mathcal{A}$, and if R' is also an algebra (resp. σ algebra) such that $R' \supset \mathcal{A}$, then

$$R \subset R'.$$

The class R is referred to as the algebra (resp. σ algebra) generated by the class $\mathcal{A}$.

A particularly important class of sets is given by the smallest σ algebra containing the open sets of some topological space. Formally, we have:

Definition 3

For a topological space X the *class of Borel sets* is the σ algebra $\mathfrak{B}$ generated by the open sets of X. The reader will have no difficulty in showing that the Borel sets are also generated by the closed sets of X.

It will be useful later to work with the σ algebra generated by a class of sets. Although it is unfortunately impossible to give a constructive procedure for obtaining this σ algebra, this will not, at the level of presentation here, present any difficulty.

With these simple set structures in mind, we now pass on to consider set functions, and in particular those set functions which are called measures.

2.3. *Set functions*

We will confine our attention to set functions which will be defined on a non-empty class $\mathcal{C}$ of subsets of some set E. Thus μ associates with a set $A \in \mathcal{C}$ a real number or $\pm\infty$. The empty set $\varnothing$ is always a member of $\mathcal{C}$. If we denote by R^* the compactification of the real line by the addition of the two points $+\infty$ and $-\infty$, then the operations represented by $+$ and $\times$ are extended in the conventional way, for example,

$$0 \times \pm\infty = 0.$$

The purpose of this chapter is not to consider arbitrary functions of abstract interest, but to tie ourselves to those which will be of use in economic theory. A first condition that the functions must satisfy if they are to correspond to the intuitive idea of assigning "weights" or "lengths" is that the "weight" of two disjoint sets taken together should be equal to the sum of their individual weights.

Definition 4

A set function $\mu : \mathcal{C} \to R^*$ is said to be (*finitely*) *additive* if

(i) $\mu(\varnothing) = 0$,

(ii) for every finite collection $E_1, E_2, \ldots, E_n$ of disjoint sets of $\mathcal{C}$ such that $\bigcup_{i=1}^{n} E_i \in \mathcal{C}$,

then

$$\mu\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} \mu(E_i). \tag{2.1}$$

In fact, condition (i) is superfluous provided that for some set E in $\mathcal{A}$, $\mu(E)$ is finite. It is natural to define an additive set function on an algebra since we have

$$E_i \in \mathcal{A} \quad \text{implies} \quad \bigcup_{i=1}^{n} E_i \in \mathcal{A}, \qquad i = 1, \ldots, n.$$

The reader will note that if $\mathcal{A}$ is an algebra, we cannot have sets E and $F \in \mathcal{A}$ with $E \cap F = \varnothing$ and $\mu(E) = +\infty$ and $\mu(F) = -\infty$. Thus, although it is not always sufficient to confine our attention to finite valued set functions, they will not take on both the values $+\infty$ and $-\infty$.

We will now give several examples of set functions which are additive which will aid in understanding the nature of measure.

Example 1

Consider X any set with infinitely many points and the set of all subsets of X. Define μ by

$$\begin{aligned} \mu(E) &= \#E \quad \text{if } E \text{ is finite,} \\ &= +\infty \quad \text{if } E \text{ is infinite,} \end{aligned} \qquad \text{for} \quad E \in \mathcal{P}(X).$$

Thus if X represents the individuals in a large economy, this measure simply "counts" the agents in any coalition. We will encounter a more useful "counting measure" later in the chapter.

Example 2

Consider X any set and define $\mathcal{P}(X)$ as before.

For $\hat{x}$ a point of X, let

$$\begin{aligned} \mu(A) &= 1 \quad \text{if} \quad \hat{x} \in A, \\ &= 0 \quad \text{if} \quad \hat{x} \notin A, \end{aligned} \qquad \text{for} \quad A \in \mathcal{P}(X).$$

Example 3

Let $X = R$ and let $\mathcal{A}$ be the set of all finite intervals of R. Any $E \in \mathcal{A}$ is then defined by its end points a, b, and let

$$\mu(E) = b - a.$$

Thus we simply take the value of an interval to be its length.

Example 4

Let X be the half open interval $(0, 1]$ and let $\mathcal{A}$ be the class of half open intervals $(a, b]$ with $0 \leqslant a \leqslant b \leqslant 1$, and let

$$\mu(a, b) = b - a \quad \text{if} \quad a \neq 0,$$

and

$$\mu(0, b) = +\infty.$$

All these examples are of additive set functions, but we will come back to see whether they satisfy the additional criteria that we will impose.

Having defined finite additivity, we will now give a stronger requirement—that of σ additivity—that is we will ask that our set function should be additive not only on a finite union of sets but on countable unions as well. Why is this necessary? The following example from probability theory gives a clear answer.

Definition 5

Consider a set X and $\mathcal{A}$ an algebra of subsets of X. Define a finitely additive function,

$$\mu: \mathcal{A} \rightarrow [0, 1],$$

with $\mu(X) = 1$. Such a function is called a *probability distribution*. If one thinks of an experiment with a number of possible outcomes then $\mu(S)$ expresses the intuitive idea of the probability that the outcome of the experiment will be in the set S.

Now consider a map $f: X \rightarrow R$. Such a map is called a *simple random variable*.[19] Thus, it associates a real number to each possible outcome of an experiment.

Next consider an infinite sequence of independent trials of the experiment. That is, from the population X is drawn each time, according to the probability distribution μ, an element $x \in X$. An outcome then may be represented as $(x_1, x_2, \ldots)$. Let X_∞ be the space of all such outcomes.

If we wish to be able to make such statements as "the 'sample mean' of n observations converges to some number α as $n \rightarrow \infty$", we will need to construct sets which can only be obtained in a countable and not a finite number of operations. That is, to construct the set of all sequences in X_∞ whose sample mean converges to α is not possible in a finite number of operations.

[19] In fact this map must satisfy a regularity condition, that of measurability, which we will shortly define but for the purpose of the example we will ignore this requirement.

σ algebra

Once again, we assume we are interested in collections of subsets of some set X which have X itself as a member. We can thus define:

Definition 6

A collection $\mathcal{A}$ of subsets of a set X is called a σ *algebra*, if

(i) $\varnothing \in \mathcal{A}$,
(ii) $A \in \mathcal{A}$ then $A^c \in \mathcal{A}$,
(iii) $A_1, A_2, \ldots \in \mathcal{A}$ then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$.

Now we may extend out definition of an additive set function to the following:

Definition 7

A set function $\mu: \mathcal{A} \to R^*$ is σ *additive*, if

(i) $\mu(\varnothing) = 0$,
(ii) for any sequence $E_1, E_2, \ldots$ of sets of $\mathcal{A}$, where

$$E = \bigcup_{i=1}^{\infty} E_i \in \mathcal{A},$$

then

$$\mu(E) = \sum_{i=1}^{\infty} \mu(E_i).$$

Obviously any σ additive set function is also additive but the converse does not hold. Consider Example 4 given earlier. Let in that example

$$E = (0,1] \quad \text{and} \quad E_n = \left(\frac{1}{n+1}, \frac{1}{n}\right], \qquad n = 1,2,\ldots$$

Now the sequence (E_n) has each of its elements in A, and E itself is in $\mathcal{A}$, but clearly

$$\mu(E) = +\infty \quad \text{and} \quad \sum_{n=1}^{\infty} \mu(E_n) = \sum_{n=1}^{\infty} \left(\frac{1}{n} - \frac{1}{n+1}\right) = 1.$$

Having discussed various types of set functions we can now restrict the class of such functions to those which have particular interest for us; that is, those to which we will refer as measures.

Consider a set X and a σ algebra $\mathcal{A}$ of its subsets. We will refer to the couple $(X, \mathcal{A})$ as a *measurable space*.

Definition 8[20]

If $(X, \mathcal{A})$ is a measurable space then any function $: \mathcal{A} \to R_+$ (where $R_+ = \{x | x \in R^* \text{ and } x \geqslant 0\}$ which is σ additive is called a *measure*.

Definition 9

For $(X, \mathcal{A})$ a measurable space, if μ is a measure on $\mathcal{A}$ and $\mu(X) = 1$, then μ is called a *probability space*.

It will, in general, be enough for us to restrict our attention to probability measures but it is useful to have the more general definition.

The reader should now return to the examples given and check which are measures.

Before proceeding we need a further definition.

Definition 10

A set function $\mu: \mathcal{A} \to R^*$ is called σ *finite* if for each $E \in \mathcal{A}$ there exists a sequence of sets E_i $(i = 1, 2, \ldots)$ with $E_i \in \mathcal{A}$ such that $E \subset \bigcup_{i=1}^{\infty} E_i$ and $\mu(E_i) < \infty$ for all i.

We will now show that starting with a measure on an algebra $\mathcal{A}$ we can extend it uniquely to a measure on the σ algebra generated by $\mathcal{A}$.

If we start with a measure μ defined on an algebra $\mathcal{A}$ of subsets of a set X consider the function defined by

$$\mu^*(E) = \inf \sum_{i=1}^{\infty} \mu(F_i),$$

where the infimum is taken over all sequences of sets (F_i) such that $E \subset \bigcup_{i=1}^{\infty} F_i$. We can now state the following:

Theorem 2

Let $\mathcal{A}$ be an algebra of subsets of a set X and $\mu: \mathcal{A} \to R^+$ a measure on $\mathcal{A}$. Then

[20] It is worth noting that, although we start here with the natural domain of definition, a σ algebra, there is no need to do so and we could have started with some other class of subsets. In addition, we have required that a measure be positive, a restriction not generally imposed.

there is an extension of μ to a measure r where $r: S(\mathcal{C}) \to R^+$ and $S(\mathcal{C})$ is the σ algebra generated by $\mathcal{C}$. The extension is unique and σ finite on $S(\mathcal{C})$ if μ is σ finite. r is the restriction of μ^* to $S(\mathcal{C})$ where μ^* is defined as above by

$$\mu^*(E) = \inf \sum_{i=1}^{\infty} \mu(F_i).$$

We have then arrived at the point where given some arbitrary set X and a measure defined on a simple structure (an algebra) of its subsets we can extend this measure uniquely to the σ algebra generated by that algebra.

We can now give the following:

Definition 11

A *measure space* $(X, \mathcal{C}, \mu)$ is a triple where X is a set, $\mathcal{C}$ is a σ algebra of subsets of X, and μ a measure defined on $\mathcal{C}$.

An example: Lebesgue measure

In Euclidean space the notion of measure corresponds intuitively to the ideas of length, area, or volume depending upon the dimension in question. How is the measure of a set defined in this case? In R we consider the class $\mathcal{P}$ of half open intervals $(a, b]$; these generate the class $\mathcal{R}$ of all elementary figures, i.e., sets of the form

$$E = \bigcup_{i=1}^{n} (a_i, b_i] \quad \text{with} \quad b_i < a_{i+1},\ i = 1, 2, \ldots, n-1.$$

In other words, the elementary figures consist of all sets which are expressible as a finite union of disjoint sets of $\mathcal{P}$.

In R^ℓ the half open intervals are given by

$$\{(x_1, x_2, \ldots, x_\ell)\}, \quad a_i < x_i \leqslant b_i, \qquad i = 1, 2, \ldots, \ell;$$

these generate analogously the elementary figures $\mathcal{R}^\ell$.

Define now the natural set function to express length, i.e.,

$$\mu(a, b] = b - a,$$

or area or volume,

$$\mu\{(x_1, \ldots, x_\ell) : a_i < x \leqslant b_i, \quad i = 1, 2, \ldots, \ell\} = \prod_{i=1}^{\ell} (b_i - a_i).$$

Such a μ is a measure, and it can be uniquely extended by our previous results to the σ algebra generated by the elementary figures, i.e., the Borel sets $\mathfrak{B}^\ell$, and is referred to as *Lebesgue measure*.

In fact, the class of Lebesgue measurable sets $\mathfrak{L}^\ell$ is larger than $\mathfrak{B}^\ell$, but this is not of great importance for the present discussion.

The interpretation of measures in probabilistic terms is clear and the measure of a subset is the probability that the outcome of some "experiment" will fall in that subset. In economics, we often need to formalise the idea that people forecast future prices. Such a forecast by an individual of n prices would be given by a measure on the unit simplex R^n. A discussion of such forecasts and requirements imposed on them is to be found in Chapter 19 on temporary general equilibrium theory.

A wholly different approach to the use of measure theory in economics, as has been mentioned, is the idea of representing a purely competitive economy by a continuum of agents or more generally by a measure space. Having discussed the nature of measure space and the nature of measures at length, we can now look at the economic interpretation given to them.

A *measure space of economic agents* $(A, \mathcal{A}, \mu)$ can be viewed as follows: A is the set of individual names or labels. For example, in a finite economy A could be a set of integers, while in a continuum economy we could, as Aumann (1964) did, use the closed unit interval $[0, 1]$ as the underlying set of labels.

$\mathcal{A}$ is the σ algebra of subsets of A which can be thought of as corresponding to the possible coalitions of A. As we have observed, we cannot in general define the measure on all the subsets of A, but the reader can consider for practical purposes all subsets as possible coalitions. In other words those coalitions which are eliminated by confining our attention to a σ algebra, rather than all subsets, are of no special interest. In probability theory, for example, we may well be interested in the probability that a number drawn from some set falls into a certain interval or collection of intervals, but we are probably less interested in the probability of its being rational or irrational. Even this is manageable, but there are sets to which we cannot assign probabilities. However, such sets are not constructed as collections of intervals, and it is in these that our interest generally lies. μ the measure on A simply conveys the idea of the proportion of agents belonging to any subset. Thus we will have for a finite economy a measure space given by the following:

Definition 12

A measure space $(A, \mathcal{A}, \mu)$ is *simple* if A is finite, $\mathcal{A}$ is the set of all subsets of A and $\mu(E) = \#E/\#A$.

We should also note in our discussion of the continuum economy as a representation of a perfectly competitive situation we suggested that each individual had no weight. If we use Lebesgue measure on the unit interval this is clearly the case but in general we need to make an assumption that the measure space satisfies the following:

Definition 13

A measure space $(A, \mathcal{A}, \mu)$ is *atomless* if for every $E \in \mathcal{A}$ with $\mu(E) > 0$ there exists $F \in \mathcal{A}$ and $F \subset E$ such that $\mu(E) > \mu(F) > 0$.

This rules out some individual having positive "weight" or "influence". Indeed the idea of a measure space with atoms had already been used to designate situations which are not perfectly competitive that is to convey the idea of monopoly. See e.g. Shitovitz (1974).

An alternative use of the notion of an an atom would be as mentioned earlier when we wish to define the notion of what Arrow described as a "dictator" in social choice with an arbitrarily large number of members.

That is if a^* an agent is "decisive" for A in that his preferences determine those of the society as a whole then we define the Dirac measure as follows:

$$\begin{aligned} \mu(E) &= 1 \quad \text{if} \quad a^* \in E, \\ &= 0 \quad \text{otherwise}, \end{aligned} \qquad \forall E \in \mathcal{P}(A).$$

Clearly μ defines a measure and a^* may be thought of as a dictator in the Arrovian sense, if we make the rule that if for some coalition E, $x \succ_a y$, $\forall a \in E$, then x is socially preferred to y if $\mu(E) = 1$.

Incidentally, one can see from the above example that in general it is the requirement that the measure be atomless which prevents us from defining it on all subsets of A. In the example it is clear that μ is a measure defined on all subsets of A.

The notion of a measure on a set allows us, as we mentioned in the introduction, to make statements about which subsets are of no importance, that is, which are "negligible".

If $(A, \mathcal{A}, \mu)$ is a measure space then a set $B \subset A$ is said to be *negligible* (for μ) if there exists a set $E \in \mathcal{A}$ such that $\mu(E) = 0$, and $B \subset E$.

If a certain property holds for all points of A except for a set B where $\mu(B) = 0$ then we say that that property holds "*almost everywhere*". In economics such a description is useful as a way of characterising particular phenomena as exceptional or rare.

If μ is a probability measure then the term "*almost surely*" replaces almost everywhere.

Situations in economics which occur only for configurations of parameters which together have measure zero in the space of all such configurations may be thought of as "unlikely" or "rare". This is a useful idea which enables us to avoid making strong or unnatural assumptions to rule out cases which are "exceptional" and which enables us to give a precise interpretation of the word "exceptional".

Liapunov's theorem

We now give a result of considerable importance in applying measure theory to economics and one which has played a central role in the formalisation of, and the equivalence between, solution concepts for large economies.

Theorem 3 (Liapunov)[21]

Let $\mu_1, \ldots, \mu_m$ be atomless measures on $(A, \mathcal{A})$, then the set

$$\{(\mu_1(E), \mu_2(E), \ldots, \mu_m(E)) \in R^m, \quad E \in \mathcal{A}\}$$

is a closed and convex subset of R^m.

This theorem is particularly useful since when considering a continuum economy we can always find a "scaled down" version of that economy and the reader will find a discussion in Chapter 18 by Hildenbrand. We also note in passing that this theorem is fundamental in the article by Dubins and Spanier (1961) to which we referred earlier.

Measurable mappings

In economics we will frequently be concerned with mappings from one measurable space to another. Indeed, when defining an exchange economy, for example, we will be concerned with identifying with each agent his endowments and preferences. We will need a certain regularity property of such a mapping, in particular that the pre-image of every set in the σ algebra of the range shall be a set in the σ algebra of the domain. This is inconvenient but necessary for technical reasons.

Definition 14

For two measurable spaces $(A_1, \mathcal{A}_1)$ and $(A_2, \mathcal{A}_2)$ a mapping $f: A_1 \to A_2$ is *measurable* if $f^{-1}(E) = \{a \in A_1 \mid f(a) \in E\} \in \mathcal{A}_1$ for each $E \in \mathcal{A}_2$.

[21] A proof is given in Lindenstrauss (1969).

Note that the measurability of a function depends upon the σ algebras, and thus for the same underlying sets A_1 and A_2 changing the σ algebra associated with each can change whether a function is measurable or not. When A_1 and A_2 are metric spaces we will generally take $\mathcal{A}_1$ and $\mathcal{A}_2$ to be the respective Borel σ algebras.

It would seem at first sight that it might be difficult to determine whether a given function is, in fact, measurable, but in fact it is sufficient to check for any class of subsets of A_2 which generates $\mathcal{A}_2$. More formally, we have:

Remark

If for a class $\mathcal{C}$ of subsets of A_2 which generates $\mathcal{A}_2$, and a mapping f from a measurable space $(A_1, \mathcal{A}_1)$ into a measurable space $(A_2, \mathcal{A}_2)$, $f^{-1}(C) \in \mathcal{A}_1$, for every $C \in \mathcal{C}$; then f is measurable.

It is also important to note that composing two measurable mappings preserves measurability. Thus we have:

Proposition 1

Let f and g be two measurable mappings from $(A_1, \mathcal{A}_1)$ to $(A_2, \mathcal{A}_2)$ and from $(A_2, \mathcal{A}_2)$ to $(A_3, \mathcal{A}_3)$, respectively, then the composition $g \circ f$ is a measurable mapping.

In addition, the following result is frequently useful:

Proposition 2

Let g be a measurable mapping from a measurable space $(A_1, \mathcal{A}_1)$ into a measurable space $(A_2, \mathcal{A}_2)$ and f a function from A into R^m, then f is measurable with respect to the σ algebra $g^{-1}(\mathcal{A}_2)$ if and only if there exists a measurable function h of $(A_2, \mathcal{A}_2)$ into R^m such that $f = h \circ g$.

Real-valued measurable functions

In particular if we consider a mapping f from a measurable space $(A, \mathcal{A})$ to the extended real line R^*, then any of the following conditions are necessary and sufficient for f to be measurable:

(i) $\{x \mid f(x) \leqslant c) \in \mathcal{A}$ for all $c \in R$,
(ii) $\{x \mid f(x) > c\} \in \mathcal{A}$ for all $c \in R$,
(iii) $\{x \mid f(x) < c\} \in \mathcal{A}$ for all $c \in R$,
(iv) $\{x \mid f(x) \geqslant c\} \in \mathcal{A}$ for all $c \in R$.

Other useful properties of real-valued or extended real-valued measurable functions are given by the following:

Proposition 3

If $(A, \mathcal{A})$ is a measurable space and f and g two measurable functions into R. (resp. into R^*), then the functions

(i) $f+g$. (resp. $f+g$ if the function if defined),
(ii) $\sup(f, g)$,
(iii) $\inf(f, g)$,
(iv) $f.g$,
(v) αf, $\forall \alpha \in R$,

are measurable.

Examples and further properties

Consider now a generalisation of the special mapping mentioned earlier often referred to as the "indicator variable" that is $\mathcal{X}_c : A \in R$ such that

$$\begin{aligned} \mathcal{X}_c &= 1 \quad \text{if} \quad a \in C, \\ &= 0 \quad \text{if} \quad a \notin C, \end{aligned} \qquad \text{for every} \quad C \in \mathcal{A},$$

then the mapping is measurable.

If we wish to confine our attention to a restricted class of a σ algebra then it is useful to know that, if $(A, \mathcal{A})$ is a measurable space and $\mathcal{A}'$ a sub σ algebra of $\mathcal{A}$ then the *identity map*,

$$\text{id.}(A, \mathcal{A}) \to (A, \mathcal{A}') \quad \text{where} \quad \text{id.}(a) = a,$$

is measurable.

When we consider functions from a metric space M into R^* it is important to observe that:

Proposition 4

Every lower or upper semi-continuous function from a metric space M into R^* (and thus in particular every continuous function) is measurable.

Finally we give a result which will be used in the next section:

Proposition 5

Let the sequence $(f_n)(A, \mathcal{A})$ into R be such that:

(i) f_i is measurable $(i = 1, 2, \ldots)$.

Then (a) the functions $\sup_n f_n$ and $\inf_n f_n$ are measurable, and (b) the functions $\limsup_n f_n$ and $\liminf_n f_n$ are measurable.

Furthermore if the following condition is also satisfied:

(ii) $\lim f_n(a)$ exists for every $a \in A$.

Then the function g defined by $g(a) = \lim f_n(a)$ is measurable.

Note that we cannot extend these results to include non-countable operations. To see this consider the following:

Example 5

Let A be a subset of $[0, 1]$ which is not Lebesgue measurable. Let

$$f_\alpha(x) = 1 \quad \text{if} \quad x = \alpha,$$

$$f_n(x) = 0 \quad \text{if} \quad x \neq \alpha.$$

For each $\alpha \in A$ the function f is clearly measurable, but

$$\chi_A(x) = \sup_{\alpha \in A} f_\alpha$$

is obviously not Lebesgue measurable.

This creates particular problems, for example, when considering stochastic processes with a continuous time parameter.

Integration

The idea of the integral of a function plays a very important role whether we are considering the probabilistic aspect of measure theory of whether we are considering the application of measure theory directly to "idealised", "perfectly competitive" or "limit" economies. In the former case the reader will be aware that the integral gives the "mean" or "expectation" of a given function f with respect to a particular probability distribution. In this case the function is a "random variable with distribution μ" and the integral gives the familiar idea of the expected value of the random variable.

Recall that in economies with a measure space of agents we are faced with a simple definitional problem. How, with an infinite number of agents each possessing a positive bundle of goods, can we talk of an equilibrium in which the demand for these goods equals the supply of them? Since the sum is of no interest, the appropriate notion is that average, or "per capita", demand equals supply. Here again the integral will be the appropriate concept. The integral $I(f)$ will be a real number associated with a particular function f and we will

require that for suitable functions f the operator $I(f)$ should satisfy certain properties. Let $\mathfrak{F}$ be a class of functions $f: A \to R^*$ and let $I: \mathfrak{F} \to R$ define a real number for each $f \in \mathfrak{F}$, then the following properties would seem intuitively, to be required of I, particularly if one thinks of the interpretation of "the area under a curve" as the integral of a function from R into R.

(i) If for all $f \in \mathfrak{F}$ we have $f(a) \geqslant 0$ for all $a \in A$, then we have $I(f) \geqslant 0$; that is, I preserves non-negativity.
(ii) For f and $g \in \mathfrak{F}$ and α and $\beta \in R$, it holds that $\alpha f + \beta g \in \mathfrak{F}$ and $I(\alpha f + \beta g) = \alpha I(f) + \beta I(g)$; in other words I is linear on $\mathfrak{F}$.
(iii) I is continuous on $\mathfrak{F}$, in that, if (f_n) is an increasing sequence of functions in $\mathfrak{F}$ and

$$f_n(a) \to f(a) \quad \text{for all} \quad a \in A,$$

then $f \in \mathfrak{F}$ and $\lim_{n \to \infty} I(f_n) = I(f)$.[22]

Our procedure for obtaining an integral which satisfies these three conditions is, first, to restrict our attention to a particular class of functions for which the integral has an obvious intuitive definition, and then to extend this class of functions to as large a class as possible.

To do this we need first the idea of a "simple function" from a set A to R which is one which takes on a finite number of values, one for each set of a *partition* of A, and is constant on each set of the *partition*. The idea is illustrated in Figure 2.1 for a function from $[0, 1]$ into R.

Definition 15

A finite collection of sets $E_1, \ldots, E_n$ such that

$$E_i \cap E_j = \varnothing, \qquad i = 1, \ldots, n, \quad j = 1, \ldots, n,$$

and

$$\bigcup_{i=1}^{n} E_i = A,$$

is said to form a *finite partition* of A. In particular, if $E_i \in \mathfrak{A}$ $(i = 1, \ldots, n)$ then

[22] The Riemann integral with which the reader will be familiar from the integral calculus does not satisfy this property but does satisfy the following weakened version of it:

(iii*) Let (f_n) be a monotone decreasing sequence of functions with $\lim_{n \to \infty} f_n(a) = 0$ for all $a \in A$, then $\lim_{n \to \infty} I(f_n) = 0$.

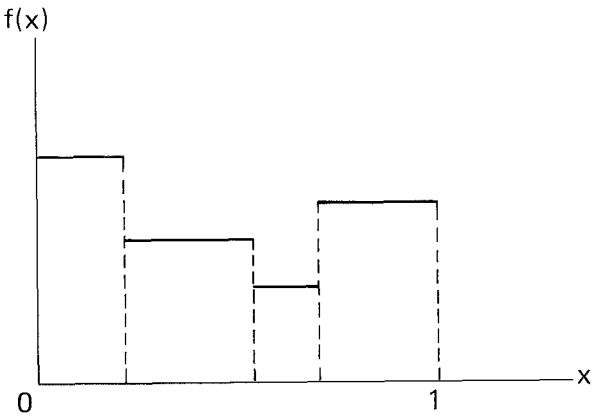


Figure 2.1

these sets form an $\mathcal{C}$ *partition* of A. With this we can proceed to the following:

Definition 16

A function $f: A \to R$ is called $\mathcal{C}$ *simple* if it can be expressed as $f(x) = \sum_{i=1}^{n} c_i \chi_{E_i}(x)$, where $E_1, E_2, \ldots, E_n$ form an $\mathcal{C}$ partition of A and $c_i \in R$ $(i = 1, 2, \ldots, n)$.

Remark

If f and g are two simple functions from A to R then the functions

$$f+g, \quad f-g, \quad fg,$$

are also simple functions.

Note also that an $\mathcal{C}$ simple function is $\mathcal{C}$ measurable.[23] Now we can start to extend our attention to measurable functions by considering the following:

Theorem 4

If a function $f: A \to R_+$ is measurable then it is the limit of a monotone increasing sequence of non-negative simple functions.

Now to move towards the desired results, we must show that any measurable function is the limit of a sequence of simple functions.

[23] We will frequently speak of measurable functions rather than $\mathcal{C}$ measurable functions when only one σ algebra is under consideration.

First for a function $f: A \to R^*$ define

$$f_+(x) = \max[0, f(x)], \qquad f_-(x) = -\min[0, f(x)],$$

Clearly,

$$f(x) = f_+(x) - f_-(x).$$

Now from a previous remark, if f is measurable so are f_+ and f_-, and since both are non-negative each is the limit of a sequence of non-negative simple functions. Applying the remark again we then have the following important theorem which provides the basis for the definition of the integral:

Theorem 5

Any measurable function $f: A \to R^*$ is the limit of a sequence of simple functions.

This link between simple and measurable functions enables us to proceed to the definition of the integral for simple functions and to extend it to measurable functions.

Thinking of measure on a set A as the distribution of mass in physical terms or as a probability distribution over a set of outcomes, it is clear that the natural notion of the integral for the particularly convenient case of a non-negative simple function is given by:

Definition 17

Given a measure space $(A, \mathcal{A}, \mu)$ and a non-negative simple function,

$$f(x) = \sum_{i=1}^{n} c_i \chi_{E_i}(x) \quad \text{with} \quad c_i \geqslant 0, \qquad i = 1, \ldots, n,$$

(with respect to μ), the *integral* $\int f \mathrm{d}\mu$ is defined by

$$\int f \mathrm{d}\mu = \sum_{i=1}^{n} c_i \mu(E_i).$$

Referring back to Figure 2.1, it is clear that the integral of such a function consists of the sum of the area of the rectangles under each step of the function. This sum is always defined since the individual terms are non-negative,[24] and it is independent of which of the possible representations of f is chosen.

[24] If we are treating general measures it is possible that $\mu(E_i) = \infty$ and $c_i = 0$; in this case we take $\mu(E_i)c_i = 0$.

Remark

It is easily shown that the integral is linear on the class of non-negative simple functions S_+, that is, if $f, g \in S_+$ and $\alpha, \beta \geqslant 0$, then

$$\int (\alpha f + \beta g)\, d\mu = \alpha \int f\, d\mu + \beta \int g\, d\mu.$$

Furthermore the integral is order preserving on the same class, i.e., if $f, g \in S_+$ and $f \geqslant g$, then

$$\int f\, d\mu \geqslant \int g\, d\mu.$$

Now we can proceed to the second step—that of extending the definition of the integral to the class of non-negative measurable functions M_+.

For f in M_+ there exists by Theorem 6 a monotone increasing sequence (f_n) of simple functions with $f_n \to f$. Now for each f_n in the sequence $\int f_n\, d\mu$ is defined, and by our previous observations the sequence $(\int f_n\, d\mu)$ is monotone increasing and has a limit.[25] Hence we define for $f \in M_+$,

$$\int f\, d\mu = \lim_{n \to \infty} \int f_n\, d\mu.$$

Clearly the monotone sequence (f_n) which converges to a given f is not unique, but the integral, as defined, is independent of the choice of sequence. Note that it follows directly from our earlier observation for functions in the class S_+ that the integral operator is linear on the class M_+, i.e., for $f, g \in M_+$ and $\alpha, \beta \geqslant 0$,

$$\int (\alpha f + \beta g)\, d\mu = \alpha \int f\, d\mu + \beta \int g\, d\mu.$$

Definition 18

A non-negative measurable function f is said to be *integrable* if

$$\int f\, d\mu \quad \text{is finite.}$$

Thus far we have been concerned with measurable functions in M_+. We now extend our definition of the integral to the class of integrable measurable functions.

[25] The limit may, of course, be $+\infty$.

First, observe again that if $f: A \to R^*$ is measurable, then so are f_+ and f_-. In particular if the two non-negative measurable functions f_+ and f_- are integrable that we say that f is integrable. More precisely, we have the following:

Definition 19

If $f: A \to R^*$ is such that f_+ and f_- are integrable, then f *is called integrable* and the integral of f with respect to μ is given by

$$\int f_+ \, d\mu - \int f_- \, d\mu.$$

Often we will be concerned with the integral of a function f over only a subset $E \in \mathcal{A}$, and in this case we define

$$\int_E f d\mu = \int f \cdot \chi_E d\mu,$$

provided that $f \cdot \chi_E$ is defined. Then there are two conditions each of which will ensure the integrability of a function over a given set. Either:

(i) $f \cdot \chi_E$ is non-negative and measurable, or
(ii) $f \cdot \chi_E$ is measurable and integrable.

f is then integrable over A if $f \cdot \chi_A$ is integrable. We denote the set of all integrable functions from $(A, \mathcal{A}, \mu)$ into R^* by $\mathcal{L}(A, \mathcal{A}, \mu)$.

Confirmation of the properties we demanded of the integral at the outset is given by the following:

Theorem 6

If $(A, \mathcal{A}, \mu)$ is a measure space, E, F are two disjoint sets in $\mathcal{A}$, and f, g are two functions belonging to $\mathcal{L}(A, \mathcal{A}, \mu)$, then

(i) f, g are integrable over E and F;
(ii) $f+g, |f|, |g|$ belong to $\mathcal{L}(A, \mathcal{A}, \mu)$;
(iii) $\int_{E \cup F} f d\mu = \int_E f d\mu + \int_F f d\mu$;
(iv) f, g are finite μ a.e.;
(v) $\int (f+g) d\mu = \int f d\mu + \int g d\mu$;
(vi) $\left| \int f d\mu \right| \leqslant \int |f| d\mu$;
(vii) for $c \in R$, $c \cdot f$ is μ integrable and $c \int f d\mu = \int cf d\mu$;
(viii) $f \geqslant 0 \Rightarrow \int f d\mu \geqslant 0 : f \geqslant g \Rightarrow \int f d\mu \geqslant \int g d\mu$;
(ix) if $f \geqslant 0$ and $\int f d\mu = 0$, then $f = 0$ μ a.e.;
(x) $f = g \mu$ a.e. $\Rightarrow \int f d\mu = \int g d\mu$;
(xi) If $h: A \to R^*$ is A measurable and $|h| \leqslant f$ then $h \in \mathcal{L}(A, \mathcal{A}, \mu)$.

From these results follows:

Corollary to Theorem 6

If a function f: $A \to R^*$ is bounded, $\mathcal{A}$ measurable and if $f(x)=0$ when $x \notin E$ for some $E \in \mathcal{A}$ with $\mu(E) < \infty$ then f is μ integrable.

As we will see in what follows, an exchange economy will be defined by a measurable mapping from the underlying measure space of agents to the space of agents' characteristics. In other words, defining an economy consists of specifying for each agent his preferences and his initial endowments. Now for many of the results in Chapter 18 of this Handbook, it will be important to show that properties of very large economies — that is economies with a measure space of agents — are, in some sense, also true for large finite economies. To do this we will need to consider sequences of economies and sequences of allocations, i.e., sequences of mappings from the space of agents to R^{ℓ}_{+}. The following three results will prove to be particularly useful, and we will later investigate in more detail different notions of convergence of measurable functions.

Proposition 6

If the sequence (f_n) in $\mathcal{L}(A, \mathcal{A}, \mu)$ is increasing (decreasing), $\lim f_n(a)$ is finite for every $a \in A$, and if $\lim_n f_n$ is finite, then

$$\lim_n f_n \in \mathcal{L}(A, \mathcal{A}, \mu) \quad \text{and} \quad \lim_n \int f_n = \int \lim_n f_n.$$

Lemma 1 (Fatou)

If (f_n) is a sequence in $\mathcal{L}(A, \mathcal{A}, \mu)$ and if $f_n \leqslant g$ where $g \in \mathcal{L}(A, \mathcal{A}, \mu)$ then

$$\int \limsup_n f_n \geqslant \limsup_n \int f_n.$$

Furthermore, if $h \leqslant f_n$ where $h \in \mathcal{L}(A, \mathcal{A}, \mu)$, then

$$\int \liminf_n f_n \leqslant \liminf_n \int f_n.$$

Theorem 7 (Lebesgue)

If (f_n) is a sequence in $\mathcal{L}(A, \mathcal{A}, \mu)$, and if $\lim_n f_n(a)$ exists for every $a \in A$ and $|f_n| \leqslant g$ $(n=1,2,\ldots)$, where $g \in \mathcal{L}(A, \mathcal{A}, \mu)$, then

$$\lim_n f_n \in \mathcal{L}(A, \mathcal{A}, \mu) \quad \text{and} \quad \int \lim_n f_n = \lim_n \int f_n.$$

2.4. *Product spaces and product measures*

Before proceeding to our discussion of convergence of measurable functions we will need to discuss the idea of product spaces and product measures. To see why we need these notions consider again the example of the exchange economy mentioned earlier. It will be defined by a mapping from the measure space of agents to the space of agents characteristics. The latter, however, is the product of two spaces, the space of preferences $\mathcal{P}$ and the space of initial endowments R^{ℓ}_{+}. Now the natural procedure is to use the structure of each space to define the product space and product measure since, in particular, this allows us to use the natural idea of "projection", for example, when we wish to concentrate on the distribution of initial endowments. Recall that the Cartesian product $A \times B$ of two spaces A and B is the set of all ordered pairs (a, b) where $a \in A$ and $b \in B$.

Definition 20

A set in $A \times B$ of the form $E \times F$ with $E \subset A$ and $F \subset B$ is called a *rectangle*.

Definition 21

Let $\mathcal{A}$ and $\mathcal{B}$ be classes of subsets in A and B, respectively, then $\mathcal{A} \times \mathcal{B}$ denotes the class of all rectangles $E \times F$ with $E \in \mathcal{A}$ and $F \in \mathcal{B}$, i.e., *the product of the classes* $\mathcal{A}$ *and* $\mathcal{B}$.

Definition 22

Let $\mathcal{A}$ and $\mathcal{B}$ be algebras (resp. σ algebras) in A and B, respectively, then the *product algebra* (resp. σ algebra) is the algebra (resp. σ algebra) generated by $\mathcal{A} \times \mathcal{B}$, and is denoted $\mathcal{A} \otimes \mathcal{B}$.

It is important to note that if $\mathcal{A}$ and $\mathcal{B}$ are σ algebras, $\mathcal{A} \times \mathcal{B}$ will not be a σ algebra.

As we mentioned above we will sometimes be interested in restricting our attention to one of the components of the product space and for this we need two definitions:

Definition 23

For any set $E \subset A \times B$ and any point $a \in A$, the set

$$E_a = \{b | (a, b) \in E\}$$

is called the *section* of E at a. Similarly for any $b \in B$ the subset $E^b = \{a | (a, b) \in E\}$ is called the *section* of E at b.

Definition 24

For any set $E \subset A \times B$ the sets $\{x | \text{there exists } y \text{ with } (x, y) \in E\}$ and $\{y | \text{there exists } x \text{ with } (x, y) \in E\}$ are called the *projections* of E into the respective spaces A and B, and are denoted $\text{proj}_A E$ and $\text{proj}_B E$.

It is again important to note that $E \in \mathcal{A} \otimes \mathcal{B}$ does not mean that $\text{proj}_A E \in \mathcal{A}$.

Now with these definitions we may proceed to consider product measures. Consider two measure spaces $(A_1, \mathcal{A}_1, \mu_1)$ and $(A_2, \mathcal{A}_2, \mu_2)$ where μ_1 and μ_2 are both σ finite. Define for any rectangle set $E_1 \times E_2$,

$$\mu(E_1 \times E_2) = \mu_1(E_1)\mu_2(E_2).$$

It is easy to show that μ is finitely additive on $\mathcal{A}_1 \times \mathcal{A}_2$; indeed μ is a measure on the semi algebra $\mathcal{A}_1 \times \mathcal{A}_2$ which can be extended uniquely to the generated algebra, and thence of course, by previous results, to the generated σ algebra which is $\mathcal{A}_1 \otimes \mathcal{A}_2$. The resulting λ^* is called the *product measure* on $\mathcal{A}_1 \otimes \mathcal{A}_2$. This may be summarised in the following:

Theorem 8

Given two measure spaces $(A_1, \mathcal{A}_1, \mu_1)$, $(A_2, \mathcal{A}_2, \mu_2)$ such that μ_1 and μ_2 are σ finite, there is an unique measure λ defined on the product σ algebra $\mathcal{A}_1 \otimes \mathcal{A}_2$ on $A_1 \times A_2$ such that

$$\lambda(E_1 \times E_2) = \mu_1(E_1)\mu_2(E_2) \quad \text{for} \quad E_1 \in \mathcal{A}_1 \quad \text{and} \quad E_2 \in \mathcal{A}_2.$$

Definition 25

If μ is a measure on the product space $A \times B$, then the *marginal distribution* of μ on A is given by

$$\mu(C) = \mu(C \times B) \quad \text{for } C \text{ a subset of } A.$$

Although this result extends immediately to any finite Cartesian product of σ finite measure spaces, more care has to be taken in defining product measures for countable products of measure since one has to make sure that the infinite products of real numbers involved converge. The problem may be avoided by sticking to probability measure spaces where the natural generalisation holds.

If we have two measure spaces $(A_1, \mathcal{A}_1, \mu_1)$ and $(A_2, \mathcal{A}_2, \mu_2)$, we can in fact show quite simply the link between the integral of a function f from $A_1 \times A_2 \to R^*$ with respect to the product measure and a two-step procedure including integrating with respect to μ_1 and μ_2. The idea is clear: we fix $x \in A_1$ then integrate f with respect to μ_2. The resulting function from A_1 to R^* is then integrated with respect to μ_1, and the result turns out, with certain restrictions, to be equivalent to having integrated with respect to the product measure λ directly.

We first state a result which shows how to obtain the product measure λ by integrating the measure of the section for each fixed x over all x with respect to μ_1.

Theorem 9

For two σ finite measure spaces $(A_1, \mathcal{A}_1, \mu_1)$, $(A_2, \mathcal{A}_2, \mu_2)$ define the product measure λ on the σ algebra $\mathcal{A}_1 \otimes \mathcal{A}_2$. Then for each $E \in \mathcal{A}_1 \otimes \mathcal{A}_2$, $\mu_2(E_x)$ is $\mathcal{A}_1$ measurable and $\mu_1(E_y)$ is $\mathcal{A}_2$ measurable, and

$$\lambda(E) = \int \mu_1(E_y)\, d\mu_2 = \int \mu_2(E_x)\, d\mu_1.$$

Incidentally, it now follows from our previous discussion that we have:

Corollary to Theorem 9

For $E \in \mathcal{A}_1 \otimes \mathcal{A}_2$, $\lambda(E) = 0$ if and only if $\mu_2(E_x) = 0$ for almost all x, and if and only if $\mu_1(E_y) = 0$ for almost all y.

We now state the following important:

Theorem 10

Under the conditions of Theorem 11 denote by $\mathcal{A}$ the σ algebra $\mathcal{A}_1 \otimes \mathcal{A}_2$. Then if h is any $\mathcal{A}$ measurable function from $A_1 \times A_2 \to R^+$, then

$$\int h\, d\lambda = \int \left(\int h_x\, d\mu_2 \right) d\mu_1 = \int \left(\int h_y\, d\mu_1 \right) d\mu_2.$$

We now move on to develop the idea of a derivative of a set function, but to do so we will need two definitions:

Definition 26

Given a measure space $(A, \mathcal{A}, \mu)$, the function $v: \mathcal{A} \to \mathfrak{R}^*$ is *absolutely continuous* with respect to μ, if for any $E \in \mathcal{A}$, $\mu(E) = 0$ implies $v(E) = 0$.

In particular, if $f: A \to R^*$ is μ integrable, then

$$v(E) = \int_E f\, d\mu \quad \text{for} \quad E \in \mathcal{A}$$

is a finite valued absolutely continuous set function.

In order to define our derivative, we will take a general σ additive set function and decompose it into an absolutely continuous part and a remainder which is

concentrated on a set which is μ null. To make this last remark more precise, we give the following:

Definition 27

For a measure space $(A, \mathcal{A}, \mu)$, a set function $v: \mathcal{A} \to R^*$ is *singular* with respect to μ, if there exists a set $E_0 \in \mathcal{A}$ with $\mu(E_0) = 0$ and

$$v(E) = v(E \cap E_0) \quad \text{for all } E.$$

We can now give the important:

Theorem 11

For a σ finite measure space $(A, \mathcal{A}, \mu)$ and a σ additive, σ finite set function $v: \mathcal{A} \to R^*$ there is a unique decomposition

$$v = v_1 + v_2,$$

where v_1 and v_2 are σ additive and σ finite, such that v_1 is singular with respect to μ and $v_2 < \mu$.

In addition, there is a finite valued measurable $f: A \to R$ such that

$$v_2(E) = \int_E f \, d\mu \quad \text{for all} \quad E \in \mathcal{A};$$

f is unique in that if there is a function g such that

$$v_2(E) = \int_E g \, d\mu \quad \text{for all} \quad E \in \mathcal{A},$$

then $f = g$ except on a set of zero measure.

This last observation is important, for it means that when we define the derivative of a set function this is not defined uniquely at any given point but as a function must coincide with any other function representing the same derivative except on a set of measure zero. With this in mind we give the following:

Definition 28

For a σ finite measure space $(A, \mathcal{A}, \mu)$, if $v(E) = \int_E f \, d\mu$ for all $E \in \mathcal{A}$, f is called the *Radon–Nikodyn derivative* of v with respect to μ and is denoted $dv/d\mu$.

2.5. *Convergence of measurable functions*

As we have said before it will be important for later economic applications such as those found in Chapter 18 of this Handbook, to study the convergence of

measurable functions. Several different types of convergence can be defined, and we will always be considering a sequence (f_n) of measurable functions from a measure space $(A, \mathcal{A}, \mu)$ to R^*.

Definition 29

If (f_n) is a sequence of measurable functions from $(A, \mathcal{A}, \mu)$ to R^*, (f_n) is said to *converge point-wise* to a measurable function f on E if for every $x \in E$ $\lim_{n\to\infty} f_n(x) = f(x)$. If $\mu(E) = \mu(A)$ then we say (f_n) *converges to f almost everywhere* (a.e.).

Furthermore if (f_n) and f are finite-valued, then we add the following:

Definition 30

If a sequence (f_n) and f are finite-valued functions from E to R then (f_n) is said to *converge uniformly* to f if for each $\varepsilon > 0$ there exists an integer N such that

$$x \in E \quad \text{and} \quad n \geqslant N \quad \text{implies} \quad |f_n(x) - f(x)| < \varepsilon.$$

The idea of convergence uniformly a.e. is self-evident.

A slightly weaker notion of convergence is given by the following:

Definition 31

Let $f_n: E \to R^*$ $(n = 1, 2, \ldots)$ and $f: E \to R^*$ be functions which are a.e. finite on E. Then f_n *converges almost uniformly* to f on E if for each $\varepsilon > 0$ there is a set $F_\varepsilon \subset E$, $F_\varepsilon \in \mathcal{A}$, $\mu(F_\varepsilon) < \varepsilon$, such that $f_n \to f$ uniformly on $(E - F_\varepsilon)$.

If μ is the Lebesgue measure on $E = [0, 1]$ it is clear that the sequence $f_n(x) = x^n$ converges almost uniformly but not uniformly a.e. From the definition it should be evident that convergence uniformly a.e. implies almost uniform convergence.

However, a less obvious implication is given by the following:

Theorem 12 (Egoroff)

Let $E \in \mathcal{A}$ with $\mu(E) < \infty$, and let (f_n) be a sequence of measurable functions from E to R^* which are finite a.e. and converge a.e. to a function $f: E \to R^*$ which is finite a.e.. Then $f_n \to f$ almost uniformly in E.

We consider next a rather different idea of proximity in which we look at the measure of the set on which two functions differ by some given number.

Definition 32

Let $f_n: A \to R^*$ and $f: A \to R^*$ be $\mathcal{A}$ measurable functions. Then f_n *converges in measure* (μ) to f if for each $\varepsilon > 0$,

$$\lim_{n\to\infty} \mu\{x : |f_n(x) - f(x)| \geqslant \varepsilon\} = 0.$$

It should be clear that the functions in question must be finite a.e. for the definition to be meaningful.

Our final notion of convergence makes use of the fact that the set $\mathfrak{L}_m$ of μ integrable functions is a linear space in which the idea of mean is defined. We have then:

Definition 33

Let (f_n) be a sequence of functions in $\mathfrak{L}_m$. Then (f_n) *converges to f in mean* if

$$\int |f_n - f| \, d\mu \underset{n\to\infty}{\to} 0.$$

The different notions of convergence are, of course, related and, as they are used very generally, in particular in studying large economies, we give the following basic results:

Theorem 13

If a sequence (f_n) of measurable functions converges almost everywhere to f, then (f_n) converges in measure to f.

For a more limited class of functions we have the following:

Theorem 14

Let (f_n) be a sequence of positive integrable functions. The sequence (f_n) converges in the mean to the integrable function f if and only if (f_n) converges to f in measure, and

$$\lim_n \int f_n = \int f.$$

One further result will complete the basic results we need on the convergence of measurable functions.

Theorem 15 (Scheffé)

If (f_n) is a sequence of positive integrable functions with

$$\int \liminf_n f_n = \lim_n \int f_n < \infty,$$

then (f_n) converges in the mean to $\lim_n \inf f_n$.

We will need these ideas of convergence for many purposes and, in particular, when discussing the notion of a sequence of economies which converges to a

limit economy. In order to give a simple and concise description of such a notion we take a sequence of finite economies for which we wish to show that certain properties true for "continuum economies" are approximately true for "large enough" economies. To avoid the problem that the space of agents and hence the associated mapping changes dimension as the number of agents increases we construct a series of parallel "equivalent" economies each with a continuum of agents and show that our results hold via the equivalent sequence. However, before discussing this problem in more detail we have to establish the meaning of this equivalence and for this we will need to discuss the idea of the convergence of measures and of a distribution.

2.6. *On metric spaces: Weak convergence*

We will focus our attention here on "weak convergence" which has been extensively used by Hildenbrand (1974) in particular.[26] This convergence may be characterised in several different ways two of which are fairly intuitive. If we consider any "well behaved" function f from a metric space T into the real line and a sequence (μ_n) of measures[27] on T then a requirement that (μ_n) converge to μ would be that the integral of f with respect to μ_n should converge to the integral of f with respect to μ. Alternatively, and perhaps more naturally, for any "convenient" subset B of T we should have $\lim_n \mu_n(B)=\mu(B)$. The basic idea is clearly that we require in a certain sense that the "weights" attached by μ_n to the various subsets should be very little different from those given by μ for n large enough. In particular, we might require for a sequence of economics that the "distribution of agents characteristics" should be "close" for n large to that of the limit economy. These ideas will be developed in detail in Chapter 18 of this Handbook. To make our previous remarks precise and to add two other equivalent definitions of weak convergence of measures we give the following:

Proposition 7

If T is a metric space and (μ_n) a sequence of probability measures on T then the following are equivalent:

(i) (μ_n) converges weakly to μ;
(ii) $\int f \mathrm{d}\mu_n \to \int f \mathrm{d}\mu$ for every bounded and uniformly continuous function f: $T \to R$;

[26] The reader is referred to Billingsley (1968) for a complete treatment of the problems mentioned here.

[27] We will be dealing exclusively with probability measures in this section; hence measure should be read as probability measure.

(iii) $\lim_n \mu_n(B) = \mu(B)$ for every subset $B \subset T$ for which the μ measure of the boundary of B is zero;
(iv) $\lim_n \sup \mu_n(C) \leqslant \mu(C)$ for every closed subset C in T;
(v) $\lim_n \inf \mu_n(D) \geqslant \mu(D)$ for every open subset D in T.

The following example cited by Hildenbrand (1974) may aid the reader's intuition.

Example 6

Let $T = R^m$. Define for a measure μ on R^m the distribution function,

$$F_\mu : R^m \to R,$$

by

$$F_\mu(x) = \mu\{z \in R^m \mid z \leqslant x\}.$$

The sequence (μ_n) of measures on R^m converges weakly to the measure μ on R^m if and only if the sequence (F_{μ_n}) of distribution functions converges to F_μ at every point x where F_μ is continuous.

Now suppose that we are concerned with a measure on a product space $A \times B$, for example, when we consider the space of agents' characteristics $\mathcal{P} \times R_+^\ell$. Then we will be concerned with marginal distributions.

The following result shows the relationship between the convergence of marginal distributions and the convergence of measures on a product space:

Theorem 16

If the sequence (μ_n) of probability measures on the separable measure space $A \times B$ converges weakly to the measure μ, then the sequences of marginal distributions (μ_n^A) and (μ_n^B) converge weakly to the marginal distributions μ^A and μ^B, respectively.

Let (μ_n) and (ν_n) be sequences of measures on the separable metric spaces A and B, respectively. Then the sequence $(\mu_n \times \nu_n)$ of product measures on $A \times B$ converges weakly to the product measure $\mu \times \nu$ on $A \times B$ if and only if (μ_n) converges weakly to μ and (ν_n) converges weakly to ν.

If in particular A is a separable metric space and we denote by $\mathfrak{M}(A)$ the set of all probability measures on A, then we have the following:

Proposition 8

There exists a metric ρ on $\mathfrak{M}(A)$ such that the space $(\mathfrak{M}(A), \rho)$ is separable and a sequence (μ_n) converges to μ in $(\mathfrak{M}(A), \rho)$ if and only if it converges weakly to μ.

Such a distance between measures enables us to endow $\mathfrak{M}(A)$ with a structure similar to that of A. An explicit example of such a metric is given by the Prohorov-metric defined as follows:[28]

$$\rho(\mu, \nu) = \inf \{\varepsilon > 0 | \nu(E) \leqslant \mu(B_\varepsilon(E)) + \varepsilon \text{ and } \mu(E) \leqslant \nu(B_\varepsilon(E)) + \varepsilon, \text{ for any } E \in \mathfrak{B}(A)\}.$$

One more notion that is important for many applications to economics is that of the support of a probability measure. Frequently we will be concerned with knowing that a measure concentrates all of its weight on a compact set, that is, that only isolated exceptions lie outside this set. For example, as we indicated in the introduction, we might require of somebody forecasting prices that with probability one he expects prices to fall within some compact set, or, more generally, we might require the following:

If μ is a probability measure on a separable metric space A then there is a closed subset B of A such that $\mu(B) = 1$ and if $F \subset A$ is closed and $\mu(F) = 1$ then $B \subset F$. Now consider:

Definition 34

The *support* of a probability measure μ denoted supp(μ), on a separable metric space A, is the smallest closed subset of A with measure one.

Then a very useful result which takes us in the right direction is the following:

Proposition 9

The set of probability measures with finite support is dense in $(\mathfrak{M}(A), \rho)$.

Now recalling an earlier discussion of price forecasting, we need to be sure that if the underlying space of outcomes is not compact that forecasts are "essentially" concentrated on some compact subset.

What is needed to formalise this requirement is the following:

Definition 35

A family of probability measures M on the metric space A is called *tight* if for every $\varepsilon > 0$ there exists a compact set $K \subset A$ such that $\mu(K) > 1 - \varepsilon$ for every $\mu \in M$.

In the light of this definition the reader should consider the example mentioned previously for the case of one good in which the family μ_n is such that the forecast of tomorrow's price attaches probability one to $p^{t+1} = n$.

[28] Here $B_\varepsilon(E)$ denotes, as usual, the ε-neighbourhood of E, i.e., $B_\varepsilon(E) = \{x \in A | \text{dist}(x, E) < \varepsilon\}$.

Alternatively consider the family of measures (μ_n) where μ_n is the uniform probability distribution on $[0, n]$. The problems such examples pose will be evident in the chapter on temporary general equilibrium theory (Chapter 19).

Two results of particular interest are given by:

Theorem 17

If the family of probability measures M on a metric space A is tight, then every sequence (μ_n) of probability measures contains a weakly converging subsequence.

Specialising to families with only one member we have:

Theorem 18

Every probability measure on a complete separable measure space is tight.

Further results may be found in Hildenbrand (1974); and for a more complete mathematical development, see Billingsley (1968).

2.7. *Distributions*

We return now to a concept which shows, in particular, as we observed earlier the true value of using measure spaces as a description of an economy. This is the idea of a distribution, and, as we suggested, it is frequently useful to work with the distribution of characteristics, for example, as the basic description of an economy.

We give now the formal version of the definition given in the introduction:

Definition 36

Let $(A, \mathcal{A}, \mu)$ be a probability space, M a metric space, and f a measurable mapping of A into m. The *distribution* v of f denoted by $\mu \circ f^{-1}$ is defined by

$$v(B) = \mu\{a \in A \mid f(a) \in B\} \quad \text{for every} \quad B \in \mathcal{B}(M).$$

As already observed, the choice of the measure space is arbitrary, and it is frequently the distribution that conveys the real information with which we are concerned. In particular, in studying "large economies" the frequent choice of the unit interval $[0, 1]$ as the space of agents is purely for convenience and has no particular significance in itself. Indeed, we know that every measure on a metric space M is the distribution of some measurable mapping on some measure space. More particularly, if M is complete and separable then for every probability measure on M there exists a measurable mapping f of the closed unit interval

$[0,1]$ into M such that $\mu=\lambda\circ f^{-1}$, where λ denotes the Lebesgue measure on $[0,1]$. Thus we could, given a suitable distribution on the space of characteristics of agents, always construct an associated economy with the unit interval as the space of agents.

We now come to a result which proved crucial in establishing general limit theorems concerning the equivalence of different solutions to the problem of allocating resources in a market. This result indicates clearly how we may overcome the problem that if in a sequence of economies the number of agents changes then so does the space of agents and the notion of convergence to a limit is unclear.

Theorem 19 (Skorokhod)

Let T be a separable metric space and (μ_n) a weakly converging sequence of measures on T with limit μ. Then there exists a measure space $(A,\mathcal{A},\nu)$ and measurable mappings f and f_n $(n=1,2,\ldots)$ of A into T such that $\mu=\nu\circ f^{-1}$, $\mu_n=\nu\circ f_n^{-1}$ and $\lim_n f_n=f$ a.e. in A.

Furthermore if T is complete then the measure space $(A,\mathcal{A},\nu)$ can be chosen to be the unit interval with Lebesgue measure.

In Chapter 18 by Hildenbrand the reader will encounter an extensive discussion of sequences of finite economies which converge to limit economies. What is important is that the preference endowment distributions of the finite economies are always defined on the same space. Thus, although each economy has a different space of agents we can by Skorokhod's theorem construct an analogous space of agents which is the same for each of the economies in the sequence. In other words, the distribution of agents characteristics will give us the information required and we can replace the original agent space by a more convenient artifact without changing any of the economic features of the model. This discussion anticipates our next section.

2.8. *Convergence in distribution*

Given a sequence of measurable mappings (f_n), each from a measure space $(A_n,\mathcal{A}_n,\mu_n)$ into a metric space T, we will want to define a sense in which these mappings converge. This leads us to:

Definition 37

A sequence (f_n) of measurable mappings with values in a metric space T *converges in distribution* to a measurable mapping f with values in T if the sequence (ν_n) of distributions of (f_n) converges weakly to the distribution ν of f.

Consider the special case in which $T=R$ and all the functions f_n and f are defined on the same measure space. In this case convergence in measure, and hence convergence almost everywhere, implies convergence in distribution. The converse is true only if f is a constant function.

We now give three results on convergence in distribution which will be of particular use when studying limit theorems for increasing sequences of economics.[29]

Proposition 10

Let (f_n) and f be a sequence of functions and a function all from a measure space $(A, \mathcal{A}, \mu)$ into a separable metric space (T, d) (where d is the metric). Then the function $w \mid \to d(f_n(w), f(w))$ from $A \to R$ is measurable and if the sequence $(d(f_n(\cdot), f(\cdot)))$, $n=1,2,\ldots$, converges in measure to zero then the sequence (f_n) converges in distribution to f.

Before proceeding to the next result we will need to extend the notion of integrability to a sequence of functions:

Definition 38

Let (f_n) be a sequence of measurable functions and $(A_n, \mathcal{A}_n, \mu_n)$ a sequence of measure spaces with $f_n: A_n \to R$. Then (f_n) is said to be *uniformly integrable* if (i)

$$\lim_{q\to\infty}\left(\sup_n \int_{|f_n|>q} |f_n| \, d\mu_n\right)=0;$$

or, equivalently, (ii)

$$\sup_n \int |f_n| \, d\mu_n < \infty;$$

or (iii)

$$\lim_n \int_{E_n} |f_n| \, d\mu_n \to 0 \quad \text{for every sequence } (E_n) \text{ for which } \mu_n(E_n)\to 0.$$

We note that:

Proposition 11

If the sequence (f_n) is uniformly integrable then the sequence of distributions of f_n is tight.

[29] These results, together with much of this section, are taken directly from Hildenbrand (1974).

Now we state a result of fundamental importance in studying sequences of economies:

Theorem 20 (Generalisation of Lebesgue's Theorem)

Let the sequence (f_n) of measurable functions converge in distribution to the measurable function f. If the sequence (f_n) is uniformly integrable then f is integrable, and furthermore,

$$\lim_n \int f_n \mathrm{d}\mu_n = \int f \mathrm{d}\mu.$$

If f and all f_n are positive and integrable, then the above equation implies that the sequence (f_n) is uniformly integrable.

The last result of this section is of interest since it shows how we may deal with the situation typically found in establishing results for growing sequences of economies. At each step we deal with a finite economy, and it is only in the limit that we are concerned with an infinite economy. To see how we may think of the infinite economy as representing the limit of the sequence of finite economies, we consider the following idea: At each step we draw from some fixed hypothetical distribution a finite sample and as these samples increase in size we would want the sample distributions to approximate more and more closely that of the underlying infinite population. With this in mind we state the following:

Theorem 21 (Glivenko–Cantelli)

Let $(A, \mathcal{A}, \mu)$ be a measure space and (x_n) an independent sequence of identically distributed measurable mappings x_n of A into a separable metric space T. For every $a \in A$, let $\nu_n(a,.)$ be the distribution of the sample $\{x_1(a), \ldots, x_n(a)\}$ of size n $(n=1,2,\ldots)$, i.e.,

$$\nu_n(a, B) = \frac{1}{n}\{i \mid x_i(a) \in B,\ i=1,\ldots,n\}.$$

Then for almost all $a \in A$ the sequence $(\nu_n(a,.))$, $n=1,2,\ldots$, of sample distributions converges weakly to the distribution of x_n.

3. Some results

We will now look at some examples in which many of the preceding concepts are used.

3.1. A large economy

Example 7[30]

Consider $\mathcal{P}$, the set of irreflexive and continuous binary relations on R^ℓ_+ with the property that for every price vector $p \gg 0$ the set $\phi(\succ, e, p)$ of maximal elements for $\succ$ in the consumer's budget set $\{x \in R^\ell_+ \mid p \cdot x \leqslant p \cdot e\}$ is non-empty. If $\mathcal{P}$ is endowed with Hausdorff's topology of closed convergence it is a separable metric space.

First we look at the case of a finite number of economic agents and define, as mentioned earlier, an exchange economy as a mapping from the space of agents to the space of characteristics, i.e., preferences and endowments,

$$\mathcal{E}: A \to P \times R^\ell_+ .$$

Now the distribution μ_ε of agents' characteristics is given by

$$\mu_\varepsilon(B) = \frac{\#\mathcal{E}^{-1}(B)}{\#A} \quad \text{for every} \quad B \text{ of } \mathcal{P} \times R^\ell_+ .$$

Again we emphasise that in the case of a large economy this second description may well be considered as more appropriate since we are not really concerned with specifying the characteristics of each individual, but are more interested in knowing what proportion of individuals fall within any given subset of characteristics.

Now for the infinite case we define the space of agents as a measure space $(A, \mathcal{A}, v)$ with $v(A) = 1$, i.e., v is a *probability measure* and an economy is a *measurable mapping* $\mathcal{E}: (A, \mathcal{A}, v) \to \mathcal{P} \times R^\ell_+$ such that $\int e \, dv < \infty$.

As will be described in detail in Chapter 18 by Hildenbrand, we may show the equivalence of two different solution concepts for the problem of allocating goods in an infinite economy. For this result to be interesting we must show it to be essentially true for large economies. To do this we must be able to describe a sequence of economies converging to a limit (infinite) economy, where, in particular, that limit economy is *atomless*. To this end we introduce the following:

Definition 39

The sequence $(\mathcal{E}_n)$, $\mathcal{E}_n: A_n \to \mathcal{P} \times R^\ell_+$, is called *purely competitive* if and only if

(i) the number $\#A_n \underset{n \to \infty}{\to} \infty$;

[30] For a full discussion of this example, see Hildenbrand (1975) from which it is taken, and Hildenbrand (1974).

(ii) the sequence (μ_{ε_n}) of preference endowment distributions *converges weakly* to a limit μ;
(iii) $\int e \, d\mu_\varepsilon \to \int e \, d\mu \gg 0$.

Note that the important idea here is that the distributions converge and that the limit economy is characterised by its distribution. This is again because the underlying measure space is arbitrary and, since it can be shown that two economies with the same preference endowment distribution are essentially the same, we can forego the micro distribution of an economy.

With Definition 39 a number of important limit theorems can be proved and details are given in Chapter 18.

3.2. *Fair division: Some results*

We return now to a problem mentioned in the introduction which has interested mathematicians for some period of time and which, with the increasing interest of economists in equitable distributions, shows clearly how measure-theoretic tools may be useful. Recall that the simplest expression of the problem, as we saw it, is that of dividing some object U amongst a finite number n of individuals such that each individual i receives in his own estimation (expressed by a measure μ_i on U) at least $1/n$ of the total "value" of U. More generally we might want to assign parts of the object to individuals such that the ith individual receives α_i and the others receive α_j, in his opinion, of the total where $\sum_{i=1}^{n} \alpha_i = 1$. The answer to this problem is to be found in Dubins and Spanier (1961). Their first result is:

Theorem 22

Let $(U, \mathfrak{A})$ be a measurable space $\mu_1, \ldots, \mu_n$ atomless probability measures on $(U, \mathfrak{A})$. Then given k and $\alpha_1, \ldots, \alpha_k > 0$ with $\sum_{i=1}^{k} \alpha_i = 1$, there exists a partition $A_1, \ldots, A_k$ of U such that $\mu_i(A_j) = \alpha_j$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, k$.

The reader will observe that this result answers the question posed for $k = n$ and in particular setting $\alpha_i = 1/n$ for all i shows that a partition can be found that not only gives the ith individual his fair share but one which everybody believes gives the others their fair share too.

Provided that at least two individuals have different measures then there exist partitions which give strictly more than α_i to the ith individual, i.e., such that $\mu_i(A_i) > \alpha_i$, $i = 1, \ldots, n$.

This result involves an extension of Liapunov's theorem (Theorem 3), and Dubins and Spanier give a proof of that Theorem in proving the proposition from which Theorem 24 is derived.

The authors go on to prove the existence of partitions which are optimal in some sense. For example, one might wish to adopt the utilitarian criterion and find a partition $A_1, \ldots, A_n$ to maximise

$$\sum_{i=1}^{n} \mu_i(A_i),$$

and indeed the maximum is shown to exist.

Perhaps of more interest is the anticipation of Rawl's criterion of maximising the welfare of the least well-off individual.

Of all partitions consider those which maximise the amount received by the person who gets least. From these select those which give the most to the person who gets next to the least and so forth. More precisely if P is a partition then arrange the members $\mu_i(A_i)$ in non-decreasing order to construct the sequence

$$a_1(P), a_2(P), \ldots, a_n(P).$$

Now construct the lexicographic ordering on partitions P. Thus P is maximal in that ordering if for any other partition P' either $a_i(P)=a_i(P')$ for all i or if j is the smallest i such that $a_j(P)\neq a_j(P')$; then $a_j(P)>a_j(P')$. Such a maximal element we will call an *optimal partition*. Dubins and Spanier prove that such partitions exist and furthermore that if each μ_i is absolutely continuous with respect to every other then every optimal partition is equitable in the sense that

$$\mu_i(A_i)=\mu_j(A_j) \quad \text{for all } i \text{ and } j.$$

3.3. *Integration of correspondences*[31]

It is often the case that we are concerned with set-valued mappings, or correspondences, in economics. For example the demand of a given individual may be a set of bundles rather than a particular bundle for some given prices. Again we may wish to associate with an individual a production technology which would be a set of possible combinations of inputs and outputs. If we wish to talk about perfect competition in such circumstances and to use a continuum economy to do so, then we will need to be able to integrate such correspondences. If, for example, we want to be able to talk about mean demand for the

[31] This brief discussion is intended to give an indication of the sort of problem encountered in an area that demands greater sophistication than the other topics mentioned. A full treatment and references may be found in Hildenbrand (1974, pp. 53–79). A standard treatment of this problem is one where the integral is considered as the expectation of a set-valued random variable.

whole continuum economy we will be forced to integrate the demand correspondences of individuals.

Consider first a function f from a measure space $(A, \mathcal{A}, \mu)$ into R^m, that is, $f=(f^1,\ldots,f^m)$ where each $f^i: A\to R$ $(i=1,\ldots,m)$. The function f is said to be *integrable* if each coordinate function f^i is integrable and the integral $\int f \mathrm{d}\mu$ is defined by

$$\left(\int f^1 \mathrm{d}\mu,\ldots,\int f^m \mathrm{d}\mu\right).$$

Now consider ϕ a set-valued mapping or correspondence of A into R^m. Denote by $\mathcal{L}_\phi$ the set of all μ integrable functions that have the property

$$f(a)\in\phi(a) \quad \text{a.e. in } A.$$

The functions in the set $\mathcal{L}_\phi$ are called integrable selections of ϕ.

Definition 40[32]

The set $\{\int f \mathrm{d}\mu \in R^m \mid f\in\mathcal{L}_\phi\}$ is called the *integral* of the correspondence ϕ and is denoted by $\int\phi \mathrm{d}\mu$ or by $\int\phi$.

Although the meaning of this definition is clear we have yet to show that there is a large class of correspondences for which the integral is non-empty, that is for which there exists an integrable selection. If a correspondence admits an integrable selection we say that the correspondence itself is integrable.

To ensure that the integral is non-empty we need first to establish that there exists a measurable selection. First, then, we give the following:

Definition 41[33]

A *correspondence* ϕ from a measure space $(A, \mathcal{A}, \mu)$ into a complete separable metric space S is *measurable* if the graph of that correspondence belongs to $\mathcal{A}\otimes\mathcal{B}(S)$.

Now the basic result is the following:

Theorem 23 (Measurable Selection)

Let ϕ be a measurable correspondence of a measure space $(A, \mathcal{A}, \mu)$ into a complete separable metric space S. Then there exists a measurable function f of A into S such that $f(a)\in\phi(a)$ a.e. in A.

[32] This is the definition given by Aumann (1965).

[33] Note that for many purposes in economics $S=R^\ell$.

We need now, of course, to make sure that the correspondence is bounded in order to ensure that it is integrable:

Definition 42

A correspondence ϕ of $(A, \mathcal{A}, \mu)$ into R^ℓ is *integrably bounded* if there exists an integrable function h of A into R^ℓ_+ such that for every $a \in A$ and for every $x \in \phi(a)$ we have

$$|x| = (|x_1|, \ldots, |x_\ell|) \leqslant h(a).$$

Now we can give the following:

Theorem 24

A measurable, integrably bounded correspondence from $(A, \mathcal{A}, \mu)$ into R^ℓ is integrable.

Now we turn to a particularly interesting aspect of large economies, that is, the "convexifying effect" of large numbers. Thus, at an individual or micro level we are obliged to make assumptions of convexity of individual preferences to guarantee that the demand correspondence is convex valued. This property then carries over to the aggregate excess demand correspondence, and one can make use of Kakutani's fixed point theorem to prove the existence of equilibrium. This problem is fully dealt with in the chapter on the existence of competitive equilibrium, Chapter 15 by Debreu.

However, if we have an atomless measure space of agents we may dispense with the requirement of convexity of preferences, since even though individual demand correspondences will not be convex valued, mean demand will necessarily be so. This follows directly from:

Theorem 25

Let ρ be a correspondence from an atomless measure space $(A, \mathcal{A}, \mu)$ into R^m then the set

$$\mathcal{Z} = \left\{ \int_E f \mathrm{d}\mu \mid f \in \mathcal{L}_\rho, E \in \mathcal{A}, \mu(E) > 0 \right\}$$

is a convex set in R^m.

The proof of this theorem depends on Liapunov's theorem, emphasising the importance of the latter. One can also proceed to develop approximation results for large finite economies.

The frequent use of correspondences in economic theory has led to the development of a substantial literature on the integration of correspondences and the interested reader will find the appropriate references in Hildenbrand (1974).

In passing, it is interesting to note that measure theory and a measure-theoretical approach to the description of large economies has in a sense diminished some of the problems posed by the use of correspondences.

In particular, we discussed earlier a standard problem, namely that for individuals with convex preferences the appropriate demand concept is a correspondence. Yet it can be shown that if we have a measure space of consumers and the support of that measure is sufficiently rich, then the aggregate or mean demand may, in fact, be considered as a function if we add some more structure to the space of preferences. That is, for any given prices, only a negligible subset of consumers will have a set rather than a single bundle as their demand; hence the demand will be a function. Thus, in a sense, the importance of correspondences as such is somewhat diminished in large economies.

4. Conclusion

In this chapter we have presented a number of results which should be useful for the reader interested in economic theory in two ways: they should provide an underpinning for many of the probabilistic approaches to problems in economic theory, and they should also provide a basis for an understanding of that part of the literature which adopts the measure-theoretic approach as a description of an economy.

No attempt has been made to give a comprehensive survey of the general literature in which measure theory is applied to problems in economic theory since this literature is already extensive and is involving an increasing number of different areas. However, with the selection of results presented here the reader should find much of this literature readily accessible.

References

Allen, R. G. D. and A. L. Bowley (1935), Family expenditure: A study of its variation. London: Staples Press.

Araujo, A. and A. Mas-Colell (1978), "Notes on the smoothing of aggregate demand", Journal of Mathematical Economics, 5:113–129.

Arrow, K. J. (1963), Social choice and individual values, 2nd ed. New York: Wiley.

Aumann, R. J. (1964), "Markets with a continuum of traders", Econometrica, 32:39–50.

Aumann, R. J. (1965), "Integrals of set-valued functions", Journal of Mathematical Analysis and Applications, 12:1–12.

Aumann, R. J. (1966), "Existence of competitive equilibria in markets with a continuum of traders", Econometrica, 32:1–17.
Billingsley, P. (1963), Convergence of probability measures. New York: Wiley.
Blaug, M. (1963), Economic theory in retrospect. Homewood, IL: Richard D. Irwin.
Brown, D. J. and A. Robinson (1972), "A limit theorem on the core of large standard exchange economies", Proceedings of the National Academy of Sciences, USA, 69:1258–1260. Correction, 69:3068.
Debreu, G. (1959), Theory of value. New York: Wiley.
Debreu, G. (1970), "Economies with a finite set of equilibria", Econometrica, 38:387–392.
Debreu, G. and H. Scarf (1963), "A limit theorem on the core of an economy", International Economic Review, 4:235–246.
Dierker, E., H. Dierker and W. Trockel (1978), "Continuous mean demand function", Discussion paper. Bonn: University of Bonn.
Dubins, L. E. and E. H. Spanier (1961), "How to cut a cake fairly", American Mathematical Monthly, 1:1–17.
Edgeworth, F. Y. (1881), Mathematical psychics. London: P. Kegan.
Fishburn, P. C. (1970), "Arrow's impossibility theorem, concise proof, and infinite voters", Journal of Economic Theory, 2:103–106.
Grandmont, J. M., A. P. Kirman and W. Neuefeind (1974), "A new approach to the uniqueness of equilibrium", Review of Economic Studies, 41:289–292.
Halmos, P. R. (1961), Measure theory, 7th ed. Princeton, NJ: Van Nostrand.
Hildenbrand, W. (1970), "Existence of equilibria for economies with production and a measure space of consumers", Econometrica, 38:608–623.
Hildenbrand, W. (1974), Core and equilibria of a large economy. Princeton, NJ: Princeton University Press.
Hildenbrand, W. (1975), "Distributions of agents' characteristics", Journal of Mathematical Economics, 2:129–138.
Hildenbrand, W. (1979), "On the uniqueness of mean demand for dispersed families of preferences", Discussion Paper. Bonn: University of Bonn.
Ichiishi, T. (1976), "Economies with a mean demand function", Journal of Mathematical Economics, 3:167–171.
Khan, M. A. (1974), "Some equivalence theorems", Review of Economic Studies, 41:549–565.
Kirman, A. P. and D. Sondermann (1972), "Arrow's theorem, many agents, and invisible dictators", Journal of Economic Theory, 4:267–277.
Lindenstrauss, J. (1966), "A short proof of Liapunov's convexity theorem", Journal of Mathematics and Mechanics, 15:971–972.
Mas-Colell, A. (1979), "Perfect competition and the core", Working paper no. IP280. Berkeley, CA: Center for Research in Management Science, University of California.
Neveu, J. (1965), Mathematical foundations of the calculus of probability. San Francisco, CA: Holden-Day.
Robinson, A. (1965), Non-standard analysis. Amsterdam: North-Holland.
Schumpeter, J. (1954), History of economic analysis. Oxford: Oxford University Press.
Shapley, L. S. (1953), "Stochastic games", Proceedings of the National Academy of Sciences, USA, 39:327–332.
Shitovitz, B. (1974), "Oligopoly in markets with a continuum of traders", Econometrica, 41:467–501.
Sondermann, D. (1975), "Smoothing demand by aggregation", Journal of Mathematical Economics, 2:201–224.

Chapter 6

THE ECONOMICS OF UNCERTAINTY: SELECTED TOPICS AND PROBABILISTIC METHODS*

STEVEN A. LIPPMAN and JOHN J. McCALL

University of California, Los Angeles

Thus Peirce conjectured that the world was not only ruled by the *strict Newtonian laws*, but that it was also at the same time ruled by *laws of chance*, or of randomness, or of disorder: by laws of statistical *probability*. This made the world an interlocking system of clouds and clocks, so that even the best clock would, *in its molecular structure*, show some degree of cloudiness. So far as I know Peirce was the first post-Newtonian physicist and philosopher who thus dared to adopt the view that to some degree *all clocks are clouds*; or in other words, that *only clouds exist*, though clouds of very different degrees of cloudiness.

Karl Popper, *Of Clouds and Clocks*

1. The economics of uncertainty

1.1. Introduction and overview

The literature on probabilistic economics is enormous and burgeoning.[1] In this chapter we will identify some areas where probabilistic analysis has enriched our understanding of economic behavior and has generated fundamental advances in economic theory.

Few would doubt that uncertainty has a decisive influence on economic behavior. Almost every phase of consumption and production is affected by uncertainty. Perhaps the most significant uncertainty is length of life. Also, individuals are uncertain about their incomes and producers are unsure of their sales and costs. The number, size of purchase, and intervals between arrivals of customers at a store are all stochastic, as are the number of employees who arrive at work on a given day. Inventory depletions, equipment breakdowns, wars, depressions, and inflations all occur unpredictably, and the results of research and technological processes are probabilistic. In short, economic agents operate in an environment permeated by stochastic phenomena, and their basic

*This research was partially supported by the National Science Foundation through Grant SOC-7808985 and the Walgreen Foundation. We acknowledge the helpful comments of A. A. Alchian, D. P. Baron, D. W. Carlton, J. M. Harrison, M. D. Intriligator, H. E. Leland, and L. G. Telser.

[1]The main references include Arrow (1971), Balch, McFadden and Wu (1974), Borch (1968), Diamond and Rothschild (1978), McCall (1971), and Hirshleifer and Riley (1979).

economic decisions are modified accordingly. Consequently, the underlying determinants of supply and demand have stochastic components, and relative prices are random variables.

At first blush, risk would appear to have no influence on economic behavior unless people were averse to risk. But, though in many circumstances risk aversion is a fact *and* is essential to understanding economic behavior, much economic behavior is a direct consequence of uncertainty and is independent of risk aversion. Human behavior adapts to uncertainty and risk aversion in a variety of ways. Insurance, futures markets, contingency clauses in contracts, and use of stock markets are among the most important institutions that facilitate adaptation to risk aversion. On the other hand, methods like inventory control, preventive maintenance, and annual physical examinations are also used by individuals and firms to cope with uncertainty, independent of risk aversion. For example, Jones and Ostroy (1976) have shown that risk aversion is not essential to the increased "flexibility" that is achieved by building plants with flat average cost curves [Stigler (1939)], engaging in parallel research and development programs [T. Marschak (1962) and Nelson (1961)], and holding substantial quantities of liquid assets [Makower and J. Marschak (1938)]. A position is defined to be "more flexible than another if the range of alternative future positions attainable from it at any given level of cost [subsumes]...that of the other".

These adaptations to uncertainty are important manifestations of rational behavior. A deterministic economic theory does not provide an adequate explanation of these responses to a stochastic environment. Similarly, the information accumulation that characterizes the decision processes of people is inexplicable by a purely deterministic model. On these grounds, it is easy to explain and to applaud the development of probabilistic economics. It is not easy, however, to explain the long time during which deterministic models have dominated economic theory.

In 1958, Arrow remarked that:

> uncertainty has been long discussed in economic literature, but usually only in a marginal way. Discussion rarely advanced beyond the point of the famous article by Daniel Bernoulli who argued that the individual acts in such a way as to maximize the mathematical expectation of his utility. Despite the elegance and simplicity of this theory, little real use was made of it in explaining the facts of the business world beyond the existence of insurance, at least until the work of Frank Knight in 1921, which was continued somewhat belatedly by J. R. Hicks and Albert G. Hart who made the first fruitful applications of the theory of uncertainty to the behavior of business firms, particularly in regard to such questions as liquidity, the holding of inventories, and flexibility of production processes.

In spite of these insights by Knight (1921), Hicks (1931), and Hart (1942), for the most part the economics profession continued to propagate the classical economic theory in which costlessly acquired perfect information accompanies all economic activity.

When confronted with stark empirical realities, how could such a theory survive? It was useful in explaining several phenomena: tax effects, demand and supply as resource allocators, explanations of relative prices, specialization in production, foreign exchange rates, etc. But an enormous set of economic activity could not be explained, and, worse yet, came to be regarded as inefficient, undesirable, or irrational. Undoubtedly another part of the answer lies in the manner in which economic theory is tested. Econometricians always include stochastic terms in their econometric models. Hard empirical facts necessitated these additions. However, they were seldom grounded in economic theory and, when present, were merely appended to economic models that had been constructed in a certainty milieu. Furthermore, the stochastic component of econometric models frequently was viewed as an inconvenience which a properly formulated deterministic model would eventually eliminate.

Another probable reason for the paucity of probabilistic analysis is that economic theory rarely attempted to explain the behavior of individual firms or individual consumers. Rather the focus has been on the behavior of the *representative* firm and the *average* consumer. Where there are so many firms and consumers, a law of large numbers was implicitly invoked to reduce uncertainty essentially to zero. While invoking the law of large numbers is appropriate for some applications, it is misleading for others. Consumers search for low prices, purchase insurance, place a positive value on information, and diversify holding of risky assets. The firm searches for productive employees, insures against fire and other "acts of God", and purchases many kinds of information. None of these actions is consistent with a certainty model.

The success of probabilistic modeling in other sciences like physics and genetics raises the question—Is there some reason to believe that economic behavior is less susceptible to probabilistic formulation than, say, genetic behavior? We think not, believing that economics is the cloudiest of the sciences and thereby agreeing with de Finetti (1974) that:

> in the field of economics, the importance of probability is, in certain respects, greater than in any other field. Not only is uncertainty a dominant feature, but the course of events is itself largely dependent on people's behavior...It is, therefore, probability theory, in the broadest and most natural sense, that best aids understanding in this area.

Before we discuss the topics covered in this chapter, a further comment is in order. First, our training, experience, and pedagogical preferences obviously

have influenced our selections from the vast body of economic literature in which probabilistic methods play a prominent role. And even though topics such as signalling,[2] the theory of agency,[3] bidding,[4] stochastic equilibria,[5] growth theory,[6] the theory of finance (e.g. option pricing),[7] and the economics of queuing[8] are not included, we hope that the end product provides a broad perspective of the field. Our goal has been to serve a hearty meal from the body of knowledge that we refer to as the "economics of uncertainty".

The economics of job search is the topic of Section 2. Many probabilistic models have been used in the job search literature. The basic problem is determining when an individual should stop searching and accept employment. As will be shown this is a problem in optimal stopping, and it can be solved using martingale arguments. We use the job search setting to illustrate the usefulness of other probabilistic constructs and concepts like Markov chains, the Borel–Cantelli lemma, Bayes' rule, first- and second-order stochastic dominance, Jensen's inequality, and uniform integrability.

Section 3 discusses what we regard as the most important topic in the economics of uncertainty, viz., the economics of insurance. We briefly describe moral hazard and adverse selection and then use a simple two-state model to study the effects of increases in risk aversion and risk on the demand for insurance.

A fundamental stochastic process is presented in Section 4 to describe security price fluctuations; it is the Brownian motion that has proven so useful in other applications. The intimate relation between the efficient market hypothesis and martingales is also discussed. Finally, both a discrete time (random walk) and a continuous time (Brownian motion) model of cash balances are analyzed, making use of the Optional Sampling Theorem for martingales.

The subject matter of Sections 5 and 6 is consumption and production under uncertainty, respectively. We begin Section 5 with a simple two-period model and then follow with several multi-period models. Because consumption streams exhibit all of the essential features of sequential decision making under uncertainty, these models, like those of Sections 2–4 are formulated as dynamic programs. But unlike those of Sections 2–4 risk aversion plays the lead role. Using the notions of stochastic dominance and contraction mappings, we study the effects of increased uncertainty, with the certainty model as the polar case. In the production models, the impact of uncertainty coupled with risk aversion yields results quite different from those of classical price theory. Here we

[2] See Spence (1973) and Riley (1975).
[3] See Harris and Raviv (1979) and Ross (1973).
[4] See Englebrecht Wiggams (1978) and Vickrey (1961).
[5] See Green (1973) and Radner (1968).
[6] See Cass and Shell (1978) and Shell (1967).
[7] See Brock (1981), Rubinstein (1979), and Ziemba and Vickson (1975).
[8] See Levhari and Sheshinski (1972) and Lippman and McCall (1979).

analyze not only the effects of increased uncertainty but also increased aversion to risk on production decisions, where a firm's aversion to risk is defined by either of the Arrow–Pratt measures discussed in Section 1.2.

We conclude this chapter in Section 7. It discusses the intimate relation between economics and biology, with evolution being the critical nexus. Stochastic processes are, of course, fundamental to the theory of evolution; in particular, Markov chains and results from branching processes are employed.

1.2. Measures of risk and risk aversion

During the course of our analysis we shall assume that the economic agent encounters a stochastic environment and acts so as to maximize his expected utility of the random outcome. The agent's utility function u is assumed to be non-decreasing and concave, whence $r_u \geqslant 0$, where the *Arrow–Pratt measure of absolute risk aversion*, r_u, is defined by

$$r_u(t) = -u''(t)/u'(t).$$

Of course, $r_u \equiv 0$ if u is linear. The appropriateness/usefulness of this measure of risk aversion will be revealed in our analysis. Finally, we say that an agent with utility function u is more risk averse than an agent with utility function v if $r_u \geqslant r_v$. On several occasions (e.g., see Section 5.2) we shall also employ the *Arrow–Pratt measure of relative risk aversion*, R_u, defined by

$$R_u(t) = tr_u(t).$$

Having specified a measure of the agent's aversion to risk, we proceed by supplying two measures of risk, each of which induces a partial order on the random variables in the agent's environment. To be of use, of course, there must necessarily be a close connection (in fact, an equivalence) between the ordering of the set of random variables and the associated ordering induced by their expected utilities.

Let X and Y be two random variables with cumulative distribution functions F and G, respectively. Our goal is to furnish criteria which assert that X is "better than" Y. We say that the random variable X is *stochastically larger* than the random variable Y, written $X \succ_1 Y$, if and only if

$$G(t) - F(t) \geqslant 0, \quad \text{for all } t. \tag{1.1}$$

When F and G satisfy (1.1), we say that X dominates Y—or F dominates G according to the criterion of first-order stochastic dominance. Because

$$\mathrm{E}(X) = -\int_{-\infty}^{0} F(t)\,\mathrm{d}t + \int_{0}^{\infty} [1 - F(t)]\,\mathrm{d}t, \tag{1.2}$$

for any random variable X [with the proviso that at least one of the two integrals in (1.2) is finite], it is clear that $\mathrm{E}(X) \geqslant \mathrm{E}(Y)$ whenever $X \succ_1 Y$. The link between first-order stochastic dominance and expected utility is provided in the next theorem.

Theorem 1

A necessary and sufficient condition for X to be stochastically larger than Y is

$$\mathrm{E}h(X) \geqslant \mathrm{E}h(Y), \quad \text{all} \quad h \in \mathfrak{L}_1, \tag{1.3}$$

where $\mathfrak{L}_1$ is the set of non-decreasing functions.

As revealed by (1.2), $X \succ_1 Y$ and $F \neq G$ imply that $\mathrm{E}(X) > \mathrm{E}(Y)$. Consequently, we seek a weaker condition, one that will enable us to distinguish between random variables with equal means. We say that the random variable X is *less risky* than the random variable Y, written $X \succ_2 Y$, if and only if

$$\int_{-\infty}^{t} [G(s) - F(s)]\, \mathrm{d}s \geqslant 0, \quad \text{for all } t. \tag{1.4}$$

When F and G satisfy (1.4), we say that X dominates Y in the sense of second-order stochastic dominance. Again, (1.2) reveals that $\mathrm{E}(X) \geqslant \mathrm{E}(Y)$ whenever $X \succ_2 Y$. However, second-order stochastic dominance includes the case $X \succ_2 Y$, $F \neq G$, and $\mathrm{E}(X) = \mathrm{E}(Y)$. Additionally, the so called "mean preserving spread" is simply a special case, namely, the one in which $F - G$ changes sign exactly once and $\mathrm{E}(X) = \mathrm{E}(Y)$.

The connection between second-order stochastic dominance and expected utility is contained in the next theorem. Of particular significance is (1.6), because it is often applied to u', the derivative of the agent's utility function; unlike u, u' is a decreasing function.

Theorem 2

Let $\mathfrak{L}_2$ be the set of non-decreasing concave functions. A necessary and sufficient condition for X to be less risky than Y is

$$\mathrm{E}h(X) \geqslant \mathrm{E}h(Y), \quad \text{all} \quad h \in \mathfrak{L}_2. \tag{1.5}$$

Moreover, if $X \succ_2 Y$ and $\mathrm{E}(X) = \mathrm{E}(Y)$, then

$$\mathrm{E}h(X) \geqslant \mathrm{E}h(Y), \quad \text{all concave functions } h. \tag{1.6}$$

2. The economics of search

2.1. Introduction

The economics of search is an important area where probabilistic modeling has played a prominent role.[9] Search is a fundamental property of economic markets. Deterministic theories of economic markets encounter obstacles when they attempt to explain such market phenomena as different prices for "identical" outputs or inputs, persistent positive levels of "unemployed" resources, and "underutilization" of employed resources. These phenomena are explicable by a search theory of economic markets. The main decisions confronting searchers are determining the appropriate amount of information and efficient methods of acquiring information before acting, where, for example, action is the acceptance of a particular job offer by a job searcher or the hiring of employees with certain identifiable characteristics by the searching employer. Of course, search is not restricted to labor markets and, indeed, many important contributions have emanated from the study of other economic models where the objective is to locate the lowest price rather than the highest wage. However, for expositional ease we will concentrate on job search in labor markets.

This section begins with an extensive discussion of an elementary search model in a labor market setting. In the next two sections we consider the impact of increased uncertainty and utilize martingales to establish rigorously the existence of a reservation wage. Adaptive search and a changing economy are briefly treated in the final section.

2.2. The elementary search model

We begin with the simplest sequential model of job search. An individual, referred to as the searcher, is seeking employment. Each and every day (until he accepts a job) he ventures out to find a job, and each day he generates exactly one job offer.[10] The offer should be interpreted as the (discounted present value of the) lifetime earnings from a job. The cost of generating each offer (which includes all out-of-pocket expenditures such as advertising and transportation that are incurred each time a job offer is obtained) is a constant c, and there is no limit on the number of offers the searcher can obtain. We consider both the case wherein offers not accepted immediately are lost and the case in which all

[9]Search models were first discussed in the pioneering work of Stigler (1961, 1962). Whereas Stigler's model was non-sequential, sequential models have predominated since McCall (1965). Much of this material is drawn from the recent survey by Lippman and McCall (1976b).

[10]To allow for the possibility that on some days he receives no offer, we permit some employers to offer a wage of zero.

offers are retained; these two cases are referred to as sampling without recall and sampling with recall, respectively. When an offer is accepted, the searcher transits to the permanent state of employment.

Whereas the searcher's skills are unvarying, prospective employers do not necessarily evaluate or value them equally; consequently, different employers tender different offers to the searcher. This "dispersion of offers" is incorporated into the model by assuming that there is a probability distribution F of wages which governs the offers tendered; additionally, the distribution is assumed invariant over time. Thus, on any given day the probability that the searcher will receive an offer of w or less is $F(w)$, independently of all past offers and of the time the offer is made. Moreover, we assume that the job searcher knows F. All participants in job search are assumed to be risk neutral (i.e., possess linear utility functions) and seek to maximize their expected net benefits. The only decision the searcher must make is when to stop searching and accept an offer.

To be more precise, a job offer X_i is presented each period, where each X_i is a random variable with cumulative distribution function $F(\cdot)$, $\mathrm{E}(X_i) < \infty$, and the X_i's are mutually independent. (To simplify the mathematical analysis, we assume that X_1 is a continuous random variable.) The job searcher is assumed to retain the highest job offer so that the return from stopping after the nth search is given by

$$Y_n = \max(X_1, \ldots, X_n) - nc,$$

where c is the (out-of-pocket) cost per period of search.

The objective is to find a stopping rule that maximizes $\mathrm{E}(Y_N)$ where N is the random stopping time, i.e., the (random) number of job offers received until one of them is accepted.

Clearly, the optimal amount of search (the period of unemployment) depends on the distribution F of wages that the individual's services command in the labor market and on c, the opportunity cost of the searching activity. If the searcher's skills are highly valued, he will reject offers that fall short of his expectations and remain unemployed. On the other hand, if the cost of search is high, the job searcher will tend to limit his searching activities. The literature in this field concentrates on situations in which the optimal policy for the job searcher is to reject all offers below a single critical number, termed the *reservation wage*, and to accept any offer above this critical number. Policies with this simple structure are said to have the *reservation wage property*. Thus, it is appealing to restrict our attention to policies (i.e., stopping rules) of the form

$$\text{Accept a job offer if and only if it is at least } y, \tag{2.1}$$

where y is the particular critical number under consideration.

In Section 2.4, we shall demonstrate (assuming that the variance of X_1 is finite) that there is an optimal rule with the form given in (2.1). For the time being, assume without proof that (a) there is an optimal rule, and (b) it has the reservation wage property [i.e., is of the form given in (2.1)].[11]

Let g_y be the expected gain from search when using a policy with the reservation wage y. Furthermore, denote by N_y the number of offers needed until an acceptable offer is found, whence N_y is a geometric random variable[12] with parameter $p \equiv 1-F(y)$ and $\mathrm{E}(N_y)=1/p$. Then [provided $1-F(y)>0$] g_y satisfies

$$g_y = -c/(1-F(y)) + \int_y^\infty x\,\mathrm{d}F(x)/(1-F(y)), \tag{2.2}$$

since N_y is the number of observations needed to find an offer greater than or equal to y and the second term on the right-hand side of (2.2) is merely the conditional expected value of an offer given that it is at least y. Rearranging (2.2) yields

$$c = \int_y^\infty (x-g_y)\,\mathrm{d}F(x). \tag{2.3}$$

From (2.3) and the fact that we seek y to maximize g_y, it is clear upon reflection[13] that if ξ is the optimal reservation wage, then $g_\xi = \xi$ so that ξ satisfies

$$c = \int_\xi^\infty (x-\xi)\,\mathrm{d}F(x). \tag{2.4}$$

Moreover, g is unimodal as shown in Figure 2.1. Defining H by

$$H(x) = \int_x^\infty (t-x)\,\mathrm{d}F(t), \tag{2.5}$$

[11] Robbins (1970) has provided a proof of this under the rather weak assumption that $\mathrm{E}(X_1)<\infty$.

[12] That is, $\mathrm{P}(N_y=k)=p(1-p)^{k-1}$, $k=1,2,\ldots$

[13] To see this, let g_y satisfy (2.3) and suppose that $y-g_y=\varepsilon\neq 0$. Whether $\varepsilon>0$ or $\varepsilon<0$, we have

$$\int_{y-\varepsilon}^\infty (x-g_{y-\varepsilon})\,\mathrm{d}F(x) = c = \int_y^\infty (x-g_y)\,\mathrm{d}F(x) < \int_{y-\varepsilon}^\infty (x-g_y)\,\mathrm{d}F(x).$$

Hence, $g_{y-\varepsilon}>g_y$ when $y>g_y$ (i.e., $\varepsilon>0$) as well as when $y<g_y$ (i.e., $\varepsilon<0$).

Alternatively, take the derivative of g with respect to y and set it equal to zero to obtain [$F'(y)\equiv f(y)$; the H function is defined in (2.5)]

$$0 = g_y' = \{H(y)-c\}f(y)/[1-F(y)]^2$$

so that $g_y'>0$ for $y<\xi$, $g_\xi'=0$, and $g_y'<0$ for $y<\xi$, whence g_y is maximized at ξ. Moreover, $g_y\to 0$ as $y\to\infty$. These facts enable us to assert that g is as depicted in Figure 2.1.

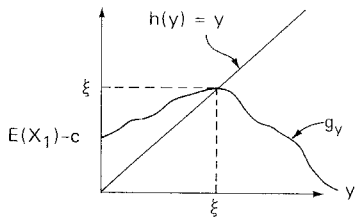


Figure 2.1. A graph of the g function.

and noting (see Figure 2.2) that H is a convex, non-negative, strictly decreasing function which approaches 0 and $E(X_1)$ as y approaches ∞ and 0, respectively, we see from (2.4) that ξ is the unique[14] solution of $H(x)=c$.

An alternative and heuristic line of reasoning implies that ξ is the unique wage for which the searcher is indifferent between accepting ξ and continuing. If he continues it will be optimal to stop with the next observation as the wage available will then be at least ξ. Thus, ξ should satisfy

$$\xi = E\max(\xi, X_1) - c, \tag{2.6}$$

as the left-hand side is what he receives upon accepting the offer of ξ and the right-hand side is the expected return of taking one more observation. It is easy to verify that (2.6) is equivalent to (2.4).

Equation (2.4) has a simple economic interpretation: the critical value ξ associated with the optimal stopping rule is chosen to equate c, the marginal cost of obtaining one more job offer, with $H(\xi)$, the expected marginal return from one more observation. Thus, it suffices for the job searcher to behave myopically; namely, he need only compare his return from accepting employment with the expected return from exactly one more observation.

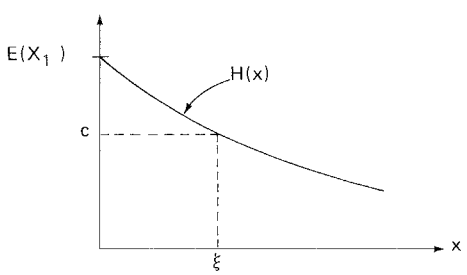


Figure 2.2. A graph of the H function.

[14] If $c > E(X_1)$ then $\xi \equiv 0$.

It is important to distinguish between the reservation wage property and the myopic property: the former tells us which offers are acceptable, specifically those exceeding ξ, whereas the latter provides us with a simple method for calculating ξ. Summing up, the structure of the optimal policy is such that it is characterized by a single number, referred to as the reservation wage. In addition, this number is obtained by comparing the value of stopping not with the value of continuing on (perhaps for a long time) in an "optimal" manner but rather with the value of taking exactly one more observation. Furthermore, the expected return from following the optimal policy is precisely equal to the reservation wage (i.e., $g_\xi = \xi$).

In the above analysis we have only considered the case of sampling with recall wherein the optimal policy has the reservation wage property; in particular, the recall option is never utilized. Consequently, it is clear that it makes no difference whether or not recall is permitted. But if, for example, the marginal cost c of search were rising with the passage of time or the searcher were risk averse or the number of search opportunities were finite, then the recall option would play a part, with offers that were previously unacceptable possibly later becoming acceptable.

2.3. *The impact of increasing uncertainty*

From Figure 2.1 it is obvious that the lower the cost of search the higher the reservation wage and the longer the duration of search. Less obvious is the impact upon the searcher's optimal policy associated with increasing the uncertainty in some facet of his environment. While appealing, the presumption that increased uncertainty always is detrimental to the searcher's welfare is unfounded. To show this we shall first consider a change in the offer distribution F and, second, let the number of job offers received per period be a random variable instead of the constant 1. (It is instructive to pause and reflect on which situation will be detrimental and which beneficial.)

To begin, let X and Z be non-negative random variables with cumulative distribution functions F and G, respectively, and suppose that the wage offer distribution G is riskier than F in the sense of second-order stochastic dominance. To ensure comparability, we assume that $\mathrm{E}(Z) = \mathrm{E}(X)$ so that the average offer tendered is not dependent upon whether the wage offer distribution is F or G.

Denote by H_F and H_G the H function associated with F and G, respectively, where H is defined in (2.5). Similarly, let ξ_F and ξ_G be the associated reservation wages, and consider the convex increasing function u_y defined by

$$\begin{aligned} u_y(x) &= 0, && x \leqslant y, \\ &= x - y, && x > y. \end{aligned} \tag{2.7}$$

Applying (1.6) to (2.7) and recalling (2.5) yields

$$H_G(y)=\mathrm{E}u_y(Z)\geqslant \mathrm{E}u_y(X)=H_F(y), \quad \text{all} \quad y\geqslant 0, \tag{2.8}$$

so that H_G lies above H_F. Coupling (2.4) and (2.8), we obtain

$$H_F(\xi_F)=c=H_G(\xi_G)\geqslant H_F(\xi_G), \tag{2.9}$$

so the decreasing nature of H_F yields

$$\xi_F\leqslant \xi_G. \tag{2.10}$$

Since the reservation wage equals the expected return from following the optimal policy, we have demonstrated that increasing the riskiness of the wage offer distribution is beneficial to the searcher. The explanation for this phenomenon turns on the fact that mean-preserving changes only in the tails of the distribution of F — i.e., on the interval $[0,\xi]$ or $[\xi,\infty)$ — have no impact whatsoever whereas a mean-preserving increase in the probability F places on the tail $[\xi,\infty)$ is strictly beneficial.

Next suppose that the number N_i of offers received on the ith day is a non-negative, integer-valued random variable and that the N_i are independent and possess the same distribution. To facilitate comparison with the standard case, i.e., $\mathrm{P}(N_i=1)=1$, we assume that $\mathrm{E}(N_i)=1$, so that the (expected) search cost per offer tendered remains c.

The searcher is allowed to consider all of the N_i offers before deciding whether or not to accept one of them. Naturally, if he decides to accept an offer, he accepts the best one among those received. Consequently, the distribution G of the best offer received when N_1 offers are tendered is given by

$$G(t)=\sum_{i=0}^{\infty} p_i[F(t)]^i, \tag{2.11}$$

where $p_i=\mathrm{P}(N_1=i)$.

Jensen's inequality states that $\mathrm{E}u(Z)\geqslant u(\mathrm{E}(Z))$ for any random variable Z and convex function u [provided $u(Z)$ is defined]. Consequently, applying Jensen's inequality with $u(i)=F(t)^i$, $i\geqslant 0$, and letting Z have distribution N_1 yields

$$G(t)\geqslant F(t), \quad \text{all} \quad t\geqslant 0. \tag{2.12}$$

Thus, the effective offer distribution G is stochastically smaller than the original

offer distribution F. Consequently,[15] $H_G \leqslant H_F$ so that $\xi_G \leqslant \xi_F$, and hence it is preferable to have exactly one offer per day rather than a random number with a mean of one per day. The intuitive explanation for this phenomenon is as follows: although the expected cost per observation is still c, the facts that (a) several acceptable offers (i.e., offers with $w > \xi$)can arrive on the same day and (b) the searcher can utilize but one acceptable offer implies that the cost per utilized acceptable offer has increased.

In summary, increasing the riskiness or variation of the offer distribution (while leaving its mean unchanged) is beneficial, whereas increasing the variation in the number of offers tendered per day (while leaving its mean unchanged) is detrimental.

2.4. *Martingales and the existence of a reservation wage*

In this section, we present a rigorous argument which demonstrates that (i) there is indeed an optimal policy and (ii) it has the form given in (2.1). In deriving these results we utilize several concepts and results from the theory of martingales and optimal stopping. To begin, consider an arbitrary sequence $X_1, X_2, \ldots$ of random variables, and for each n let Y_n be a random variable whose value is determined by the first n observations $X_1, X_2, \ldots, X_n$. The sequence $Y_1, Y_2, \ldots$ is said to be a *supermartingale* with respect to the sequence $X_1, X_2, \ldots$ if for each n, $\mathrm{E}(Y_n)$ exists and (with probability 1)

$$\mathrm{E}(Y_{n+1} | X_1, \ldots, X_n) \leqslant Y_n. \tag{2.13}$$

Moreover, if

$$\mathrm{E}(Y_N) \leqslant \mathrm{E}(Y_1), \tag{2.14}$$

for every stopping rule[16] N for which $\mathrm{E}(Y_N)$ exists, then the supermartingale is said to be *regular*. Applied to our model of job search, we take X_i to be the ith offer and $Y_i = \max\{X_1, X_2, \ldots, X_i\} - ic$.

Our objective is to choose a stopping rule N, termed "optimal", so as to make $\mathrm{E}(Y_N)$ as large as possible. This raises the question whether there exists an optimal stopping rule. Regardless of whether the sequence $Y_1, Y_2, \ldots$ forms a supermartingale, it can be shown [see DeGroot (1970, p. 347)] that among the

[15] This follows from (1.3).

[16] A non-negative integer-valued random variable N is said to be a "stopping rule" for the sequence $X_1, X_2, \ldots$ if the event $N = k$ depends only upon the observed values of $X_1, \ldots, X_k$ and not upon the (as yet unobserved) values of $X_{k+1}, X_{k+2}, \ldots$ In other words, the decision to stop is not based on knowledge of the future.

class of stopping rules that actually stop with probability 1 there is, in fact, an optimal stopping rule if both of the following two conditions hold:

$$\lim_{n\to\infty} Y_n = -\infty, \quad \text{with probability 1,} \tag{2.15}$$

and

$$\mathrm{E}(|Z|) < \infty, \tag{2.16}$$

where $Z \equiv \sup_n Y_n$. It is helpful to think of Z as the payoff one would receive if one possessed perfect foresight (in regard to the X_i's).

Assuming $\mathrm{E}(X_1^2) < \infty$ in the job search model, a direct application of the Borel–Cantelli lemma[17] establishes that (2.15) and (2.16) hold, so that there is an optimal stopping rule for the job search model. In addition, the assumption $\mathrm{E}(X_1^2) < \infty$ applied to the job search model implies

$$\mathrm{E}(Y_n^2) \leqslant M < \infty, \quad \text{for all } n, \tag{2.17}$$

so that the sequence $Y_1, Y_2, \ldots$ is uniformly integrable.[18] Our interest in (2.17) stems from the fact that a uniformly integrable supermartingale is regular.

The fruit of this detour is the following important result [due to Chow and Robbins (1961) and referred to as the "monotone case"]: Consider a stopping problem in which an optimal stopping rule exists. Suppose that for any set of observed values $X_1 = x_1, \ldots, X_n = x_n$ which satisfies

$$\mathrm{E}(Y_{n+1} | x_1, \ldots, x_n) \leqslant y_n, \tag{2.18}$$

the sequence $Y_{n+1}, Y_{n+2}, \ldots$ is a regular supermartingale with respect to the sequence $X_{n+1}, X_{n+2}, \ldots$. Then there is an optimal stopping rule which stops if (2.18) holds and continues otherwise.

[17] With any sequence $\langle A_i \rangle$ of events we associate the event $\bar{A}$ defined by

$$\bar{A} = \bigcap_{j=1}^{\infty} \bigcup_{i=j}^{\infty} A_i.$$

One often encounters the notation $\{A_i, \text{i.o.}\}$ in place of $\bar{A}$ as $\bar{A}$ is the occurrence of an infinite number (i.o. stands for infinitely often) of the A_i. The Borel–Cantelli lemma asserts that if the $\langle A_i \rangle$ are independent, then $P(\bar{A})$ is zero or one according to whether $\Sigma_{i=1}^{\infty} \mathrm{P}(A_i)$ is finite or infinite.

[18] A sequence $Y_1, Y_2, \ldots$ of random variables is said to be uniformly integrable if and only if

$$\lim_{z\to\infty} \int_{\{|Y_n| > z\}} |y_n| \, \mathrm{d}P = 0, \quad \text{uniformly in } n.$$

In other words, if the hypotheses of this theorem hold, then the myopic stopping rule is optimal. Of course, the myopic stopping rule is the rule which stops if and only if the current return from stopping [i.e., the right-hand side of (2.18)] exceeds the expected value of stopping after taking exactly one more observation [the left-hand side of (2.18)].

In order to apply this theorem to our job search model, we need to verify all of its hypotheses. Assuming that $E(X_1^2)<\infty$, it only remains to verify that $Y_{n+1}, Y_{n+2},\ldots$ is a supermartingale whenever (2.18) holds. Toward this end, define ξ to be the unique solution of $H(y)-c=0$ so that $y\geqslant\xi$ implies $H(y)-c\leqslant 0$.

Suppose (2.18) holds for $X_1=x_1,\ldots,X_n=x_n$, and define $Z_n=Y_n+nc=\max\{X_1, X_2,\ldots, X_n\}$. Then we have

$$\begin{aligned}E(Y_{n+1}|X_1=x_1,\ldots,X_n=x_n)&=-(n+1)c+\int_0^\infty \max[z_n,x]\,dF(x)\\&=-(n+1)c+z_n+H(z_n)\\&=y_n+[H(z_n)-c],\end{aligned}\tag{2.19}$$

so that $H(z_n)-c\leqslant 0$ since (2.18) holds. Since H is a strictly decreasing function and Z_n increases in n, the desired inequalities (2.13) hold. Moreover, (2.19) reveals that (2.18) holds if and only if $z_n\geqslant\xi$.

2.5. *Variations of the basic model*

The elementary search model has been modified in numerous ways, including consideration of (i) quits and layoffs, (ii) finite time horizon, (iii) on-the-job-search, (iv) discounting, (v) risk aversion, (vi) unemployment insurance, and (vii) equilibrium models.[19] Among those modifications possessing the most attractive probabilistic content are search in a dynamic economy and adaptive search. We very briefly present these two modifications.

2.5.1. *Search in a dynamic economy*

Let $\{1,2,\ldots,K\}$, $K\leqslant+\infty$, be the set of states of the economy and denote the distribution function of wages associated with state i by F_i. The economy changes according to a discrete time Markov chain, with a one-step transition

[19]For discussions of these topics, the reader is referred to the following articles: (i) Mortensen (1978), Wilde (1979), Lippman and McCall (1981); (iii) Burdett (1978); (v) Danforth (1979), Hall, Lippman and McCall (1979); (vi) Marston (1975), Classen (1979); (vii) Kormendi (1979), Diamond and Maskin (1979), Wilde and Schwartz (1979).

matrix $P=(P_{ij})$, which is independent of the sequence of offers drawn from $\{F_i\}$.

The job offer available is the value of the random draw from last period's distribution function. If the searcher accepts an offer of x, the process terminates and the searcher is absorbed into the employment state for the n periods remaining in his working life. If the searcher rejects x, he must pay a fixed price c for a draw from the distribution F_i. The economy then moves to a new state j according to P_{ij}. After this transition to state j, the searcher must decide whether to accept or to reject the new offer y. The searcher is not allowed to retain rejected offers. The goal is to find a stopping rule which maximizes the β-discounted expected net benefits.

Let $V_n(i, x)$ denote the maximal β-discounted expected return attainable when n periods remain and the economy is in state i and the currently available job offer is x. Then $V_n(i, x)$ satisfies the recursive equation ($V_0 \equiv 0$)

$$V_{n+1}(i,x) = \max\left\{x, -c+\beta \sum_{j=1}^{K} P_{ij} \int_0^\infty V_n(j,y)\,\mathrm{d}F_i(y)\right\}$$
$$\equiv \max\{x, R_n(i)\}. \tag{2.20}$$

Clearly, $R_n(i)$ is the reservation wage rate when $n+1$ periods remain and the economy is in state i. Furthermore, it can be shown [see Lippman and McCall (1976a)] that these reservation wage rates satisfy

$$R_n(1) \leqslant R_n(2) \leqslant \cdots \leqslant R_n(K), \tag{2.21}$$

and

$$R_{n+1}(i) \geqslant R_n(i). \tag{2.22}$$

All that is necessary to ensure that (2.21) and (2.22) hold is to make the larger numbered states preferable. This is accomplished by assuming that

$$F_1(t) \geqslant F_2(t) \geqslant \cdots \geqslant F_K(t), \quad \text{all} \quad t \geqslant 0, \tag{2.23}$$

(i.e., the distribution functions $\{F_i\}$ are stochastically increasing) and

$$\sum_{j=k}^{K} P_{ij} \text{ is non-decreasing in } i \text{ for each fixed } k. \tag{2.24}$$

2.5.2. Adaptive search

Perhaps the most restrictive and least palatable assumption of the elementary search model is the supposition that the offer distribution F is known. Instead, suppose that F is known except for the fact that the searcher is given a prior distribution on θ, one of the parameters of F. In this case a wage offer represents not only an employment opportunity but also a piece of information that is used to revise (in Bayesian fashion) the prior distribution.

As is true when the searcher is risk averse, the fact that F is not completely known can drastically alter the character of the optimal policy. For example, (a) even without recall there may be no reservation wage [see Rothschild (1974b)] and (b) with a finite horizon, the recall option might be used *before* reaching the end of the horizon and the optimal policy need not be myopic [see Rosenfield and Shapiro (1981)].

When recall is not allowed Rosenfield and Shapiro show the existence of reservation wages under conditions that require (roughly) that (in the sense of first-order stochastic dominance) higher offers lead the searcher to expect future offers to be higher, but not too much higher. Presumably distributions with an unknown mean-related parameter would satisfy this condition, and one example of this type is when F is normal with mean θ and θ is itself normal. Their conditions are also satisfied when F is exponential with unknown mean parameter θ and the prior on θ is exponential. Using completely different techniques, Rothschild (1974b) also shows the existence of a reservation wage when F is a multinomial distribution with Dirichlet prior. With recall, Rosenfield and Shapiro give a simple condition which ensures that the optimal policy is myopic. Again, examples involving multinomial, exponential, and normal distributions are given.

3. The economics of insurance

3.1. Introduction

The prevalence of risk aversion has led to a variety of institutional forms enabling individuals and firms to transfer risks among themselves.[20] The purpose of these arrangements is to reallocate risk away from individuals and firms whose livelihood is threatened by uncertainty to firms whose livelihood is based (via the law of large numbers) on the pooling of uncertainties. The most apparent and familiar of these transfers is the ordinary insurance policy. The

[20] Extensive discussions of optimal insurance policies under a variety of circumstances are contained in Borch (1968), Buhlmann (1970), and Seal (1969). Pioneering work on the economics of insurance was accomplished by Arrow (1963, 1971) and Borch (1960).

essence of insurance contracts is the payment of a fixed fee by the insuree in exchange for the insurer's promise to pay a certain amount of money provided a stipulated event occurs. This well-known contractual arrangement is only one of a multitude of devices that have been created for coping with the risks that afflict any economic system. These risks include not only fire, theft, sickness, and death but also fluctuating prices, equipment malfunctions, zero inventory levels causing unsatisfied demands, and failure of basic research ranging from falsely "proved" theorems to unisolated viruses. The existence of futures contracts permits the farmer or food processor to specialize in production, while the speculator specializes in risk bearing. The risk of equipment failure can be reduced by improved design and maintenance procedures like redundancy and frequent inspection. The probability of an unfulfilled demand can be diminished by maintenance of larger inventories. The costs of research failure are frequently insured against by initiation of a large number of relatively independent projects (self-insurance), or, where the costs are large and uncertain, by adoption of inefficient contractual procedures like the cost-plus, fixed-fee contract (government insurance).

The basic institution for shifting the risks of business from entrepreneurs to the general public is the securities market.[21] Individuals can diversify their portfolio of stocks to achieve an acceptable level of expected return for a given level of risk. This ability of individuals to spread risks thereby permits firms to engage in projects which otherwise would be unacceptable. Consequently, society is better off.

These insurance arrangements are, however, far from ideal. It is usually impossible for a firm to transfer *only* rights to the outcomes of its highly risky ventures. In contrast with the futures market, the stock market is usually incapable of separating production and risk, leaving the former to the entrepreneur and transferring the latter to the general public. Instead, the stock certificate is a relatively blunt instrument for disentangling risk and production. The fact that society has not created a sharper instrument attests to the refractory nature of this problem.

From this description it is clear that insurance is a phenomenon that permeates economic institutions. Indeed, it is our belief that the economics of insurance is *the* most important topic in the economics of uncertainty. Accordingly, a rather extensive treatment is in order. Because of space limitations, however, only succinct descriptions of three of the most significant problems in the economics of insurance — moral hazard, adverse selection, and equilibrium analysis of insurance markets — will be presented.

Fundamental to any risk transfer is its effect on the incentives of the insured. These incentive effects are commonly referred to as the *moral hazard problem*.

[21] For a theoretical discussion of this aspect of the securities market see Arrow (1971, ch. 4).

The problem is to design insurance contracts that share risk *and* preserve incentives. Its roots stem from the inability of the insurer to observe costlessly the actions of the insured. These actions together with the state of nature determine the outcome. Moral hazard can be diminished by requiring the insured to bear some of the costs of the contingency and/or by monitoring his behavior. Determining optimal incentive and monitoring contracts is a flourishing activity in the economics of insurance.[22]

Adverse selection is similar to moral hazard in that the problem arises because insurance companies do not have costless access to the information possessed by buyers and vice versa. For example, some purchasers of health insurance have much more information about their health status than the insurance companies. Because of imperfect information individuals who are quite different will be treated as if they were identical. Presumably, the reason for identical treatment is that the cost of separating individuals into homogeneous subgroups is "prohibitive". Akerlof (1970) illustrates this by "observing" that individuals over 65 have difficulty obtaining health insurance. The reason why price does not increase to cover the additional risk is that

> ... as the price level (sic) rises the people who insure themselves will be those who are increasingly certain that they will need insurance ... The result is that the average medical condition of insurance applicants deteriorates as the price level rises—with the result that no insurance sales may take place at any price.

Clearly, insurance companies are not as helpless as this example suggests. They can and do cope with this informational asymmetry by (a) experience rating (i.e., continually adjusting the premiums to reflect the size and incidence of the individual insuree's claims) and (b) designing policies so as to elicit the information necessary for partitioning buyers into distinct categories.[23]

Equilibrium analysis of insurance markets has produced many insights. For example, equilibrium in insurance markets is equivalent to the Arrow–Debreu contingent claims competitive equilibrium.[24] Furthermore, the implications of moral hazard and adverse selection can be correctly perceived only in an equilibrium setting.[25]

In the remainder of this section we present a simple framework for analyzing insurance. This is the two-state model that has illuminated many of insurance's attractive and fascinating features. In particular, we utilize this framework to

[22] See Alchian and Demsetz (1972), Harris and Raviv (1976, 1978), Jensen and Meckling (1976), Ross (1973, 1974), Shavell (1976), Spence and Zeckhauser (1971), Stiglitz (1974), and Wilson (1977).

[23] For discussions of adverse selection see Akerlof (1970) and Rothschild and Stiglitz (1976).

[24] See Kihlstrom and Pauly (1971).

[25] See Rothschild and Stiglitz (1976) and Shavell (1979).

investigate how the optimal amount of insurance responds to changes in both the risk aversion of the insured and the riskiness of his endowments.

3.2. *A two-state model*

For simplicity in exposition, assume that only one of two states of nature prevails. In state 1, the individual is endowed with an income (or, perhaps, wealth) of w, whereas his income in state 2, the disaster state, is y, with $y<w$. The probabilities of these states are $1-p$ and p, respectively.[26]

Before the state of nature is known the individual can guard against the low endowment in state 2 by purchasing insurance. The rate of exchange between state 1 and state 2 income is π, that is, an increase of s in state 2 income can be purchased by a reduction of πs in state 1 income. We refer to π as the cost of insurance. Note that the actuarially fair price of insurance (i.e., a shift of state 1 to state 2 income) is simply $p/(1-p)$, the odds that state 2 occurs. The most practical as well as interesting case is that for which $\pi > p/(1-p)$.

The individual possesses an increasing and strictly concave utility function u. His objective is to select that amount of insurance, termed optimal, so as to maximize the expected utility of his income. Thus, he seeks the optimal level s^* of insurance, where s^* satisfies

$$U(s^*) = \max_{s \geq 0} U(s), \tag{3.1}$$

and

$$U(s) = (1-p)u(w-\pi s) + pu(y+s). \tag{3.2}$$

The strict concavity of u induces U to be strictly concave. Consequently, insurance will be purchased (i.e., $s^* > 0$) if and only if $U'(0) > 0$. Assuming that the endowments w and y, the probability of disaster p, the cost of insurance π, and the utility function u satisfy $U'(0) > 0$, the strict concavity of U implies that s^* is the unique solution of the first-order equilibrium condition

$$\pi = \frac{p}{1-p} \frac{u'(y+s)}{u'(w-\pi s)}. \tag{3.3}$$

If $\pi = p/(1-p)$, then $y + s^* = w - \pi s^*$; that is, the individual has fully insured against risk, and, accordingly, he is completely indifferent between the occur-

[26] This two-state framework employed here has been utilized by Ehrlich and Becker (1972), Hirshleifer (1970), and Rothschild and Stiglitz (1976). The extensions presented here, as embodied in Theorems 1–4, are new and are further elaborated upon in Lippman and McCall (forthcoming).

rence of states 1 and 2. If $\pi < p/(1-p)$, then $y+s^* > w-\pi s^*$ and the individual therefore prefers disaster, whereas $y+s^* < w-\pi s^*$ and the individual eschews disaster if the more reasonable case $\pi > p/(1-p)$ obtains. Henceforth, we assume that $\pi > p/(1-p)$.

It is obvious from (3.3) that s^* increases as w increases and that s^* increases to $(w-y)/(\pi+1)$, the amount that renders the individual fully insured, as $\pi/[p/(1-p)]$ decreases to 1. Furthermore, s^* is unchanged if p itself is a random variable with known distribution. But how does s^* vary with changes in the utility function and changes in the endowments? And how will s^* change if the endowments themselves are random?

3.2.1. The effect of increased risk aversion

We can now give a qualitative answer to the question of how the optimal amount of insurance varies with changes in the utility function, namely, the more risk averse individual buys more insurance. To distinguish between the optimal levels of insurance for various utility functions and to indicate its dependence on the particular utility function v under consideration we shall replace s^* by s_v.

Theorem 1

If $r_u > r_v$, then $s_u > s_v$.

Proof

Since $r_u > r_v$, it follows [see Theorem 1.e or (20) of Pratt] that

$$u'(x)/u'(t) > v'(x)/v'(t) \quad \text{for} \quad x < t. \tag{3.4}$$

Because v' is strictly decreasing, we can assert that $v'(y+s)/v'(w-\pi s)$ increases (strictly) as s decreases. Coupling this fact with (3.3) and (3.4) yields the desired result as illustrated in Figure 3.1. Q.E.D.

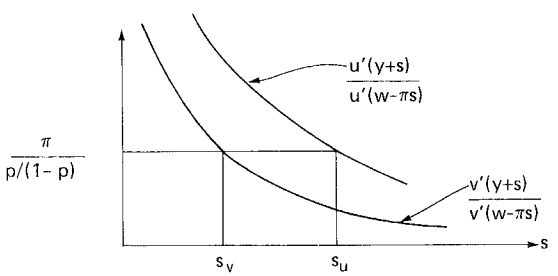


Figure 3.1

Theorem 1 tells us that s_v increases with r_v. However, it is difficult to display just how fast the level of insurance increases with the level of risk. Moreover, no matter how risk averse he is, the individual will not fully insure, i.e., $s^* < (w-y)/(\pi+1)$ as $\pi > p/(1-p)$. But are there individuals who come "close" to fully insuring? With the aid of the following example we give a precise answer to this question as well as provide some sense of how fast s_u increases with r_u.

Suppose an individual has constant risk aversion $\lambda > 0$, in which case his utility function u_λ is given by

$$u_\lambda(x) = a + b\mathrm{e}^{-\lambda x} \quad \text{for} \quad b < 0. \tag{3.5}$$

Of course $u'_\lambda(x) = |b|\lambda \mathrm{e}^{-\lambda x}$ and $r_{u_\lambda}(x) \equiv \lambda$. From (3.3) and (3.5) it follows that s_λ, the optimal level of insurance for an individual with constant risk aversion λ, is simply [assuming $U'(0) > 0$]

$$s_\lambda = \frac{w-y}{\pi+1} - \frac{1}{\lambda(\pi+1)} \ln \frac{\pi}{p/(1-p)}. \tag{3.6}$$

From (3.6) we see that no insurance is purchased for small values of λ, s_λ increases (as required by Theorem 1) with the rate of increase being inversely proportional to λ, and s_λ converges to $(w-y)/(\pi+1)$, the level at which the individual is fully insured. These observations are not only suggestive but also useful in establishing our next result which states that the optimal quantity of insurance purchased increases to the level at which an individual is fully insured against risk as the individual's aversion to risk increases without bound.

Theorem 2

Let $\langle u_i \rangle$ be a sequence of utility functions such that $\langle r_{u_i} \rangle$ increases (uniformly on the interval $[y, w]$) without bound. Then s_{u_i} converges to $(w-y)/(\pi+1)$.

Proof

By hypothesis, the sequence $\langle \eta_i \rangle$ converges to ∞, where $\eta_i \equiv \inf_x r_{u_i}(x)$. Observing that $r_{u_i} \geqslant r_{v_i}$, where v_i is given by (3.5) with $\lambda = \eta_i$, we can conclude from Theorem 1 and (3.6) that $s_{u_i} \geqslant (w-y)/(\pi+1) - [\ln(\pi(1-p)/p]/\eta_i$. Hence $s_{u_i} \to (w-y)/(\pi+1)$. Q.E.D.

3.2.2. The effect of increased wealth

Denote by s_w the optimal quantity of insurance, a function of the individual's state 1 endowment w, with u, p, y, and π fixed. For ease in presentation we shall assume that $w \geqslant y$ and $\hat{\pi} \equiv \pi/[p/(1-p)] \geqslant 1$ so that s_w is (eventually) strictly increasing. It was our presumption that s_w would be a strictly concave function.

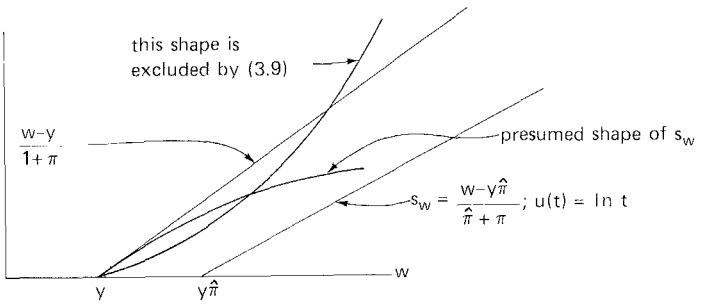


Figure 3.2. The shape of s_w.

This fails to be the case, particularly for the most familiar utility functions.

To begin, consider

$$f(s_w, w) = \frac{u'(y+s_w)}{u'(w-\pi s_w)} - \hat{\pi} = 0. \tag{3.7}$$

Applying the implicit function theorem to (3.7) yields ($r \equiv r_u$)

$$s_w' = \left[\frac{r(y+s_w)}{r(w-\pi s_w)} + \pi \right]^{-1} > 0, \tag{3.8}$$

under the proviso that $s_w > 0$, from which we can conclude that (wherever $r_u' \leqslant 0$)

$$s_w' \leqslant 1/(1+\pi). \tag{3.9}$$

[Equality obtains if $\hat{\pi} = 1$ (in which case $y + s_w = w - \pi s_w$) or r is constant on some interval.[27]] The upper bound on s_w provided by (3.9) reveals that s_w cannot be strictly convex on $[y, \infty)$. All of this, including the demand function induced by logarithmic utility, is depicted in Figure 3.2. Differentiating (3.8), we obtain

$$\operatorname{sign} \frac{d^2 s_w}{dw^2} = \operatorname{sign} \left\{ \frac{r'(y+s_w)}{r'(w-\pi s_w)} - \left[\frac{r(y+s_w)}{r(w-\pi s_w)} \right]^2 \right\}. \tag{3.10}$$

[27] If $u(t) = -1 + \ln t$ for $t \leqslant 1$ and $u(t) = -(1-e^{-\lambda}) - e^{-\lambda t}$ for $t \geqslant 1$, then u is concave and $r(t) = \lambda$ for $t > 1$. From (3.8), we see that $s_w' = 1/(1+\pi)$ except when $w \leqslant y + (1/\lambda)\ln \hat{\pi}$, in which case it is 0.

Defining m_ε for $\varepsilon > 0$ by

$$m_\varepsilon(t) = \frac{r'(t)}{r'(t+\varepsilon)} - \left[\frac{r(t)}{r(t+\varepsilon)}\right]^2, \quad \text{all } t, \tag{3.11}$$

it is clear that (3.10) provides the following characterization of s_w.

Theorem 3

The demand s_w for insurance is concave or convex depending on whether $m_\varepsilon \leqslant 0$ for all $\varepsilon > 0$, or $m_\varepsilon \geqslant 0$ for all $\varepsilon > 0$.

Below we provide five examples. In the first three s_w is linear. This occurs for the cases of (a) constant absolute risk aversion, and (b) and (c) constant relative risk aversion. The linearity of s_w persists even when a is strictly positive in example (c) so that the relative risk aversion is decreasing. Concavity is obtained in example (d) and convexity (on a limited interval) in example (e), wherein decreasing and increasing absolute risk aversion are encountered, respectively.

Examples

(a) $u(t) = -e^{-\lambda t}$, $\lambda > 0$: $r(t) \equiv \lambda$, and $m_\varepsilon \equiv 0$.

(b) $u(t) = \ln t$: $r(t) = 1/t$, $r'(t) = -1/t^2$, and $m_\varepsilon \equiv 0$.

(c) $u(t) = -(a+t)^{-q}$, $a \geqslant 0$, $q > 0$: $r(t) = (q+1)/(a+t)$, $r'(t) = -(q+1)/(a+t)^2$, and $m_\varepsilon \equiv 0$.

(d) $u(t) = -e^{-\lambda t} + bt$, $\lambda > 0$, $b > 0$: $r(t) = (\lambda^2 e^{-\lambda t})/(\lambda e^{-\lambda t} + b)$, and $m_\varepsilon(t) < 0$ for all t and all $\varepsilon > 0$.

(e) $u(t) = \sin t$, $0 \leqslant t \leqslant \pi/2$: $r(t) = \sin t/\cos t$, and $r'(t) = 1/\cos^2 t$, so $m_\varepsilon(t) > 0$ for $\varepsilon > 0$ and $0 \leqslant t \leqslant \pi/2$.

These examples suggest that the demand for insurance is concave (including linear) whenever u exhibits decreasing absolute risk aversion.

3.2.3. *The effect of increased risk*

Finally, we consider the impact of replacing the endowment w by a random variable W. (As before, $w > y$.) We interpret W as wages and y as (the certain) death benefits. To begin the analysis, consider two individuals with respective endowment pairs (w, y) and (W, y), where $EW = w$, and denote by s_w and s_W the amount of insurance each of them purchases.[28] Naive intuition might suggest

[28] It is convenient and perhaps instructive, to interpret w or W as labor income and refer to the individuals with endowments (w, y) and (W, y) as the civil servant and the farmer, respectively.

that $s_W > s_w$, as (W, y) is riskier and therefore needs more protection. More reasonable, however, is the conjecture that $s_w > s_W$ based on the following argument: Since $u(w) > Eu(W)$, the individual with endowments (w, y) has more to protect against the possibility of disaster and thus it is he who needs more insurance.

Though quite appealing, this argument is incorrect. While $u(w) > Eu(W)$ does mean that he has more endowment to protect, it doesn't follow that taking more insurance protects him more than the (W, y) individual. Presumably, taking insurance protects both individuals. The crux of the matter revolves around the obvious question of which one of them benefits more (on the margin) from taking additional insurance.

By the strict concavity of u, if an amount s of insurance is purchased, then $u(w-\pi s) > Eu(W-\pi s)$. We interpret $-\pi s$ as the decrease in the wealth position of the two individuals. If the individuals' utility function exhibits decreasing risk aversion, then it would seem intuitive that since the second individual is the one who experiences uncertainty in state 1 (uncertainty that cannot be controlled), the random nature of $W-\pi s$ causes him to suffer more than the fellow with certain endowments as the effective wealth level in state 1 is reduced.

We are arguing that the diminution in state 1 utilities associated with purchasing an incremental unit of insurance satisfies

$$Eu(W-\pi s) - Eu(W-\pi(s+\varepsilon)) \geqslant u(w-\pi s) - u(w-\pi(s+\varepsilon)); \tag{3.12}$$

that is, insurance is more costly to the second individual. On the other hand, the increase in state 2 utility associated with purchasing an incremental unit of insurance is the same for both individuals. Thus, the logical consequence of (3.12) is $s_w > s_W$. Dividing by $-\pi\varepsilon$ and letting ε approach zero yields

$$Eu'(W-\pi s) \geqslant u'(w-\pi s). \tag{3.13}$$

As $w \succ_2 W$, (1.6) implies that (3.13), or equivalently, (3.12), holds if and only if u' is convex. Our next lemma states that if an individual exhibits decreasing absolute risk aversion, then his utility function does indeed have a convex derivative. (The converse of Lemma 1 is not true.)[29]

Lemma 1

If r_u is non-increasing and u''' exists, then $u''' > 0$; that is, u has a strictly convex derivative.

[29]Knowing that r_u is increasing does not enable us to conclude that $u''' \leqslant 0$. For example, if $u(x) = -(b-x)^c$ for $x < b$ and $c > 1$, then r_u is increasing, but $u''' > 0$ when $c > 2$ and $u''' < 0$ when $1 < c < 2$.

Proof

Since r_u is non-increasing, $r_u'(x) = -\{u'''(x)u'(x) - u''(x)^2\}/u'(x)^2 \leqslant 0$. Hence, $u'''(x) \geqslant u''(x)^2/u'(x) > 0$. Q.E.D.

We now formalize the above argument and generalize it to cover the case where both individuals have random state 1 endowments. Using Lemma 1, we can now show that an increase in the riskiness of the state 1 endowment leads to the purchase of less insurance if r_u is decreasing.

Theorem 4

Let s_W and s_Z be the optimal quantities of insurance for two individuals with endowments (W, y) and (Z, y). If $Z \succ_2 W$, then

$$s_Z \geqslant s_W \quad \text{if} \quad r_u \text{ is non-increasing,}^{30} \tag{3.14}$$

and

$$s_Z \leqslant s_W \quad \text{if} \quad u' \text{ is concave and } \mathrm{E}(Z) = \mathrm{E}(W). \tag{3.15}$$

Proof

If r_u is non-increasing, then u' is convex by Lemma 1 so that we can conclude from (1.5) that $\mathrm{E}u'(Z - \pi s) \leqslant \mathrm{E}u'(W - \pi s)$. Hence,

$$\frac{u'(y+s)}{\mathrm{E}u'(Z-\pi s)} \geqslant \frac{u'(y+s)}{\mathrm{E}u'(W-\pi s)}. \tag{3.16}$$

Coupling (3.3), (3.16), and the fact that $u'(y+s)/\mathrm{E}u'(W-\pi s)$ is strictly decreasing in s, yields $s_Z \geqslant s_W$. If u' is concave, then the inequality in (3.16) is reversed so that $s_Z \leqslant s_W$. The proviso $\mathrm{E}(Z) = \mathrm{E}(W)$ is needed to ensure $\mathrm{E}u'(Z - \pi s) \geqslant \mathrm{E}u'(W - \pi s)$ because u' is a decreasing function [see (1.6)]. Q.E.D.

4. Optimal consumption under uncertainty

4.1. *Introduction*

The amount of goods to consume is a daily decision confronting all economic agents. It is a sequential decision that must be made under conditions of

[30] If $F_Z \neq F_W$, then $\mathrm{E}u'(Z - \pi s) < \mathrm{E}u'(W - \pi s)$ by strict concavity of $-u'$, so that $s_Z > s_W$. Similarly, $F_Z \neq F_W$ implies $s_Z < s_W$ if $u''' < 0$ and $\mathrm{E}(Z) = \mathrm{E}(W)$.

uncertainty. Our concern is how uncertainty, combined with the sequential aspects of the problem, affects the consumer's allocation between immediate consumption and saving. Presumably the introduction of uncertainty does alter the allocation decision, but does it lead to an increase or a decrease in immediate consumption? One might argue that immediate consumption should increase as a hedge against the uncertain future as did Marshall (1920) who asserted that "the thriftlessness of early times was in great measure due to the want of security that those who made provision for the future would enjoy it: only those who were already wealthy were strong enough to hold what they had saved;..." On the other hand, a decrease in immediate consumption enhances future consumption prospects, or, in the words of Boulding (1966): "Other things being equal, we should expect a man with a safe job to save less than a man with an uncertain job".[31]

In his pioneering article Phelps (1962) developed an optimal sequential procedure for allocating initial wealth and a known income stream between daily consumption and investment in a single risky asset. The procedure is optimal in that it maximizes expected utility of lifetime consumption. [This model has been generalized by Hakansson (1970) to include investment in any number of risky assets and to permit borrowing and lending.] Phelps shows that for some utility functions the optimal level of immediate consumption decreases when uncertainty is introduced in the rate of return associated with saving, but it increases for other utility functions. Thus, neither the argument of Marshall nor Boulding's countervailing argument carries the unqualified force of logic; there is merit in both. Instead, we shall see that the direction of change in this allocation crucially depends on the third derivative of the utility function as well as the source of the uncertainty itself.

In Section 4.2 we present the two-period models that have received some considerable attention in the literature. These two-period consumption models fall into one of two classes, the division being based on the source of uncertainty. In the first class uncertainty is caused by stochastic labor income in the second period. In these labor income models, the rate of return on first period savings is assumed to be non-random. The problem is to choose first period's consumption so as to maximize the expected utility of the two-period consumption process. The second class of models focuses on the randomness inherent in the first period investment: the rate of return on capital is assumed to be a random variable.[32] The problem then is to allocate a fixed amount of capital between first period consumption and investment so as to maximize the expected

[31] See Sandmo (1970) for more complete recitations.

[32] The work of Drèze and Modigliani (1972), Leland (1968), and Sandmo (1970) falls into the first category while that of Kihlstrom and Mirman (1974), Mirman (1971), Rothschild and Stiglitz (1971), and Sandmo (1969) belongs in the second.

utility of the two-period consumption process. Thus, the objective as well as the control variable is the same for these two classes of problems.

As will be evidenced in the ensuing discussion, the models with capital risk are considerably simpler to analyze and yield sharper results.[33] Real difficulties, however, are encountered in extending these models to multiperiod settings. This is done in Section 4.3. However, it is only carried out under the assumption that the one-period utility functions exhibit constant relative risk aversion or u' is convex in the case of capital and income risk, respectively. In the former case, the infinite horizon model is precisely that of the two-period model. In the latter case our treatment is that of Miller (1976). We also present Levhari and Mirman's (1977) model in which labor income and return on investment are certain but the length of the decision maker's lifetime (i.e., the horizon length) is random.

4.2. *Two-period models*

Starting with an endowment w, the consumer selects an amount c to consume in period 1, so that his savings or investment is $w-c$. His second period endowment, which he consumes in its entirety, consists of his random labor income Y plus $r(w-c)$, the appreciated value of his investment, where $r-1$ is the certain rate of interest. The consumer's objective in this model of pure income risk is to maximize the expected value $V(w)$ of his utility for the two-period consumption stream subject to the constraint that $0 \leqslant c \leqslant w$. Thus, the problem is to determine the optimal level c^* of period 1 consumption; that is,

$$V(w) \equiv \max_{0 \leqslant c \leqslant w} \mathrm{E}U(c, Y + r(w-c)). \tag{4.1}$$

As is usual, we assume that the consumer's utility function U is concave increasing in each argument, i.e., $U_i = \partial U(c_1, c_2)/\partial c_i > 0$ and $U_{ii} = \partial^2 U(c_1, c_2)/\partial c_i^2 < 0$, $i = 1, 2$.

The first-order condition for solving (4.1) is (totally differentiating U with respect to c)[34]

$$\mathrm{E}U_1 = r\mathrm{E}U_2. \tag{4.2}$$

If $Y = \mathrm{E}Y$, then (4.2) reduces to the familiar condition

$$U_1 = rU_2. \tag{4.3}$$

[33] This is attributable to the fact that the source of randomness is additive in the state variable for the models with income risk and multiplicative for those with capital risk.

[34] If U is additive, it is readily apparent from (4.2) that $0 < \mathrm{d}c^*/\mathrm{d}w < 1$; see the discussion preceding footnote 40.

For ease of reference, let c_d be the optimal consumption level [which solves (4.3)] for this deterministic version of the problem. Utilizing (4.3), expanding EU_1 and EU_2 in a Taylor series around $\langle c_d, EY+r(w-c_d)\rangle$, and ignoring terms of the form $\partial^k U_i/\partial^k c_2 E(Y-E(Y))^k/k!$ for $k>2$ [presumably they will be $o(\sigma^2)$, where $\sigma^2=E\{(Y-E(Y))^2\}$], we obtain

$$\begin{aligned} EU_1 - rEU_2 &= U_1 + U_{122}\sigma^2/2 - r\left[U_2 + U_{222}\sigma^2/2\right] \\ &= \frac{\sigma^2}{2}\left[U_{122} - rU_{222}\right] \\ &= \frac{\sigma^2}{2}\left[U_{122} - \frac{U_1}{U_2}U_{222}\right] \equiv \frac{\sigma^2}{2}\mathfrak{A}. \end{aligned} \tag{4.4}$$

Assuming an interior solution, $0<c_d<w$, $(d^2/dc^2)U|_{c_d}<0$, implying that

$$\frac{d}{dc}\{EU_1 - rEU_2\}|_{c_d} = \frac{d^2}{dc^2}EU|_{c_d} < 0, \tag{4.5}$$

as $(d^2/dc^2)EU|_{c_d}=(d^2/dc^2)U|_{c_d}+O(\sigma^2)$. Consequently, if (4.4) is negative, (4.5) implies that $c^*<c_d$ when Y entails sufficiently small risks.[35]

As revealed above, we can ensure that the precautionary demand for savings is positive (i.e., $c^*<c_d$) when the uncertainty of the labor income Y is small if the quantity $\mathfrak{A}$ is negative.[36] Now if U has decreasing absolute risk aversion, then $U_{iii}>0$; furthermore, if U is additive (separable), then its cross partials are zero, yielding $\mathfrak{A}<0$ as desired. Improving upon this result, Sandmo (1970) demonstrated that if period 2 labor income is $\alpha Y+(1-\alpha)E(Y)$, $0\leqslant\alpha\leqslant k$, where $-E(Y)(1-k)/k$ is the minimal level of period 2 income, then c_α, the optimal level of consumption, is a decreasing function of α whenever U is separable and exhibits decreasing absolute risk aversion.

Leland (1968) also provides an intuitively appealing extension of decreasing absolute risk aversion that ensures $\mathfrak{A}<0$ when U is not separable. He assumes that the consumer (with utility function U) becomes less averse to risk in a given variable as that variable becomes increasingly predominant in a constant utility bundle.[37]

Next we turn our attention to a variant of the model with pure capital risk rather than pure (labor) income risk. Specifically, the individual receives no

[35] Replace Y by $EY+\varepsilon Z$, were $P(Z=1)=P(Z=-1)=1/2$. Then $c_\varepsilon<c_d$ for all ε sufficiently small.

[36] If U is quadratic, all derivatives of order higher than two are zero so that the Taylor series is exact and $c^*=c_d$ regardless of the distribution of Y. Thus, mere risk avoidance does not produce the precautionary demand for savings.

[37] This is equivalent to $(dU_{ii}/dc_i)|_{U_{const.}}>0$ which, upon expansion, yields $U_{222}-(U_2/U_1)U_{221}>0$ for $i=2$, as U_2/U_1 is the marginal rate of substitution of c_2 for c_1. Also see Sandmo (1970).

labor income but rather invests the difference between his initial wealth w and c, his period 1 consumption. The rate of return $X-1 \geqslant 0$ on his investment or saving is random so that the amount available for period 2 consumption is $(w-c)X$.[38] For simplicity we assume that the two-period utility function is additive,[39] so the consumer's problem is to find the optimal level c_X of consumption so as to

$$\max_{0 \leqslant c \leqslant w} \{u(c)+\mathrm{E}v[X(w-c)]\}, \tag{4.6}$$

where we use u and v to represent the strictly concave utility functions for consumption in periods 1 and 2, respectively.

In analogy to the model with pure income risk, the first-order condition for solving (4.6) is simply

$$u'(c)=\mathrm{E}\{Xv'[X(w-c)]\}. \tag{4.7}$$

In constrast to that model, however, the presence of pure capital risk enables us to delineate conditions on v such that increasing the riskiness of X causes c_X, the optimal level of period 1 consumption, to decrease — without restriction upon the amount of uncertainty (riskiness) embodied in the random variable X. We hasten to note that when we say that X is riskier or more uncertain than Z we mean that $Z \succ_2 X$ [see (1.4)].

Theorem 1

Suppose X is riskier than Z and $\mathrm{E}(X)=\mathrm{E}(Z)$. Then $c_X \leqslant c_Z$ if f is convex and $c_X \geqslant c_Z$ if f is concave, where

$$f(t)=tv'(t) \quad \text{for} \quad t \geqslant 0. \tag{4.8}$$

Proof

To begin, assume that f is convex. Because c is constrained to be less than or equal to w, $w-c \geqslant 0$ and thus $X(w-c)$ is riskier than $Z(w-c)$. Consequently, we

[38] Mirman (1971) generalizes this by assuming a concave production function f, so that $Xf(w-c)$ is consumed in period 2. Of course, a deterministic period 2 labor income could be appended without disturbing the analysis.

[39] Our analysis is essentially that contained in Mirman (1971) and in Rothschild and Stiglitz (1971). The more general non-additive case is analyzed by Sandmo (1969) and by Kihlstrom and Mirman (1974).

have

$$\begin{aligned} \mathrm{E}\{Xv'(X(w-c))\} &= \frac{1}{w-c}\mathrm{E}f(X(w-c)) \\ &\geqslant \frac{1}{w-c}\mathrm{E}f(Z(w-c)) \\ &= \mathrm{E}\{Zv'(Z(w-c))\}, \end{aligned} \tag{4.9}$$

where the equalities follow from the definition of f and the inequality from (1.6), $Z \succ_2 X$, and the convexity of f. Now v' is a decreasing function (for $v'' < 0$) so that if c_X where greater than c_Z it would follow from (4.7) and (4.9) that

$$\begin{aligned} u'(c_X) &= \mathrm{E}\{Xv'(X(w-c_X))\} \\ &\geqslant \mathrm{E}\{Zv'(Z(w-c_X))\} \\ &> \mathrm{E}\{Zv'(Z(w-c_Z))\} \\ &= u'(c_Z). \end{aligned}$$

But this contradicts the fact that u is concave. Hence, we must have $c_X \leqslant c_Z$ as desired. The same argument applies when f is concave. Q.E.D.

An immediate consequence of Theorem 1 is that the uncertainty in the rate of return vis-à-vis a known rate of return leads to a decrease in period 1 consumption—or equivalently an increase in saving—provided that f is convex.

Here, as in the model with pure income risk, the positivity of the third derivative of the utility function also plays a crucial role. Since $f''(c) = 2v''(c) + cv'''(c)$, it is clear that a necessary, though not sufficient, condition for the convexity of f is that the third derivative of v be positive. (Of course, a negative third derivative is sufficient to ensure the concavity of f.) Now the isoelastic utility functions, i.e., the ones with constant relative risk aversion, $v(c) = c^\gamma/\gamma$, $\gamma < 1$ and $\gamma \neq 0$, all have positive third derivatives, but f is convex when $\gamma < 0$ and concave when $\gamma > 0$, as $f''(c) = \gamma(\gamma-1)c^{\gamma-2}$. [On the boundary (i.e., $\gamma = 0$) we set $v(c) = \ln c$. Here, $f'' \equiv 0$ so that f is both convex and concave so that c_X depends on X only through $\mathrm{E}(X)$; in fact, $c_X = w\mathrm{E}(X)/(E(X)+\beta)$ when $u = \ln$ and $v = \beta \ln$, with β being the discount factor.]

Finally note that $0 < \mathrm{d}c_X/\mathrm{d}w < 1$. To see this notice that an increase in w causes the right-hand side of (4.7) to decrease while the left-hand side is unchanged; an increase in c causes the left-hand side to decrease and the right-hand side to increase. Hence, $\mathrm{d}c_X/\mathrm{d}w > 0$. If $\mathrm{d}c_X/\mathrm{d}w \geqslant 1$, then $X(w-c_X)$ decreases with w, thereby increasing the right-hand side of (4.7), whereas

$dc_X/dw>0$ implies that the left-hand side decreases with w. This violates (4.7) so we must have $dc_X/dw<1$.[40]

4.3. *Multiperiod models*

In multiperiod models of consumption, analytical tractability is only[41] obtained in exchange for the somewhat restrictive assumptions placed upon the multiperiod utility function U of the consumption stream $\boldsymbol{c}=\langle c_i\rangle_{i=1}^{\infty}$. In all cases we assume that U is additive and that U is of the form

$$U(\boldsymbol{c})=\sum_{i=1}^{\infty}\beta^{i-1}u(c_i), \tag{4.10}$$

where β is the discount factor, $0<\beta<1$, and the one-period utility function u is, as always, concave. Furthermore, we assume that all pertinent random variables are independent.

If, in addition to (4.10) and the independence of the random variables, we assume that the one-period utility function has constant relative risk aversion, i.e., u is of the form

$$\begin{aligned} u(c)&=c^{\gamma}/\gamma, \qquad \gamma\neq 0, \quad \gamma<1,\\ &=\ln c, \qquad \gamma=0, \end{aligned} \tag{4.11}$$

then, as noted by Mirman (1971), the two-period and the multiperiod models with pure capital risk are equivalent; as before increasing risk causes consumption to decrease when $\gamma<0$ and increase when $\gamma>0$.

In his multiperiod models with pure income risk, Miller (1976) considers a slightly broader class of one-period utility functions than that represented by (4.11). In his model, u is restricted to the class $\mathcal{F}$, where

$$\mathcal{F}=\{u: u'>0 \text{ and } u' \text{ is convex}\}. \tag{4.12}$$

Thus, $u'''>0$ is required and, as will be stated, consumption decreases as risk

[40] It is also true that dc^*/dw lies strictly between 0 and 1 in the case of pure income risk provided that

$$r(d/dc)U_2=r[U_{21}-rU_{22}]>U_{11}-rU_{12}=(d/dc)U_1.$$

Of course, this condition holds whenever U is additive.

[41] Merton's (1971) continuous time model is an exception to this rule. There he uses an exponential utility function, but the income stream is a Poisson process.

increases if $u \in \mathfrak{F}$. Since $u''' > 0$ if u is as given in (4.11), the direction of the change in the optimal level of consumption does not depend upon the sign of γ. This difference rather sharply delineates the distinction between the two models.

Let Y_j be the non-negative income received at the end of period j, so $\{Y_j\}$ are independent. For simplicity in exposition, assume that the Y_j's are identically distributed and that borrowing is not allowed. The latter assumption implies that current consumption c is constrained to lie between 0 and w, the current level of wealth. Denote the certain return on investment by $r-1$ so current wealth w in conjunction with current consumption c yields $Y+r(w-c)$ as next period's wealth.

The assumption $r^{\gamma}\beta < 1$ must be added to guarantee that the utility of the optimal consumption stream remains appropriately bounded. The analysis begins by demonstrating that the operator A is a contraction mapping[42] on the space $\mathcal{V}$ of functions, where A is defined by

$$Av(w) = \sup_{0 \leqslant c \leqslant w} \{u(c) + \beta \mathrm{E} v[Y + r(w-c)]\}, \qquad w > 0, \quad v \in \mathcal{V}, \tag{4.13}$$

and $\mathcal{V}$ is the (Banach) space of continuous, increasing, concave functions with domain $(0, \infty)$ which, for technical reasons, satisfy the boundedness condition

$$\|v\| \equiv \sup_{w>0} |v(w)| / \max(|u(w)|, 1) < \infty. \tag{4.14}$$

Knowing that A is a contraction enables us to conclude, among other things, that A has a unique fixed point (in $\mathcal{V}$) so that $V(w)$, the utility from following an optimal consumption stream when initial wealth is w, is the unique solution to

$$V(w) = \sup_{0 \leqslant c \leqslant w} \{u(c) + \beta \mathrm{E} V(Y + r(w-c))\}, \qquad w > 0, \tag{4.15}$$

and $c_Y(w)$, the optimal level of consumption when wealth is w and Y is the random labor income, is the unique value of c which attains the maximum in (4.15). Moreover, since V is the fixed point of A, V is concave. The concavity of u and V enable us to show, employing the same proof used at the end of Section 4.2, that $c_Y(w)$ and $w - c_Y(w)$ are increasing functions of w, that is, $0 < \mathrm{d}c_Y(w)/\mathrm{d}w < 1$.

[42] The operator A is a contraction on $\mathcal{V}$ if there is a number $\alpha < 1$ such that

$$\|Av_1 - Av_2\| \leqslant \alpha \|v_1 - v_2\|, \quad \text{all} \quad v_1, v_2 \in \mathcal{V},$$

where $\|\cdot\|$ is the norm on $\mathcal{V}$. Here, $\|\cdot\|$ is as given in (4.14).

The concavity of V also permits a simple demonstration of the fact that a more risky income[43] Z is less desirable than Y in that $V_Z(w) \leqslant V_Y(w)$ for all w, where we have subscripted V and A to distinguish between the return associated with Z and Y. To show this, note that for any c, $0 \leqslant c \leqslant w$, the fact that Z is more risky than Y (i.e., $Y \succ_2 Z$) implies that

$$u(c)+\beta \mathrm{E}V_Y(Z+r(w-c)) \leqslant u(c)+\beta \mathrm{E}V_Y(Y+r(w-c)), \tag{4.16}$$

as V_Y is a concave increasing function. Now maximizing the left-hand side of (4.16) over c yields

$$A_Z V_Y(w) \leqslant u(c_Z(w))+\beta \mathrm{E}V_Y(Y+r(w-c_Z(w))) \leqslant V_Y(w)$$

or simply

$$A_Z V_Y \leqslant V_Y. \tag{4.17}$$

Because A_Z is monotone (i.e., $v \leqslant \hat{v}$ implies $A_Z v \leqslant A_Z \hat{v}$), (4.17) can be iterated to obtain

$$A_Z^{n+1} V_Y \leqslant A_Z^n V_Y \leqslant \cdots \leqslant V_Y. \tag{4.18}$$

Finally, the fact that A_Z is a contraction means that the method of successive approximations can be employed to find its fixed point; that is

$$\lim_{n\to\infty} A_Z^n v = V_Z \quad \text{for all} \quad v \in \mathcal{V}. \tag{4.19}$$

In particular, utilizing (4.19) with $v = V_Y$ in conjunction with (4.18) yields

$$V_Z = \lim_{n\to\infty} A_Z^n V_Y \leqslant V_Y,$$

as desired.

Finally, by demonstrating that $u \in \mathcal{F}$ implies that $V \in \mathcal{F}$, Miller (1976) verifies that $c_Z(w) \leqslant c_Y(w)$ whenever Z is riskier than Y.

Rather than having income or capital risk be the source of the uncertainty, Levhari and Mirman (1977) consider the case wherein uncertainty arises as a result of the consumer's unknown lifespan. Let p_i be the probability that he lives exactly i years, $i = 0, 1, \ldots, n$, so that n is his maximum lifespan. We assume that the consumption decision is made at the beginning of each time period prior to discovering whether he will survive beyond the current period; moreover, we attach no value to bequests. For simplicity, assume that there is no labor

[43] Because V_Y is an increasing function we can employ (1.5) and we need not assume $\mathrm{E}(Z) = \mathrm{E}(Y)$.

income. Denote the certain return on investment by $r-1$ and define P_i to be the probability that he lives at least i years so that

$$P_i = \sum_{j=i}^{n} p_j. \tag{4.20}$$

Letting $V_i(w)$ be the maximum expected present value of utility of consumption in periods i through n inclusive when wealth at the beginning of period i is w, we have

$$V_i(w) = \max_{0 \leq c \leq w} \{P_i u(c) + \beta V_{i+1}(r(w-c))\}, \qquad i=0,1,\ldots,n-1, \tag{4.21}$$

and

$$V_n(w) = P_n u(w). \tag{4.22}$$

Clearly the value $c_i(w)$ of c which attains the maximum is the optimal amount to consume in period i if wealth at the beginning of period i is w.

As stated earlier, we will restrict our attention to one-period utility functions that satisfy (4.11). When $\gamma \neq 1$ it can be shown via induction that there are constants K and f such that[44]

$$V_i(w) = u(w)K_i, \tag{4.23}$$

and

$$c_i(w) = wf_i. \tag{4.24}$$

The optimal proportion f_i to consume in period i satisfies

$$f_i = \frac{P_i^{1/(1-\gamma)}}{P_i^{1/(1-\gamma)} + \alpha k_i},$$

where $\alpha = (\beta r^{\gamma})^{1/(1-\gamma)}$ and $k_i = [P_i f_i^{\gamma} + P_{i+1}\alpha^{1-\gamma}(1-f_i)^{\gamma}]^{1/(1-\gamma)} = K_i^{1/(1-\gamma)}$.

The analog of (4.23) and (4.24) for the simpler case $\gamma = 1$ is

$$V_i(w) = \sum_{j=0}^{n-i} P_{i+j}\beta^j \ln w + K_i \tag{4.25}$$

[44] Equation (4.24) differs from the one given on page 270 of Levhari and Mirman (1977).

and

$$c_i(w) = w \frac{P_i}{\sum_{j=0}^{n-i} P_{i+j}\beta^j}. \tag{4.26}$$

Now consider two individuals with different random lifetimes X and Y, $\mathrm{P}(X=i) = p_i$, $\mathrm{P}(Y=i) = q_i$, $P_i = \sum_{j=i}^{n} p_j$, and $Q_i = \sum_{j=i}^{n} q_j$. We will compare their consumption in period 0 when Y is seen as riskier.

Theorem 2

Suppose $Y \neq X$ and Y is riskier than X so that

$$\sum_{i=0}^{j} P_i \geqslant \sum_{i=0}^{j} Q_i, \qquad j=0,1,\ldots,n. \tag{4.27}$$

Then the individual with lifetime Y will consume strictly more in the initial period, without regard to whether $\mathrm{E}(X) > \mathrm{E}(Y)$ or $\mathrm{E}(X) = \mathrm{E}(Y)$.[45]

Proof

Let $\langle \delta_i \rangle_0^n$ and $\langle \beta_i \rangle_0^n$ be two sequences of numbers with the property that

$$\Delta_i = \sum_{j=0}^{i} \delta_j \geqslant 0, \qquad i=0,1,\ldots,n \quad (\Delta_{-1}=0)$$

and

$$\beta_i \geqslant \beta_{i+1} \geqslant 0, \qquad i=0,1,\ldots,n-1.$$

Summing by parts yields

$$\begin{aligned}\sum_{i=1}^{n} \delta_i\beta_i &= \sum_{i=0}^{n} (\Delta_i - \Delta_{i-1})\beta_i \\ &= \sum_{i=0}^{n} \Delta_i\beta_i - \sum_{i=0}^{n-1} \Delta_i\beta_{i+1} \\ &= \Delta_n\beta_n + \sum_{i=0}^{n-1} \Delta_i(\beta_i - \beta_{i+1}) \geqslant 0.\end{aligned}$$

[45] Theorem 2 would be a special case of a result of Levhari and Mirman (1977), but we do not require $\mathrm{E}(X) = \mathrm{E}(Y)$.

Setting $\delta_i = P_i - Q_i$ and $\beta_i = \beta^{i-1}$, we obtain

$$\sum_{i=1}^{n} P_i \beta^{i-1} \geqslant \sum_{i=1}^{n} Q_i \beta^{i-1}. \tag{4.28}$$

Combining (4.26) and (4.28) yields the desired result. [Note that $X \neq Y$ ensures that at least one of the inequalities in (4.27) is strict, in which case (4.28) is strict.] Q.E.D.

We hasten to point out that if the discount factor were one, the initial, consumption would then depend on X only through its mean, for it would be proportional to $1/(\mathrm{E}(X)+1)$.

5. Competitive firm under uncertainty

5.1. Introduction

Our initial intention was to effect a comprehensive presentation of the recent literature on production under uncertainty. However, the enormity of this literature gave us pause and led us to pursue a less ambitious objective. This is reflected in this section by restricting attention to the behavior of competitive firms in an uncertain environment.[46] In particular, neither the production literature pertaining to regulation nor the literature which explicitly incorporates financial markets is treated.[47] Furthermore, within this class of problems, we focus on single-period models.[48] In this setting it is shown that uncertainty matters even when firms are risk neutral. We view this as a crucial result for probabilistic economics. It would be most unfortunate if the vitality of the stochastic theory of the firm relied solely on the controversial assumption of risk aversion.

For a reminder, we begin with an analysis of risk averse firms producing in a competitive single-period setting. The optimal output decision is compared with the classical risk indifferent regime, and it is shown that when combined with risk aversion uncertainty causes output to be reduced. Moreover, the greater the aversion to risk, in the sense of Arrow and Pratt, the greater the reduction in output. On the other hand, new results reveal that the change in output induced

[46]Other aspects of production under uncertainty are addressed in Section 7.2 of this chapter and in Section VII.5 of Lippman and McCall (1979).

[47]For a treatment of these two topics the reader is directed to Baron and Taggart (1977), Myers (1973), and Leland (1974a) and the references therein, and to Diamond (1967), Leland (1974b), and Baron and Forsythe (1979).

[48]This theory has been developed by Baron (1970, 1971, 1973), Horowitz (1970), Leland (1972), McCall (1967), Mills (1959, 1962), Penner (1967), Sandmo (1971), and Sheshinski and Dreze (1976). The multiperiod case has been investigated by Zabel (1967, 1971). A survey of the earlier literature on stochastic production can be found in McCall (1971).

by a mean preserving increase in risk depends upon the shape of the cost curve as well as the sign of the third derivative of the utility function. Unlike the certainty case, the effect of increasing the fixed cost or proportional profits tax rate will have an impact on output, and the direction of the impact varies inversely with the absolute and relative risk aversion, respectively.

The implications of risk aversion for the factor market are then presented. This is followed by an evaluation of "optimal" competitive firm behavior when safety-first criteria are employed instead of expected utility maximization. Finally, we present an analysis of competitive industry behavior under uncertainty. In order to isolate the effect of uncertainty, firms are assumed to be risk neutral. The main results are: expected output for each firm is less than the output that minimizes industry average cost, competitive equilibrium is efficient, and the number of firms in the industry increases as either the mean or variance of output increases.

5.2. *Competitive production for risk averse firms*

In this subsection we determine the optimal output for a competitive firm with risk aversion. The price $P \geqslant 0$ is a non-degenerate random variable (with known distribution function F and mean μ), and the firm has no control whatsoever over it. Accordingly, as this is a competitive environment, it sells all of its output q at the going price p. The value p of P is made known after the firm has decided upon its production quantity q.[49] Ours is a one-period model, so no storage is permitted from one selling period to the next.

The relationship between the firm's profit π and its output q is given by

$$\pi(q) = Pq - C(q), \tag{5.1}$$

where $C(q)$, the total cost of producing q units of the product, consists of a fixed cost B and a variable cost $c(q)$. Naturally C is an increasing function. Because we assume that the marginal cost of production is non-decreasing (so C' is positive and non-decreasing), π is a concave function of q. The firm has a strictly concave utility function u and it seeks q to maximize the expected utility U of profits. Thus, the firm seeks the optimal production level q^* where q^* satisfies[50]

$$U(q^*) = \max_{q \geqslant 0} U(q), \tag{5.2}$$

[49]See Leland (1972) for a discussion of the case in which the firm controls the price but not the quantity sold.

[50]If $\mathrm{E}\pi(q) \leqslant -B$ all $q \geqslant 0$, then by Jensen's inequality

$$U(q) = \mathrm{E}u(\pi(q)) < u(\mathrm{E}\pi(q)) \leqslant u(-B) = U(0),$$

so that $q^* = 0$. To avoid trivialities, assume $\mathrm{E}\pi(q) > -B$ for some q. Because C' is non-decreasing, this means that $\mathrm{E}P > C'(0)$.

and

$$U(q)=\mathrm{E}u(\pi(q)). \tag{5.3}$$

Throughout this section we shall make use of the following result:

Lemma 1

If $h>0$ is a decreasing function and Z is a non-degenerate random variable, then

$$\mathrm{E}\{Zh(Z)\}<h(0)\mathrm{E}(Z).$$

Proof

$$\begin{aligned}\mathrm{E}\{Zh(Z)\}&=\int_{-\infty}^{0} th(t)\,\mathrm{d}F_Z(t)+\int_{0}^{\infty} th(t)\,\mathrm{d}F_Z(t)\\&<h(0)\left[\int_{-\infty}^{0} t\,\mathrm{d}F_Z(t)+\int_{0}^{\infty} t\,\mathrm{d}F_Z(t)\right]\\&=h(0)\mathrm{E}(Z). \qquad \text{Q.E.D.}\end{aligned}$$

The first-order condition is (the concavity of u and π guarantees that of U)

$$\mathrm{E}\{u'(\pi(q))P\}=C'(q)\mathrm{E}u'(\pi(q)). \tag{5.4}$$

As u' is a decreasing function and π increases in P, it follows from Lemma 1 that $\mathrm{E}\{u'(\pi)(P-\mu)\}<0$. Upon subtracting $\mu\mathrm{E}u'(\pi)$ from both sides of (5.4), this inequality asserts that

$$(C'(q)-\mu)\mathrm{E}u'(\pi)=\mathrm{E}\{u'(\pi)(P-\mu)\}<0,$$

from which the positivity of $\mathrm{E}u'(\pi)$ implies that

$$C'(q^*)<\mu. \tag{5.5}$$

In view of (5.5), C' non-decreasing, and the fact that the optimal output either for a risk neutral firm or deterministic demand has marginal cost equal to price, it follows that *uncertainty combined with risk aversion leads to decreased output*. (If the firm were risk preferent, i.e., u is convex, then the same argument reveals that output increases.)

Other issues that we would like to address are: (a) How do (mean preserving) increases in the riskiness of P (again in the sense of second-order stochastic dominance) affect output? (b) What is the impact on output associated with an

increase in the individual's aversion to risk? (c) Will changes in the fixed cost B or the introduction of a profits tax cause output to vary?

5.2.1. *Impact of an increase in risk*

Because a mean preserving increase in the riskiness of P leaves μ unchanged, it comes as no surprise that such a change has an impact upon q^*. However, the direction of the change is unclear; in fact, increases and decreases in output are both possible. We begin by verifying the anticipated result that an increase in risk can lead to a decrease in production. As in our discussion of optimal consumption strategies, the function f defined by

$$f(t)=tu'(t), \qquad t\geqslant 0, \tag{5.6}$$

as well as the sign of u''' play a role; in addition, the presence of a fixed cost enters.

Theorem 1

Let q_i denote the optimal output when the price P has the same distribution as the random variable P_i, $i=1,2$, and suppose that P_1 is strictly riskier than P_2 with $\mathrm{E}(P_1)=\mathrm{E}(P_2)$. Output decreases (i.e., $q_1<q_2$) if f is concave and either u' is convex and marginal cost exceeds average cost or u' is concave and average cost exceeds marginal cost. Output increases (i.e., $q_1>q_2$) if f is convex and either u' is convex and average cost exceeds marginal cost or u' is concave and marginal cost exceeds average cost.[51]

Proof

Let U_i play the role of U in (5.3) when $P=P_i$ and write E_i to indicate that the expectation is with respect to P_i, $i=1,2$. By hypothesis $\mathrm{E}P_1=\mathrm{E}P_2$ and f is concave, so applying (1.6) to f yields

$$\begin{aligned} U_1'(q)-U_2'(q) &= \mathrm{E}_1\{Pu'(\pi)\}-\mathrm{E}_2\{Pu'(\pi)\}-C'(q)[\mathrm{E}_1u'(\pi)-\mathrm{E}_2u'(\pi)] \\ &= \frac{1}{q}[\mathrm{E}_1f(\pi)-\mathrm{E}_2f(\pi)] \\ &\quad +\left[\frac{C(q)}{q}-C'(q)\right][\mathrm{E}_1u'(\pi)-\mathrm{E}_2u'(\pi)] \\ &\leqslant \left[\frac{C(q)}{q}-C'(q)\right][\mathrm{E}_1u'(\pi)-\mathrm{E}_2u'(\pi)]. \end{aligned}$$

[51] The strict concavity or convexity of f is sufficient to ensure $q_1\neq q_2$. This result appeared in Lippman and McCall (1981). A similar result is contained in Ishii (1977).

Since marginal cost exceeds average cost, $C(q)/q < C'(q)$. Applying (1.6) to the convex function u', we obtain

$$U_1'(q) < U_2'(q),$$

from which $q_1 < q_2$. The other results are obtained in the same manner. Q.E.D.

The most familiar utility functions satisfying f concave and u' convex are the isoelastic utility functions with parameter $\gamma \geqslant 0$, i.e., $u(c) = \ln c$ and $u(c) = c^\gamma/\gamma$, $0 < \gamma < 1$; f is convex if $\gamma \leqslant 0$. The assumption that marginal cost exceeds average cost necessitates that there be no fixed cost (i.e., $B = 0$), whereas there must be a fixed cost if average cost exceeds marginal cost. Theorem 1 can easily be extended to cover the case of a U-shaped average cost curve.[52]

Even though it does not yield a particularly useful result in this particular context, Theorem 1 of Diamond and Stiglitz (1974) is of interest in discerning the impact of changes in risk. In particular let $v(p, q)$ be the utility associated with the decision q when the random variable P_r assumes the value p. The decision maker seeks to maximize

$$U_r(q) \equiv \int v(p, q)\, \mathrm{d}F_r(p), \tag{5.7}$$

where F_r is the cumulative distribution function of the random variable P_r, and increases in r represent mean-preserving increases in risk. Denote the optimal decision by q_r^* and assume that v increases in p and is strictly concave in q. Diamond and Stiglitz (1974, p. 340) have shown that q^* increases [decreases] with r if $\partial v/\partial q$ is a strictly convex [concave] function of p, i.e., $\partial^3 v/\partial q \partial p^2 > [<]$ 0.

In our competitive model, we have $v(p, q) = u(pq - C(q))$. Now v is increasing in p; $\partial^2 v(p, q)/\partial q^2 = (p - C'(q))^2 u''(\pi) - C''(q)u'(\pi) < 0$ because u is strictly concave; and $\partial^3 v/\partial q \partial p^2 < 0$ provided that

$$-\frac{u'''(\pi)}{u''(\pi)}\left[q(p - C'(q))\right] > 2. \tag{5.8}$$

Thus, if u satisfies (5.8) on the *relevant* range of p and q, an increase in (mean-preserving) risk will lead to reduced output.

[52] See Lippman and McCall (1981) for a full discussion of this point.

5.2.2. *Changes in risk aversion*

In an early paper explicitly considering production under uncertainty, McCall (1967) showed that for firms with constant absolute risk aversion, the risk averse firm produces less than the risk neutral firm and it, in turn, produces less than the risk preferant firm. Unfortunately, this result only applies to a limited class of utility functions; moreover, it allows no comparison amongst risk averse firms. Our next result overcomes these limitations.[53] To distinguish between the optimal value of q for the utility functions u and v, replace q^* by q_u and q_v, respectively.

Theorem 2

If $r_u > r_v$, then $q_u < q_v$.

Proof

To begin, observe that if there is a strictly positive and strictly increasing function g such that the positive functions h and k satisfy $k = gh$, then $\mathrm{E}[Xk(X)] > g(0)\mathrm{E}[Xh(X)]$ for any random variable X for which $\mathrm{var}(X) > 0$ and $\mathrm{E}[Xh(X)] < \infty$. This follows as $\mathrm{E}[Xk(X)] = \mathrm{E}[Xg(X)h(X)]$ and $tg(t) > tg(0)$ when $t > 0$ and when $t < 0$, as g is positive and increasing.

Letting $X = P - C'(q)$, $k(t) = v'(tq + qC'(q) - C(q))$, and $h(t) = u'(tq + qC'(q) - C(q))$ and observing that $(\mathrm{d}/\mathrm{d}t)[v'(t)/u'(t)] > 0$, we have

$$V'(q) > g(0)U'(q), \quad \text{all} \quad q \geqslant 0, \tag{5.9}$$

where $V(q) \equiv \mathrm{E}v(\pi(q))$. The desired result directly follows from (5.9), $g(0) > 0$, and the concavity (uniqueness of the local maxima) of U and V. Q.E.D.

5.2.3. *Influence of fixed costs and a profits tax*

Baron (1970), Leland (1972), and Sandmo (1971) have shown[54] that the competitive firm's response to fixed costs is radically different under uncertainty than under certainty. It is, of course, well established that short-run production decisions are invariant to fixed costs. This is not true for a firm possessing a concave utility function and operating in the risk environment described here: increases in fixed costs lead to decreases in output whenever the firm has decreasing absolute risk aversion.

[53] David Baron was the first to obtain this result; see Baron (1970, proposition 1; 1973, theorem 1). In addition, Hartley (1978) demonstrated that theorem 4 of Diamond and Stiglitz (1974) can be applied to obtain this result. The proof presented here is shorter and more straightforward.

[54] Also see Rothschild and Stiglitz (1971, pp. 82–83).

The reason for this (paradoxical) behavior is fairly clear. It has already been shown that risk averse firms produce less than risk indifferent firms. If the firm has a decreasing (increasing) absolute risk aversion function, then output should increase (decrease) with wealth. But fixed costs are akin to negative wealth, and hence the result follows.

Theorem 3

If r_u is decreasing [increasing], then $\mathrm{d}q^*/\mathrm{d}B<0$ [>0].

Proof[55]

Consider the utility function v defined by ($\varepsilon>0$)

$$v(t)=u(t-\varepsilon), \quad \text{all } t. \tag{5.10}$$

Then increasing the fixed cost B by ε induces the firm to seek the output q_v that maximizes $\mathrm{E}v(\pi(q))$. If r_u is decreasing then $r_v>r_u$ and $q_v<q_u$ by Theorem 2. The inequalities are reversed if r_u is increasing. Q.E.D.

Next suppose there is a proportional profits tax at rate t,[56] so that after-tax profit π is given by

$$\pi(q)=(1-t)(Pq-C(q)), \tag{5.11}$$

and as before the firm seeks q^* to maximize

$$U(q)=\mathrm{E}u(\pi(q)). \tag{5.12}$$

In the absence of uncertainty output does not vary with t. In the presence of uncertainty, however, it decreases or increases in accord with the decreasing or increasing nature of the firm's relative risk aversion. The explanation for this behavior is simple: an increase in t reduces the scale of the gamble (π) for each level of output and such a reduction in scale enhances the gamble for a decision maker with increasing relative risk aversion.

[55]An alternative proof consists in differentiating the first-order condition and applying Lemma 1 in conjunction with r_u decreasing and $\pi(q)$ increasing in P to yield

$$\frac{\mathrm{d}}{\mathrm{d}B}U'(q)<r_u(C'(q)q-C(q))U'(q),$$

so that $U'(q^*)=0$ implies $(\mathrm{d}/\mathrm{d}B)U'(q^*)<0$. Consequently, an increase in B necessitates an increase in U', which, by the decreasing nature of U', induces a decrease in q^*.

[56]See Penner (1967) and Horowitz (1970, pp. 389–391) for a related but distinct discussion. Theorem 4 is not new; see Horowitz for various versions.

Theorem 4

If R_u is decreasing [increasing], then $dq^*/dt<0$ [>0].

Proof

Verification of this result is, mutatis mutandis, the same as for the change in fixed costs. In particular, if R_u is decreasing, then

$$\begin{aligned}\frac{d}{dt}U'(q) &= -U'(q)/(1-t)+(1-t)E\{(P-C'(q))(-u''(\pi))(Pq-C(q))\}\\ &= -U'(q)/(1-t)+E\{(P-C'(q))u'(\pi)R_u(\pi)\}\\ &<(U'(q)/(1-t))(R_u-1),\end{aligned}$$

so that $(d/dt)U'(q^*)<0$. Q.E.D.

If price were to increase by ε with probability 1 and r_u is non-increasing, then (as per the proof of Theorem 4)

$$\begin{aligned}\frac{d}{d\varepsilon}U'(q) &= E\{u'(\pi(q))+q(P-C'(q))u''(\pi)\}\\ &> -E\{(P-C'(q))u'(\pi)r_u(\pi)\}/q\\ &> -U'(q)r_u(C'(q)q-C(q))/q.\end{aligned}$$

Hence, $dU'(q^*)/d\varepsilon>0$ so that $dq^*/d\varepsilon>0$. Here, however, the converse is not true; r_u increasing does not ensure $dq^*/d\varepsilon<0$.

5.3. *Factor demand under price uncertainty*

This section analyzes factor market responses to price uncertainty.[57] We assume the same model as in Section 5.2, i.e., the firm is a price taker and must produce before the output price is known. The firm knows the distribution of prices and maximizes the expected utility of profits.

Letting q, K, and L be output, capital, and labor, the (non-decreasing) production function is given by

$$q=f(K, L), \tag{5.13}$$

and profit by

$$\pi=Pq-wL-rK-B, \tag{5.14}$$

[57]This section is based on Batra and Ullah (1974) and Hartman (1975).

where w is the wage rate, r the cost of capital, and B the fixed cost. The firm's utility function u is strictly concave, and it seeks factor inputs K^* and L^* to maximize the expected utility U of profits, where

$$U(K, L) = \mathrm{E}u(Pf(K, L) - wL - rK - B). \tag{5.15}$$

To ensure that U is strictly concave, we assume that f is concave.

Because f is concave (so the firm cannot experience any increasing returns to scale) and costs are linear in the factors, $C(q)$, the total cost of producing q units induced by employing the optimal levels K_q and L_q of factor inputs, has non-decreasing marginal cost. Therefore, the analysis of the previous section reveals that uncertainty results in a smaller output.[58]

Moreover, the smaller output will, of course, necessitate changes in K and L. In particular, K will decrease[59] [increase] if $f_K f_{LL} - f_L f_{KL} < 0$ [>0]; correspondingly, L will decrease [increase] if $f_L f_{KK} - f_K f_{KL} < 0$ [>0]. Presumably f is well behaved in that $f_{KL} > 0$ in which case uncertainty causes both factor inputs to decrease.

Uncertainty also causes the (expected) value of the marginal product of each factor to exceed its marginal cost. To show this, note the first-order conditions:

$$f_K \mathrm{E}\{Pu'(\pi)\} = r\mathrm{E}u'(\pi), \tag{5.16}$$

and

$$f_L \mathrm{E}\{Pu'(\pi)\} = w\mathrm{E}u'(\pi). \tag{5.17}$$

[58] Determining the impact on output of increased riskiness of P is difficult. Decreasing absolute risk aversion is not sufficient to guarantee that output decreases as asserted by Batra and Ullah (1974). To see this take $u(t) = t^\gamma/\gamma$, $\gamma \leqslant 0$, $f(K, L) = K + L$, $w = r$, and $B > 0$. Then Theorem 1 of Section 5.2 shows output to be increasing with risk.

[59] Given the level q of output, the first- and second-order conditions for minimizing the cost $wL + rK$ of the inputs are

(a) $g(L, K) \equiv f_K/f_L - r/w = 0$,

and

(b) $2f_{KL}f_K f_L - f_L^2 f_{KK} - f_K^2 f_{LL} > 0$.

The implicit function theorem ensures the existence of a function h such that $g(L, h(L)) = 0$ on an interval containing L^* such that $h'(L^*) = -g_L(L^*, K^*)/g_K(L^*, K^*)$. Consequently, at (L^*, K^*) we have

(c) $\mathrm{d}q/\mathrm{d}L = f_L + f_K h' = (2f_{KL}f_K f_L - f_L^2 f_{KK} - f_K^2 f_{LL})/-[f_L f_{KK} - f_K f_{KL}]$.

The numerator is positive by (b) so the sign of $\mathrm{d}q/\mathrm{d}L$, and hence that of $\mathrm{d}L/\mathrm{d}q$, is positive if and only if the term in brackets is negative.

As demonstrated in the previous section, $\mathrm{E}\{(P-\mu)u'(\pi)\}<0$. Consequently, subtracting $\mu f_K \mathrm{E}u'(\pi)$ from (5.16) yields $(r-\mu f_K)\mathrm{E}u'(\pi)<0$, yielding

$$r<\mu f_K, \tag{5.18}$$

as $u'>0$. By the same argument,

$$w<\mu f_L. \tag{5.19}$$

Clearly, we ought to interpret $f_i \mathrm{E}\{Pu'(\pi)\}/\mathrm{E}u'(\pi)$ as the value of the marginal product of factor i, $i=K, L$.

Finally, consider a monopolist facing the (downward sloping) demand curve X given by

$$X(q)=D(q)+Z_q, \tag{5.20}$$

where $Z_q \equiv Z$ and $\mathrm{E}Z=0$. Then

$$\begin{aligned} 0=U_L&=\mathrm{E}\{u'(\pi)(\mathrm{d}\pi/\mathrm{d}L)\} \\ &=\mathrm{E}\{u'(\pi)[-w+f_L(D(q^*)+qD'(q^*))+f_L Z]\}, \end{aligned}$$

from which we obtain (if u is linear, $\mathrm{E}\{Zu'(\pi)\}=0$)

$$\begin{aligned} w&=f_L MR_{q^*}+f_L(\mathrm{E}\{Zu'(\pi)\}/\mathrm{E}u'(\pi)) \\ &<f_L MR_{q^*}=MRP_{q^*}, \end{aligned} \tag{5.21}$$

where the inequality follows from Lemma 1 of Section 5.2. Again, uncertainty causes the marginal revenue product to exceed the factor price.

5.4. *Alternatives to the expected utility hypothesis: Safety-first criteria*

Preferences based on expected profits and the probability of loss (safety-first criteria) have been considered as contenders to the expected utility hypothesis. Following the recent paper by Arzac (1976), we consider two of them and their implications for optimal output.

The demand function D facing the firm is given by

$$D(q, Z)=g(q)+h(q)Z, \tag{5.22}$$

where q is output, $h>0$, and Z is a random variable (that does not depend upon q) with $\mathrm{E}(Z)=1$. This formulation includes additive uncertainty ($h\equiv 1$) and

multiplicative uncertainty ($g \equiv 0$) as special cases. To ensure that the monopolistic as well as the competitive environment is included, assume that the expected price $\mu(q) \equiv \mathrm{E}D(q, Z)$ is non-increasing, that is,

$$\mu'(q) = g'(q) + h'(q) \leqslant 0, \quad \text{all} \quad q \geqslant 0. \tag{5.23}$$

As before, the firm's profit π is given by

$$\pi(q) = D(q, Z)q - C(q), \tag{5.24}$$

with C representing total cost. Finally, we assume that there is a level of output for which expected profit is positive and that $\mathrm{E}\pi(q)$ is concave.

In view of the above, the optimal level $\bar{q}$ in the certainty case (i.e., $Z \equiv 1$) is the unique solution to

$$\overline{MR}(q) = q\mu'(q) + \mu(q) = C'(q) = MC(q). \tag{5.25}$$

5.4.1. The Roy criterion

Roy (1952) proposed setting output so as to minimize the probability R of loss. According to Roy's criterion, the firm seeks the level q_R that satisfies

$$R(q_R) = \min_{q > 0} R(q), \tag{5.26}$$

where

$$R(q) = P\{\pi(q) \leqslant 0\}. \tag{5.27}$$

Utilizing (5.22) and (5.24), we obtain

$$R(q) = P\{Z \leqslant (AC(q) - g(q))/h(q)\},$$

where $AC(q) \equiv C(q)/q$ is average cost. Thus, we seek to minimize $H(q) = (AC(q) - g(q))/h(q)$. Differentiating H and re-arranging terms leads to

$$H' = \left[(MC - \overline{MR})h - (AC - \mu)(h + qh')\right]/qh^2. \tag{5.28}$$

By evaluating H' at q_R, (5.28) yields

$$\overline{MR} = MC + (\mu - AC)(h + qh')/h. \tag{5.29}$$

With additive uncertainty this reduces to

$$\overline{MR} = MC + \mu - AC, \tag{5.30}$$

whereas multiplicative uncertainty results in ($\mu = h$)

$$\overline{MR} = MC + (\mu - AC)\,\overline{MR}\,/h = MC(\mu/AC). \tag{5.31}$$

As $E\pi(q_R) > 0$, $\mu_{q_R} > AC(q_R)$ so $\overline{MR}(q_R) > C'(q_R)$ in both cases and, in turn,

$$q_R < \bar{q} \tag{5.32}$$

by the concavity of $E\pi(q)$. In the general case $h + qh' > 0$ is sufficient to ensure $q_R < \bar{q}$.

Under competition $D(q, \cdot)$ is constant in q for each value of Z so $g' = h' = 0$. Consequently, as anticipated, (5.29) reduces to $MC = AC$.

5.4.2. *The Telser criterion*

Following Telser (1955) output is optimal when the firm maximizes expected profit subject to the condition that the probability of loss does not exceed some predetermined level α, $0 \leqslant \alpha < 1$. Thus, the problem is to find the output q_T which maximizes $E\pi(q)$ subject to the constraint $P\{Z \leqslant H(q)\} \leqslant \alpha$. If $P\{\pi(\bar{q}) \leqslant 0\} \leqslant \alpha$, then $q_T = \bar{q}$; otherwise we must have $H(q_T) < H(\bar{q})$. But $\bar{q} > q_R$ and $H'(q) > 0$ for $q > q_R$ so this implies $q_T < \bar{q}$. Furthermore, $q_T > q_R$ if $P\{\pi(q_R) \leqslant 0\} < \alpha$.

Arzac considers two other safety first criteria. In addition, he obtains interesting results by considering the effect of changes in fixed costs, the proportional tax rate, and an upward shift in the demand curve.

5.5. *The competitive industry under uncertainty*

In their recent paper, Sheshinski and Dreze (1976) studied competitive industry behavior under uncertainty. The industry is composed of s identical firms producing a single product. Industry demand is a random variable $Q \geqslant 0$ with $EQ = \mu > 0$; thus, the price elasticity is zero. After Q has been observed, the demand is divided equally amongst the s firms. As there is no storage, each firm produces Q/s. The total cost, average cost, and marginal cost associated with a firm's producing q units is denoted by $T(q)$, $A(q)$, and $M(q)$, respectively. We assume (1) there are no barriers to entry, (2) the average cost curve is U-shaped and continuously differentiable, and (3) the marginal cost is increasing and convex. As the industry is competitive, the price p will be equal to the firms' marginal cost.

Because A is U-shaped, there is a number $\bar{q}$ such that

$$\begin{aligned} A'(q) &< 0 \quad \text{for} \quad q < \bar{q}, \\ &> 0 \quad \text{for} \quad q > \bar{q}, \end{aligned} \tag{5.33}$$

and

$$\begin{aligned} M(q) &< A(q) \quad \text{for} \quad q < \bar{q}, \\ &> A(q) \quad \text{for} \quad q > \bar{q}, \end{aligned}$$

for

$$M(q) = A(q) + qA'(q). \tag{5.34}$$

Coupling (5.34) and the fact that price equals marginal cost, reveals that $\pi(q)$, the profit due to the firm's producing q units, satisfies

$$\pi(q) = qM(q) - T(q) = q[\, M(q) - A(q)\,] = q^2 A'(q). \tag{5.35}$$

The first problem is to find the industry size s^* that minimizes the expected cost C for the industry. Thus, we seek the optimal industry size s^* satisfying[60]

$$C(s^*) = \min_{s>0} C(s), \tag{5.36}$$

$$C(s) = \mathrm{E}\{sT(Q/s)\}. \tag{5.37}$$

The mean output per firm in an industry of optimal size is, of course, $q^* = \mu/s^*$. The first result demonstrates that uncertainty causes the output for an industry with risk neutral firms to be smaller than the output $\bar{q}$ that minimizes average cost.

Theorem 5

If $\mathrm{var}Q > 0$, then $q^* < \bar{q}$ so that $A(q^*) > M(q^*)$.

Proof

The first-order condition is

$$0 = C'(s) = -\mathrm{E}\pi(Q/s). \tag{5.38}$$

Because M is increasing and convex, π is increasing and strictly convex so that

$$\pi(q) > \pi(\bar{q}) + \pi'(\bar{q})(q - \bar{q}) \quad \text{for} \quad q \neq \bar{q}. \tag{5.39}$$

[60] For convenience we do not restrict s to be an integer. Accordingly, we need not have $\mu > \bar{q}$.

Inspection of (5.33) and (5.35) reveals that $\pi(\bar{q})=0$. Consequently, taking the expectation of (5.39) with $q=Q/s^*$ yields

$$0=\mathrm{E}\pi(Q/s^*)>\pi(\bar{q})+\pi'(\bar{q})\mathrm{E}\{Q/s^*-\bar{q}\}=\pi'(\bar{q})(q^*-\bar{q}),$$

so that $q^*<\bar{q}$ since π' is strictly positive. Q.E.D.

From (5.28) and the fact that there is free entry, it is clear that the number of firms for which expected profit equals zero is precisely s^*. Thus, the competitive equilibrium is efficient in that total expected cost is at its minimum and it is characterized by excess capacity in that $q^*<\bar{q}$.

Sheshinski and Dreze (1976) have shown that increasing either the mean or the variance of the industry demand will cause the optimal number of firms to increase. Similarly, Lippman and McCall (1981) have shown that this change holds for any mean-preserving increase in the riskiness of industry demand.

6. Brownian motion, martingales, and their economic applications

6.1. *Introduction*

Although it arose (1827) in an attempt to explain the motion exhibited by small particles immersed in a liquid, Brownian motion has proved extraordinarily useful in describing the behavior of many economic variables, the most noteworthy being price. The copious literature on the efficient market hypothesis had its origin in the supposed Brownian motion of stock market prices.[61] Later it was recognized that the martingale property of Brownian motion was the crucial ingredient of the efficient market hypothesis.[62] However, Brownian motion has recently reappeared as a central actor in the fundamental research on option pricing.[63] Brownian motion has also played a key role in the design of models for the optimal control of (i) inventories, (ii) the maintenance of equipment, (iii) the demand for cash balances, and (iv) the analysis of index bonds.[64]

We begin this section with a brief description of Brownian motion. This is followed by a discussion of the efficient market hypothesis in which the relation

[61] See the articles contained in Cootner (1967), especially those of Bachelier and Osborne.

[62] See Fama (1970).

[63] For an historical and up-to-date treatment of this subject, see Rubenstein (1979).

[64] We do not discuss the maintenance literature here. McCall (1965) and Pierskalla and Voelker (1976) have summarized this literature. The importance of Brownian motion in maintenance modeling is clearly revealed in the paper by Anderson and Friedman (1977). Cash balances and index bonds are treated in Sections 6.4 and 6.5. See Whitt (1973) and the references therein for applications to inventory and production.

between Brownian motion and martingales is exhibited. A random walk model of cash balances is then presented along with a derivation of the demand for cash balances. This is done in an inventory-theoretic setting and is generalized to the continuous time Brownian motion model. Finally, the demand for index bonds is studied when the inflation rate follows an Ito diffusion process. This provides a glimpse of the stochastic calculus.

6.2. Brownian motion

In this section we begin by presenting a simple limiting argument for Brownian motion, then state the axioms of Brownian motion, and mention some of its properties.

6.2.1. A limiting argument for Brownian motion

A particle, which may, for example, represent a stock market price or an individual's wealth position, moves to the right or the left a distance Δx with equal probability, viz., one-half. These moves occur every Δt units of time, i.e., at times $\Delta t, 2\Delta t, \ldots$ Let $Y_1, Y_2, \ldots$ be a sequence of independent random variables such that

$$\mathrm{P}(Y_i = \Delta x) = \mathrm{P}(Y_i = -\Delta x) = \tfrac{1}{2} \quad \text{for} \quad i = 1, 2, \ldots \tag{6.1}$$

At time t the total number of moves is simply $[t/\Delta t]$, where $[w]$ denotes the greatest integer less than or equal to w, and the position of the particle (total wealth) at time t is given by

$$X(t) = Y_1 + Y_2 + \cdots + Y_{[t/\Delta t]}. \tag{6.2}$$

To obtain Brownian motion $B(t)$ as a limiting stochastic process, Δx and Δt must approach zero in such a way that $X(t) \to B(t)$. To begin, note that $\mathrm{E}(X^2(t)) = \Delta x^2 [t/\Delta t]$ so that

$$\mathrm{E}X^2(t) \simeq (\Delta x)^2 t/\Delta t. \tag{6.3}$$

In order for this variance to be strictly positive and less than infinity, Δx must be of order of magnitude $(\Delta t)^{\frac{1}{2}}$. Thus, let $\Delta x = (\Delta t)^{\frac{1}{2}}$ and $\Delta t = 1/n$. Then for all i, Y_i equals $\pm(n)^{-\frac{1}{2}}$, each with probability one-half. Consequently, $X(t)$ has the same distribution as

$$X^{(n)}(t) = \frac{Z_1 + Z_2 + \cdots + Z_{[nt]}}{\sqrt{n}} = \sqrt{t}\, \frac{Z_1 + Z_2 + \cdots + Z_{[nt]}}{\sqrt{nt}}, \tag{6.4}$$

where the Z's are ± 1 with probability one-half. By the central limit theorem[65] $X^{(n)}(t)$ converges to a normal distribution with zero mean and variance t as $n \to \infty$. It can also be shown that all the joint distributions of $X^{(n)}(t)$ are asymptotically identical to those of Brownian motion.

6.2.2. The axioms of Brownian motion

The three axioms of Brownian motion are:

(i) *Independence*: The random variable $B(t+\Delta t)-B(t)$ is independent of the sigma algebra generated by all random variables up to time t; that is, the change in position during $(t, t+\Delta t)$ is independent of anything that has happened till time t.

(ii) *Stationarity*: The distribution of the random variable $B(t+\Delta t)-B(t)$ is independent of t.

(iii) *Continuity*: For all $\delta > 0$, $\lim_{\Delta t \to 0} \mathrm{P}(|B(t+\Delta t)-B(t)| \geqslant \delta)/\Delta t = 0$.

A physical interpretation of the first axiom is that the momentum incurred by the particle from molecular bombardment during $(t, t+\Delta t)$ is independent of anything that occurred before t. This is reasonable if the displacement of the particle due to its initial velocity at the beginning of $(t, t+\Delta t)$ is negligible relative to the motion induced by molecular attack during $(t, t+\Delta t)$. The stationarity assumption requires that the process (i.e., motion of the particle) be homogeneous over time — the probability of change over any time interval depends only on the length of the interval and not on its location relative to the origin. The third axiom is precisely what is needed to guarantee that the motion of the particle be continuous, which, of course, it should be. Specifically, each sample path $B(t, \omega)$ ought to be a continuous function of t — except perhaps for a set of ω's with probability zero. Demonstrating that the sample paths of Brownian motion are continuous is an arduous task. Instead we shall show that (iii) is nearly equivalent to having $B(t)$ continuous in t with probability one. To do so fix $\delta > 0$ and define

$$Y_n = \max_{1 \leqslant k \leqslant n} |B(k/n) - B((k-1)/n)|. \tag{6.5}$$

If $Y_n \to 0$, then $B(t)$ would, in fact, be continuous on $[0, 1]$ and it would follow that

$$\lim_{n \to \infty} \mathrm{P}(Y_n \geqslant \delta) = 0. \tag{6.6}$$

[65] Let $\{Z_k\}$ be a sequence of independent, identically distributed random variables with mean 0 and variance 1. Then by the Central Limit Theorem, $X^{(n)}(t) \to N(0, t)$.

But the independence and stationarity of $B(1/n)-B(0), B(2/n)-B(1/n),\ldots$ yield

$$\begin{aligned}\mathrm{P}(Y_n \geqslant \delta) &= 1-\mathrm{P}(Y_n<\delta)\\ &= 1-\mathrm{P}(|B(1/n)-B(0)|<\delta)^n\\ &= 1-\left[1-\mathrm{P}(|B(1/n)-B(0)|\geqslant\delta)\right]^n\\ &\approx 1-\exp -n\,\mathrm{P}(|B(1/n)-B(0)|\geqslant\delta),\end{aligned}$$

as $1-t\approx \mathrm{e}^{-t}$ (in fact $1-t\leqslant \mathrm{e}^{-t}$ for all $t\geqslant 0$). Thus $\mathrm{P}(Y_n\geqslant\delta)\to 0$ if and only if $n\,\mathrm{P}(|B(1/n)-B(0)|\geqslant\delta)\to 0$, which is precisely the statement of the continuity axiom with $\Delta t=1/n$ so that (6.6) is equivalent to the continuity axiom.

6.2.3. *Properties of Brownian motion*

Of fundamental importance is the fact that when $B(0)=0$ there are numbers μ and σ such that, for each t, $B(t)$ has a normal distribution with mean μt and variance $\sigma^2 t$. Furthermore, the finite dimensional distributions are multivariate normals. When $\mu=0$ and $\sigma=1$, we refer to B as standard Brownian motion. [It is conventional to assume that $B(0)=0$.]

Given the discussion of axiom (iii), it should come as no surprise that there is a version of Brownian motion on $[0,\infty)$ such that *all* sample paths are continuous. However, it is rather remarkable that almost every Brownian path is nowhere differentiable. While we are not in a position to formally verify this, we can attempt to make it credible.

To begin, it can be demonstrated [see Karlin and Taylor (1975)] that for each path the total squared deviation of standard Brownian motion on $[0,t]$ is simply t as

$$\lim_{n\to\infty}\sum_{k=1}^{2^n}\Delta_{n,k}^2=t, \tag{6.7}$$

where $\Delta_{n,k}\equiv|B(kt/2^n)-B((k-1)t/2^n)|$. Define δ_n to be the maximum of the $\Delta_{n,k}$, $1\leqslant k\leqslant 2^n$, and note that $\Delta_{n,k}\geqslant\Delta_{n,k}^2/\delta_n$, so that

$$\sum_{k=1}^{2^n}\Delta_{n,k}\geqslant\sum_{k=1}^{2^n}\Delta_{n,k}^2/\delta_n. \tag{6.8}$$

Equation (6.7) asserts that the numerator on the right-hand side of (6.8) converges to t whereas δ_n converges to 0 because Brownian paths are continuous and hence uniformly continuous on the interval $[0,t]$. Consequently, taking the

limit on n in (6.8) reveals that the total variation of each Brownian path on $[0, t]$ is infinite; the infinite variation, which is itself of interest, suggests that the paths are nowhere differentiable.[66]

The infinite total variation on finite intervals might lead one to believe that little can be said about the oscillations of Brownian motion. The oscillations do, however, follow the law of the iterated logarithm:

$$\mathrm{P}\left\{\limsup_{t\downarrow 0} \frac{B(t)}{\sqrt{2t\log(\log 1/t)}} = 1\right\} = 1. \tag{6.9}$$

6.2.4. *Computation of operating characteristics*

Brownian motion is probably the most tractable of all stochastic processes. Using standard mathematical methods—and there are several distinct approaches—one can compute explicit formulas for virtually every operating characteristic of the policies having the form described in Section 6.4. In particular, because the problem is discounted, the Laplace transforms (with variable α) of various first passage times are of interest. To illustrate this we shall make use of martingale arguments in calculating two operating characteristics for the simple S policy found in Section 6.4.2.

First, we wish to calculate the probability p that the process reaches S before it reaches zero when at time zero it starts at w [i.e., $X(0)=w$]. This is equivalent to the probability that the process reaches $S-w$ before it reaches $-w$ if $X(0)=0$. With $a=-w$ and $b=S-w$, the first passage time T is the first time the process hits a or b; that is,

$$T=\inf\{t\geqslant 0: X(t)=a \text{ or } X(t)=b\}. \tag{6.10}$$

Related but distinct arguments are needed to treat the cases of $\mu=0$ and $\mu\neq 0$, so for simplicity we shall assume that $\mu=0$. As $X(0)=0$ and $\mu=0$, $\mathrm{E}(X(t))=0$; thus, we might guess that $\mathrm{E}(X(T))$ also equals zero, in which case we have

$$\begin{aligned} 0&=\mathrm{E}(X(T))\\ &=a\,\mathrm{P}(X(T)=a)+b\,\mathrm{P}(X(T)=b)\\ &=a(1-p)+bp. \end{aligned} \tag{6.11}$$

[66]Alternatively, define $D_n(t)$ by $D_n(t)=[B(t+1/n)-B(t)]/[1/n]$ so that $\mathrm{E}(D_n^2(t))=\sigma^2/(1/n)$ and $\mathrm{E}D_n^2(t)\to\infty$ as $n\to\infty$. Then it can be shown that D_n converges in quadratic mean to $B'(t)$ for a differentiable stochastic process. Thus, if Brownian motion were in fact differentiable then the convergence of $\mathrm{E}\{(D_n(t)-B'(t))^2\}$ to 0 would imply that $\mathrm{E}D_n^2(t)$ converges to $\mathrm{E}(B'(t)^2)$ rather than to infinity, a contradiction.

Solving (6.11) for p yields

$$p = -a/(b-a) = w/S. \tag{6.12}$$

We now proceed to verify that $\mathrm{E}(X(T)) = 0$. To do so we need only verify that T satisfies the following version of the

Optional Sampling Theorem

Let $\{X(t)\}$ be a martingale and T a Markov or stopping time.[67] If

$$\mathrm{P}(T < \infty) = 1, \quad \mathrm{E}(|X(T)|) < \infty, \quad \int_{\{T>t\}} X(t)\,\mathrm{P}(\mathrm{d}\omega) \to 0 \quad \text{as} \quad t \to \infty, \tag{6.13}$$

then $\mathrm{E}(X(T)) = \mathrm{E}(X(0))$.

Because $|X(t)| \leqslant |a| + b$ for all $t \leqslant T$, the second condition in (6.13) holds. This boundedness would also enable us to verify the third condition if $\mathrm{P}(T > t) \to 0$ as $t \to \infty$. But this is merely equivalent to knowing $\mathrm{P}(T < \infty) = 1$. To show that the hitting time T of a or b is finite we proceed by noting that $V(t) \equiv X^2(t) - \sigma^2 t$ is a martingale, for ($v^2 \equiv u + \sigma^2 t$)

$$\begin{aligned}
\mathrm{E}\{X^2(t+s) | X^2(t) = v^2\} &= \mathrm{E}\{[X(t+s) - X(t) + X(t)]^2 | X^2(t) = v^2\} \\
&= \mathrm{E}\{(X(t+s) - X(t))^2 | X^2(t) = v^2\} \\
&\quad \pm 2v\,\mathrm{E}\{X(t+s) - X(t)\} + v^2 \\
&= s\sigma^2 + 0 + v^2,
\end{aligned}$$

so that

$$\mathrm{E}\{V(t+s) | V(t) = u\} = s\sigma^2 + v^2 - \sigma^2(t+s) = u.$$

As $T \wedge t$[68] satisfies (6.13), $\mathrm{E}\{V(T \wedge t)\} = \mathrm{E}V(0) = 0$, so that

$$\sigma^2 \mathrm{E}(T \wedge t) = \mathrm{E}\{X^2(T \wedge t)\} \leqslant a^2 + b^2, \quad \text{all} \quad t \geqslant 0. \tag{6.14}$$

Passing to the limit in (6.14) reveals that $\mathrm{E}(T)$ is finite and, hence, $\mathrm{P}(T < \infty) = 1$.

[67] The random variable T is a Markov (stopping) time relative to the stochastic process $\{Z_n\}$ if it takes on non-negative integer values and if for every $n = 0, 1, \ldots$, the event $\{T = n\}$ depends only on $(Z_0, Z_1, \ldots, Z_n)$. A similar definition holds for continuous time stochastic processes. In particular, the hitting time of any closed set (e.g., $\{a, b\}$) is a stopping time if the process has continuous sample paths as does Brownian motion.

[68] $x \wedge y \equiv \min(x, y)$.

Next, we calculate $E(T)$, the expected time till a transfer occurs (from cash to securities if S is hit first or from securities to cash if 0 is hit first). Since $V(t)$ is a martingale satisfying the conditions of the optional sampling theorem, we obtain

$$0 = E[V(0)] = E[V(T)] = E(X^2(T)) - \sigma^2 E(T), \tag{6.15}$$

$$E(T) = E(X^2(T))/\sigma^2 = [a^2(1-p) + b^2 p]/\sigma^2 = w(S-w)/\sigma^2. \tag{6.16}$$

6.3. *The efficient market hypothesis*

Many stochastic models have been devised to study the motion of security prices. The goal of the early work in this area was to design a scheme for predicting future movements of security prices.[69] While this quest has not been abandoned — indeed, a success would, like a perfect crime, go unnoticed — the models that have been most successful in describing stock market behavior possess the martingale property. And as we will see, it is impossible to design a system that will beat the market when the underlying stochastic process is a martingale.

In the study of capital markets, security prices play a crucial allocative role. Firms use them to guide their investment decisions. Similarly, the allocation of an investor's funds across securities (the solution to the portfolio selection problem) is based on these prices. The question that immediately arises is: how informative are these prices with respect to these key decisions? In answering this question, Fama (1970) distinguishes among three different types of information: (i) information contained in past prices of the securities in question, $\mathcal{F}^1$; (ii) information contained not only in past prices but also in all past events that have been publicly reported, $\mathcal{F}^2$; and (iii) information contained in *all* past events, $\mathcal{F}^3$. Clearly $\mathcal{F}^1 \subset \mathcal{F}^2 \subset \mathcal{F}^3$, and the question can be more formally stated as: after allowance for a "normal" rate of return, is the sequence of prices a martingale with respect to $\mathcal{F}^1$, $\mathcal{F}^2$ or $\mathcal{F}^3$? If the martingale property holds for $\mathcal{F}^1$, the market is said to be "weakly efficient"; if it also holds for $\mathcal{F}^2$ the market is "semi-strongly efficient"; finally, if the martingale property holds for $\mathcal{F}^3$, the market is "strongly efficient". There are several reasons for defining market efficiency in terms of martingales. First, it is simple and parsimonious. The definition of a martingale does not require that the sequence $\{X_n\}$ of random variables be identically distributed or independent. The assumptions are merely

$$E(|X_n|) < \infty \quad \text{for all } n, \tag{6.17}$$

[69] The literature on this topic is immense. While our treatment is brief it is sufficient to reveal the rich probabilistic structure that permeates this subject. For extensive discussions, see Fama (1970), Granger (1972), Jensen (1972), Sharpe (1970), and Ziemba and Vickson (1975).

and

$$\mathrm{E}(X_{n+1}|\mathcal{F}_n)=X_n, \tag{6.18}$$

where $\mathcal{F}_n$ is the σ-algebra generated by the random variables $X_1,\ldots,X_n$. There has been some confusion on this point. Early studies of security price behavior assumed that the underlying stochastic process was either a random walk or Brownian motion. We know that a random walk has the martingale property (and is Markovian) with the sequence of random variables being independent and identically distributed. Thus every random walk is a martingale, but not conversely. Similarly, with Brownian motion. Standard Brownian motion, $B(t)$, is normally distributed with mean 0 and variance t. Standard Brownian motion does have the martingale property, but clearly not all martingales are Brownian motions. Thus the stochastic behavior of stock prices need not be normal; in fact, there are infinite variance distributions that satisfy the martingale property.

A second reason for the martingale assumption is its informational implications for prices. As discussed in Section 6.2.4, the Optional Sampling Theorem states that when the underlying process has the martingale (fair game) property it is impossible to design a (gambling) system that is favorable. Under the weak form of the martingale assumption, current stock prices contain all past information and no system can, on average, yield a rate of return in excess of the "normal" rate.

A third reason for the martingale assumption is that its implications can be empirically tested. The empirical studies surveyed by Fama showed that the hypothesis of weak efficiency could not be rejected. Serial correlation between lagged prices tended to be close to zero. Also trading systems (filter rules) were not superior to the simple buy and hold policy. There was evidence that large daily price changes are followed by large changes, suggesting that the independence assumption of the random walk model is violated. However, the signs of these large changes are random and so this is not evidence against the martingale hypothesis. Moreover, Fama (1970) tells us that:

> Semi strong form tests, in which prices are assumed to fully reflect all obviously public available information, have also supported the efficient markets hypothesis.

The strong form of the martingale hypothesis was tested by observing the behavior of corporate insiders and specialists. They do indeed have special information and profit from it. Fama conjectures that they are the only ones who violate the efficient market model. This imperfection would appear to reside more in the market for information than in the capital market itself.

6.4. *Stochastic models of inventory and the demand for cash balances*

One of the first applications of probabilistic economics was the use of inventory models to explain the demand for cash balances. Arrow's review[70] of the pre-1958 literature contains an excellent discussion of the three motives for holding cash: transactions, precautionary, and speculative.

In this section we derive the demand for cash using the Miller and Orr (1966) discrete time model and then present the Harrison and Taylor (1978) continuous time Brownian version of this model.

6.4.1. *A discrete time model of the demand for cash balances*

Miller and Orr (1966) assume that the firm has two assets: cash balances and a portfolio of liquid assets. The return on the portfolio is r dollars per day and a fixed cost of γ is incurred whenever transfers are made from one account to the other. Once it has been decided to make a transfer, it occurs without delay. Fluctuations in cash balances are governed by a symmetric random walk; during a specified time interval cash balances increase or decrease by m dollars with probability p or $1-p$, respectively.

The firm wishes to minimize the long-run average cost of managing cash balances. This objective function is applied to the following simple policy:[71] Cash balances are allowed to fluctuate according to the above random walk provided they are greater than 0 and less than h. When the upper bound h is hit, an amount $h-z$ of cash is transferred to the portfolio of liquid assets; when the lower boundary 0 is reached, z dollars are transferred from the liquid portfolio to cash balances. Miller and Orr refer to this as an (h, z) policy. They calculate the optimal values of h and z and derive the long-run average demand for money.

The steady-state distribution of cash holdings is triangular with mean $(h+z)/3$, and they identify this mean as the long-run average demand for cash balances. The optimal values of h and z are

$$h^* = 3z^* \quad \text{and} \quad z^* = \left(\frac{3\gamma}{4r}\sigma^2\right)^{\frac{1}{3}}, \tag{6.19}$$

so the demand for money is

$$M = \frac{h^*+z^*}{3} = \frac{4}{3}\left(\frac{3\gamma}{4r}\sigma^2\right)^{\frac{1}{3}}, \tag{6.20}$$

[70] See Arrow, Karlin and Scarf (1958, ch. 1).

[71] Eppen and Fama (1969) and Vial (1972) show that the optimal policy does have this simple form. Vial gives an excellent account of the current status of this problem.

where σ^2 $(=4m^2p(1-p))$ is the variance of daily cash flows. The demand for money varies directly with γ and σ^2 and inversely with the opportunity cost r of holding cash.

In an entirely different approach, Arrow (1971, p. 103) shows that the wealth elasticity of the demand for cash balances is at least one if the firm has increasing relative risk aversion.

6.4.2. A continuous time model of the demand for cash balances

We now present a continuous time version of the stochastic cash management problem as formulated and solved in Harrison and Taylor (1978). The firm has a reservoir of cash balances which is augmented by sales revenue and diminished by operating expenses. Assume that the level of cash balances $X(t)$ generated by these stochastic additions and deletions is describable by Brownian motion. In particular $X=\{X(t),\ t\geqslant 0\}$ is Brownian motion with starting state $x\geqslant 0$, drift μ, and variance σ^2 so that $\mathrm{E}[X(t)]=x+\mu t$ and $\mathrm{var}[X(t)]=\sigma^2 t$. The level of cash balances is controllable by moving funds back and forth from a portfolio of liquid assets. These transfers occur instantaneously at a cost of k per dollar transferred into cash balances and c per dollar transferred out of cash balances. Money held in the form of cash balances earns nothing, whereas the per unit return on the portfolio of liquid assets is h per unit time. Thus h is the per unit opportunity cost of holding cash balances. Let $Y(t)$ denote the total amount conveyed from the liquid asset portfolio to cash balances by time t. Similarly, let $Z(t)$ be the total amount transferred from cash balances to the liquid asset portfolio by time t. The objective is to determine input and output controls Y and Z to minimize expected discounted costs subject to the non-negative constraint on cash balances, i.e.,

$$W(t)=X(t)+Y(t)-Z(t)\geqslant 0,\quad \text{all}\quad t\geqslant 0. \tag{6.21}$$

Letting α be the discount rate this is equivalent to finding an admissible policy (Y, Z) that minimizes

$$h\mathrm{E}\int_0^\infty \mathrm{e}^{-\alpha t}W(t)\,\mathrm{d}t+k\left\{Y(0)+\mathrm{E}\int_0^\infty \mathrm{e}^{-\alpha t}\,\mathrm{d}Y(t)\right\}$$
$$+c\left\{Z(0)+\mathrm{E}\int_0^\infty \mathrm{e}^{-\alpha t}\,\mathrm{d}Z(t)\right\}. \tag{6.22}$$

Defining $R(g)=\int_0^\infty \mathrm{e}^{-\alpha t}g(t)\,\mathrm{d}t$ for non-decreasing, integrable functions g, (path-wise) integration by parts in (6.22) coupled with (6.21) enable us to demonstrate that minimizing (6.22) is equivalent to minimizing

$$(h/\alpha+k)\mathrm{E}R(Y)+(-h/\alpha+c)\mathrm{E}R(Z), \tag{6.23}$$

as (6.22) and (6.23) only differ by the uncontrollable cost $ER(X)h/\alpha$, a term that involves neither Y nor Z. Harrison and Taylor show that the optimal policy is characterized by the minimal pair of controls (Y, Z) such that $0 \leqslant W(t) \leqslant S$ for all $t \geqslant 0$, where S is the unique positive solution to a specific transcendental equation. Therefore the manager should convert cash balances into liquid securities so that $W(t) \leqslant S$, and he should convert the minimum number of securities into cash to keep $W(t)$ positive. The controlled process $W(t)$ follows a Brownian motion X with reflecting barriers at zero and S.

If in addition to the proportional charge k a fixed cost K is incurred whenever liquid assets are converted to cash, then the optimal policy is to increase cash balances to s whenever they hit zero. This is accomplished by augmenting Y by s units where s is the solution to another transcendental equation. The manager reduces cash balances whenever they exceed $s+S$. This is achieved by appropriate increases in Z whenever W exceeds $s+S$. Finally, Taylor (1978) has shown that with a fixed cost for transfers in both directions the optimal policy is characterized by three critical numbers: increase $Y(t)$ by s whenever the cash balance hits zero and increase $Z(t)$ by q whenever $W(t)$ strikes an upper critical value S, where the three values for s, q, and S are unique solutions to three independent transcendental equations. Except for the added generalization that the two fixed costs can be unequal, this is the obvious analog of Miller and Orr's discrete time model.

6.5. *The demand for index bonds*

We conclude this section with one of the recent applications of diffusion processes to economics.[72] In his study of index bonds Fischer (1975) assumes that an individual's portfolio contains three assets: a real bond, an equity, and a nominal bond. Let w_1, w_2, and w_3 be the proportions invested in the real bond, equity, and nominal bond, respectively. These proportions are subject to instantaneous and costless adjustment by the individual. The returns on these assets are given by diffusion processes. There are two sources of uncertainty, each a Brownian motion. The first emanates from the diffusion process for inflation, while the second originates in the stock market. In deriving the demand for index bonds, Fischer assumes that households choose consumption, w_1, w_2, and w_3 so as to maximize expected utility. We do not present this derivation here, but instead focus only on the diffusion process describing the rate of inflation.

A basic stochastic assumption of this study is that the rate of inflation satisfies the stochastic differential equation

[72] The most celebrated application is the option pricing model by Black and Scholes (1973) and Merton (1976). For a discussion of this model, see Merton (1981) and Rubinstein (1979).

$$dP/P=\pi dt+\sigma dB, \tag{6.24}$$

where P is the price level, and the drift coefficient π is the expected rate of inflation; the diffusion coefficient σ is the variance of the process per unit time. The term dB is the stochastic ingredient of this differential equation with B being standard Brownian motion. At first sight analysis of this equation looks hopeless since

$$(B(t+\Delta t)-B(t))/\Delta t$$

has mean zero and variance $1/\Delta t$, and hence has no probabilistic limit. Since $B(t)$ has no derivative, the differential $dB(t)$ is meaningless. Part of this problem is resolved by interpreting (6.24) in the integrated form

$$P(t)=P(u)+\int_u^t \pi P(r)\,dr+\int_u^t \sigma P(r)\,dB(r). \tag{6.25}$$

Now, however, the paths of B being of unbounded variation, we must define what is meant by the stochastic integral with respect to Brownian motion. This is the point of departure for the development of stochastic calculus.[73] We ignore these problems and simply interpret (6.24) as stating that over a short period of time the proportionate change in the price level is normal with mean πdt and variance $\sigma^2 dt$. Now rewrite (6.24) as

$$dP=P\pi dt+P\sigma dB, \tag{6.26}$$

and let

$$y(t)=P(0)\exp\left[(\pi-\sigma^2/2)t+\sigma\int_0^t dB\right]. \tag{6.27}$$

The major result of stochastic calculus, Ito's lemma, shows that $y(t)$ is a solution to (6.26). Taking the logarithm of both sides of (6.27) gives

$$\ell\equiv\log\frac{P(t)}{P(0)}=(\pi-\sigma^2/2)t+\sigma\int_0^t dB. \tag{6.28}$$

hence, $P(t)$ is lognormal with

$$E\ell=(\pi-\sigma^2/2)t \quad \text{and} \quad \text{var}\,\ell=\sigma^2 t. \tag{6.29}$$

[73] The interested reader should consult Malliaris and Brock (1982) and Gikhmann and Skorokhod (1969).

7. Evolutionary processes in economics

7.1. *Introduction*

We now turn to evolution, a subject that has not yet received much attention in probabilistic economics. However, this condition is changing,[74] and we think the topic is so important to probabilistic economics that it will not be ignored in the future.

Malthusian concepts had a profound influence on the genesis of Darwin's theory of evolution. It is lamentable that after this auspicious beginning, the intellectual interplay between economics and evolution remained relatively dormant. However, this situation is now changing and both economists and biologists are once again recognizing the symbiotic relation between their two sciences.[75] Alchian's article "Uncertainty, Evolution and Economic Theory" reawakened economists to the usefulness of evolutionary ideas in reformulating the theory of the firm. The title of this influential article indicates the central role of probability theory in both economics and evolution.

The fundamental entities of evolutionary theory are the genes that are transmitted across generations, mutations, and natural selection. In the contest for survival, nature chooses those organisms that are best suited environmentally; poorly designed organisms are eliminated. In essence, nature ruthlessly applies the law of large numbers and after innumerable trials arrives at an "optimal" combination of organisms, i.e., a set of organisms each of which has a structure that has passed nature's trial by selection.

Alchian claims that the economic entities corresponding to organisms, genes, mutations, and natural selection are, respectively, firms, imitation, innovation, and positive profits.[76] Innovations may result from bungled imitations, as well as conscious efforts to improve, when the current routines run into trouble.[77] Thus a subtle combination of luck and intelligence determines whether a particular firm prospers.

Time plays essentially the same role in Darwinian evolution and the evolution of firms; however, the former has been operating on biological systems for millions of years, whereas, in comparison, the firm is a mere infant. Organisms that were unacceptable were leisurely terminated. Nature, having an abundance of time, merely waited until "suitable" organisms were generated. Rational man, of which the business firm is an obvious manifestation, responds less passively to the threat of elimination. Firms deliberate and decide which procedures are

[74]See Alchian (1950), Becker (1976), Day and Groves (1975), Farrell (1970), Nelson and Winter (1975), and Winter (1964, 1975).

[75]See Hirshleifer (1977) for an insightful discussion of this renaissance.

[76]In his critique of the methodological foundations of economics, Winter (1964) uses a different nomenclature and refers to routines or "rule of thumb" procedures as the firm's genes.

[77]See Winter (1964).

most likely to survive, i.e., yield positive profits. The results of these deliberations are "tried out" and, if found erroneous, are quickly eliminated. Thus, evolution has become endogenous to the firm itself. Firms that continue to err or who are unable to adapt will of course cease to exist. But because firms can rapidly adapt *via* the trial and error process the business environment might be viewed as benign relative to the Darwinian environment. Furthermore, the ability to survive mistakes (the mistake is eliminated instead of the entrepreneur) implies that an efficient firm can evolve more rapidly than an efficient organism.[78]

7.2. *Farrell's evolutionary speculator*

As an application of evolutionary logic to economics, consider the Farrell (1970) speculator model. At the start of the process, the speculator has one dollar. If his first speculation is successful he acquires two dollars; if unsuccessful he is ruined. Let p be the probability of winning and assume that the speculator continues to speculate each dollar in independent ventures. Each venture has a probability p of success. Using branching processes it can be shown that the probability of ruin or extinction is the smallest non-negative root of [79]

$$g(s)=s,$$

where $g(s)$, the generating function for the speculator's individual gamble, is given by

$$g(s)=1-p+ps^2.$$

Therefore, the extinction probability is the solution to

$$(s-1)(s-(1-p)/p)=0,$$

and equals $(1-p)/p$ if $p>\frac{1}{2}$ and unity if $p\leqslant\frac{1}{2}$.

Farrell notes that if, initially, there are n individuals with ability $p_i>\frac{1}{2}$, then the probability of group extinction is $[(1-p_i)/p_i]^{n_i}$. As a consequence, all groups with $p_i<\frac{1}{2}$ do become extinct, but a large group with $p_i=\frac{1}{2}+\varepsilon$ has a large survival probability relative to a single individual with p_i close to unity.

[78]Karl Popper (1972) is an advocate of this perspective: "It is different with primitive man, and with the amoeba. Here there is no critical attitude, and so it happens more often than not that natural selection eliminates a mistaken hypothesis or expectation by eliminating those organisms which hold it, or believe in it. So we can say that the critical or rational method consists in letting our hypothesis die in our stead."

[79]See Karlin and Taylor (1975).

7.3. *Nelson and Winter's evolutionary model of the factor market*

Nelson and Winter have made a number of contributions to evolutionary economics. In Nelson and Winter (1975), they designed an evolutionary model in which essentially neoclassical conclusions regarding the effect of factor prices on factor ratios were obtained without recourse to the notions either of profit maximization or of equilibrium. A direct consequence of the neoclassical assumption of profit maximization by each firm is that a competitive industry does, in fact, achieve an equilibrium. Moreover, in equilibrium only efficient firms survive, with each firm earning a normal rate of return. In the evolutionary paradigm, however, competition and maximization are viewed only as tendencies. Of course, competitive pressures are present and exert influence so that they do (tend to) regulate firms in the neoclassical sense.

The major advantage of the evolutionary approach is that it explicitly admits the possibility of disequilibrium behavior from which stems its ability to model disequilibrium phenomena. The major concern rests not with the difficulty of formulating an appropriate model but rather with the analytic power associated with such a model. We shall now present Nelson and Winter's model and, while deriving the standard neoclassical relationship between factor prices and factor ratios without invoking either profit maximization or industry equilibrium, indicate that this concern about analytic power is not well founded.[80]

Their model incorporates two dynamic mechanisms: search and selection. The profit motive induces firms to look (search) for better (less costly) production techniques. Selection also operates through the profit motive, for the more profitable firms are "selected" to produce more output. More precisely, we assume that each firm is completely characterized by four quantities: its capital–labor ratio r, its technology efficiency coefficient t, its size S, and its constant marginal cost c. As there are to be no economies of scale, the marginal cost c is not a function of the firm size S. When convenient, we shall replace the four-tuple (r, t, S, c) for a firm by f. For convenience we shall also replace the per unit cost c by the per unit profit π as $\pi = p - c$, where p is the sales price. The industry is characterized by three quantities: the prevailing wage rate w, the cumulative distribution function D of firms in the industry (so D has r, t, S, c as its arguments), and a "Darwinian" selection function s. The selection function s is simply a non-negative non-decreasing function such that the size S' of the firm $f = (r, t, S, c)$ in the next period is given by

$$S' = Ss(\pi). \tag{7.1}$$

It is in this sense that selection operates, and there are no barriers to entry.

[80] Whereas a number of economists are applying the tools of neoclassical analysis to biological phenomena, the direction of the flow, and the one we prefer, in the work of Nelson and Winter is in the other direction, viz., the evolutionary logic is applied to economic phenomena.

Given the wage rate w, the firm $f=(r, t, S, c)$ will undertake search for a new and more efficient technology (r, t). It is important to remember that in this search the firm might change both aspects of its technology. That is, it may change its capital–labor ratio r as well as its efficiency t. Let $F_{w,r}$ be the cumulative distribution of tomorrow's capital–labor ratio if the firm's current ratio is r and the industry's prevailing wage today is w. (Note that changes in a firm's ratio are independent not only of its efficiency t but also of its size S and profitability π.) We assume that

$$F_{w,r} \geqslant F_{w+\varepsilon,r} \quad \text{for} \quad \varepsilon>0. \tag{7.2}$$

That is, for any given capital–labor ratio employed by a firm today, tomorrow's capital–labor ratio stochastically increases with the wage rate. Furthermore, it is assumed that

$$F_{w,r} \geqslant F_{w,r+\varepsilon}, \quad \text{for} \quad \varepsilon>0. \tag{7.3}$$

That is, tomorrow's ratio tends to increase with today's ratio. From (7.2) and (7.3) we see that search is "local" in the sense that the newly discovered technique is likely to resemble the one in use.

In order to separate the effects of search and selection, Nelson and Winter assume that:

> the expected unit cost saving achieved by a firm as a result of today's search process is independent both of its capital–labor ratio today and the capital–labor ratio it adopts tomorrow, and also independent of firm size. This means that cost reduction is as easy for a firm at any one capital–labor ratio as at any other, and does not depend on the change in the capital–labor ratio. And it is as easy, or hard, for small firms as for big.

Continuing, they acknowledge that:

> this assumption is quite brazen (as is the assumption of neutrality of technical change in neoclassical models) and its only justification is that it is a powerful plank in building the overall theorem proving structure.

Consonant with this "brazen" assumption, denote by T_t the cumulative distribution function of tomorrow's efficiency coefficient given that today's efficiency coefficient is t. Implicit in our notation is the idea that the new coefficient is found independently of r, S, and w. Thus, $F_{w,r}T_t$ is the distribution of tomorrow's technology for a firm whose current technology is (r, t) when the prevailing wage rate is w. Finally, it would appear most reasonable to assume that

$$T_{t+\varepsilon} \leqslant T_t \quad \text{for all} \quad \varepsilon>0 \quad \text{and} \quad t>0, \tag{7.4}$$

and

$$T_t(x)=0 \quad \text{for} \quad x<t. \tag{7.5}$$

Equation (7.4) expresses the fact that even after search the efficient firms tend to remain more efficient, whereas (7.5) merely relates the fact that if a production method which is less efficient than the one currently being used is uncovered by the search then it will not be adopted. (The unpleasant aspect of this is that part but not all of the newly found technology—namely the part relating to the factor ratio—will be implemented.)

In separating the forces of search and selection, Nelson and Winter (1975, p. 477) assume "that the expected unit cost saving achieved by a firm as a result of today's search processes is independent both of its capital–labor ratio today and the capital–labor ratio it adopts tomorrow and also independent of firm size." While the (implicit) assumption of constant returns to scale is palatable, the per unit costs are inextricably entwined in the interplay between the ratio of the factor inputs and the ratio of the factor costs. Specifically, the per unit cost is a function of the technology and the wage rate (the cost of capital does not change in the ensuing analysis). Consequently, we adopt the weaker assumption that $c(r, t, w)$, the per unit cost as a function of the firm's technology (r, t) and the prevailing wage rate w, satisfies

$$\pi(r+\varepsilon, t, w)-\pi(r, t, w) \quad \text{is non-decreasing in } w \text{ for} \quad \varepsilon>0, \tag{7.6a}$$

or, more conveniently,

$$\frac{\partial^2 c(r, t, w)}{\partial w\, \partial r}<0. \tag{7.6b}$$

According to (7.6), capital intensive technologies are preferable to labor intensive technologies for high wage rates and inferior for low wage rates.

Suppose factor inputs of K units of capital and L units of labor produce $K^{\alpha}L^{1-\alpha}/t$ units of output, where α, the elasticity of output with respect to capital, is strictly between 0 and 1. Then there are constant returns to scale, and

$$c(r, t, w)=\frac{wL+\gamma K}{K^{\alpha}L^{1-\alpha}/t}=t\frac{w+\gamma r}{r^{\alpha}} \tag{7.7}$$

satisfies (7.6), where γ is to be interpreted as the cost of capital. Interpreting γ as before, another example of a cost function satisfying (7.6) is given by

$$c(r, t, w)=\frac{wL+\gamma K}{(K+L)/t}=t\frac{w+\gamma r}{1+r}. \tag{7.8}$$

This cost function obtains when inputs of K and L induce an output of $(K+L)/t$.[81] One advantage of having c given by (7.8) as opposed to (7.7) is that there exists a critical wage $\bar{w}$ such that ($\varepsilon > 0$)

$$\begin{aligned} \pi(r+\varepsilon, t, w) - \pi(r, t, w) &< 0 \quad \text{for all } r \quad \text{if} \quad w < \bar{w} \\ &= 0 \quad \text{for all } r \quad \text{if} \quad w = \bar{w} \\ &> 0 \quad \text{for all } r \quad \text{if} \quad w > \bar{w}. \end{aligned} \tag{7.9}$$

The (indifference) wage rate $\bar{w}$ is the wage rate at which profit is independent of the ratio.[82]

Whereas the prevailing wage rate w is an exogenously given constant, the distribution D_i of firms in the industry at time i evolves in accord with the dynamics of the model as expressed by (7.1) to (7.5). Nelson and Winter (1975, p. 478) pose the following question: "What would be the effect of a (permanent) increase in the wage rate?"

Using only very elementary facts about finite state Markov chains, Nelson and Winter obtain surprisingly strong conclusions. First, by assuming that there are but a finite set of possible ratios and each distribution $F_{w,r}$ attaches positive probability to each of these ratios, the resultant finite state Markov chain is irreducible and aperiodic so that there is a steady state or equilibrium. It easily follows from (7.2) that an increase in the prevailing wage rate will (stochastically) increase the steady state distribution of capital–labor ratios for each firm. An even stronger statement can be made: an increase in the wage rate will cause the distribution of the ratio to increase for each firm at each point in time. [It is apparent that the strength of the above relationships is further augmented by adding the assumptions embodied in (7.8) to the model.]

Finally, the size of the labor force used by a firm f whose ratio and size are r_f and S_f is simply $S_f/(1+r_f)$. Consequently, given the distribution D of firms in the industry, the size L_D of the labor force for that particular period is given by

$$L_D = \int S_f/(1+r_f)\,\mathrm{d}D(f). \tag{7.10}$$

Hence, given the labor supply curve $\mathbb{S}$ and the distribution D of firms, the prevailing wage rate could be endogenously defined by $\mathbb{S}(L_D)$ rather than exogenously specified.

[81] If c is as given in (7.7) or (7.8), then (7.5) could be replaced by the assumption that the new technology (r', t') is adopted in favor of (r, t) only if $c(r', t', w) < c(r, t, w)$.

[82] Notice that the analytical attractiveness of (7.8) as exhibited in the existence of an indifference wage is purchased at the expense of empirical realism in that the production function $q = K + L$ associated with (7.8) has infinite elasticity of substitution. On the other hand, the Cobb–Douglas production function associated with (7.7) is more agreeable, for changes in factor prices would not lead from specialization in one factor to specialization in the other. In this case the elasticity of substitution between factors is unity.

References

The list of references has been partitioned in accord with the seven sections of this chapter so as to facilitate location of specific items during reading. Moreover, because each section is self-contained, this partition provides a concise yet complete guide for each topic.

References for Section 1

Arrow, K. J. (1958), "Utilities, attitudes, choices: A review note", Econometrica, 26:24–36.

Arrow, K. J. (1971), Essays in the theory of risk bearing. Amsterdam: North-Holland.

Balch, M. S., D. L. McFadden and S. Y. Wu, eds. (1974), Essays in economic behavior under uncertainty. Amsterdam: North-Holland.

Black, F. and M. Scholes (1973), "The pricing of options and corporate liabilities", Journal of Political Economy, 81:637–654.

Borch, K. H. (1968), The economics of uncertainty. Princeton, NJ: Princeton University Press.

Brock, W. (1981), "Asset prices in a production economy", in: J. J. McCall, ed., The economics of information and uncertainty. Chicago, IL: University of Chicago Press for the National Bureau of Economic Research.

Cass, D. and K. Shell, eds. (1976), The Hamiltonian approach to dynamic economics. New York: Academic Press.

de Finetti, B. (1974), Theory of probability, Vol. 1. New York: Wiley.

Diamond, P. and M. Rothschild, eds. (1978), Uncertainty in economics. New York: Academic Press.

Englebrecht-Wiggams, R. (1978), "Auctions and bidding models: A survey", Cowles discussion paper no. 496. New Haven, CT: Yale University.

Green, J. R. (1973), "Temporary general equilibrium in a sequential trading model with spot and futures transactions", Econometrica, 41:1103–1124.

Harris, M. and A. Raviv (1979), "Optimal incentive contracts with imperfect information", Journal of Economic Theory, 20:231–259.

Hart, A. G. (1942), "Risk, uncertainty, and unprofitability of compounding probabilities", in: O. Lange, F. McIntyre and T. O. Yntema, eds., Studies in mathematical economics and econometrics, pp. 110–118. Chicago, IL: University of Chicago Press.

Hicks, J. R. (1931), "The theory of uncertainty and profit", Economica, 11:170–189.

Hirshleifer, J. and J. G. Riley (1979), "The analytics of uncertainty and information—An expository survey", Journal of Economic Literature, 17:1375–1421.

Jones, R. and J. Ostroy (1976), "Liquidity as flexibility", Department of Economics discussion paper no. 73. Los Angeles, CA: University of California.

Knight, F. H. (1921), Risk, uncertainty, and profit. New York: Houghton-Mifflin.

Levhari, D. and E. Sheshinski (1974), "The economics of queues: A brief survey", in: M. S. Balch, D. L. McFadden and S. Y. Wu, eds., Essays in economic behavior under uncertainty, pp. 195–212. Amsterdam: North-Holland.

Lippman, S. A. and J. J. McCall (1979), "The economics of uncertainty: Selected topics and probabilistic methods", Western Management Science Institute working paper no. 281. Los Angeles, CA: University of California.

Makower, H. and J. Marschak (1938), "Assets, prices, and monetary theory", Economica, 5:261–288.

Marschak, T. (1962), "Strategy and organization in a system development project", in: National Bureau of Economic Research, The rate and direction of inventive activity. Princeton, NJ: Princeton University Press.

McCall, J. J. (1971), "Probabilistic microeconomics", Bell Journal of Economics and Management Science, 2:403–433.

Nelson, R. (1961), "Uncertainty, learning, and the economics of parallel research and development efforts", Review of Economics and Statistics, 43:351–364.

Popper, K. (1972), Objective knowledge: An evolutionary approach. Oxford: Clarendon Press.

Pratt, J. W. (1964), "Risk aversion in the small and in the large", Econometrica, 32:122–146.
Radner, R. (1968), "Competitive equilibrium under uncertainty", Econometrica, 36:31–58.
Riley, J. G. (1975), "Competitive signalling", Journal of Economic Theory, 10:175–186.
Ross, S. A. (1973), "The economic theory of agency: The principal's problem", American Economic Review, 63:134–39.
Rubinstein, M. (1979), Option markets. Englewood Cliffs, NJ: Prentice Hall.
Shell, K., ed. (1967), Essays on the theory of optimal economic growth. Cambridge, MA: M.I.T. Press.
Spence, M. (1973), "Job market signaling", Quarterly Journal of Economics, 87:355–374.
Stigler, G. (1939), "Production and distribution in the short run", Journal of Political Economy, 47:305–328.
Vickrey, W. (1961), "Counterspeculation, auctions and competitive sealed tenders", Journal of Finance, 41:8–37.
Ziemba, W. T. and R. G. Vickson, eds. (1975), Stochastic optimization models in finance. New York: Academic Press.

References for Section 2

Alchian, A. (1970), "Information costs, pricing, and resource employment", in: Edmund S. Phelps, ed., Microeconomic foundations of employment and inflation theory, pp. 27–52. New York: W. W. Norton.
Burdett, K. (1978), "A theory of employee job search and quit rates", American Economic Review, 68:212–220.
Chow, Y. S. and H. Robbins (1961), "A martingale systems theorem and applications", Proceedings of the 4th Berkeley symposium, pp. 93–104. Berkeley, CA: University of California Press.
Chow, Y. S., H. Robbins and D. Siegmund (1971), Great expectations: The theory of optimal stopping. New York: Houghton-Mifflin.
Chung, K. L. (1968), A course in probability theory. New York: Hartcourt, Brace and World.
Classen, K. (1979), "Unemployment insurance and job search", in: S. A. Lippman and J. J. McCall, eds., Studies in the economics of search, Chapter 10. Amsterdam: North-Holland.
Danforth, J. (1979), "On the role of consumption and decreasing absolute risk aversion in the theory of job search", in: S. A. Lippman and J. J. McCall, eds., Studies in the economics of search, Chapter 6. Amsterdam: North-Holland.
DeGroot, H. (1970), Optimal statistical decision. New York: McGraw-Hill.
DeGroot, H. (1968), "Some problems of optimal stopping", Journal of the Royal Statistical Society, B 30:108–122.
Diamond, P. and E. Maskin (1979), "An equilibrium analysis of search and breach of contract, I: Steady states", Bell Journal of Economics, 10:282–318.
Gordon, D. F. (1974), "A neo-classical theory of Keynesian unemployment", Economic Inquiry, 12:431–469.
Hall, J., S. A. Lippman and J. J. McCall (1979), "Expected utility maximizing job search", in: S. A. Lippman and J. J. McCall, eds., Studies in the economics of search, Chapter 7. Amsterdam: North-Holland.
Karni, E. and A. Schwartz (1977), "Search theory: The case of search with uncertain recall", Journal of Economic Theory, 16:38–52.
Kohn, G. and S. Shavell (1974), "The theory of search", Journal of Economic Theory, 9:93–123.
Kormendi, R. (1979), "Dispersed transactions prices in a model of decentralized exchanges", in: S. A. Lippman and J. J. McCall, eds., Studies in the economics of search, Chapter 4. Amsterdam: North-Holland.
Landsberger, M. and D. Peled (1977), "Duration of offers price structure and the gain from search", Journal of Economic Theory, 16:17–37.
Lippman, S. A. (1975), "On dynamic programming with unbounded rewards," Management Science, 21:1225–1233.
Lippman, S. A. and J. J. McCall (1976a), "Job search in a dynamic economy", Journal of Economic Theory, 12:365–390.

Lippman, S. A. and J. J. McCall (1976b), "The economics of job search: A survey", Economic Inquiry, 14:155–189, 347–368.
Lippman, S. A. and J. J. McCall, eds. (1979), Studies in the economics of search. Amsterdam: North-Holland.
Lippman, S. A. and J. J. McCall (1981), "The economics of belated information", International Economic Review.
Marston, S. (1975), "The impact of unemployment insurance on job search", Brookings Papers on Economic Activity, 1:13–60.
McCall, J. J. (1965), "The economics of information and optimal stopping rules", Journal of Business, 38:300–317.
McCall, J. J. (1970), "Economics of information and job search", Quarterly Journal of Economics, 84:113–126.
Mortensen, D. T. (1970), "Job search, the duration of unemployment and the Phillips curve", American Economic Review, 60:847–862.
Mortensen, D. T. (1978), "Specific capital and labor turnover", Bell Journal of Economics, 9:572–586.
Phelps, E. S., ed. (1970), Microeconomic foundations of employment and inflation theory. New York: W. W. Norton.
Robbins, H. (1970), "Optimal stopping", American Mathematical Monthly, 77:333–343.
Rosenfield, D. B. and R. D. Shapiro (1981), "Optimal price search with Bayesian extensions", Journal of Economic Theory.
Rothschild, M. (1974a), "Models of market organization with imperfect information: A survey", Journal of Political Economy, 81:1283–1308.
Rothschild, M. (1974b), "Searching for the lowest price when the distribution of prices is unknown", Journal of Political Economy, 82:689–711.
Salop, S. C. (1973), "Systematic job search and unemployment", Review of Economic Studies, 40:191–201.
Stigler, G. J. (1961), "The economics of information", Journal of Political Economy, 69:213–225.
Stigler, G. J. (1962), "Information in the labor market", Journal of Political Economy, 70:94–104.
Wilde, L. (1979), "An information theoretic approach to job quits", in: S. A. Lippman and J. J. McCall, eds., Studies in the economics of search, Chapter 3. Amsterdam: North-Holland.
Wilde, J. and A. Schwartz (1979), "Equilibrium comparison shopping", Review of Economic Studies, 46:543–553.

References for Section 3

Akerlof, G. (1970), "The market for lemons: Qualitative uncertainty and the market mechanism", Quarterly Journal of Economics, 90:488–500.
Alchian, A. A. and H. Demsetz (1972), "Production, information costs, and economic organization", American Economic Review, 62:777–795.
Arrow, K. J. (1971), Essays in the theory of risk-bearing. Amsterdam: North-Holland. In 1963, Chapter 8, "Uncertainty and the economics of medical care", already published in: American Economic Review, 53:941–973.
Borch, K. H. (1960), "Equilibrium in a reinsurance market", Econometrica, 30:162–184.
Borch, K. H. (1968), The economics of uncertainty. Princeton, NJ: Princeton University Press.
Buhlman, H. (1970), Mathematical methods in risk theory. New York: Springer-Verlag.
Ehrlich, I. and G. S. Becker (1972), "Market insurance, self-insurance and self-protection", Journal of Political Economy, 80:623–648.
Harris, M. and A. Raviv (1978), "Some results of incentive contracts with applications to education and employment, health insurance, and law enforcement", American Economic Review, 68:20–30.
Harris, M. and A. Raviv (1979), "Optimal incentive contracts with imperfect information", Journal of Economic Theory, 20:231–259.
Hirshleifer, J. (1970), Investment, interest and capital. Englewood Cliffs, NJ: Prentice Hall.
Jensen, M. and W. Meckling (1976), "Theory of the firm: Managerial behavior, agency costs and ownership structure", Journal of Financial Economics, 3:305–360.

Kihlstrom, R. and M. Pauly (1971), "The role of insurance in the allocation of risk", American Economic Review, 61:371–379.
Lippman, S. A. and J. J. McCall (forthcoming), The economics of insurance.
Pratt, J. W. (1964), "Risk aversion in the small and in the large", Econometrica, 32:122–146.
Ross, S. A. (1973), "The economic theory of agency: The principal's problem", American Economic Review, 63:134–139.
Ross, S. A. (1974), "On the economic theory of agency and the principle of similarity", in: M. S. Balch, D. L. McFadden and S. Y. Wu, eds., Essays in economic behavior under uncertainty, pp. 215–237. Amsterdam: North-Holland.
Rothschild, M. and J. E. Stiglitz (1976), "Equilibrium in competitive insurance markets: An essay on the economics of imperfect information", Quarterly Journal of Economics, 90:629–650.
Seal, H. L. (1969), Stochastic theory of a risk business. New York: Wiley.
Shavell, S. (1979), "On moral hazard and insurance", Quarterly Journal of Economics, 93:541–563.
Spence, A. M. and R. Zeckhauser (1971), "Insurance, information and individual action", American Economic Review, 61:380–387.
Stiglitz, J. E. (1974), "Incentives and risk sharing in sharecropping", Review of Economic Studies, 41:219–255.
Wilson, C. (1977), "A model of insurance with incomplete information", Journal of Economic Theory, 16:167–207.

References for Section 4

Boulding, K. E. (1966), Economic analysis, Vol. I: Microeconomics, 4th ed. New York: Harper and Row.
Diamond, P. and J. Stiglitz (1974), "Increases in risk and in risk aversion", Journal of Economic Theory, 8:337–360.
Drèze, J. H. and F. Modigliani (1972), "Consumption decisions under uncertainty", Journal of Economic Theory, 5:308–335.
Fama, E. F. (1970), "Multiperiod consumption–investment decisions", American Economic Review, 60:163–174.
Hahn, F. H. (1970), "Savings and uncertainty", Review of Economic Studies, 37:21–24.
Hakansson, N. H. (1970), "Optimal investment and consumption strategies under risk for a class of utility functions", Econometrica, 38:587–607.
Kihlstrom, R. E. and L. J. Mirman (1974), "Risk aversion with many commodities", Journal of Economic Theory, 8:361–388.
Leland, H. E. (1968), "Savings under uncertainty: The precautionary demand for saving", Quarterly Journal of Economics, 82:465–473.
Levhari, D. and L. J. Mirman (1977), "Savings and uncertainty with an uncertain horizon", Journal of Political Economy, 85:265–281.
Levhari, D. and T. N. Srinivasan (1969), "Optimal savings under uncertainty", Review of Economic Studies, 36:153–164.
Levhari, D. and T. N. Srinivasan (1972), "Optimal savings and portfolio choice under uncertainty", in: G. Szegö and K. Shell, eds., Mathematical methods in investment and finance, pp. 34–48. Amsterdam: North-Holland.
Marshall, A. (1920), Principles of economics, 8th ed. London: Macmillan.
Merton, R. C. (1971), "Optimum consumption and portfolio rules in a continuous-time model", Journal of Economic Theory, 3:373–413.
Miller, B. L. (1974), "Optimal consumption with a stochastic income stream", Econometrica, 42:253–266.
Miller, B. L. (1976), "The effect on optimal consumption of increased uncertainty in labor income in the multiperiod case", Journal of Economic Theory, 13:154–167.
Mirman, L. J. (1971), "Uncertainty and optimal consumption decisions", Econometrica, 39:179–185.
Mukherjee, R. and E. Zabel (1974), "Consumption and portfolio choices with transaction costs", in: M. S. Balch, D. L. McFadden and S. Y. Wu, eds., Essays in economic behavior under uncertainty, pp. 157–184. Amsterdam: North-Holland.

Phelps, E. S. (1962), "The accumulation of risky capital: A sequential utility analysis", Econometrica, 30:729–743.
Raiffa, H. and R. Schlaifer (1964), Applied statistical decision theory. Cambridge, MA: Harvard University Press.
Rothschild, M. and J. Stiglitz (1971), "Increasing risk II: Its economic consequences", Journal of Economic Theory, 3:66–84.
Sandmo, A. (1969), "Capital risk, consumption, and portfolio choice", Econometrica, 37:586–599.
Sandmo, A. (1970), "The effect of uncertainty on savings decisions", Review of Economic Studies, 37:353–360.
Yaari, M. E. (1964), "Uncertain lifetime, life insurance and the theory of the consumer", Review of Economic Studies, 32:137–158.
Yaari, M. E. (1976), "A law of large numbers in the theory of consumer's choice under uncertainty", Journal of Economic Theory, 12:202–217.

References for Section 5

Arzac, E. R. (1976), "Profits and safety in the theory of the firm under price uncertainty", International Economic Review, 17:163–171.
Baron, D. P. (1970), "Price uncertainty, utility, and industry equilibrium in pure competition", International Economic Review, 11:463–480.
Baron, D. P. (1971), "Demand uncertainty in imperfect competition", International Economic Review, 12:196–208.
Baron, D. P. (1973), "Point estimation and risk preferences", Journal of the American Statistical Association", 68:944–950.
Baron, D. P. and R. Forsythe (1979), "Models of the firm and international trade under uncertainty", American Economic Review, 69:565–574.
Baron, D. P. and R. A. Taggart, Jr. (1977), "A model of regulation under uncertainty and a test of regulatory bias", Bell Journal of Economics, 8:151–167.
Batra, R. N. and A. Ullah (1974), "Competitive firm and the theory of input demand under price uncertainty", Journal of Political Economy, 82:537–548.
Blair, R. D. (1974), "Random input prices and the theory of the firm", Economic Inquiry, 12:214–226.
Diamond, P. A. (1967), "The role of a stock market in a general equilibrium model with technological uncertainty", American Economic Review, 57:759–776.
Diamond, P. A. and J. E. Stiglitz (1974), "Increases in risk and risk aversion", Journal of Economic Theory, 8:337–360.
Hartley, P. (1978), "Models of the firm under uncertainty". Chicago, IL: University of Chicago.
Hartman, R. (1975), "Competitive firm and the theory of input demand under pure uncertainty: Comment", Journal of Political Economy, 83:1289–1290.
Horowitz, I. (1970), Decision making and the theory of the firm. New York: Holt, Rinehart and Winston.
Ishii, Y. (1977), "On the theory of the competitive firm under price uncertainty: Note", American Economic Review, 67:768–769.
Leland, H. (1972), "Theory of the firm facing random demand", American Economic Review, 62:278–291.
Leland, H. (1974a), "Regulation of natural monopolies and the fair rate of return", Bell Journal of Economics and Management Science, 5:3–15.
Leland, H. (1974b), "Production theory and the stock market", Bell Journal of Economics and Management Science, 5:125–144.
Lippman, S. A. and J. J. McCall (1979), "The economics of uncertainty: Selected topics and probabilistic methods", Western Management Science Institute working paper no. 281. Los Angeles, CA: University of California.
Lippman, S. A. and J. J. McCall (1981), "Competitive production and increases in risks", American Economic Review, 71.

McCall, J. J. (1971), "Probabilistic microeconomics", Bell Journal of Economics and Management Science, 2:403–433.
McCall, J. J. (1967), "Competitive production for constant risk utility functions", Review of Economic Studies, 34:417–420.
Mills, E. S. (1959), "Uncertainty and price theory", Quarterly Journal of Economics, 73:116–130.
Mills, E. S. (1962), Price, output, and inventory policy. New York: Wiley.
Myers, S. C. (1973), "A simple model of firm behavior under regulation and uncertainty", Bell Journal of Economics and Management Science, 4:304–315.
Nelson, R. (1961), "Uncertainty, prediction and competitive equilibrium", Quarterly Journal of Economics, 75:41–62.
Penner, R. G. (1967), "Uncertainty and the short-run shifting of the corporation tax", Oxford Economic Papers, 19:99–110.
Pratt, J. S. (1964), "Risk aversion in the small and in the large", Econometrica, 32:122–136.
Rothenberg, T. and K. Smith (1971), "The effect of uncertainty on resource allocation", Quarterly Journal of Economics, 82:440–453.
Rothschild, M. and J. E. Stiglitz (1970), "Increasing risk: I: A definition", Journal of Economic Theory, 2:225–243.
Rothschild, M. and J. E. Stiglitz (1971), "Increasing risk: II: Its economic consequences", Journal of Economic Theory, 3:66–84.
Roy, A. D. (1952), "Safety first and the holding of assets", Econometrica, 20:431–449.
Sandmo, A. (1971), "On the theory of the competitive firm under price uncertainty", American Economic Review, 61:65–73.
Sheshinski, E. and J. H. Drèze (1976), "Demand fluctuations, capacity utilization, and costs", American Economic Review, 66:731–742.
Telser, L. G. (1955), "Safety first and hedging", Review of Economic Studies, 23:1–16.
Turnovsky, S. J. (1973), "Production flexibility, price uncertainty and the behavior of the competitive firm", International Economic Review, 14:395–413.
Zabel, E. (1967), "A dynamic model of the competitive firm", International Economic Review, 8:194–208.
Zabel, E. (1971), "Risk and the competitive firm", Journal of Economic Theory, 3:109–133.

References for Section 6

Anderson, R. F. and A. Friedman (1977), "Optimal inspections in a stochastic control problem with costly information", Mathematics of Operations Research, 2:155–190.
Arrow, K. J. (1971), Essays in the theory of risk bearing. Amsterdam: North-Holland.
Arrow, K. J., S. Karlin and H. Scarf (1958), Studies in the mathematical theory of inventory and production. Stanford, CA: Stanford University Press.
Black, F. and M. Scholes (1973), "The pricing of option and corporate liabilities", Journal of Political Economy, 81:637–654.
Breiman, L. (1968), Probability. Reading, MA: Addison-Wesley.
Cootner, P., ed. (1967), The random character of stock market prices. Cambridge, MA: M.I.T. Press.
Eppen, G. D. and E. F. Fama (1969), "Cash balances and simply dynamic portfolio problems with proportional costs", International Economic Review, 10:119–133.
Fama, E. F. (1970), "Efficient capital markets: A review of theory and empirical work", Journal of Finance, 25:383–417.
Fischer, S. (1975), "The demand for index bonds", Journal of Political Economy, 83:509–534.
Gikhmann, I. I. and A. V. Skorokhod (1969), Introduction to the theory of random processes. Philadelphia, PA: Saunders.
Granger, C. W. J. (1972), "Empirical studies of capital markets: A survey", in: G. Szegö and K. Shell, eds., Mathematical methods in investment and finance, pp. 464–519. Amsterdam: North-Holland.
Harrison, J. M. and A. J. Taylor (1978), "Optimal control of a Brownian storage system", Stochastic Processes and Their Applications, 6:179–194.

Jensen, M. C., ed. (1972), Studies in the theory of capital markets. New York: Praeger.
Karlin, S. and H. M. Taylor (1975), A first course in stochastic processes, 2nd ed. New York: Academic Press.
Lamperti, J. (1966), Probability. New York: Benjamin.
Malliaris, A. and W. Brock (1982), Stochastic calculus with applications in economics and finance. Amsterdam: North-Holland.
McCall, J. J. (1965), "Maintenance policies for stochastically failing equipment: A survey", Management Science, 11:493–524.
Merton, R. C. (1976), "Option pricing when underlying stock returns are discontinuous", Journal of Financial Economics, 3:125–144.
Merton, R. C. (1981), "On the microeconomic theory of investment under uncertainty," in: K. J. Arrow and M. D. Intriligator, eds., Handbook of mathematical economics. Amsterdam: North-Holland.
Miller, M. H. and D. Orr (1966), "A model of the demand for money by firms", Quarterly Journal of Economics, 80:413–435.
Mossin, J. (1968), "Optimal multiperiod portfolio policies", Journal of Business, 41:215–229.
Pierskalla, W. P. and J. A. Voelker (1976), "A survey of maintenance models: The control and surveillance of deteriorating systems", Naval Research Logistics Quarterly, 23:353–388.
Rubinstein, M. (1979), Option markets. Englewood Cliffs, NJ: Prentice Hall.
Sharpe, W. F. (1970), Portfolio theory and capital markets. New York: McGraw-Hill.
Smith, C. W. (1976), "Option pricing: A review", Journal of Financial Economics, 3:3–51.
Taylor, A. S. (1978), "Optimum impulse control of a Brownian storage system", School of Business working paper no. 77-22. Kingston, Ont.: Queen's University.
Vial, J. P. (1972), "A continuous time model for the cash balance problem", in: G. Szegö and K. Shell, eds., Mathematical methods in investment and finance. Amsterdam: North-Holland.
Whitt, W. (1973), "Diffusion models for inventory and production systems". New Haven, CT: Yale University.
Ziemba, W. T. and R. G. Vickson, eds. (1975), Stochastic optimization models in finance. New York: Academic Press.

References for Section 7

Alchian, A. A. (1950), "Uncertainty, evolution, and economic theory", Journal of Political Economy, 58:211–221.
Becker, G. S. (1976), "Altruism, egoism and genetic fitness: Economics and sociobiology", Journal of Economic Literature, 14:817–826.
Day, R. H. and T. Groves, eds. (1975), Adaptive economic models. New York: Academic Press.
Farrell, M. J. (1970), "Some elementary selection processes in economics", Review of Economic Studies, 37:305–319.
Hirshleifer, J. (1977), "Economics from a biological viewpoint", Journal of Law and Economics, 20:1–52.
Karlin, S. and H. M. Taylor (1975), A first course in stochastic processes, 2nd ed. New York: Academic Press.
Nelson, R. R. and S. G. Winter (1975), "Factor price changes and factor substitution in an evolutionary model", Bell Journal of Economics, 6:466–486.
Popper, Karl (1972), Objective knowledge: An evolutionary approach. Oxford: Clarendon Press.
Winter, S. G. (1964), "Economic 'Natural selection' and the theory of the firm", Yale Economic Essays, 4:225–272.
Winter, S. G. (1975), "Optimization and evolution in the theory of the firm", in: R. H. Day and T. Groves, eds., Adaptive economic models, pp. 73–118. New York: Academic Press.

Chapter 7

GAME THEORY MODELS AND METHODS IN POLITICAL ECONOMY*

MARTIN SHUBIK

Yale University

An overview of different models and solution concepts of game theory is given together with a sketch of the major areas of application to political economy, as well as with some indications of major open problems. Although many references are given the reader may find more detail and references in a series of Rand reports on game theory in economics.[1]

1. Modelling methods

Perhaps the most important aspect of the theory of games as applied to political economy is that the methodology it provides for constructing the mathematical models for the study of conflict and cooperation forces is of an explicitness found rarely even in the many mathematical investigations of political economy. In particular, the *extensive form* of a game calls for a complete process description.

A fully described game should be playable without too much difficulty by a group of students. If the game is well defined but difficult to play because of unreasonable demands on the time and data processing abilities of the individuals it well may be that it is not a good model of the economic process it purports to represent. Thus the discipline called for in building a well defined and playable game may enable the economist to formulate an operationally valuable critique of his initial model and to isolate factors which were not deemed to be important until the criterion of "playable" was applied.

A *game of strategy* is one with two or more players each with partial control over the environment, where, in general, the payoff to each player depends not only upon his actions but also upon the actions of others.

*This work relates to Department of the Navy Contract N0014-77-C-0518 issued by the Office of Naval Research Under Contract Authority NR 277-239. However, the content does not necessarily reflect the position or the policy of the Department of the Navy or the Government, and no official endorsement should be inferred.

[1]Much of this work is based upon portions of an unpublished manuscript by Shapley and Shubik (1971–74).

There are three highly different representations of a game of strategy, the *extensive form*, the *strategic form*,[2] and the *cooperative form*.[3] Each serves a different purpose, i.e., they are designed to answer different questions and hence utilize different descriptions of the phenomena being studied. A game in extensive form can be used to specify the game in strategic form, and one in strategic form may be used to define the cooperative form, but the reverse does not hold true. There may be many different games in strategic form which give rise to the same cooperative form. It is best to think of the three forms as three independent formulations designed for different purposes, which, when the occasion calls for it, can be related to each other.

Before we consider these forms we must note the assumptions made concerning preferences, utility, and payoffs.

1.1. Preferences, utility, and payoffs

Von Neumann and Morgenstern (1944) presented axioms for the existence of a utility function defined up to a linear transformation, based upon considering gambles among a set of outcomes over which an individual has an ordering of preference. This utility measure was utilized by von Neumann and Morgenstern in their evaluation of the employment of mixed strategies in games in strategic form. Totally independent of this construction and its use was the assumption concerning the existence of some form of "U-money" or transferable utility which served to enable them to provide a particularly simple description of a game in cooperative form.

In many of the earlier critiques of the applicability of game theory to economics and other disciplines doubts were raised concerning the value of game theory to the behavioral sciences because of the two assumptions which were deemed to be highly unrealistic. As a greater understanding of the importance of decision making under uncertainty and the strength of the various axiom systems which lead to a utility measure has come about, so has the acceptance of the von Neumann–Morgenstern position.

Concerning cooperative games and transferable utility, as was noted by von Neumann and Morgenstern, the assumption of transferable utility was a preliminary simplification made in order to start to open up analysis in what promised to be an extremely complex domain of mathematics. The conceptual framework of the various cooperative theories for the *solution*[4] of games in no way depends critically upon this assumption. Although Shapley and Shubik

[2] Also known as the *normal form*.

[3] Also known as the *characteristic function form*.

[4] Solutions are discussed in detail in Section 2. A "solution" is essentially a selection of a subset of the outcomes with special properties.

(1971–74) pointed out the possibility of studying solutions to cooperative games without sidepayments, and Nash (1953) and Harsanyi (1959) presented analyses of bargaining, it was not until the work of Aumann and Peleg (1960) that this became a practical possibility.

In general, a game leads to some set of outcomes, and the individuals are assumed to have some sort of preferences with respect to those outcomes. The specialization of game theory in application to specific disciplines comes about to a great extent in making the appropriate assumptions which provide an appropriate structure to the set of outcomes. For example, *market games*[5] reflect the special structure of a barter economy with individualistic preferences, and *simple games*[6] have a natural interpretation in terms of voting.

1.2. The extensive form

The literature on oligopoly, auctions, bargaining, and international trade especially, whether verbal or mathematical, is replete with partial or complete descriptions of processes. Offers, counteroffers, threats, promises, demands, etc. are all critical features in the description of processes. Game theory provides a formal language for dealing with the description of the rules of the game which enables us to lay out the details of process with great precision. This is the language used to describe a game in extensive form.

The earliest descriptions of games in extensive form were given by von Neumann and Morgenstern (1944) and then Kuhn (1953). Both deal with finite games, i.e., games in which the number of players, moves, and choices are all finite. Chess or Poker serve as examples. Many of the situations faced in economics or in politics are only crudely modelled as finite games. Usually they have continuous strategic possibilities, continuous time, and the possibility of an indefinite continuation into the future.

A simple duopolistic market is illustrated by a Kuhn game tree as a finite approximation.[7] Suppose that two firms must each select one among three levels of production simultaneously. Their levels of production determine the market outcome and the payoffs to both. Figure 1.1 presents an extensive form description of this game. The diagram shows a *rooted tree* with the initial node marked by R. Each node represents a state in which the game might be found by an observer. Each node or vertex is labelled with a P_i or O_j indicating that it is either a decision point for a player or an outcome of the game. Any node labelled P_i is a decision point for player i. He must select one of the branches leading out of that node.

[5]Shapley and Shubik (1969b).
[6]Shapley (1962a).
[7]Production is assumed to yield discrete rather than continuous levels of output.

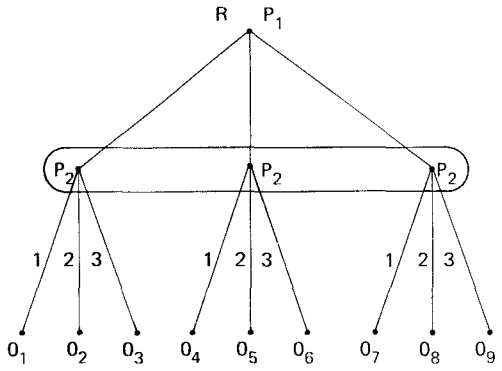


Figure 1.1

In the example above Player 1 has the first choice; he must select as his move one of the three branches leading out of the node marked P_1. After his move the play progresses to one of the three nodes labelled P_2. Player 2 makes his choice and the game will then reach one of the final nine nodes labelled O_j which are the outcomes. Any path from the initial node of the tree to a terminal node represents a possible *play* of the game.

In many situations we may require that players move simultaneously. In general, our concern is not with the formality that they select their moves at the same moment, but that they select their moves without information of what the other is doing. It does not matter who goes first as long as the other is not informed. We can illustrate this lack of information on the game tree by encasing all of the nodes among which a player cannot distinguish in a closed contour which indicates that these choice points belong to the same information set (or more descriptively the same "lack-of-information set"). In Figure 1.1 the three nodes of Player 2 are encased in a single set which means that when he is called upon to choose he does not know what Player 1's move has been.

We may enlarge our description to take care of exogenous uncertainty by adding an extra player called "Nature", distinguished by the name P_0. Whenever this player is called upon to move it selects a branch with given probabilities. A simple example is given in Figure 1.2. The single player P_1 must choose after which Nature determines the outcome.

A game with one point information sets only is said to be a game with *perfect information*. Chess is such a game. At any point in the play all players know all the details of the path to that point. This is not true for Poker, or for sealed bid auctions.

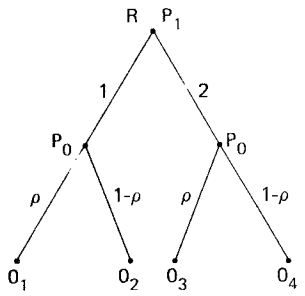


Figure 1.2

A *strategy* of a player is a complete plan of action which specifies what he is to do under all contingencies. In terms of our description of a game tree we can describe it as follows:

> A strategy is a function that associates with each of a player's information sets one of the alternatives issuing from that set.

A brief contemplation of the size of the game tree for the game of chess and the size of the set of strategies for the players in chess[8] will quickly convince anyone that except for games with few moves and choices the game tree is not in general of direct applied use. Furthermore the game-theoretic definition of strategy is clearly different from the definition that would be used by a general or political strategist in the sense that they implicitly have in mind a considerable aggregation of detail, as well as delegation of decision making to agents.

Even though it may not be possible to draw the game tree of a complex market process in detail, the formal method provided should serve to guide the modeller in making explicit the nature of his simplifications and abbreviations in his descriptions of process.

1.3. Games in strategic form

When a game is modelled in strategic form the details concerning moves and information are suppressed. Strategies are treated as primitive elements without any attempt being given to explain their genesis. The strategic form of an n-person game which in extensive form is representable by a finite game tree is

[8]In chess, for example, black has 20 alternatives for his first move thus white's contingent planning for his first move involves 20^{20} plans. Even for the game of Tic-Tac-Toe, leaving aside symmetry, the first player has between $9 \cdot 7^8 \cdot 5^6$ and $9 \cdot 7^8 \cdot 5^6 \cdot 3^4$ strategies or at least over 810 billion strategies.

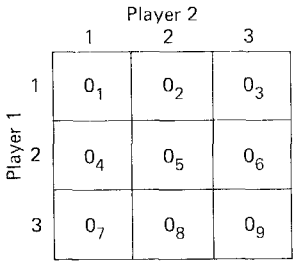


Figure 1.3

given by a set of n payoff matrices of dimension n. An example based on the game illustrated in Figure 1.1 will serve to illustrate the related form.

In Figure 1.3 the numbers on the left of the matrix are the strategies of Player 1 (which, because he has no information when he moves, coincide with his moves). The numbers at the top of the matrix are the strategies for Player 2. The entries in the nine cells are the payoffs. We may consider that O_j is generally a vector of n dimensions, indicating the payoffs to each player. Thus, in Figure 1.3, $O_1 = (5,4)$ would be interpreted as a payoff of 5 obtained by Player 1 and 4 by Player 2 if each uses his first strategy.

Suppose that the information set for Player 2 in Figure 1.1 were replaced by two information sets indicating that if Player 1 chooses 1 Player 2 is informed, but otherwise he does not know whether Player 1 chooses 2 or 3. The strategic form associated with this game is a matrix of size 3×9. The moves and outcomes are all the same as before, but the strategies for Player 2 now depend upon his extra knowledge. In particular he has 9 strategies which can be described as follows:

If Player selects 1 then Player 2 selects i,

if Player 1 selects 2 or 3 then Player 2 selects j.

Any $i = 1,2,3$ and $j = 1,2,3$ can be selected to give a strategy for Player 2 in this game.

Much of the experimental work on games has been devoted to experimentation with 2×2 matrix games. In particular many experiments have been run with the "Prisoners' Dilemma Game"[9] which is a 2×2 matrix game where the payoffs are as indicated in Figure 1.4, and where $a_i > b_i > c_i > d_i$ and $a_i + d_i < 2b_i$,[10] for $i = 1,2$.

Rapoport and Guyer (1966) have calculated that, confining themselves to strict orderings and eliminating symmetries, there are 78 strategically different

[9] Rapoport and Chammah (1965) and Rapoport, Guyer, and Gordon (1975).
[10] This last condition is needed to avoid mixed strategy equilibria.

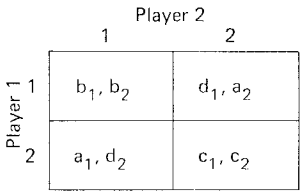


Figure 1.4

ordinal representations of a 2×2 matrix game. All of these games have been used for experimental purposes [Rapoport, Guyer and Gordon (1975)].

Simple 2×2 or 3×3 matrix games have been used considerably for expository and exploratory purposes, as is evinced by the work of Luce and Raiffa (1957) and Schelling (1960).

Most duopoly models or other economic models tend to use compact strategy sets, where in the simplest instances strategies and moves coincide. For example, a Cournot (1838) duopoly model calls for each player to select simultaneously a level of production, where the level of production may be any number within an interval. Thus if Player 1 selects x where $0 \leqslant x \leqslant A$, and Player 2 selects y where $0 \leqslant y \leqslant B$, then the payoffs to 1 and 2 are given by the two functions $f_1(x, y)$ and $f_2(x, y)$.

The use of compact strategy sets, especially in economics, comes about because frequently there is a natural structure present that is not present in most games in general. Chess, for example, cannot be modelled with continuous moves, but a wheat market can. Furthermore in many instances in economics there are natural ways to aggregate moves. Thus in a wheat market individual i may offer q_i units as his move but the outcome to him may depend only upon his offer and the total volume of wheat, or $q=\sum_{j=1}^{n} q_j$. In many games the addition of moves has no operational meaning.[11]

1.4. Games in cooperative or coalitional form

The stress in the presentation of a game in strategic form is upon the power of individuals in the sense that what they can obtain is a function of their strategies and the strategies of others. No particular attention is paid to explicit patterns of cooperation.

When we wish to study cartel formation, international trade or bargaining, or other group or sociological phenomena, the focus of attention may be upon the

[11] One way in which we might approach the economic description of a mass market is by axioms such as aggregation of moves.

possible gains from coalition formation without paying particular attention to information conditions, the details of why or how various strategic options are available, and the details and costs of coalition formation (provided they are deemed to be sufficiently low). Our attention may be focussed on the critical questions of how much groups have to gain from cooperation. This attention leads to formulating or presenting the game in cooperative or coalitional form.

As a simple illustration we may use the game given in strategic form in Figure 1.4. Two forms of this game in coalitional form are presented: the first makes use of an assumption of transferable utility, while the second does not use this assumption.

Let $v(S)$ stand for the amount that a coalition S of players can obtain together if they play as one. We denote the *characteristic function* by the letter v. It is a function from the subsets of players onto the real numbers. For an n-person game there are $2^n - 1$ coalitions that are non-empty.

The notation $v(\overline{ij})$ is used to denote a specific coalition consisting of players i and j. The characteristic function for the Prisoners' Dilemma game shown in Figure 1.4 is as follows [with $\emptyset$ being the set of no players]:

$$v(\emptyset)=0, \quad v(\overline{1})=c_1, \quad v(\overline{2})=c_2,$$
$$v(\overline{12})=\max[(b_1+b_2),(a_1+d_2),(a_2+d_1)].$$

The characteristic function may be regarded as a "presolution" to a game inasmuch as the act of calculating it provides considerable insight into the structure of the game. In this example the values have been calculated by asking what is the most that any coalition can achieve by itself on the assumption that the remaining players will try to minimize its payoff. The best that Player 1 or Player 2 can do alone is to employ his second strategy (see Figure 1.4) and obtain c_1 or c_2. Together the players can obtain b_1+b_2. In this instance it is fairly easy to see that it is reasonable to evaluate $v(\overline{1})$ at c_1 because Player 2, while minimizing the score of Player 1, is simultaneously optimizing his own score. This is not generally true as is shown in the game in Figure 1.5.

		Player 2	
		1	2
Player 1	1	5, 5	0, -100
	2	10, 5	-1, -1000

Figure 1.5

Here the characteristic function is given by

$$v(\bar{0})=0, \qquad v(\bar{1})=0, \qquad v(\bar{2})=5, \qquad v(\overline{12})=15.$$

It seems somewhat odd that this appears to portray Player 2 as the most favored player. The paradox is in the treatment of threats. The calculation of the characteristic function for Player 1 does not take into account the high cost to Player 2 incurred if he employs his strategy 2. A more detailed discussion of the problem of threats in the evaluation of the characteristic function is given in Shapley and Shubik (1971–74).

The possibility for cooperation is changed but not eliminated if comparisons of utility and sidepayments are not permitted. We may define a generalized characteristic function or a "characterizing function" $V(S)$ which defines for every set of players S a set of optimal achievable payoffs. [This contrasts with a single number for $v(S)$.]

A way of defining the generalized characteristic function is illustrated by Figure 1.6 which shows a three-person example, the axes $\alpha_1, \alpha_2, \alpha_3$ indicating the payoffs obtained by individuals 1, 2, and 3. We treat each $V(S)$ as though it were a cylinder which will "punch out" a part of the Pareto optimal surface of the n-person game as a whole. For example, the coalition $\overline{12}$ can obtain at least as much as they are offered at any point on that part of the Pareto optimal set[12] ABC delineated by EFC.[13]

Starting from the game in strategic form there are two ways of defining the effectiveness of a coalition in a game without sidepayments. We may specify either that which a *coalition can achieve* or that which it cannot be *prevented from achieving*. A simple example of this distinction made by Jentsch (1964) and by Aumann and Peleg (1960), is provided by Shapley and Shubik (1971–74).

A fairly natural restriction to place upon a characteristic function with sidepayments is that of *superadditivity*, i.e.,

$$v(S\cup T)\geqslant v(S)+v(T) \quad \text{where} \quad S\cap T=\emptyset,^{14}$$

and, for the no sidepayments case,

$$V(S\cup T)\supset V(S)\cap V(T).$$

The argument for this condition is at the center of economic modelling and the modelling of any societal behavior. The presumption is that trade, exchange,

[12] See definition in Section 2.
[13] See Aumann (1961), Billera (1970), Scarf and Shapley (1973).
[14] Von Neumann and Morgenstern (1944).

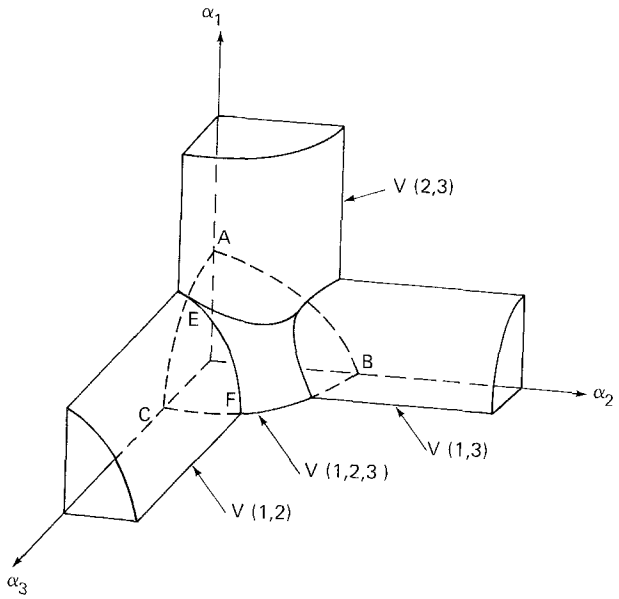


Figure 1.6

and social interaction take place when all parties have something to gain from cooperation, in contrast with opting for no cooperation.

When one attempts to obtain a more process-oriented view of cooperation than is provided by the characteristic function this assumption is by no means as natural as it may seem. The costs of organization appear to be critical in determining the formation of coalitions, groups, and institutions. Game theory techniques *per se* do not provide us with an adequate way to make these distinctions.

In spite of the many difficulties and limitations in the defining of a characteristic function, a game cast in this form is suited for the answering of several questions of interest concerning the power of individuals and coalitions, methods of fair division, and patterns of social stability.

In recent years there has been a growth in experimentation with games in coalitional form. A survey of much of this work is given in Shubik (1975a).

1.5. Continua of strategies, time, players, and goods

The first development of the theory of games concentrated upon situations with fixed numbers of players making choices from finite sets of alternatives in games

with a finite end. Most of human affairs can be modelled to a good approximation by these conditions. But, both for reasons of obtaining better approximations and of exploiting deeper and more appropriate mathematical methods other assumptions are of use.

In particular, there are many games in which continuous strategies and continuous time appear to be called for. The Cournot duopoly model already noted provides an example where even if mass production occurs in integral outputs, we may be able to construct models of greater mathematical tractability by assuming continuous and differentiable production functions.

Duels and pursuit problems provide examples of games where it is natural to think of events occurring in continuous time. There is a large literature on games on the unit square,[15] duels, pursuit, and differential games in general.[16] The application of these to economic problems has been relatively small.[17]

Possibly the most important simplification to game theoretic theorizing in application to economics has come about in the development of games with a continuum of players. One of the key underlying assumptions in the study of mass market economies, polities, and societies is the idea that, although the individual may have freedom of choice, for many purposes his influence on the economy or society as a whole is negligible.

Many of the most paradoxical features in the understanding of the relationship between micro- and macroeconomics rest upon the fallacy of composition that distinguishes individual from mass behavior.

The first attempts to mathematize the relationship between the influence of the single individual and the number of individuals in the market were made by Cournot (1838) and Edgeworth (1881). The method of replication of players was essentially clearly spelled out by them. In the context of game theory applied to economics Shubik (1955b, 1959a), Shapley (1954–60), and Debreu and Scarf (1963) formulated and developed the method of replication; while Hildenbrand (1974) and others have generalized this approach.[18]

Aumann (1964) first treated the set of traders in a closed economy as a continuum with individual nonatomic traders, thus offering a direct mathematical approach to the study of small traders in the market. Milnor and Shapley (1961) and Shapley (1962b) had previously applied the concept of a continuum of players to voting processes in their treatment of oceanic games.

Most of the applications of game theory to economics to date have utilized a description of trade and production with a finite set of commodities. The classification and taxonomy of commodities is somewhat arbitrary. For some purposes two items may be regarded as perfect substitutes whereas for other

[15] Dresher (1961).
[16] Berkovitz and Dresher (1959), Fox and Kimeldorf (1969), and Friedman (1971).
[17] Clemhout, Lietman, and Wan (1973).
[18] See also Chapter 18 in this Handbook.

purposes they may differ. It is clear thus that any result in economic theory which appears to depend in any critical manner upon the relative number of commodities and traders must be suspect. Although this problem is by no means one confined alone to game theory applications to economics it is nevertheless of importance in understanding monopolistic competition.[19]

2. Solutions

2.1. Presolutions

The mathematical representation of a game in and of itself is a step towards answering the question that is being posed. Thus the descriptions of a game in extensive, strategic, or cooperative form can be regarded as presolutions in the sense that the labor involved in translating and modelling may yield all the insights that are required. For example, the characteristic function provides an indication of the potential gains from cooperation of various groups. This information alone may be all that is needed to understand what is at stake in a negotiation.

A natural presolution is the *Pareto optimal surface*. Consider a payoff vector x in an n-person game; x is feasible if

$$x \in V(N),$$

where N is the set of all players. It is Pareto optimal if feasible and

$$\sum_{i \in S} x_i \notin D(S), \qquad S \subseteq N.$$

where $D(S)$ is the interior of $V(S)$.

The Pareto optimal surface satisfies our concepts of efficiency and societal rationality, but it does not include any conditions on individual rationality. There may be points on the Pareto optimal surface where an individual obtains less than he could get by acting by himself.

How much an individual can maintain without the cooperation of others is a matter of modelling political, economic, and social reality. In economic models of trade it is usually assumed that an individual can maintain ownership over his initial endowments.

If we add a condition of individual rationality to the conditions for Pareto optimality we restrict ourselves to payoffs in the *imputation set* which is part of

[19]Chamberlin (1950), Shapley and Shubik (1969c), and Bewley (1972).

the Pareto optimal set. The additional condition is

$$x \notin D(\bar{i}) \quad \text{all} \quad i \in N,$$

so that each player is as well off as if he acted alone.

We may define an *imputation* x as a vector of n numbers $(x_1, x_2, \ldots, x_n)$ where each x_i represents the payoff to player i. For games with sidepayments $\sum_{i=1}^{n} x_i = v(N)$.

For games with sidepayments the imputation set can be represented by a simplex which provides a particularly convenient geometric representation. Figure 2.1 shows a diagram analogous to that of Figure 1.6 for a three-person game with sidepayments represented by the characteristic function

$$v(\bar{1}) = v(\bar{2}) = v(\bar{3}) = 0,$$
$$v(\overline{12}) = 1, \quad v(\overline{13}) = 2, \quad v(\overline{23}) = 3,$$
$$v(\overline{123}) = 4.$$

ABC in Figures 1.6 and 2.1 denote the imputation sets in the three-person

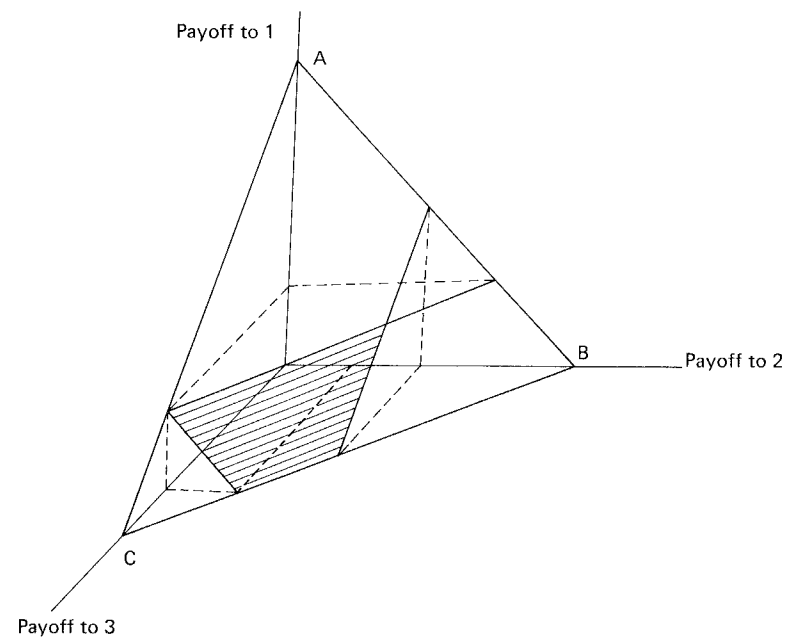


Figure 2.1

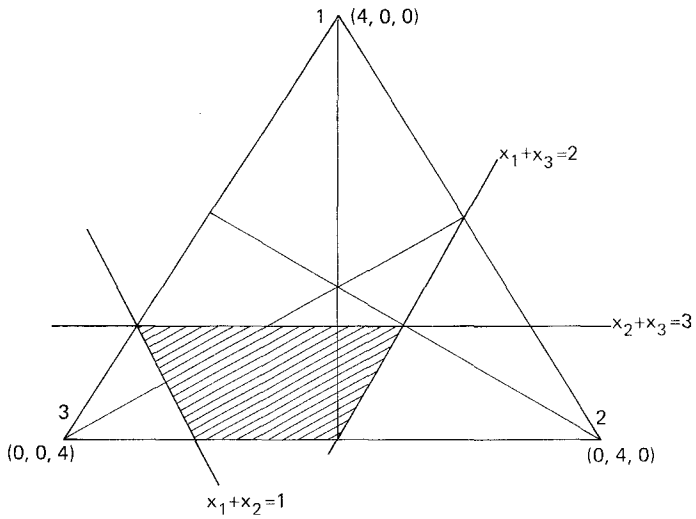


Figure 2.2

sidepayment and no-sidepayment games, respectively. Figure 2.2 shows the sidepayment game imputations as a simplex; dispensing with the higher dimensional representation in Figure 2.1. The lines $x_1+x_2=1$, $x_1+x_3=2$, and $x_2+x_3=3$ are drawn on the simplex. An imputation $x=(x_1, x_2, x_3)$ that can be blocked by the coalition $\overline{12}$ acting alone has to satisfy

$$x_1 \geqslant 0, \quad x_2 \geqslant 0, \quad x_3 \geqslant 0,$$

and

$$x_1+x_2 \leqslant 1 \quad \text{and} \quad x_1+x_2+x_3=4.$$

One additional kind of presolution has been suggested by Milnor (1952). A payoff vector $x=(x_i : i \in N)$ is defined as *reasonable* iff (if and only if) it satisfies

$$x_i \leqslant \max_{i \in S} [\, v(S)-v(S-\{i\})] \quad \text{for all} \quad i \in N.$$

This condition states that no individual should ever obtain more than the most he contributes to any coalition. Most popular solution concepts indeed involve only imputations which are Milnor-reasonable.

A caveat

The idea of presolution is a link between modelling and analysis; i.e., certain reasonable conditions are loaded onto the model before the heavier analysis begins. In particular, it cannot be overstressed that the dangers in blindly accepting the characteristic function and derivative concepts are extremely large. Shapley and Shubik (1971–74) have suggested the term *c-game* to stand for a game whose characteristic function adequately reflects the underlying structure of the behavioral situation.

2.2. *Cooperative solutions*

A basic dichotomy has been made in the development of static solution concepts. This is the dichotomy between cooperative and non-cooperative solutions. For cooperative solutions Pareto optimality is assumed. This is not so for non-cooperative solutions. When we contemplate dynamics this facile dichotomy breaks down. We return to this point in Section 2.4.

Cooperative solution theories in general use as their basis the characteristic function for sidepayment games or the extended characteristic function for no-sidepayment games. The descriptions and definitions given below are for sidepayment games but subsequent comments note the differences of importance between sidepayment and no-sidepayment solutions.

The eight solution concepts we consider are:

(1) core,
(2) value,
(3) von Neumann–Morgenstern stable set,
(4) bargaining set,
(5) kernel,
(6) nucleolus,
(7) ε-core,
(8) inner core.

Others have been suggested, but this list certainly includes the major cooperative solutions.

2.2.1. *The core*

The core was originally defined by Gillies (1959), and suggested as an independent solution concept by Shapley (1953a). Essentially it consists of the set of imputations which leave no coalition in a position to improve the payoffs to all of its members.

Formally the *core* consists of all imputations x such that

$$\sum_{i \in S} x_i \geqslant v(S), \qquad S \subseteq N.$$

It is easy to observe that many games may have no core. In Figures 1.6 and 2.1 the cores of the three-person no-sidepayment and sidepayment games respectively are indicated by the shaded parts of the imputation sets.

A key link between game theory and economics comes in the defining of a class of games known as *market games*,[20] originally considered by Shapley and Shubik in 1953, and in the recognition of the important link between the price system and the existence of cores in a game and all of its subgames. An n-person market game has the property that every one of the 2^n subgames which can be formed from all subsets of players has a core.

A class of games more suited to the analysis of voting problems known as *simple games* has the property that the values of the characteristic function are only 0 or 1 or "lose" and "win". Most of the games of this variety have no core.

Simple games can be defined directly via four basic assumptions:

(1) Every coalition is either winning or losing.
(2) The empty set is losing.
(3) The all-player set is winning.
(4) No losing set contains a winning subset.

Two extra assumptions which we may require are:

(5) The complement of any winning set is losing.
(6) The complement of any losing set is winning.

The last two assumptions provide, respectively, for superadditivity and that the game be constant sum. A game having all six properties is said to be a *decisive* simple game.

Intuitively it appears that as one moves from nicely structured economic markets to markets with externalities to political means for distributing resources the chances for conditions for the existence of a price system and then for a core diminish.

Balanced games

From the superadditivity property of a characteristic function we know that, for every family $\{S_j\}$ of coalitions which forms a partition of S,

$$v(S_1) + \cdots + v(S_m) \leqslant v(S).$$

[20] Shapley and Shubik (1969b).

In a market game we consider the possibility that S is broken up into groups which may overlap but for which each set S_j uses only a fraction f_j of the resources (or time) of each of its members. If it is possible to select the f_j such that each player is used so that his fraction of weights sums to 1 and

$$f_1 v(S_1) + \cdots + f_m v(S_m) \leqslant v(S),$$

then the $\{S_j\}$ is said to be a *balanced family of subsets*. A game in characteristic function form is *totally balanced* if for every S it is possible to satisfy the balancing conditions.

Consider the characteristic function of the three-person game illustrated in Figure 2.1. The two-person coalitions form a balanced family of subsets with weights $\frac{1}{2}$ each,

$$\tfrac{1}{2}v(\overline{12}) + \tfrac{1}{2}v(\overline{23}) + \tfrac{1}{2}v(\overline{13}) = 3 < v(\overline{123}).$$

It has been shown by Shapley and Shubik (1969b) that every market game is totally balanced and, for sidepayment games, vice versa. Shapley (1973) and Billera and Bixby (1973) have considered the no-sidepayment games.

The intuitive appeal of the core as a possible solution to problems in political economy is that if it exists it implies that there are ways of imputing wealth which not only satisfy individual and total group rationality but also satisfy all subgroup rationality, i.e., no subgroup is offered less than it could obtain by itself.

2.2.2. *The value*

The core picks up the claims of groups, but offers no fair or equitable manner for resolving these claims. A completely different approach to a solution is offered by the value (or "Shapley value"). Here a direct attempt is made to characterize or axiomatize a concept of fair division. Paradoxically, these attempts not only succeeded in producing several fair division schemes, but they also showed the intimate relationship between considerations of fair division and power. In particular, a key element where these considerations come together is in the definition of the *status quo* point needed to fix the initial conditions from where the fair division is to take place.

Using essentially four axioms — (1) efficiency, (2) a dummy player gets nothing, (3) symmetry, and (4) additivity — Shapley (1953a) was able to deduce a unique value for a sidepayment game. The first three axioms are fairly evident; the fourth axiom is that if we consider two strategically independent games played by the same players, the value calculated by considering the games as one will be the same as that calculated by assigning values to each and then

adding them. Under these axioms the payoff to player i for the *value* solution is given by

$$\phi_i = \sum_{\substack{S \subset N \\ i \in S}} \frac{(n-s)!(s-1)!}{n!} [v(S) - v(S - \{i\})].$$

There is a simple economic interpretation for this value. Each individual is assumed to enter every possible coalition in every way randomly, and he is then assigned the expected value of the incremental gain he brings to all. The value provides a combinatoric marginal evaluation.

Banzhaf (1965) has suggested a different weighting to coalition formation, and Shapley (1977) and Dubey and Shapley (1978) have developed and given mathematical precision to the Banzhaf value.

Nash (1953) developed a two-person bargaining scheme for no-sidepayment games using a symmetry axiom, Pareto optimality, measurable utility, and a construction to evaluate threats, which was generalized for n-person games, with some difficulties remaining, by Harsanyi (1959). Shapley (1964) has suggested a value solution for n-person no-sidepayment games which differs somewhat from that of Harsanyi.

The fundamental difficulties to be overcome in the development of the value were how to treat variable threats to fix the *status quo* point and how to cope with the no-sidepayment game.

Owen (1972) has suggested a natural extension of Shapley's model which reflects the possibility that the likelihood of players joining coalitions may be biased. Aumann and Shapley (1974) and Dubey (1975) have considered values and generalized values for games with a continuum of players.

2.2.3. *The stable set solution*

Von Neumann and Morgenstern (1944) offered a rather sophisticated concept of solution which, in my estimation, turned out to be not as fruitful or general as had been originally hoped.

The essential idea behind the stable set solution is that the collection of imputations comprising a stable set must exhibit the properties of *internal stability* and *external stability*. In order to illustrate these it is first necessary to define *domination* and *effective set*.

An imputation x *dominates* y if there exists a coalition S such that

$$x_i > y_i \quad \text{for all} \quad i \in S,$$

and

$$\sum_{i \in S} x_i \leqslant v(S).$$

This last condition states that the coalition S is an *effective set* for x; i.e, by itself it could obtain what its members obtain in x.

A set of imputations is (i) *internally stable* if no member of the set is dominated by another member of the set; (ii) *externally stable* if any imputation not in the set is always dominated by some imputation in the set; and (iii) a *stable set solution* if it is both internally and externally stable.

A large bibliography on stable set solutions is provided in Shapley and Shubik (1971–74). Originally it had been conjectured by von Neumann that all sidepayment games had stable set solutions. Lucas (1969) was able to prove that this was false, giving a 10-person game counterexample. Shapley and Shubik (1969b) were able to show that the Lucas counterexample could be regarded as a market game, and hence there would exist economies without stable set solutions:

2.2.4. The bargaining set

This is a solution concept originally due to Aumann and Maschler (1964). It has been defined in several slightly different ways, i.e, Peleg (1967). Its genesis was inspired by observing players in an experimental bargaining game.

A *bargaining point* of the game (N, v) has the property that for each pair $i, j \in N$ any objection of i against j can be met by a counterobjection by j against i.

An *objection* consists of a coalition S containing i but not j together with an imputation for which S is effective which is preferred to the given imputation by all members of S.

A *counterobjection* consists of a different coalition T containing j but not i together with an imputation for which T is effective that is (weakly) preferred to the objection by every member of $T \cap S$ and is (weakly) preferred to the original imputation by every member of $T - S$.

If x is the original imputation, (S, y) the objection and (T, z) the counterobjection then

$$\sum_{k \in S} y_k \leqslant v(S) \quad \text{and} \quad y_k > x_k \quad \text{for all} \quad k \in S,$$

and

$$\sum_{k \in T} z_k \leqslant v(T) \quad \text{and} \quad z_k \geqslant y_k \quad \text{for all} \quad k \in T \cap S,$$

$$\text{and} \quad z_k \geqslant x_k \quad \text{for all} \quad k \in T - S.$$

The *bargaining set* is the set of all bargaining points.

Although some use of the bargaining set has been made in experimental studies, little application of the bargaining set has been made to economics. The computational difficulties for games larger than 3 or 4 players make it somewhat unattractive.

2.2.5. *The kernel*

Davis and Maschler (1965) have suggested a solution which is contained within the bargaining set.

In order to define the kernel it is convenient to first define the excess and the surplus.

By the *excess* of a coalition S at an imputation x we mean:

$$e(S, x) = v(S) - \sum_{i \in S} x_i,$$

i.e., it is the amount by which the worth of the coalition exceeds its preferred payoff.

The *surplus* of a player i against another player j with respect to a given imputation is the largest excess of any coalition that contains i and not j.

The surplus basically measures a potential bargaining pressure of i against j. The *kernel solution* consists of all imputations x such that for any two players i and j

$$\max_{\substack{i \in S \\ j \notin S}} e(S, x) = \max_{\substack{j \in T \\ i \notin T}} e(T, x).$$

It picks up the idea of symmetry or equalization of bargaining pressure.

2.2.6. *The nucleolus*

The nucleolus is a solution concept introduced by Schmeidler (1969) which is a unique outcome in the kernel of a sidepayment game. Although it should be noted that no totally satisfactory definition of the nucleolus for a no-sidepayment game exists.[21]

The *nucleolus* is the imputation for which the maximal excess is minimal. Intuitively it is the point that minimizes dissatisfaction. It seems that it should have a natural application in the design of taxation and subsidies, yet, although

[21] One could use a λ-transfer approach, but it does not seem to be satisfactory. The λ-transfer method provides a way for finding intrinsic utility comparisons in order to utilize the point of tangency of a sidepayment hyperplane with the no-sidepayment game's Pareto set to construct a solution for the no-sidepayment game. Further explanation can be found in Shapley and Shubik (1971–74).

there have been some applications of the nucleolus[22] to operations research, the lack of an adequate no-sidepayment nucleolus has limited its application. For that matter the bargaining set, kernel, and nucleolus all appear to have been underemployed in the context of economic models.

2.2.7. *The ε-core*

The core of a game can be "fat" or for that matter non-existent. We may consider a way to uniformly tax or subsidize the players so that cores can be made to appear or shrink. Two ways of doing this are suggested here.

The *strong ε-core*[23] consists of the set of Pareto optimal outcomes x such that

$$\sum_{i \in S} x_i \geqslant v(S) - \varepsilon \quad \text{for all} \quad S \subseteq N.$$

The *weak ε-core* consists of the set of Pareto optimal outcomes x such that

$$\sum_{i \in S} x_i \geqslant v(S) - s\varepsilon \quad \text{for all} \quad s \subseteq N,$$

where s is the number of players in coalition S. We may regard the ε as an overall cost or a per capita cost to the formulation of coalitions or a frictional threshold below which it is not worth acting.

By increasing the size of ε we can eventually produce a core in any sidepayment game without a core. We can define the *least core* or *near core* to be the smallest strong ε-core. It is evident that this is close to, but not the same as the idea underlying the nucleolus.

2.2.8. *The inner core*

Although many results which hold true for sidepayment games also have their analogues for no-sidepayment games, this is by no means always the case. Among the important problems in game theory and its applications to political economy is the characterization of the differences. For example, a brief contemplation of Figures 1.6 and 2.1 should suffice to indicate that although the core of a sidepayment game will always remain simply connected this is not true for a no-sidepayment game and a counterexample can be produced for $n=3$.

A difference between sidepayment and no-sidepayment games leads us to the definition of the *inner core* of a game. Consider the core of the no-sidepayment game illustrated in Figure 1.6. Suppose that at any point in the core we

[22] Littlechild (1974).
[23] Shapley and Shubik (1971–74).

constructed a tangent hyperplane and used the direction cosines of this hyperplane to define an intrinsic comparison of utility among the players in an associated sidepayment game which has only the point of tangency in common with the no-sidepayment game.[24]

Using the comparison of utility we can describe the feasible sets of all coalitions in terms of hyperplanes. A natural question to ask is whether the point of tangency, which is a point in the core of the no-sidepayment game is also a point in the core of the associated sidepayment game. The answer is "not necessarily".

We define the *inner core* of a no-sidepayment game to be those imputations in the core which are also in the core of the associated sidepayment games.

The construction for obtaining the inner core is essentially cardinal. Yet it is of interest to note that the inner core is contained within the core of a no-sidepayment game defined ordinally. Furthermore for a market game the core shrinks under replication, and the inner core is non-empty; hence the core and inner core approach the same limit, the competitive equilibria.

2.3. *Non-cooperative solutions*

The cooperative solutions dealt with games in characteristic function form; the non-cooperative solutions are basically applied to games in strategic form. In fact, much of the basic interest in game models in oligopoly and other aspects of economics is focussed on games which are played many times over. A discussion of solutions to such games is deferred to Section 2.4. It is precisely here that the nice distinctions between cooperative and non-cooperative solution concepts start to evaporate.

2.3.1. *Two-person constant sum games*

Two-person zero sum games and their strategically equivalent constant sum games are games of *pure opposition*. The goals of the players are diametrically opposed.

Both two-person zero sum games and the minimax theorem are well known[25] and need not be presented again here. It is important, however, to note that two-person zero sum game theory, although of considerable importance in the study of military tactical problems, is of extremely limited value to the study of political economy. Extremely few situations in political economy, if any, meet the conditions of pure opposition.

[24] This is in the spirit of λ-transfer.
[25] Von Neumann and Morgenstern (1944).

2.3.2. Non-cooperative equilibrium points

Consider an n-person game in strategic form where each player i has a set of strategies S_i, $i = 1, \ldots, n$.

Let $P_i(s_1, s_2, \ldots, s_n)$ be the payoff function to player i, then an equilibrium point is a vector of strategies $(s_1^*, s_2^*, \ldots, s_n^*)$ such that for each $i = 1, \ldots, n$,

$$P_i(s_1^*, \ldots, s_n^*) = \max_{s_i \in S_i} P_i(s_1^*, \ldots, s_i, \ldots, s_n^*).$$

The general concept of a *non-cooperative equilibrium* and its existence for matrix games was given by Nash (1950), although the basic idea and its relevance to economics was given by Cournot (1838).

It is possible to show the existence of non-cooperative equilibria for games with a continuum of players.[26] This is of direct relevance in attempts to model the nuances of meaning in the concept of competitive markets.

In general, the difficulties with the non-cooperative equilibrium solution concept come far less in problems of existence than in the multiplicity of equilibria. Furthermore, it appears to be relatively easy to produce models where many of the equilibrium points appear to be quite unreasonable. A brief rogue's gallery of 2×2 matrix games is given in Figure 2.3. In the first there is a single equilibrium yielding (0,0). In the second there are two pure strategy equilibria, one favoring the first and the other the second player; there is also a bad mixed strategy equilibrium. In the third all outcomes are equilibrium points.

When we contemplate a game in strategic form the unsatisfactory static feature of the equilibrium point becomes clear. There is undoubtedly a circular stability to an equilibrium: "If A knew what B was doing then he would do such-and-such and vice-versa!" Unfortunately there is no indication of how or why the players will generate expectations to bring about an equilibrium. In short, the non-cooperative equilibrium theory is static, it does not indicate how communication is to be handled, and the equilibrium points are frequently not unique. These comments apply not only to n-person non-cooperative games in general but to economic markets viewed as non-cooperative games.

	1	2
1	5,5	-1,10
2	10,-1	0,0

	1	2
1	10,1	-20,-20
2	-20,-20	1,10

	1	2
1	1,6	10,6
2	1,3	10,3

Figure 2.3

[26] Dubey and Shapley (1977).

2.4. *Other solutions*

So much of economic analysis in one form or the other has apparently depended upon solution concepts which implicitly or explicitly amount to non-cooperative equilibria that a full understanding of what knowledge is conveyed by the solution is critical.

In particular, our problem is more in modelling and in the understanding of behavior than it is in the mathematics. The definition of an equilibrium point is purely static. The mathematics tells us what the equilibrium point is, not how it came about. Dissatisfaction with this has forced many game theorists to consider the extensive form of the game as a key to the understanding of the process.

In trying to model in the extensive form an immediate difficulty is encountered: Are communications, negotiations, and messages modelled as a part of the game? If so, how are messages, language and other forms of communication put into the extensive form? At this time there is no adequate answer to this basic coding problem. It is, however, clear that if everything is in the game then the distinction between cooperative and non-cooperative theories becomes blurred. Binding commitments and coalitions outside of the game are ruled out, and commitments within the game become the art of the possible. It can be shown that in games of indeterminate length virtually every outcome in a subgame can be converted into an equilibrium, even using behavior strategies. It is easy to see from the example in Figure 2.4 that the outcomes (5, 5) and (0, 0) and $(-11, -11)$ can all be enforced as equilibria, yet it is hard to believe that they are all equally plausible.

Selten (1965, 1975) introduced and refined the concept of a perfect equilibrium point for games in extensive form. In his original definition a *perfect equilibrium point* had the property that the players are in equilibrium in each subgame attained. A quick example helps to illustrate a perfect and a not perfect equilibrium. Consider the game shown in Figure 2.4 played twice. The strategy pairs (some strategy for each player) are

I play 1 to start, if he plays 1, then
I play 2 next; otherwise I play 3.

	1	2	3
1	5,5	-1,8	-30,-12
2	8,-1	0,0	-30,-12
3	-12,-30	-12,-30	-11,-11

Figure 2.4

and

I play 2 on both occasions,
regardless of what he does.

Both give equilibria; the first is not perfect because in the last play even if the other player has failed to play 1 to begin with there is no *ex post* motivation beyond "desire to punish" to play 3. In contrast the second is perfect.

The original definition of a subgame perfect equilibrium leaves problems with that part of the game tree not attained in the equilibrium path. An example of a 3-person game provided by Selten illustrates this. Each player has but two choices L or R hence a behavior (or mixed) strategy for i can be characterized by a probability p_i for selecting R. The two types of equilibria are

Type 1: $p_1 = 1, \quad p_2 = 1, \quad 0 \leqslant p_3 \leqslant \frac{1}{4},$

Type 2: $p_1 = 0, \quad \frac{1}{3} \leqslant p_2 \leqslant 1, \quad p_3 = 1.$

Note that in the equilibrium points of type 2 player 2's information set is not reached, hence his expected payoff is independent of his choice. In particular, consider the equilibrium point (0, 1, 1) which is of type 2. If by "accident", say a random perturbation, player 2 were reached, is it reasonable for him to play 1 given that he believes that player 3 will also play 1 giving him 4 if he switches to 0? Clearly type 2 equilibria are unreasonable; a player's choices should be conditional upon the information set reached. Selten develops a model with a perturbed game where, with some very small probability, a player will make a mistake. In this manner all information sets may be reached.

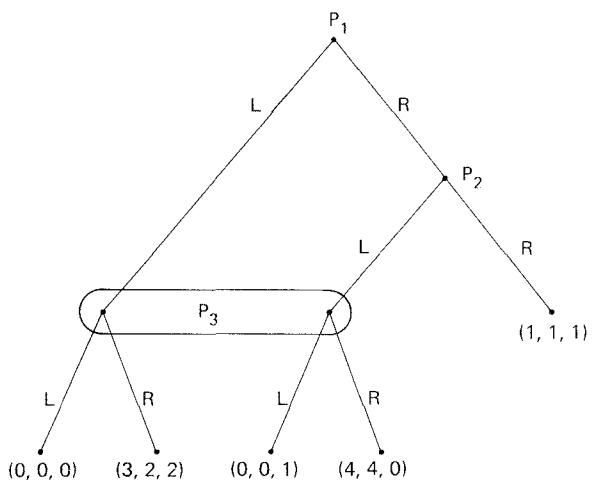


Figure 2.5

Harsanyi (1975) has been concerned with the possibility of selecting a single preferred non-cooperative equilibrium point out of the myriad which may exist.

In general, games of interest to political economy have a surfeit of equilibrium points. Dubey and Shubik (1977b) have shown that for any game representable by a finite game tree if player information sets are refined the new game will contain as pure strategy equilibrium points those of the old game and probably more. In particular, this means that if the general equilibrium trading model is not interpreted as a single simultaneous move game then it has more non-cooperative equilibria than the competitive equilibria.

Consider two games with the same game tree, differing only in the structure of their information sets. One game may be said to have *more refined information* than the other if all of its information sets either coincide with the information sets of the other or a proper subset of those information sets.

The result noted above holds for games with random moves by nature only if information concerning these moves is the same for all games.

Many economic models can be interpreted as games that are more or less repeated many times or indefinitely through time. Thus, it is attractive to contemplate some form of dynamic programming approach as suggested in several of the papers in Blaquiere (1973). This super-rationalistic approach, which can be characterized by a backward induction for finite length games yielding perfect equilibrium points, can be contrasted with the type of expectation updating devices which move forward in time in macroeconomic behavioral models. The work of Sobel (1971, 1973) is devoted to considering the links between these two types of equilibrium; i.e., when will myopic behavior be equivalent to the other? See also Friedman (1977).

The discipline of game theory has answered few questions concerning dynamics, information, and communication, but at least it provides a way to formulate precisely many of the critical unsolved problems.

Under the catchall of "other solutions" there still are more cooperative and non-cooperative solutions that we have not dealt with, mainly because of space limitations and the lack of many economic applications. These solutions include the competitive equilibrium of a game [Baudier (1973)], the Banzhaf (1965) and other values; and game payoff transformations such as maxmin the difference in a two-person non-constant sum game beat the average in certain n-person games and games of status [Shubik (1965, 1971)].

3. Applications

This brief review is offered as a somewhat short and possibly too compact survey and reference guide to the development of the applications of game theory to economics. It is not meant to be complete nor completely selfcon-

tained. A lengthier review of oligopoly theory is provided by J. Friedman in Chapter 11 of this Handbook.

3.1. Oligopolistic markets

The major applications of game theory to date have been to the study of various aspects of oligopolistic competition and collusion and to the study of the emergence of price in a closed economic system. The discussion of the latter is given in Section 3.2.

A reasonable division of the study of oligopolistic markets is:

(1) duopoly,
(2) non-cooperative and "quasi-cooperative" oligopoly,
(3) bilateral monopoly and bargaining,
(4) experimental gaming,
(5) auctions and bidding.

3.1.1. Duopoly

Models of duopoly have always held a fascination for mathematically inclined economists. The literature is so vast that it merits a separate study. A good set of references has been provided by Chamberlin (1950). The first well known model which can be clearly identified as a well defined mathematical description of competition in a duopolistic market is that of Cournot (1838). It is a game in strategic form where the competitors, selling identical products, are assumed to select a level of production each in ignorance of the other's action. The solution suggested is that of a non-cooperative equilibrium. A simple example illustrates Cournot's problem. Let q_i be the amount produced to be sold by firm i, $i=1,2$. Let $C_i(q_i)$ be total cost of production of i. Let price be given by $p=D(q_1+q_2)$, then each firm maximizes $\Pi_i=q_iD(q_1+q_2)-C_i(q_i)$. A pair of production strategies $(\bar{q}_1,\bar{q}_2)$ give rise to a non-cooperative equilibrium if given $\bar{q}_j$,

$$\max_{q_i} q_iD(q_i+\bar{q}_j)-C_i(q_i)\Rightarrow q_i=\bar{q}_i \quad \text{for} \quad j=1,2, \quad i=2,1.$$

Bertrand's (1883) critique of Cournot was primarily directed towards the former's selection of production as the strategic variable. He suggested that price would be a more natural variable. Edgeworth's (1925) model of duopoly introduced rising costs, or the equivalent of a capacity restraint; Hotelling (1929), also working with a non-cooperative model, introduced transportation costs as a form of product differentiation. Chamberlin's (1950) model introduced product differentiation in a more general manner.

A variety of duopoly models have been proposed and analyzed by authors such as Coase (1935), Harrod (1934), Kahn (1937), Nichol (1934), Stigler (1940), and many others. These and many subsequent works cannot be properly regarded as explicitly game theoretic models inasmuch as the authors do not concern themselves with the specification in detail of the strategy spaces of the two firms. Indeed the careful mathematical statement of a duopoly model calls for considerable care and detail as is shown by Wald's (1951) analysis of equilibrium.

Beckmann (1965), Levitan and Shubik (1971), Mayberry, Nash and Shubik (1953), Shapley and Shubik (1969c), and Shubik (1973) have presented a series of explicitly game theoretic models of duopoly studying the effects of inventory carrying costs, fluctuations in demand, capacity constraints, and simultaneous decisions on both price and production. Other models have compared non-cooperative solutions with other types of solution; stockout conditions and variation in information conditions have also been considered.

There is a growing literature on dynamic models of duopoly. Frequently, in teaching even the simple Cournot duopoly model, a dynamic process entailing action and reaction is sketched. An important early work was that of von Stackelberg (1934), in which a series of quasi-dynamic models were suggested. A detailed discussion of von Stackelberg is given by Fellner (1949). The papers of Smithies and Savage (1940), Shubik and Thompson (1959), and Cyert and De Groot (1973) provide more mathematical, game theoretical, and behavioral models of duopolistic behavior.

3.1.2. Oligopoly

A more or less standard way of considering oligopolistic markets is to construct a model which can be studied for duopoly and which can be compared with duopoly or with a competitive market as numbers are increased. Frequently, the assumption of symmetrically related firms enables us to make comparisons among markets of different size. Thus, for example, Cournot proceeds from an analysis of two firms to many, and Chamberlin considers small and large groups of competitors.

It is important to bear in mind that three different skills are called for in the investigation of oligopolistic markets. They are the skills of the economist at describing economic institutions and activities and selecting the relevant variables and relationships; the skills of the modeler in formulating a mathematical structure that reflects the pertinent aspects of the economic phenomenon; and the skills of the analyst in deducing the properties of the mathematical system that has been formulated. Thus, for example the work of Chamberlin may be regarded as a considerable step forward over that of Cournot in terms of its

greater relevance and reality; however, it was no advance at all (and possibly a retrogression) in terms of rigorous mathematical formulation and analysis when compared with Cournot. Both the Cournot and Chamberlinian large group analyses are based on a non-cooperative equilibrium analysis.

When competitors are few Chamberlin suggests that threats must be taken into account. When the group is large his analysis can be regarded as tantamount to a one-stage non-cooperative game.

All the mathematical rigor in the world cannot make up for lack of economic insight and understanding in the creation of the model to be analyzed; thus the development of an adequate theory depends heavily upon verbal description and less fully formal models such as those suggested by von Stackelberg (1934) and Fellner (1949). Special variables must be considered. Brems (1951) introduces technological change; Bain (1956) considers entry; Marris (1964), Baumol (1959), and Shubik (1961) stress managerial structure; Levitan and Shubik (1971) and Kirman and Sobel (1974) consider the role of inventories; and there are many other works dealing with other important and special variables such as transportation, advertising, production change costs, multiple products, financing, and so forth.

Where does the theory of games fit into these considerations beyond being a mathematical tidying up device which merely translates the insights of others into a more heavily symbolic language? The answer to this can be best seen when it is understood that the discipline called for in specifying in detail the strategic options of the individual actors leads to the discovery of gaps in the logic of less formally defined models. Many of these models are "quasi-dynamic" in description. In other words they describe some sort of adjustment process in terms which gloss over the information conditions (precisely who knows what, when, and how much does it matter?). In general in the description of action and reaction, time is not explicitly accounted for, not even by at least a rate of interest such as in the work of Cross (1969).

The Chamberlinian analysis of large group behavior and the large literature by Sweezy (1939) and many others[27] on the "kinked oligopoly curve" provide important examples of both the power and danger of an informal mix of verbal and diagrammatic modeling. This is easily shown when one tries to formulate the structure of the market as a well defined model. The kinked oligopoly curve has no objective existence, and it presupposes an extremely limited set of reactions by all competitors. It is obtained by implicitly assuming symmetry in strategic power, in the structure of moves, and in information conditions for all firms (or by not knowing enough to see that explicit or implicit assumptions must be made if the robustness of the conclusions is to be examined).

[27]See also Stigler (1947, sect. 3.2) and Chapter 11 of this Handbook by Friedman.

Furthermore, the arguments describing equilibrium or a tendency toward equilibrium using either the kinked oligopoly curve or Chamberlin's large group analysis depend only upon the local properties of these subjective curves. Edgeworth's (1925) analysis of duopoly led him to conclude that no equilibrium need exist, but his results were obtained by considering the objective structure of oligopolistic demand over all regions of definition. In other words, the analysis requires that we should be able to state what the two firms will find their demand to be, given every pair of prices (p_1, p_2).

Shubik (1959b) suggested the term "contingent demand" to describe the demand faced by an individual firm given the actions of the others as fixed. It is possible to show that the contingent demand structure may easily depend upon details of marketing involving the manner in which individual demands are aggregated. Levitan (1964) showed the relationship between the description of oligopolistic demand and the theory of rationing. Based on the study of the shape of contingent demand curves, Levitan and Shubik were able to show that the Chamberlin large group equilibrium may easily be destroyed for much the same reasons as indicated in Edgeworth's analysis.

A full understanding of the problems posed by oligopoly calls for a clear distinction to be made separating aspects of market structure, the intent of the firms, and the behavior of the firms. Then a study of the interrelationships among these factors is called for. In game theoretic terms this amounts to making a clear distinction among *the rules of the game*, the *solution concept* employed, and the *solution set* obtained.

Perhaps the most important aspect of game theoretic modeling for the study of oligopoly comes in describing information conditions and providing formal dynamic models which depend explicitly upon the information conditions. There is a growing interest in *state strategy* models, i.e., models in which the system dynamics are dependent upon only the state that the system is in currently. The work of Shubik and Thompson (1959), Miyasawa (1962) and others, working on sequential game models of economic processes, provides examples.

The sensitivity of an oligopolistic market to changes in information has been studied in Shubik (1973b). When information is relatively high there is no strong reason to suspect that a few firms in an oligopolistic market will employ state strategies. Instead, we may expect that they will use *historical strategies* where previous history, threats, and counterthreats play an important role. Marschak and Selten (1974), Selten (1973), and Shubik (1959b) have considered this possibility.

Among the books directly devoted to a game theoretic investigation of oligopoly are those of Friedman (1977), Jacot (1963), Shubik (1959b), and Telser (1972).

3.1.3. Bilateral monopoly and bargaining

Whereas most of the models of oligopolistic behavior have either offered solutions based on the non-cooperative equilibrium or have sketched quasi-dynamic processes, the work on bilateral monopoly and bargaining has primarily stressed high levels of communication with a cooperative outcome or a dynamic process which leads to an optimal outcome. A few of the models suggest the possibility of non-optimal outcomes, such as strikes which materialize after threats are ignored or rejected.

Many of the models of bargaining arise from highly different institutional backgrounds. The major ones are bilateral trade among individual traders as characterized by Böhm–Bawerk's horse market or Bowley's model. Frequently however, the model proposed refers to international trade or to labor and employer bargaining. Edgeworth's famous initial model was cast in terms of the latter, as was the work of Zeuthen.

The work of Edgeworth (1881) is clearly related to the game theory solution of *the core*, as has been noted by Shubik (1959a), Scarf (1967), and Debreu and Scarf (1963). Böhm–Bawerk's analysis may be regarded as an exercise in determining the core and price in a market with indivisibilities [Shapley and Shubik (1972a)].

Zeuthen's (1930) analysis of bargaining is closely related to the various concepts of *value* as a solution. This includes the work of Nash (1953), Harsanyi (1959), Shapley (1953b), and Selten (1964).

Another area of importance in application which is possibly closer to political science than economics is international strategic bargaining. Reference to this type of application is made in Shapley and Shubik (1971–74).

A "solution" to an economic problem may attempt to do no more than cut down the feasible set of outcomes to a smaller set. No specific outcome is predicted. The solution narrows down the set of outcomes but does not tell us exactly what will or should happen. The *contract curve* of Edgeworth and the core are solutions in this sense.

Other solutions may be used in an attempt to single out one final outcome as that which should or which will emerge. In their static versions most of the various *value* solutions and other fair division solutions which have been proposed may be regarded as normative in their suggestions and abstracted from any particular institutional background in their presentation. These remarks hold for the works of Nash (1953), Shapley (1953b), Harsanyi (1956), Braithwaite (1955), Kuhn (1966), Steinhaus (1949), and others.

Still other solutions which may be used to select a single outcome are phrased in terms of the dynamics of the bargaining process. These include the works of

Cross (1969), Harsanyi (1956), Pen (1952), Raiffa (1953), Shubik (1952), Zeuthen (1930), and others.

3.1.4. Gaming

One result of the development of the theory of games and the high-speed digital computer has been a growth in interest in using formal mathematical models of markets for gaming for teaching and/or experimental purposes. The earliest published article on an informal economic game experiment was by Chamberlin. This, however, appears in isolation from the rest of the literature. The first "business game" built primarily for training purposes was constructed by Bellman et al. (1957) several years later; this was followed by a deluge of large computerized business games. The use of these games has been broadly accepted in business schools and in some economics faculties. References on gaming are given in Shubik (1975c) and in Shapley and Shubik (1971–74).

It is important to stress the difference between business games and experimental games. The former lay stress upon teaching, the latter upon validating game theoretic, economic, or other behavioral conjectures.

Much of the earlier experimental work with games in economics did not use the computer. The games were frequently presented in the form of matrices or diagrams. Siegal and Fouraker (1960), Fouraker and Siegal (1963), and Fouraker, Shubik, and Siegal (1961) were concerned with bilateral monopoly under various information conditions and duopoly and triopoly. Stern (1966), Dolbear (1968), and others investigated the effect of numbers of competitors in a market. Friedman (1967) has considered the effect of symmetry and lack of symmetry in duopoly as well as several other aspects of oligopolistic markets, and Smith (1967) has considered the effect of market organization.

Experiments using computerized games have been run by Hoggatt (1967), Hoggatt and Selten (1973), Friedman and Hoggatt (1973), Shubik, Wolf and Eisenberg (1972), McKenney (1967), and several others. These games provide advantages in control and in ease of data processing that the non-computer games do not offer.

Several of the games noted above can be and have been solved for various game theoretic and other solutions. This means that, for instance, in the duopoly games studied by Friedman it is possible to calculate the Pareto optimal surface and the non-cooperative equilibrium points. Similarly, in the duopoly or oligopoly investigations of Hoggatt, Stern, Fouraker, Siegal, Shubik and others, it is usually possible to calculate the joint optimum, the non-cooperative equilibria, and the competitive price system.

The game designed by Levitan and Shubik (1961a–c, 1967 a–d) was specifically designed to be amenable to game theoretic analysis. Thus, the joint

maximum, the price non-cooperative equilibrium, the quantity non-cooperative equilibrium, the range of the Edgeworth cycle, the beat-the-average, and several other solutions have been calculated for this game.

It is possibly too early to attempt a critical survey of the implications of all of the experimental work for oligopoly theory, however a general pattern does seem to be emerging. All other things being equal, an increase in the number of competitors does appear to lower price, as does an increase in cross-elasticities between products. However, with few competitors, time lags, details of structure, and information and communication conditions appear to be far more critical than a reading of oligopoly theory would indicate.

In all of the experiments and in games for teaching such as those of the Carnegie Tech and Harvard Business School, the importance of considering a richer behavioral model of the individual emerges when the way in which the players attempt to deal with their environment is observed. This observation by no means runs counter to a game theory approach. It is complimentary with it. As yet there exists no satisfactory dynamic oligopoly solution provided by either standard economic theory or game theory. This appears to be due to the difficulties in describing the role of information processing and communication.

A different set of games have been used for experimentation with a certain amount of economic content, but it is far less identified with an economic market than the oligopoly games. This set includes simple bidding and bargaining exercises used by Flood, Kalish, Milnor, Nash and Nering, Stone, Maschler, Riker, Shubik, and others. In all of these instances there is a direct interest in comparing the outcomes of the experiments with the predictions of various game theory solutions.

3.1.5. Auctions and bidding

Auctions date back to at least Roman times. In many economies they still play an important role in financial markets and in commodity markets. Sealed bids are used in the letting of large systems contracts and in the sale of government property. Their history as economic market mechanisms is a fascinating subject by itself [see Cassady (1967)]. Furthermore, as an auction or a bidding process is usually quite well defined by a set of formal rules (together, on occasions, with customs or other informal rules) it lends itself naturally to formal mathematical modeling.

The mathematical models of auctions and bids fall into two major groups: those in which the role of competition is modeled by assuming a Bayesian mechanism, and those in which the model is solved as a game of strategy using the solution concept of the non-cooperative equilibrium or some other solution.

There is also a considerable literature on problems encountered in different types of bidding and on features such as problems in evaluation, risk minimization, and incentive systems.

The study of auctions and bidding lies in the zone between theory and application as can be seen by observing the tendency of the publications to appear in journals such as the *Operations Research Quarterly* or *Management Science*. A useful bibliography on bidding has been supplied by Stark and Rothkopf (1979).

In general, game theoretic work on auctions and bidding has been useful in two ways: descriptive and analytic. The careful specification of the mathematical models has forced attention to be paid to understanding the actual mechanisms, including informal rules and customs. The attempts at solution have shown that the models are extremely sensitive to information conditions and that many of the important features of auctions involve the individual's ability to evaluate what an object is worth to him and to others. This is a far cry from the economic models where the assumption is made that all have complete knowledge of all individual's preferences.

3.2. General equilibrium

An important area of application of the theory of games to economic analysis has been to the closed general equilibrium model of the economy. The solution concepts which have been explored are primarily cooperative solutions. Noncooperative solutions appear to be intimately related with monetary economies, and this work is discussed in Section 3.4.

In much economic literature it has been claimed that the study of economics requires only the assumption of a preference ordering over the prospects faced by an individual. Apart from the fact that such a strong assumption immediately rules out economic consideration topics like bargaining and fair division where virtually no analysis can be made without stronger assumptions on the measurement of utility, even if we restrict our investigation to the free functioning of a price system, the assumption of only a preference ordering is not sufficient. At least one must restrict transformations to those which preserve concavity of utility functions, otherwise markets involving gambles would emerge [Shubik (1973a)].

The investigations of game theoretic solutions in application to economic problems have been devoted primarily to the core and secondarily to other cooperative solutions. Because of the predominance of the former we deal with it separately.

3.2.1. The core

The first economist to consider bargaining and market stability in terms of the power of all feasible coalitions was Edgeworth (1881). He was dealing with a structure which can be described as a market game; hence he did not define the core solution concept in general. Shubik (1959a) observed that the Edgeworth analysis was essentially an argument that could be described in terms of the core for this class of games. He constructed a two-sided market model to demonstrate this, and used a method of replication to illustrate the emergence of a price system. This was done for the sidepayment game. He conjectured this to be true in general for no-sidepayments and proposed this problem to Scarf. The replication method of studying the limiting behavior of a many-player game involves starting with a given set of different types of traders, where a "type" is defined by a utility function and endowment. Suppose we start with a game with k players, one of each of k types The nth replication of such a market game consists of a market with nk players with n of each type.

Essentially the replication method boils down to considering an economy with thousands of butchers, bakers, and candlestick makers. Debreu and Scarf (1963) and Scarf (1967), using the method of replication, were able to generalize the previous results considerably. They showed that under replication, in a market with any number of different traders, the core "shrinks down" (under the appropriate definition which takes into account the increasing dimensions) to a set of imputations which can be interpreted in terms of a price system emerging as its limit. Hildenbrand (1974) has generalized the method of replication, doing away with the rigidity of maintaining identical types; see also Chapter 18 in this Handbook.

It must be stressed that the core consists of a set of undominated imputations and may have nothing whatsoever to do with an economy. For an economy, however, as the numbers grow in a market the limit points of the core can be interpreted in terms of a price system without specifying any price mechanism.

A different way of considering markets with many traders is by imagining what we can "chop up" traders into finer and finer pieces. Going directly to measure theory, we may consider a continuum of traders where the individual trader whose strategic power is of no significance to the market is described as having a measure of zero. Aumann (1964) first developed this approach.[28]

In the past fifteen years there has been a proliferation of literature on the core of a market. Although many of these writings have been devoted to the relationship between the core and the competitive equilibrium (for example, given a continuum of small traders of all types, the core and competitive equilibrium can be proved to be identical), some of the work has been directed

[28]See also Kirman's chapter on measure theory in this Handbook (Chapter 5).

towards other problems; thus Shapley and Shubik (1966) considered the effect of non-convex preference sets, and Aumann (1973), Shitovitz (1973), and others have been concerned with the economics of imperfect competition. Caspi (1972) and Shubik (1973a) also considered the effect of uncertainty on the core. Debreu (1975) has considered the speed of convergence of the core.

A difficulty with the cooperative game formulation appears when one tries to model production. Are production sets jointly available or individually owned? Dubey (1975) has considered the latter case for non-atomic sidepayment games and has established the coincidence of the core, value, and competitive equilibrium. Hildenbrand (1974), assuming a production possibility set for each coalition, has extended the definition of a Walrasian equilibrium and proved the coincidence of the core for the non-atomic game with the extended Walrasian equilibrium.

In summary, the import of this work to economic analysis is that it extends the concept of economic equilibrium and stability to many dimensions, and it raises fundamental questions concerning the role and the nature of coalitions in bringing about economic stability.

3.2.2. Other solutions

The core can be regarded as characterizing the role of countervailing power among groups. There are other solution concepts which reflect other views for the determination of the production and distribution of resources. In particular, the family of solutions which can be described as *the value* of an *n*-person game stress fair division, where the division is based both upon the needs or wants of the individual and his basic productivity and ownership claims.

The value

There are various differences among the different value solutions which have been suggested, which depend upon three major factors: (1) whether there are two or more individuals, (2) whether or not a sidepayment mechanism is present, and (3) whether threats play an important role and the status quo is difficult to determine.

Leaving aside the finer points, all of the value solutions are based in one form or the other upon a symmetry axiom and an efficiency axiom. Describing them loosely, if individuals have equal claims they should receive equal rewards and the outcome should be Pareto optimal.

These solutions clearly appear to have no immediate relationship with a price system, yet Shapley and Shubik (1969d), using the method of replication, were able to show that under the appropriate conditions as the number of individuals in a market increases the value approaches the imputation selected by the price

system. Shapley and Aumann and Shapley have also considered other models. In the latter work markets with a continuum of traders have been investigated and the coincidence of the value with the competitive equilibrium has been established.

The bargaining set, nucleolus, and kernel

Rather than appeal to countervailing power arguments or to considerations of fairness, one might try to delimit the outcomes by bargaining considerations. Aumann and Maschler (1964) suggested a bargaining set, and Peleg (1966) established that such a set always exists. Shapley and Shubik (1972b) were able to show that under the appropriate conditions the bargaining set lies within an arbitrarily small region of the imputation selected by the price system when the number of traders in an economy is large, however this was an extremely restricted result.

As has already been noted, neither the nucleolus nor kernel appear to have yielded significant economic applications yet.

3.2.3. Solutions, market games, and the price system

The class of games known as "market games" provides a representation of a closed economic system for which a price system exists.

The study of market games with large numbers of participants has shown a remarkable relationship between the imputations selected by a price system and the core, value, bargaining set, kernel, and nucleolus of large market games. Each solution concept models or picks up an extremely different aspect of trading. The price system may be regarded as stressing decentralization (with efficiency); the core shows the force of countervailing power; the value offers a "fairness" criterion; the bargaining set and kernel suggest how the solution might be delimited by bargaining conditions; and the nucleolus provides a way to select a point at which dissatisfaction with relative tax loads or subsidies is minimized.

If for a large market economy these many different approaches call for the same imputation of resources then we have what might be regarded as a nineteenth century laissez faire economist's dream. The imputation called for by the price system has virtues far beyond that of decentralization; it cannot be challenged by countervailing power, it is fair, and it satisfies certain bargaining conditions.

Unfortunately in most economies as we know them these euphoric conclusions do not hold for two important reasons. The first is that there is rarely if ever enough individuals of all types that oligopolistic elements are removed from all markets. The second is that the economies frequently contain elements that

modify or destroy the conditions for the existence of an efficient price system. In particular, these include external economies and diseconomies, indivisibilities, and public goods.

3.3. *Public goods, externalities, and welfare economics*

When we examine the literature on public goods it is difficult to make a completely clear distinction between economic analysis and political science studies. The basic nature of the problems is such that their investigation requires an approach based on political economy. More or less arbitrarily, even though it is related to welfare economics, we do not discuss the work on voting systems.

It is well known that, when externalities are present in an economy, an efficient price system may not exist. It might be that a tax and subsidy system can be designed to make it possible for an efficient price system to function. Shapley and Shubik (1969a) and Foley (1970) have been able to show that a tax system, essentially the one suggested by Lindahl serves the purpose. However, in the former work it is noted that such a system may not be effective when external diseconomies are present. Klevorick and Kramer (1973) have worked on a specific taxation scheme for pollution management using prices. Aumann and Kurz (1977) have applied a mixed model of threats, prices, and the value solution to taxation.

The role of threats is of considerable importance when studying many of the problems posed by externalities and public goods. Features such as forcing an individual to share a public good and preventing him from using a good unless he pays his share lead to differentiating many types of public goods. Shubik (1966) suggested a taxonomy of public goods based on these considerations.

Considerable application of game theory to tax problems and public finance has been made by Schleicher (1971).

Indivisibilities and other features which may cause non-convexities to be present in the consumption or production possibility sets of the individual have been studied by Shapley and Scarf, Shapley and Shubik, Shubik, and others.

Externalities caused by different ownership arrangements as well as pecuniary externalities caused by the presence of markets have been studied.[29] Although not primarily game theoretic in content, the work of Buchanan and Tullock, Davis and Whinston, Zeckhauser, and others is closely related to the game theory approach to public goods and welfare economics.

Another aspect of welfare economics where game theory analysis is of direct application involves the study of lump-sum taxation, subsidies, and compensa-

[29]See Shapley and Shubik (1967) and Shubik (1971c), respectively.

tion schemes. These depend delicately on assumptions made concerning the availability of a sidepayment mechanism and the relationship between social and economic prospects and the structure of individual preferences.

3.4. Money and financial institutions

In recent years a considerable interest has been evinced in the construction of an adequate microeconomic theory of money. In general, this work has taken as its basis the general equilibrium non-strategic model of the price system. The writings of Foley (1970), Hahn (1971), Starr (1974), and others serve as examples of the statics, and Grandmont (1977) provides an excellent coverage of the dynamics.

In contrast with the non-strategic approaches Dubey and Shubik (1977a, 1978), Shapley (1976), Shapley and Shubik (1977), Shubik (1972, 1975b), and others[30] have considered non-cooperative game models of trading economies using a commodity or a fiat money. The major thrust of this work has been to suggest that strategic modelling calls for the introduction of rudimentary structures and rules of the game which can be interpreted in terms of markets, financial institutions, and laws. In specifying the use of money a distinction must be made between money and credit. Bankruptcy laws must be specified.[31] The way money and credit enters the system must be noted. A decision must be made whether or not to model bankers as separate distinguished players in the "money game" [Shubik (1976)].

Information conditions clearly are of considerable importance in a mass economy. Dubey and Shubik (1977c) have noted the sensitivity of market models to changes in information conditions and have obtained a non-cooperative equilibrium in markets with non-symmetric information conditions.

A simple example of non-cooperative game with trade in a commodity money and bids and offers by traders is provided here. Let there be n traders and $m+1$ commodities. Each trader i has a utility function of form $\mathcal{O}_i(x_1^i, x_2^i, \ldots, x_{m+1}^i)$ and initial resources of $(a_1^i, a_2^i, \ldots, a_{m+1}^i)$ where $a_j^i \geqslant 0$, $a_j \ (=\sum_{i=1}^n a_j^i) > 0$. The $(m+1)$st commodity is distinguished as a money in the sense that all transactions of the first m commodities are paid for using the $(m+1)$st commodity.

A strategy by a trader i is a vector of $2m$ numbers, $s^i = (q_1^i, b_1^i, q_2^i, b_2^i, \ldots, q_m^i, b_m^i)$ where $0 \leqslant q_j^i \leqslant a_j^i$ is an offer by trader i to sell an amount q_j^i of good j; b_j^i is a bid of the amount of money trader i is willing to spend to purchase good j; and $\sum_{j=1}^m b_j^i \leqslant a_{m+1}^i$ is a cash constraint. The price of the jth good, $j=1,\ldots,m$, is

[30] See also Postlewaite (1978), and Shubik and Whitt (1973).
[31] Shubik and Wilson (1977).

given by

$$p_j = \sum_{i=1}^{n} b_j^i \Big/ \sum_{i=1}^{n} q_j^i.$$

The price of the $(m+1)$st good is fixed at 1.

The final holding of good j by i, for $j=1,\ldots,m$, is given by

$$x_j^i = a_j^i - q_j^i + b_j^i/p_j.$$

The final holding good $m+1$ by i is given by

$$x_{m+1}^i = a_{m+1}^i - \sum_{j=1}^{m} b_j^i + \sum_{j=1}^{m} p_j q_j^i.$$

Thus the game calls for each player i to select a strategy s^i to maximize $\Pi_i(s^1, s^2, \ldots, s^n) = \phi_i(x_1^i, x_2^i, \ldots, x_{m+1}^i)$.

The full details of the additional conditions required to establish the existence of non-cooperative equilibria and their relationship to competitive equilibria are given in Dubey and Shubik (1978).

Some results have been obtained using cooperative game theoretic analysis. In particular, it has been shown that if trade is assumed to take place via markets then pecuniary externalities are real.

Game theoretic aspects of insurance have been considered by Borch (1968).

3.5. *Other applications*

There has only been slight application of game theory to macroeconomic problems beyond the work of Nyblen (1951) and Faxen (1957). These developments are nevertheless suggestive of the possibility of treating aggregated units as players in a game of strategy.

Beyond the applications noted above, there have been some scattered papers in economics or topics closely allied to economics in the form of the work of Schleicher (1971) on public finance, Shapley (1962a) on bureaucracy and organization design, Shubik (1962) on the design of incentive systems, Littlechild (1974) on operations research costing and pricing problems, and Gately and Kyle (1976) on cartel problems. These is also a large literature of applications to political science which is related to work in economics, but is sufficiently far afield that it is not covered here.

References

Aumann, R. J. (1961), "The case of a cooperative game without sidepayment", Transactions of the American Mathematical Society, 98:539–552.

Aumann, R. J. (1964), "Markets with a continuum of traders", Econometrica, 32:39–50.

Aumann, R. J. (1973), "Disadvantageous monopolies", Journal of Economic Theory, 6:1–11.

Aumann, R. J. and M. Kurz (1977), "Power and taxes", Econometrica, 45:1137–1163.

Aumann, R. J. and M. Maschler (1964), "The bargaining set of cooperative games", in: M. Dresher, L. S. Shapley and A. W. Tucker, eds., Advances in game theory, pp. 443–447. Princeton, NJ: Princeton University Press.

Aumann, R. J. and B. Peleg (1960), "von Neumann–Morgenstern solutions to cooperative games without sidepayments", Bulletin of American Mathematical Society, 66:173–179.

Aumann, R. J. and L. S. Shapley (1974), Values of nonatomic games. Princeton, NJ: Princeton University Press.

Bain, J. (1956), Barriers to new competition. Cambridge, MA: Harvard University Press.

Banzhaf, J. F. (1965), "Weighted voting doesn't work: A mathematical analysis", Rutgers Law Review, 19:317–343.

Baudier, E. (1973), "Competitive equilibrium in a game", Econometrica, 41:1049–1068.

Baumol, W. J. (1959), Business behavior value and growth. New York: Macmillan.

Beckman, M. J. (1965) (with the assistance of Dieter Hochstadter), "Edgeworth–Bertrand duopoly revisited", in: R. H. Sonderdruck, ed., Operations Research—Verfahren III, pp. 55–67. Meisenheim: Anton Hain.

Bellman, R., C. E. Clark, C. J. Craft, D. G. Malcolm and F. M. Ricciardi (1957), "On the construction of a multistage multiperson business game", Journal of Operations Research, 5:469–503.

Berkovitz, L. D. and M. Dresher (1959), "A game theory analysis of tactical airwar", Operations Research, 7:599–620.

Bertrand, J. (1883), "Théorie mathématique de la richesse sociale" (review), Journal des Savants (Paris), 499–508.

Bewley, T. (1972), "Existence of equilibria in economies with infinitely many commodities", Journal of Economic Theory, 4:514–540.

Billera, L. J. (1970), "Some theorems on the core of an n-person game without sidepayments", SIAM Journal of Applied Mathematics, 18:567–579.

Billera, L. J. and R. E. Bixby (1973), "A characterization of polyhedral market games", International Journal of Game Theory, 2:252–261.

Blaquiere, A. (1973), Topics in different games. Amsterdam: North-Holland.

Borch, K. (1968), The economics of uncertainty. Princeton, NJ: Princeton University Press.

Braithwaite, R. B. (1955), Theory of games as a tool for the moral philosopher. Cambridge, MA: Cambridge University Press.

Buchanan, J. M. and G. Tullock (1962), The calculus of consent. Ann Arbor, MI: University of Michigan Press.

Caspi, Y. (1972), "A limit theorem on the core of an economy under uncertainty", RR 43. Jeresalem: Hebrew University.

Cassady, R. (1967), Auctions and auctioneering. Berkeley, CA: University of California Press.

Chamberlin, E. H. (1950), The theory of monopolistic competition, 6th ed. Cambridge, MA: Harvard University Press.

Clemhout, S., G. Lietman and H. Wan, Jr. (1973), "A differential game model of duopoly", Econometrica, 39:911–938.

Coase, R. H. (1935), "The problem of duopoly reconsidered", Review of Economic Studies, 2:137.

Cournot, A. A. (1838). In 1897 translated as: Researches into the mathematical principles of the theory of wealth. New York: Macmillan.

Cross, J. G. (1969), The economics of bargaining. New York: Basic Books.

Cyert, R. M. and H. M. De Groot (1973), "An analysis of cooperation and learning in a duopoly context", American Economic Review, 63:24–37.

Davis, M. and M. Maschler (1965), "The kernel of a cooperative game", Naval Research Logistics Quarterly, 12:223–295.
Davis, O. A. and A. B. Whinston (1965), "Welfare economics and the theory of second best", Review of Economic Studies, 32:1–14.
Debreu, G. (1975), "The rate of convergence of the core of an economy", Journal of Mathematical Economics, 2:1–7.
Debreu, G. and H. E. Scarf (1963), "A limit theorem on the core of an economy", International Economic Review, 4:235–246.
Dolbear, F. T., L. A. Lave, G. Bowman, A. Lieberman, E. Prescott, F. Reuter and R. Shepman (1968), "Collusion in oligopoly: An experiment on the effect of numbers and information", Quarterly Journal of Economics, 82:240–259.
Dresher, M. (1961), Games of strategy: Theory and applications. Englewood Cliffs, NJ: Prentice Hall.
Dubey, P. (1975), "Some results on values of finite and infinite games", Ph.D. thesis. Ithaca, NY: Cornell University.
Dubey, P. and L. S. Shapley (1977), "Noncooperative exchange with a continuum of traders", P-5964. Santa Monica, CA: Rand.
Dubey, P. and L. S. Shapley (1978), "Mathematical properties of the Banzhaf power index", Mathematics of Operations Research (forthcoming).
Dubey, P. and M. Shubik (1977a), "A closed economic system with production and exchange modelled as a game of strategy", Journal of Mathematical Economics, 4:253–278.
Dubey, P. and M. Shubik (1977b), "Information conditions, communication and general equilibrium", Cowles Foundation discussion paper no. 467. New Haven, CT: Yale University.
Dubey, P. and M. Shubik (1977c), "Trade and prices in a closed economy with exogenous uncertainty, different levels of information, money and compound futures markets", Econometrica (forthcoming).
Dubey, P. and M. Shubik (1978), "The noncooperative equilibria of a closed trading economy with market supply and bidding strategies", Journal of Economic Theory, 17:1–20.
Edgeworth, F. Y. (1881), Mathematical psychics. London: Kegan Paul.
Edgeworth, F. Y. (1925), Papers relating to political economy I, pp. 111–142. London: MacMillan.
Faxen, K. O. (1957), Monetary and fiscal policy under uncertainty. Stockholm: Almqvist and Wiksell.
Fellner, W. (1949), Competition among the few. New York: Alfred A. Knopf.
Foley, D. K. (1970a), "Economic equilibrium with costly marketing", Journal of Economic Theory, 2:276–291.
Foley, D. K. (1970b), "Lindahl's solution and the core of an economy with public goods", Econometrica, 38:66–74.
Fouraker, L. E., M. Shubik and S. Siegel (1961), "Oligopoly bargaining: The quantity adjuster models", RB no. 20. University Park, PA: Pennsylvania State University. Also reported in: Fouraker and Siegel (1963).
Fouraker, L. E. and S. Siegel (1963), Bargaining behavior. New York: McGraw-Hill.
Fox, M. and G. S. Kimeldorf (1969), "Noisy duels", SIAM Journal of Applied Mathematics, 17:353–361.
Friedman, A. (1971), Differential games. New York: Wiley.
Friedman, J. W. (1967), "An experimental study of cooperative duopoly", Econometrica, 35:379–397.
Friedman, J. W. (1977), Oligopoly and the theory of games. Amsterdam: North-Holland.
Friedman, J. W. and A. C. Hoggatt (1973), Unpublished manuscript, available in part from the authors.
Gately, D. and J. F. Kyle (1976), "Optimal strategies for OPEC's pricing decisions", Discussion paper no. 76-13. New York: Center for Applied Economics, New York University.
Gillies, D. (1959), "Solutions to general nonzero sum games", Annals of Mathematics Study, 40:47–85.
Grandmont, J. M. (1977), "Temporary general equilibrium theory", Econometrica, 45:535–572.
Hahn, F. H. (1971), "Equilibrium with transactions costs", Econometrica, 39:417–439.
Harrod, R. F. (1934), "The equilibrium of duopoly", Economic Journal, 44:335.

Harsanyi, J. C. (1956), "Approaches to the bargaining problem before and after the theory of games: A critical discussion of Zeuthen's, Hicks's and Nash's theories", Econometrica, 24:144–157.
Harsanyi, J. C. (1959), "A bargaining model for the cooperative *n*-person game", in: A. W. Tucker and R. D. Luce, eds., Contributions to the theory of games, IV, pp. 325–356. Princeton, NJ: Princeton University Press.
Harsanyi, J. C. (1975), "The tracing procedure: A Bayesian approach to defining a solution for *n*-person noncooperative games", International Journal of Game Theory, 4:61–94.
Hildenbrand, W. (1974), Core and equilibria of a large economy. Princeton, NJ: Princeton University Press.
Hoggatt, A. C. (1967), "Measuring the cooperativeness of behavior in quantity variation duopoly games", Behavioral Science, 12:109–121.
Hoggatt, A. C. and R. Selten (1973), Unpublished manuscript, available in part from the authors.
Hotelling, H. (1929), "Stability in competition", Economic Journal, 39:41–57.
Jacot, S.-P. (1963), Strategie et concurrence. Paris: SEDES.
Jentzch, G. (1964), "Some thoughts on the theory of cooperative games", Annals of Mathematics Study, 52:407–442.
Kahn, R. F. (1937), "The problem of duopoly", Economic Journal, 47:1.
Kirman, A. P. and M. J. Sobel (1974), "Dynamic oligopoly with inventories", Econometrica, 42:279–287.
Klevorick, A. K. and G. H. Kramer (1973), "Social choice on pollution management: The genossenschaften", Journal of Public Economics, 2:101–146.
Kuhn, H. W. (1953), "Extensive games and the problem of information", Annals of Mathematics Study, 28:193–216.
Kuhn, Harold W. (1966), "On games of fair division", in: M. Shubik, ed., Essays in mathematical economics in honor of Oskar Morgenstern, pp. 29–37. Princeton, NJ: Princeton University Press.
Levitan, R. E. (1964), "Oligopoly demand", RC 1239. Yorktown Heights, NY: IBM Research.
Levitan, R. E. and M. Shubik (1961a, b, c, 1967a, b, c, d), "A business game for teaching and research purposes: Parts I, II, IV–II", Cowles Foundation discussion papers no. 115, 219, 224, 225, and 227. New Haven, CT: Yale University.
Levitan, R. E. and M. Shubik (1971), "Price variation duopoly with differentiated products and random demand", Journal of Economic Theory, 3:23–39.
Levitan, R. E. and M. Shubik (1972), "Price duopoly and capacity constraints", International Economic Review, 13:111–122.
Lindahl, E. (1919), "Just taxation—A positive solution". Reprinted in part in: R. A. Musgrave and A. E. Peacock, eds., Classics in the theory of public finance, pp. 168–176. London: MacMillan. In 1919 published in German.
Littlechild, S. C. (1974), "A simple expression for the nucleolus in a special case", International Journal of Game Theory, 3:21–29.
Lucas, W. F. (1969), "The proof that a game may not have a solution", Transactions of the American Mathematical Society, 137:219–229.
Luce, R. D. and H. Raiffa (1957), Games and decisions. New York: Wiley.
Marris, R. (1964), The economic theory of managerial capitalism. New York: Free Press.
Marschak, T. and R. Selten (1974), General equilibrium with price making firms. Berlin: Springer.
Mayberry, J., J. F. Nash, Jr. and M. Shubik (1953), "A comparison of treatments of a duopoly situation", Econometrica, 21:141–155.
McKenney, J. L. (1967), Simulation game for management development. Cambridge, MA: Harvard Business School.
Milnor, J. W. (1952), "Reasonable outcomes for *n*-person games", RM-916. Santa Monica, CA: Rand.
Milnor, J. W. and L. S. Shapley (1961), "Values of large games II: Oceanic games", RM-2649. Santa Monica, CA: Rand.
Miyasawa (1962), "An economic survival game", Journal of the Operations Research Society of Japan, 4:95–113.
Nash, J. F., Jr. (1950), "Equilibrium points in *n*-person games", Proceedings of the National Academy of Science U.S.A., 36:48–49.

Nash, J. F., Jr. (1953), "Two-person cooperative games", Econometrica, 21:128–140.
Nichol, A. J. (1934), "A re-appraisal of Cournot's theory of duopoly price", Journal of Political Economy, 42:80–105.
Nyblen, G. (1951), The problem of summation in economic science. Lund: Gleerup.
Owen, G. (1972), "Multilinear extensions of n-person games", Management Science, 18:64–79.
Peleg, B. (1966), "Existence theory for the bargaining set $M_1^{(i)}$", in: M. Shubik, ed., Essays in mathematical economics in honor of Oskar Morgenstern, pp. 53–56. Princeton, NJ: Princeton University Press.
Pen, J. (1952), "A general theory of bargaining", American Economic Review, 46:24–42.
Postlewaite, A. W. and D. Schmeidler (1978), "Approximate efficiency of non-Walrasian Nash equilibria", Econometrica, 46:127–136.
Raiffa, H. (1953), "Arbitration schemes for generalized two-person games", in: H. Kuhn and A. W. Tucker, eds., Contributions to the theory of games, pp. 361–387. Princeton, NJ: Princeton University Press.
Rapoport, A. and A. M. Chammah (1965), Prisoner's dilemma. Ann Arbor, MI: University of Michigan Press.
Rapoport, A. and M. J. Guyer (1966), "A taxonomy of 2×2 games", General Systems, 11:203–214.
Rapoport, A., M. J. Guyer and D. J. Gordon (1975), The 2×2 game. Ann Arbor, MI: University of Michigan Press.
Scarf, H. E. (1967), "The core of an n-person game", Econometrica, 35:50–69.
Scarf, H. E. and L. S. Shapley (1973), "On cores and indivisibility", in: G. Dantzig and C. Eaves, eds., Studies in optimization. Washington, DC: Mathematical Association of America.
Schelling, T. C. (1960), The strategy of conflict. Cambridge, MA: Harvard University Press.
Schleicher, H. (1971), Staatshaushalt und Strategie. Berlin: Dunker and Humblot.
Schmeidler, D. (1969), "The nucleolus of a characteristic function game", SIAM Journal of Applied Mathematics, 17:1163–1170.
Selten, R. (1964), "Valuation of n-person games", in: M. Dresher, L. S. Shapley and A. W. Tucker, eds., Advances in game theory, pp. 577–626. Princeton, NJ: Princeton University Press.
Selten, R. (1965), "Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit", Zeitschrift für die gesamte Staatswissenschaft, 121:301–324, 667–689.
Selten, R. (1973), "A simple model of imperfect competition where 4 are few and 6 are many", International Journal of Game Theory, 2:141–201.
Selten, R. (1975), "Reexamination of the perfectness concept for equilibrium points in extensive games", International Journal of Game Theory, 4:25–56.
Shapley, L. S. (1953a), Unpublished lecture notes. Princeton, NJ: Princeton University.
Shapley, L. S. (1953b), "A value for n-person games", in: H. Kuhn and A. W. Tucker, eds., Contributions to the theory of games, Vol. II, no. 28, pp. 307–317. Princeton, NJ: Princeton University Press.
Shapley, L. S. (1954–60), Unpublished notes and discussions with M. Shubik.
Shapley, L. S. (1955), "Markets as cooperative games", P-B29. Santa Monica, CA: Rand. Introduced earlier in unpublished lecture notes of a 1953–54 seminar at Princeton, NJ.
Shapley, L. S. (1962a), "Simple games: An outline of the descriptive theory", Behavioral Science, 7:59–66.
Shapley, L. S. (1962b), "Values of games with infinitely many players", in: Conference on recent advances in game theory, pp. 113–118. Princeton, NJ: Princeton University.
Shapley, L. S. (1964), "Values of large market games: Status of the problem", RM-3957-PR. Santa Monica, CA: Rand.
Shapley, L. S. (1973), "On balanced games without sidepayments", in: T. C. Hu and S. M. Robinson, eds., Mathematical programming. New York: Academic Press.
Shapley, L. S. (1976), "Noncooperative general exchange", in: A. Y. Lin, ed., Theory and measurement of economic externalities. New York: Academic Press.
Shapley, L. S. (1977), "A comparison of power indices and a nonsymmetric generalizations", P-5872. Santa Monica, CA: Rand.
Shapley, L. S. and H. Scarf (1972), "On cores and indivisibility", Journal of Mathematical Economics, 1:23–38.

Shapley, L. S. and M. Shubik (1953), "Solutions of n-person games with ordinal utilities" (abstract), Econometrica, 21:348–349.
Shapley, L. S. and M. Shubik (1966), "Quasi-cores in an economy with nonconvex preferences", Econometrica, 34:805–827.
Shapley, L. S. and M. Shubik (1967), "Ownership and the production function", Quarterly Journal of Economics, 81:88–111.
Shapley, L. S. and M. Shubik (1969a), "On the core of an economic system with externalities", American Economic Review, 59:678–684.
Shapley, L. S. and M. Shubik (1969b), "On market games", Journal of Economic Theory, 1:9–25.
Shapley, L. S. and M. Shubik (1969c), "Price strategy oligopoly with production variation", Kyklos, 22:30–44.
Shapley, L. S. and M. Shubik (1969d), "Pure competition, coalitional power and fair division", International Economic Review, 10:337–362.
Shapley, L. S. and M. Shubik (1971–74), "Game theory in economics", R-904/1,2,3,4,6-NSF. Santa Monica, CA: Rand.
Shapley, L. S. and M. Shubik (1972a), "The assignment game I: The core", International Journal of Game Theory, 2:111–130.
Shapley, L. S. and M. Shubik (1972b), "Convergence of the bargaining set for differentiable market games", Mimeographed working notes.
Shapley, L. S. and M. Shubik (1977), "Trade using one commodity as a means of payment", Journal of Political Economy, 85:937–968.
Shitovitz, B. (1973), "Oligopoly in markets with a continuum of traders", Econometrica, 41:467–501.
Shubik, M. (1952), "A business cycle model with organized labor considered", Econometrica, 20:284–294.
Shubik, M. (1955a), "A comparison of treatments of a duopoly problem, Part II", Econometrica, 23:417–431.
Shubik, M. (1955b), "Edgeworth market games", Seminar notes. Palo Alto, CA: Center for Advanced Study in Behavioral Sciences.
Shubik, M. (1959a), "Edgeworth market games", in: A. W. Tucker and D. R. Luce, eds., Contributions to the theory of games, Vol. IV, pp. 267–278. Princeton, NJ: Princeton University Press.
Shubik, M. (1959b), Strategy and market structure. New York: Wiley.
Shubik, M. (1961), "Objective functions and models of corporate optimization", Quarterly Journal of Economics, 73:345–375.
Shubik, M. (1962), "The assignment of joint costs and internal pricing", Management Science, 8:325–342.
Shubik, M. (1966), "Notes on the taxonomy of problems concerning public goods", Cowles Foundation discussion paper no. 208. New Haven, CT: Yale University.
Shubik, M. (1971a), "The 'bridge game' economy", Journal of Political Economy, 79:909–912.
Shubik, M. (1971b), "Games of status", Behavioral Science, 16:117–129.
Shubik, M. (1971c), "Pecuniary externalities: A game theoretic analysis", American Economic Review, 61:713–718.
Shubik, M. (1972), "Commodity money, oligopoly, credit and bankruptcy in a general equilibrium model", Western Economic Journal, 10:24–38.
Shubik, M. (1973a), "The core of a market with exogenous risk and insurance", New Zealand Economic Papers, 7:121–127.
Shubik, M. (1973b), "Information duopoly and competitive markets: A sensitivity analysis", Kyklos, 26:736–761.
Shubik, M. (1975a), Games for society, business, and war. Amsterdam: Elsevier.
Shubik, M. (1975b), "The general equilibrium model is incomplete and not adequate for the reconciliation of micro and macroeconomic theory", Kyklos, 28:545–573.
Shubik, M. (1975c), The uses and methods of gaming. Amsterdam: Elsevier.
Shubik, M. (1976), "A noncooperative model of a closed economy with many traders and two bankers", Zeitschrift für Nationalökonomie, 36:10–18.
Shubik, M. and G. Thompson (1959), "Games of economic survival", Naval Logistics Research Quarterly, 6:111–123.

Shubik, M. and W. Whitt (1973), "Fiat money in an economy with one nondurable good and on credit (a noncooperative sequential game)", in: A. Blaquiere, ed., Topics in differential games, pp. 401–448. Amsterdam: North-Holland.
Shubik, M. and C. Wilson (1977), "The optimal bankruptcy rule in a trading economy using fiat money", Zeitschrift für Nationalökonomie (forthcoming).
Shubik, M., G. Wolf and H. Eisenberg (1972), "Some experiences with an experimental oligopoly business game", General Systems, 13:61–75.
Siegel, S. and L. E. Fouraker (1960), Bargaining and group decision making. New York: McGraw-Hill.
Smith, V. L. (1967), "Experimental studies of discrimination vs. competition in sealed-bid auction markets", Journal of Business, 40:56–84.
Smithies, A. and L. J. Savage (1940), "A dynamic problem in duopoly", Econometrica, 8:130.
Sobel, M. J. (1971), "Non-cooperative stochastic games", Annals of Mathematical Statistics, 42: 1095–1100.
Sobel, M. J. (1973), "Continuous stochastic games", Journal of Applied Probability, 10:597–604.
Stackelberg, H. von (1934), Marktform und Gleichgewicht. Berlin: Springer.
Stark, R. M. and M. H. Rothkopf (1977), "Competitive bidding: A comprehensive bibliography", Operations Research 27:364–390.
Starr, R. (1974), "The price of money in a pure exchange monetary economy", Econometrica, 42:45–54.
Steinhaus, H. (1949), "Sur la division pragmatique", Econometrica (Suppl.), 17:315–319.
Stern, D. (1966), "Some notes on oligopoly theory and experiments", in: M. Shubik, ed., Essays in mathematical economics in honor of Oskar Morgenstern, pp. 255–281. Princeton, NJ: Princeton University Press.
Stigler, G. J. (1940), "Notes on the theory of duopoly", Journal of Political Economy, 48:521.
Stigler, G. J. (1947), "The kinky oligopoly demand curve and prices", Journal of Political Economy, 55:434–449.
Sweezy, P. (1939), "Demand under conditions of oligopoly", Journal of Political Economy, 47:568–573.
Telser, L. (1972), Competition, collusion, and game theory. Chicago, IL: University of Chicago Press.
Von Neumann, J. and O. Morgenstern (1944), Theory of games and economic behavior. Princeton, NJ: Princeton University Press.
Wald, A. (1951), "On some systems of equations of mathematical economics", Econometrica, 19:368–403.
Zeckhauser, R. (1973), "Determining the qualities of a public good", Western Economic Journal, 11:39–60.
Zeuthen, F. (1930), Problems of monopoly and economic welfare. London: G. Routledge.

Chapter 8

GLOBAL ANALYSIS AND ECONOMICS

STEVE SMALE

University of California, Berkeley

One main goal of this work is to show how the existence proof for equilibria can be based on Sard's theorem and calculus foundations. At the same time, equations such as "supply equals demand", are used rather than fixed points methods. The existence proofs given here are constructive in some reasonable and practical sense. These equilibria can be found on a machine using numerical analysis methods.

Our motivation for providing a proof of the Arrow–Debreu theorem (Appendix A) is to show that calculus can be used for the foundations of equilibrium theory.

Also in the paper optimization and the fundamental theorems of welfare economics are developed via the calculus. Abstract optimization theorems are proved in Section 3 and applied in Section 4 to pure exchange economies. Debreu's finiteness of equilibria theorem is proved in Section 5. In this section a manifold structure is put on the set of optima and on a certain set of equilibria as well.

The reader can see Smale (1976b) for a general motivation for a calculus approach to equilibrium theory (as well as references to other topics in Global Analysis and Economics). Furthermore some justification is given in this reference for the continued study of classical equilibrium theory in spite of its deep inadequacies for analyzing the problems of our day.

The account here could be used as a basis for a short course and in fact it was written when giving such a course at Berkeley in the winter of 1977. Much of the background for this exposition is to be found in our papers in the Journal of Mathematical Economics.

1. The existence of equilibria

The basic idea of equilibrium theory is to study solutions of the equation; supply equals demand or $S(p)=D(p)$. For the simple case of one market, where prices are measured in terms of some extra market standard, the familiar diagram below gives some justification for existence of the equilibrium price p^*.

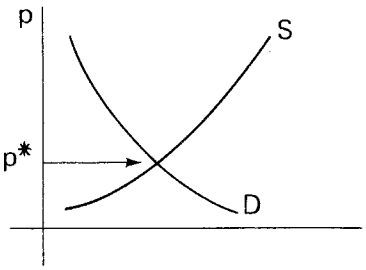


Figure 1.1

General equilibrium theory treats this problem for several markets. Let us be more precise: Suppose an economy with ℓ commodities is given. Then the space $R^\ell_+ = \{(x^1,\ldots,x^\ell) \in R^\ell;\ x^i \geqslant 0, \text{ each } i\}$ will play two roles for us: The first is as *commodity space*; so $x \in R^\ell_+$ will be interpreted as a *commodity bundle*. Thus x is the ℓ-tuple $(x^1,\ldots,x^\ell)$ with the first coordinate measuring the units of good number one, etc. But also R^ℓ_+ with the origin 0 removed will be the space of *price systems*; thus if $p \in R^\ell_+ - 0$, $p = (p^1,\ldots,p^\ell)$ represents a set of *prices* of the ℓ goods, p^1 being the price of one unit of the first good, etc.

We suppose that the economy under study presents us (axiomatically) with *demand and supply functions* $D, S: R^\ell_+ - 0 \to R^\ell_+$, from the set of price systems to commodity space. Thus $D(p)$ will be the commodity bundle demanded by the economy (or its agents in sum) at prices p. In other words, at prices $(p^1,\ldots,p^\ell)$, the vector of goods that would be purchased is $D(p)$. The *equilibrium problem* is to find (and study) under suitable conditions on D, S a price system $p^* \in R^\ell_+ - 0$ such that $D(p^*) = S(p^*)$ (equality as vectors).

Let us write $Z(p) = D(p) - S(p)$ so that the *excess demand* is a map $Z: R^\ell_+ - 0 \to R^\ell$, and we look for solution $p^* \in R^\ell_+ - 0$ of

$$Z(p^*) = 0. \tag{1.1}$$

The goal of this section is to put conditions on Z which are reasonable from economics and then to show the existence of solutions of (1.1) by a constructive method through the differential calculus. This will be done without passing to the micro-foundations of the excess demand. Then in Section 2 we will give a classical micro-foundational development for the excess demand via aggregation of demand functions of individual economic agents for a pure exchange economy. In Appendix A, we prove the full Arrow–Debreu theorem this way.

Also in this exposition, existence will at first be shown under strong hypotheses, so that one can see the methods in their simplest form. Later the hypotheses will be relaxed.

The conditions on the excess demand Z are

$$Z: R^{\ell}_{+} - 0 \to R^{\ell} \quad \text{is continuous,} \tag{1.2}$$

$$Z(\lambda p) = Z(p) \quad \text{for all} \quad \lambda > 0. \tag{1.3}$$

Thus Z is homogeneous; if the price of each good is raised or lowered by the same factor, the excess demand is not changed. This supposes we are in a complete or self-contained economy so that the prices of the commodities are not based on a commodity lying outside the system,

$$p \cdot Z(p) = 0 \quad \left(\text{using the dot product,} \quad \sum_{i=1}^{\ell} p^i Z^i(p) = 0 \right). \tag{1.4}$$

This expression states that the value of the excess demand is zero and (1.4) is called *Walras Law*. One can think of this as asserting that the demand in an economy is consistent with the assets of that economy. It is a budget constraint. The total value demanded is equal to the total value of the supply of the agents. Walras Law is no doubt the most subtle of the conditions we impose on Z here, and a micro-foundational justification will be given subsequently.

Before we state our final condition on the excess demand we give a geometric interpretation of the preceding conditions. Let $S^{\ell-1}_{+} = \{p \in R^{\ell}_{+} \mid \| p \|^2 = \sum (p^i)^2 = 1\}$ be the space of normalized price systems. By homogeneity, it is sufficient to study the restriction $Z: S^{\ell-1}_{+} \in R$. By Walras Law Z is *tangent* to $S^{\ell-1}_{+}$ at each point; $p \cdot Z(p) = 0$ says that the vector $Z(p)$ is perpendicular to p. Thus one can interpret Z as a field of tangent vectors on $S^{\ell-1}_{+}$.

The final condition on the excess demand Z is the boundary condition

$$Z^i(p) \geqslant 0 \quad \text{if} \quad p^i = 0. \tag{1.5}$$

Here $Z(p) = (Z^1(p), \ldots, Z^{\ell}(p)) \in R^{\ell}$ and $p = (p^1, \ldots, p^{\ell})$. Condition (1.5) can be interpreted simply as: if the ith good is free then there will be a positive (or at least non-negative) excess demand for it. Goods have a positive value in our model.

Theorem 1.1

If an excess demand $Z: R^{\ell}_{+} - 0 \to R^{\ell}$ is continuous, homogeneous, and satisfies Walras Law and the boundary condition [i.e., (1.2), (1.3), (1.4) and (1.5)], then there is a price system $p^* \in R^{\ell}_{+} - 0$ such that $Z(p^*) = 0$. This price system p^* is given constructively.

The last sentence will be elucidated in the proof.

The proof of Theorem 1.1 is proved via Theorems 1.2 and 1.3. These theorems are general, purely mathematical theorems about solutions of equations systems.

Theorem 1.2

Let $f: D^\ell \to R^\ell$ be a continuous map satisfying the boundary condition:

(B_D) if $x \in \partial D^\ell$ then $f(x)$ is not of the form μx for any $\mu > 0$.

Then there is $x^* \in D^\ell$ with $f(x^*) = 0$.

Here $D^\ell = \{x \in R^\ell \mid \|x\| \leqslant 1\}$ and $\partial D^\ell = \{x \in D^\ell \mid \|x\| = 1\}$.

We use for the proof of this theorem two results that have been central to global analysis and its applications to economics, the inverse mapping theorem (or implicit function theorem) and Sard's theorem. To state these results, one uses the idea of a singular point (a critical point) of a differentiable map, $f: U \to R^n$ where U is some open set of a Cartesian space, say R^k. We will say that f is C^r if its rth derivatives exist and are continuous. For x in U, the derivative $Df(x)$ (i.e., matrix of partial derivatives) is a linear map from R^k to R^n. Then x is called a *singular point* if this derivative is not surjective ("onto"). Note that if $k < n$ all points are singular. The *singular values* are simply the images under f of all of the singular points; and y in R^n is a *regular value* if it is not singular.

Inverse Mapping Theorem

If $y \in R^n$ is a regular value of a C^1 map $f: U \to R^n$, U open in R^k, then either $f^{-1}(y)$ is empty or it is a submanifold V of U of dimension $k-n$.

Here V is a *submanifold* of U of dimension $m = k-n$ if given $x \in V$, one can find a differentiable map $h: N(x) \to \mathcal{O}$ with the following properties:

(a) h has a differentiable inverse.
(b) $N(x)$ is an open neighborhood of x in U.
(c) $\mathcal{O}$ is an open set containing 0 in R^k.
(d) $h(N(x) \cap V) = \mathcal{O} \cap C$ where C is a coordinate subspace of R^k of dimension m.

Sard Theorem

If $f: U \to R^n$, $U \subset R^k$ is sufficiently differentiable (of class C^r, $r > 0$ and $r > k-n$), then the set of singular values has measure zero.

For a proof see, for example, Abraham and Robbin (1967); general background material can be found here. We say in this case that the set of regular values has *full measure*. Both of these theorems apply directly to the case of maps $f: U \to C$ where U is a submanifold of dimension k of Cartesian space of some dimension and V is a submanifold of dimension n (perhaps of some other Cartesian space). In that case the derivative $Df(x): T_x(U) \to T_{f(x)}(V)$ is a linear map on the tangent spaces.

The above summarizes the basic mathematics that one uses in the application of global analysis to economics.

Toward the proof of Theorem 1.2 consider the following problem of finding a zero of a system of equations. Suppose $f: D^\ell \to R^\ell$ is a C^2 map satisfying the very strong boundary condition:

(SB) $f(x) = -x$ for all $x \in \partial D^\ell$.

The problem is to find $x^* \in D^\ell$ with $f(x^*) = 0$. We are following Smale (1976a), influenced by a modification of Varian (1977); for history see the paper by Smale.

To solve this problem define an auxiliary map $g: D^\ell - E \to S^{\ell-1}$ by $g(x) = f(x)/\|f(x)\|$ where $E = \{x \in D^\ell \mid f(x) = 0\}$ is the solution set. Since g is C^2, Sard's theorem yields that the set of regular values of g is of full measure in $S^{\ell-1}$ (using a natural measure on $S^{\ell-1}$). Let y be such a value. Then by the inverse function theorem $g^{-1}(y)$ is a 1-dimensional submanifold which must contain $-y$ by the boundary condition. Let V be the component of $g^{-1}(y)$ starting from $-y$. So V must be a non-singular arc starting from $-y$ and open at the opposite end. Also V does not meet ∂D^ℓ at any point other than $-y$ by the boundary condition and meets $-y$ only at its initial point, since it is non-singular at $-y$. Now V is a closed subset of $D^\ell - E$ and so all its limit points lie in E. In particular E is not empty and by following along V starting from $-y$, one must eventually converge to E. This gives a geometrically constructive proof of the existence of $x^* \in D^\ell$ with $f(x^*) = 0$.

We remark that to further explicate the constructive nature of this solution, one can show that V is a solution curve of the "Global Newton" ordinary differential equation $Df(x)(dx/dt) = -\lambda f(x)$ where $\lambda = \pm 1$ is chosen according to the sign of the determinant of $Df(x)$ and changes with x. If $Df(x)$ is non-singular, then Eulers method of discrete approximation yields

$$x_n = x_{n-1} \mp Df(x_{n-1})^{-1} f(x_{n-1}),$$

which, with fixed sign, is Newton's method for solving $f(x) = 0$. Thus the "Global Newton" indeed is a global version of Newton's method in some reasonable sense. M. Hirsch and I have had some success with the computer

using the Global Newton as a tool of numerical analysis in solving systems of equations.

Now suppose only that $f\colon D^\ell \to R^\ell$ is only continuous and still satisfies $f(x) = -x$ for $x \in \partial D^\ell$. Define a new continuous map $f_0\colon D_2^\ell \to R^\ell$ by

$$f_0(x) = f(x) \quad \text{for} \quad \|x\| \leqslant 1,$$
$$f_0(x) = -x \quad \text{for} \quad \|x\| \geqslant 1.$$

Take a sequence of $\varepsilon_i \to 0$. For each i we construct a C^∞ approximation f_i of f_0, so $\|f_i(x) - f_0(x)\| < \varepsilon_i$, all $x \in D_2$. One can use "convolution" here. See Lang (1969) for details. Let φ_r be a C^∞ function on R^ℓ such that $\int \varphi_r = 1$ and the support of φ_r is contained in the disk D_r of radius r.

Then define $f_i(y) = \int f_0(y-x)\varphi_r(x)\,\mathrm{d}x = \int f_0(x)\varphi_r(y-x)\,\mathrm{d}x$ with r small enough relative to ε_i, and always $r < \frac{1}{2}$. Then f_i approximates f_0 and $f_i(x) = -x$ for $x \in \partial D_2$. We can apply the result proved above to obtain $x_i \in D_2^\ell$ with $f_i(x_i) = 0$. Clearly $x_i \in D^\ell$ and also $x_i \to \{x \in D^\ell \mid f(x) = 0\}$ as $i \to \infty$. This proves Theorem 1.2 in case of the strong boundary condition (SB). Finally, suppose only $f\colon D^\ell \to R^\ell$ is continuous and satisfies $(\mathrm{B_D})$ as in the Theorem 1.2.

We will define a continuous map $\hat{f}\colon D_2^\ell \to R^\ell$ such that $\hat{f}(x) = -x$ for $x \in \partial D_2^\ell$, as follows:

$$\hat{f}(x) = f(x) \quad \text{for} \quad \|x\| \leqslant 1,$$
$$f(x) = (2 - \|x\|) f(x/\|x\|) + (\|x\| - 1)(-x) \quad \text{for} \quad \|x\| \geqslant 1.$$

Now by the preceding result there is $x^* \in D_2^\ell$ with $\hat{f}(x^*) = 0$. But $\|x^*\| \leqslant 1$, for otherwise the boundary condition $(\mathrm{B_D})$ would be violated. Thus $f(x^*) = 0$ and the proof of Theorem 1.2 is finished.

For the main result on the existence of equilibria, we need to modify Theorem 1.2 from disks to simplices. Define

$$\Delta_1 = \left\{ p \in R_+^\ell \mid \sum p^i = 1 \right\}, \quad \partial\Delta_1 = \{ p \in \Delta_1 \mid \text{some } p_i = 0 \},$$
$$\Delta_0 = \left\{ z \in R^\ell \mid \sum z^i = 0 \right\},$$

and

$$p_c = (1/\ell, \ldots, 1/\ell) \in \Delta_1, \; p_c \text{ being the center of } \Delta_1.$$

We will deal with continuous maps $\phi\colon \Delta_1 \to \Delta_0$ which satisfy the boundary condition:

(B) $\phi(p)$ is not of the form $\mu(p - p_c)$, $\mu > 0$ for $p \in \partial\Delta_1$.

If one thinks of $\phi(p)$ as a vector based at p in ∂D_1, then $\phi(p)$ does not point radially outward in Δ_1 according to condition (B).

Theorem 1.3

Let $\phi: \Delta_1 \to \Delta_0$ be a continuous map satisfying the boundary condition (B). Then there is $p^* \in \Delta_1$ with $\phi(p^*) = 0$.

For the proof of Theorem 1.3. we will construct a "ray" preserving homeomorphism into the situation of Theorem 1.2 and apply that theorem. Define $h: \Delta_1 \to \Delta_0$ by $h(p) = p - p_c$; let $\lambda: \Delta_0 - 0 \to R^+$ be the map $\lambda(p) = -(1/\ell)(1/\min_i p_i)$. Then let $D = D^\ell \cap \Delta_0$; $\psi: D \to h(\Delta_1)$ defined by $\psi(p) = \lambda(p/\| p \|)_p$ is a ray preserving homeomorphism.

Consider the composition $\alpha: D \to \Delta_0$,

$$D \xrightarrow{\psi} h(\Delta_1) \xrightarrow{h^{-1}} \Delta_1 \xrightarrow{\phi} \Delta_0.$$

We assert that α satisfies the boundary condition (B_D) of Theorem 1.2. To that end, consider $q \in \partial D$ and let $p = \psi(q) + p_c = h^{-1}\psi(q)$. Now by (B) there is no $\mu > 0$ with $\phi(p) = \mu(p - p_c)$ or with $\mu(p - p_c) = \alpha(q)$. Equivalently there is no $\mu > 0$ with $\alpha(q) = \mu\psi(q)$, and since ψ is ray preserving that means $\alpha(q) \neq \mu q$, $\mu > 0$. This proves our assertion.

We conclude from Theorem 1.2 that there is $q^* \in D$ with $\alpha(q^*) = 0$; or if $p^* = \psi(q^*) + p_c$ then $\phi(p^*) = 0$. This proves Theorem 1.3.

To obtain Theorem 1.1, define from $Z: R^\ell_+ - 0 \to R^\ell$ of that theorem, a new map $\phi: \Delta_1 \to \Delta_0$ by $\phi(p) = Z(p) - (\sum Z^i(p))p$. Note $\sum \phi^i(p) = \sum Z^i(p) - \sum Z^i(p) \sum p^i = 0$, so that ϕ is well-defined; ϕ is clearly continuous. Also if $p \in \partial\Delta_1$, $p^i = 0$ for some i and so $\phi^i(p) = Z^i(p) \geqslant 0$. Thus (B) of Theorem 1.3 is satisfied for ϕ. Thus by Theorem 1.3 there is $p^* \in \Delta_1$ with $\phi(p^*) = 0$ or $Z(p^*) = \sum Z^i(p^*)p^*$. Take the dot product of both sides with $Z(p^*)$ to obtain, using Walras Law, that $\| Z(p^*) \|^2 = 0$ or that $Z(p^*) = 0$. This proves Theorem 1.1.

There can be natural equilibrium situations where $D(p^*) \neq S(p^*)$ as in the following one-market example for $p = 0$.

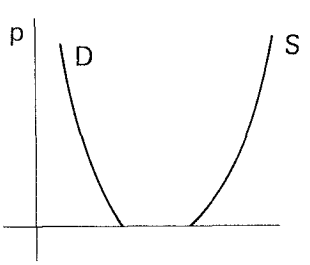


Figure 1.2

Thus for an excess demand $Z: R^\ell_+ - 0 \to R^\ell$, any p^* in $R^\ell_+ - 0$ with $Z(p^*) \leqslant 0$, i.e., $Z^i(p^*) \leqslant 0$ all i, is sometimes called an equilibrium, e.g. as in Arrow–Hahn (1971). One might also think of such a p^* as a *free disposal equilibrium* for after destroying excess supplies, one has an equilibrium with $Z(p) = 0$.

Proposition

If $Z: R^\ell_+ - 0 \to R^+$ satisfies Walras law, $p \cdot Z(p) = 0$, and $Z(p^*) \leqslant 0$, then for each i, either $Z^i(p^*) = 0$ or $p^{*i} = 0$.

Otherwise for some i, $Z^i(p^*) < 0$ and $p^{*i} > 0$; and for all i, $p^{*i} Z^i(p^*) \leqslant 0$ which contradicts Walras Law.

With weaker hypotheses than those of Theorem 1.1 one can obtain a free disposal equilibrium.

Theorem 1.4 (Debreu–Gale–Nikaidô)

Let $Z: R^\ell_+ - 0 \to R^\ell$ be continuous and satisfy this weak form of Walras Law, namely, $p \cdot Z(p) \leqslant 0$. Then there is $p^* \in R^\ell_+ - 0$ with $Z(p^*) \leqslant 0$. See Debreu (1959).

Note first that Theorem 1.4 implies Theorem 1.1. For let Z satisfy the hypotheses of Theorem 1.1, then by Theorem 1.4 there is p^* with $Z(p^*) \leqslant 0$. By the above proposition, for each i, $Z^i(p^*) = 0$ or $p^{*i} = 0$. But by the boundary condition of Theorem 1.1, if $p^{*i} = 0$ then $Z^i(p^*) \geqslant 0$, so in fact $Z^i(p^*) = 0$ and thus $Z(p) = 0$.

For the proof of Theorem 1.4, let $\beta: R \to R$ be the function $\beta(t) = 0$ for $t \leqslant 0$, and $\beta(t) = t$ for $t \geqslant 0$. Define $\bar{Z}: R^\ell_+ - 0 \to R^\ell_+$ by $\bar{Z}_i(p) = \beta(Z^i(p))$ for all i, p. Now just as in the proof of Theorem 1.1 above, define $\phi: \Delta_1 \to \Delta_0$ by $\phi(p) = \bar{Z}(p) - \left(\sum \bar{Z}^i(p)\right) p$. This ϕ satisfies the hypotheses of Theorem 1.3 and so there is $p^* \in \Delta_1$ with $\phi(p^*) = 0$ or $\bar{Z}(p^*) = \sum Z^i(p^*) p^*$. Take the inner product of both sides by $Z(p)$ and use the weak Walras to obtain $\sum Z^i(p^*) \beta(Z^i(p^*)) \leqslant 0$. But $t\beta(t) > 0$ unless $t \leqslant 0$ in which case $t\beta(t) = 0$. Therefore $Z^i(p^*) \leqslant 0$ all i. This proves Theorem 1.4.

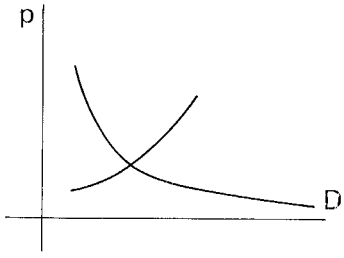


Figure 1.3

Another generalization of Theorem 1.1, and Theorem 1.4 as well, will be proved to account for $Z^i(p)\to\infty$ as $p^i\to 0$, e.g. the phenomenon illustrated in Figure 1.3. This theorem, Theorem 1.5 below, is a slight generalization of a theorem in Arrow–Hahn (1971, ch. 2, theorem 3).

Suppose now that the excess demand Z is defined only on a subset $\mathcal{D}$ of $R^\ell_+ - 0$ where $\mathcal{D}$ contains all of the interior of R^ℓ_+ and if $p\in\mathcal{D}$, so does λp for each $\lambda>0$. Consider

$$Z:\mathcal{D}\to R^\ell \quad \text{is continuous,} \tag{1.2'}$$

$$Z(\lambda p)=Z(p), \quad \text{all} \quad p\in\mathcal{D}, \quad \lambda>0, \tag{1.3'}$$

$$p\cdot Z(p)\leqslant 0, \quad \text{all} \quad p\in\mathcal{D}, \tag{1.4'}$$

$$\sum Z^i(p_k)\to\infty \quad \text{if} \quad p_k\to\bar{p}\notin\mathcal{D}. \tag{1.5'}$$

Theorem 1.5

Let $Z:\mathcal{D}\in R^\ell$ satisfy (1.2'), (1.3'), (1.4') and (1.5'). Then there is a $p^*\in\mathcal{D}$ with $Z(p^*)\leqslant 0$.

Let $\beta: R\to R$ be as in the previous proof and define $\alpha: R\to R$ by fixing $c>0$ and letting

$$\begin{aligned}\alpha(t)&=0 \quad \text{for} \quad t\leqslant 0,\\ &=1 \quad \text{for} \quad t\geqslant c,\\ &=t/c \quad \text{otherwise.}\end{aligned}$$

Define $\bar{Z}: R^\ell_+ - 0\to R^\ell_+$ by

$$\bar{Z}^i(p)=1 \quad \text{if} \quad p\notin\mathcal{D},$$

$$=\Big(1-\alpha\Big(\sum Z^i(p)\Big)\Big)\beta\big(Z^i(p)\big)+\alpha\Big(\sum Z^i(p)\Big) \quad \text{otherwise.}$$

Then $\bar{Z}$ is continuous.

Just as in the proof of Theorems 1.1 and 1.4 above, define $\phi:\Delta_1\to\Delta_0$ by $\phi(p)=\bar{Z}(p)-\sum\bar{Z}^i(p)p$. Then ϕ satisfies the hypotheses of Theorem 1.3, and so there is $p^*\in\Delta_1$ with $\phi(p^*)=0$ or

$$\bar{Z}(p^*)=\sum\bar{Z}^i(p^*)p^*.$$

First suppose that $p^*\in\mathcal{D}$. Take the inner product of both sides with $Z(p^*)$ to

obtain $Z(p^*)\cdot\bar{Z}(p^*)\leqslant 0$ (using the weak Walras Law). Then

$$\sum_i \left(1-\alpha\Big(\sum_i Z^i(p^*)\Big)\right) Z^i(p^*)\beta(Z^i(p^*)) + \alpha\Big(\sum Z^i(p^*)\Big)\sum Z^i(p^*)\leqslant 0.$$

Since for any t, $t\alpha(t)\geqslant 0$, we have as a consequence that

$$\Big(1-\alpha\Big(\sum Z^i(p^*)\Big)\Big)\sum Z^i(p^*)\beta(Z^i(p^*))\leqslant 0,$$

and even

$$\sum Z^i(p^*)\beta(Z^i(p^*))\leqslant 0.$$

But $t\beta(t)$ is strictly positive unless $t\leqslant 0$. Therefore $Z^i(p^*)\leqslant 0$ all i.

On the other hand if $p^*\notin\mathcal{D}$, it follows from the above equation on $\bar{Z}$ that p^* is $(1,\ldots,1)1/\ell$ which is in $\mathcal{D}$. So in fact p^* can't be outside $\mathcal{D}$. This proves Theorem 1.5.

2. Pure exchange economy: Existence of equilibria

This section has two parts; in the first we make stronger hypotheses and emphasize differentiability, while the second is more general. The two are pretty much independent. The existence theorems are special cases of the Arrow–Debreu theorem; see Debreu (1959) and Appendix A.

To start with, consider a single trader with commodity space $P=\{x\in R^\ell \mid x=(x^1,\ldots,x^\ell),\ x^i>0\}$. Thus x in P will represent a commodity bundle associated with this economic agent. It will be supposed that a preference relation on P is represented by a "utility function" $u\colon P\to R$ so that the trader prefers x to y in P exactly when $u(x)>u(y)$. The sets $u^{-1}(c)$ in P for c in R are called the *indifference surfaces*. Strong hypotheses of classical type are postulated:

$$u\colon P\to R \text{ is } C^2. \tag{2.1}$$

Now let $g(x)$ be the oriented unit normal vector to the indifference surface $u^{-1}(c)$ at x, $c=u(x)$. One can express $g(x)$ as $\operatorname{grad}u(x)/\|\operatorname{grad}u(x)\|$ where $\operatorname{grad}u=(\partial u/\partial x^1,\ldots,\partial u/\partial x^n)$. Then g is a C^1 map from P to $S^{\ell-1}$, $S^{\ell-1}=\{p\in R^\ell \mid \|p\|=1\}$. It plays a basic role in the analysis of consumer preferences and demand theory.

Our second hypothesis is a strong differentiable version of free disposal, "more is better", or monotonicity,

$$g(x)\in P\cap S^{\ell-1}=\operatorname{int}S^{\ell-1}_+ \quad \text{for each} \quad x\in P. \tag{2.2}$$

The word interior is shortened to int. So (2.2) means that all of the partial derivatives $\partial u/\partial x^i$ are positive.

Our third hypothesis is one of convexity, again in a strong and differentiable form. For $x \in P$, the derivative $Dg(x)$ is a linear map from R^ℓ to the perpendicular hyperplane $g(x)^\perp$ of $g(x)$. One may think of $g(x)^\perp$ as either the tangent space $T_{g(x)}(S^{\ell-1})$ or as the tangent plane of the indifference surface at x. The restriction of $Dg(x)$ to $g(x)^\perp$ is a symmetric linear map of $g(x)^\perp$ into itself,

$$\begin{aligned} &Dg(x) \text{ restricted to } g(x)^\perp \text{ has} \\ &\text{strictly negative eigenvalues.} \end{aligned} \tag{2.3}$$

We have sometimes called this condition (2.3) "differentiably convex". One can restate (2.3) equivalently as

$$\begin{aligned} &\text{The second derivative } D^2u(x) \text{ as a symmetric bilinear form} \\ &\text{restricted to the tangent hyperplane } g(x)^\perp \text{ of the indif-} \\ &\text{ference surface at } x \text{ is negative definite.} \end{aligned} \tag{2.3'}$$

We can see the equivalence of (2.3) and (2.3') as follows: Let $Du(x): R^\ell \to R$ be the first derivative of u at x with kernel denoted by $\operatorname{Ker} Du(x)$. Then since $v \cdot g(x) = Du(x)(v)/\|\operatorname{grad} u(x)\|$, $v \in \operatorname{Ker} Du(x)$ is the same condition as $v \cdot \operatorname{grad} u(x) = 0$ or $v \cdot g(x) = 0$ or yet $v \in g(x)^\perp$. Let $v_1, v_2 \in \operatorname{Ker} Du(x)$. Then $v_1 \cdot g(x) = Du(x)(v_1)/\|\operatorname{grad} u(x)\|$ and $v_1 \cdot Dg(x)(v_2) = D^2u(x)(v_1, v_2)/\|\operatorname{grad} u(x)\|$. This implies that (2.3) and (2.3') are equivalent.

Next we show:

Proposition 2.1

If $u: P \to R$ satisfies (2.3) then $u^{-1}[c, \infty)$ is strictly convex for each c.

Proof

We show that the minimum of u on any segment can not be in the interior of that segment. More precisely let $x, x' \in P$ with $u(x) \geqslant c$, $u(x') \geqslant c$. Let S be the segment $\{\lambda x + (1-\lambda)x' | 0 < \lambda < 1\}$. Let $x^* = \lambda^* x + (1-\lambda^*)x'$ be a minimum for u on S. Then $Du(x^*)(v) = 0$ where $v = x' - x$; since x^* is a minimum, $D^2u(x^*)(v, v) \geqslant 0$. This contradicts our hypothesis (2.3') that $D^2u(x^*) < 0$ on $\operatorname{Ker} Du(x^*)$. Therefore u is greater than c on S.

The final condition on u is a boundary condition and has the effect of avoiding problems associated with the boundary of R^ℓ_+:

$$\begin{aligned} &\text{The indifference surface } u^{-1}(c) \\ &\text{is closed in } R^\ell \text{ for each } c. \end{aligned} \tag{2.4}$$

This may be interpreted as the condition that the agent desires to keep at least a little of each good. It is used in Debreu (1959).

We derive now the *demand* function from the utility function of the trader. For this suppose given a *price* system $p \in \operatorname{int} R^{\ell}_{+}$ (of course $\operatorname{int} R^{\ell}_{+} = P$) and a *wealth* $w \in R_{+} = \{w \in R \mid w > 0\}$. This definition of R_{+} is convenient though maybe not consistent. Consider the *budget set* $B_{p,w} = \{x \in P \mid p \cdot x = w\}$. One thinks of $B_{p,w}$ as the set of goods attainable at prices p with wealth w. The demand $f(p,w)$ is the commodity bundle maximizing satisfaction (or utility) on $B_{p,w}$. Note that $B_{p,w}$ is bounded and non-empty, and that u restricted to $B_{p,w}$ has compact level surfaces. Therefore u has a maximum x on $B_{p,w}$ which is unique by our convexity hypothesis (2.3) (Proposition 2.1).

Then $x = f(p,w)$ is the *demand* of our agent at prices p with wealth w. It can be seen that the demand is a continuous map $f: \operatorname{int} R^{\ell}_{+} \times R_{+} \to P$. Since $x = f(p,w)$ is a maximum for u on $B_{p,w}$, the derivative $Du(x)$ restricted to $B_{p,w}$ is zero or $g(x) = p/\|p\|$. From the definition $p \cdot f(p,w) = w$ and $f(\lambda p, \lambda w) = f(p,w)$ for all $\lambda > 0$. Thus:

Proposition 2.2

The individual demand $f: \operatorname{int} R^{\ell}_{+} \times R_{+} \to P$ is continuous and satisfies

(a) $g(f(p,w)) = p/\|p\|$,
(b) $p \cdot f(p,w) = w$,
(c) $f(\lambda p, \lambda w) = f(p,w)$ if $\lambda > 0$.

Furthermore we will show the following classical fact with a modern version in Debreu (1972).

Proposition 2.3

The demand is C^1 (and will have the class of differentiability of g in general).

For the proof, note that from Proposition 2.2, we can obtain

$$\varphi: P \to (\operatorname{int} S^{\ell-1}_{+}) \times R_{+}, \qquad \varphi(x) = (g(x), x \cdot g(x)),$$

which is an *inverse* to the restriction of f to $(\operatorname{int} S^{\ell-1}_{+}) \times R_{+}$. Since φ is C^1, by a version of the inverse function theorem, f will be C^1 if the derivative $D\varphi(x)$, of φ at an arbitrary $x \in P$ is non-singular. To show that $D\varphi(x)$ is non-singular, it is sufficient to prove, $D\varphi(x)(\eta) = 0$ implies $\eta = 0$. For $\eta \in R^{\ell}$, we may write

$$D\varphi(x)(\eta) = (Dg(x)(\eta), \eta \cdot g(x) + x \cdot Dg(x)(\eta)).$$

So if $D\varphi(x)(\eta) = 0$, by this expression surely $Dg(x)(\eta) = 0$, so $\eta \in \operatorname{Ker} Dg(x)$.

But also $\eta \cdot g(x)=0$, $\eta \in g(x)^{\perp}$ and we know (3) that $Dg(x)$ restricted to $g(x)^{\perp}$ is non-singular. In other words $g(x)^{\perp} \cap \operatorname{Ker} Dg(x)=0$. This proves Proposition 2.3.

Let us elucidate this a bit. From what we have just said we may write R^{ℓ} as a direct sum $R^{\ell}=g(x)^{\perp} \oplus \operatorname{Ker} Dg(x)$ or write $\eta \in R^{\ell}$ uniquely as $\eta=\eta_1+\eta_2$ with $\eta_1 \cdot g(x)=0$, $Dg(x)(\eta_2)=0$. See Figure 2.1.

Here we are basing vectors at x. We may orient the line $\operatorname{Ker} Dg(x)$ by saying $\eta \in \operatorname{Ker} Dg(x)$ is positive if $\eta \cdot g(x)>0$. The following interpretation can be given to this line: Since $Dg(x)$ is always non-singular, the curve $g^{-1}(p)$ with $p=g(x)$, p fixed in $S_{+}^{\ell-1}$ is non-singular. It is called the *income expansion path*. At $x \in P$, the tangent line to $g^{-1}(p)$ is exactly $\operatorname{Ker} Dg(x)$ (from the definition). This curve may be interpreted as the path of demand increasing with wealth as long as prices are fixed. One may consider wealth as a function $w: P \to R$ defined by $w(x)=x \cdot g(x)$. Then w is strictly increasing along each income expansion path, and in fact $g^{-1}(p)$ can be differentiably parameterized by w.

Suppose now that the trader's wealth comes from an endowment e in P, and is the function $w=p \cdot e$ of p. Then the last property of the demand is given by:

Proposition 2.4

Let p_i be a sequence of price vectors in $\operatorname{int} R_{+}^{\ell}$ tending to p^* in ∂R_{+}^{ℓ} as $i \to \infty$. Then $\| f(p_i, p_i \cdot e)\| \to \infty$ as $i \to \infty$.

Proof

If the conclusion were false, by taking a subsequence and re-indexing we have $f(p_i, p_i e) \to x^*$. Since $u(f(p_i, p_i \cdot e)) \geqslant u(e)$ all i, by use of (2.4), x^* is in P. Therefore $g(x^*)$ is defined and equals p^*. But since $p^* \in \partial R_{+}^{\ell}$, we have a contradiction with our monotonicity hypothesis (2.2). This proves Proposition 2.4.

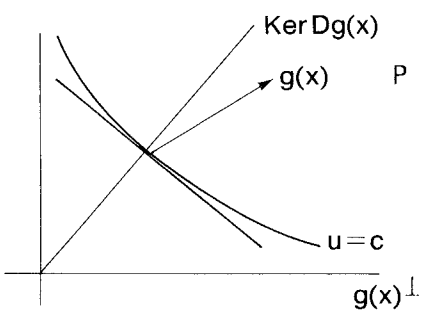


Figure 2.1

A *pure exchange economy* consists of the following: there are m agents, who are traders, and to each is associated the same commodity space P. Agent number i for $i=1,\ldots,m$ has a preference represented by a utility function $u_i: P\to R$ satisfying the conditions (2.1)–(2.4). We suppose also that to the ith agent is associated an endowment $e_i\in P$. Thus at a price system, $p\in R^\ell_+ - 0$, the income or wealth of the ith agent is $p\cdot e_i$.

One may interpret this model as a trading economy where each agent would like to trade his endowed goods for a commodity bundle which would improve or even maximize his/her satisfaction (constrained by the budget). The notion of economy may be posed as follows:

A *state* consists of an *allocation* $x\in(P)^m$, $x=(x_1,\ldots,x_m)$, $x_i\in P$ together with a price system $p\in S^{\ell-1}_+$. An allocation is called *feasible* if $\sum x_i=\sum e_i$. Thus the total resources of the economy impose a limit on allocations; there is no production. The state $(x,p)\in(P)^m\times S^{\ell-1}_+$ will be called a competitive or Walras *equilibrium* if it satisfies conditions (A) and (B):

(A) $\sum x_i=\sum e_i$.

This is the feasibility condition mentioned above.

(B) For each i, x_i maximizes u_i on the *budget set* $B=\{y\in P \mid p\cdot y=p\cdot e_i\}$.

Note that by the monotonicity condition (2.2) above, (B) does not change if in the definition of the budget set $p\cdot y=p\cdot e_i$ is replaced by $p\cdot y\leqslant p\cdot e_i$.

Note that (B) can be replaced by conditions (B_1) and (B_2):

(B_1) $p\cdot x_i=p\cdot e_i$ for each i.

(B_2) $g_i(x_i)=p$ for each i.

With (A), (B_1), and (B_2), equilibrium is given explicitly as the solution of a system of equations. We will show:

Theorem 2.5

Suppose given a pure exchange economy. More precisely let there be m traders with endowments $e_i\in P$, $i=1,\ldots,m$, and preferences represented by utilities $u_i: P\to R$, each satisfying conditions (2.1)–(2.4). Then there is an equilibrium; i.e., there are $x_i\in P$, $i=1,\ldots,m$, and $p\in S^{\ell-1}_+$ satisfying (A) and (B).

We may translate the equilibrium conditions (A) and (B) into a problem of supply and demand. Let $S: R^\ell_+ - 0\to R^\ell_+$ be the constant map, $S(p)=\sum e_i$. Let $D:\operatorname{int} R^\ell_+\to R^\ell_+$ be defined by $D(p)=\sum f_i(p,p\cdot e_i)$ where $f_i(p,p\cdot e_i)$ is the demand generated by u_i (Proposition 2.2). Define the excess demand $Z:\operatorname{int} R^\ell_+\to R^\ell$ by $Z(p)=D(p)-S(p)$. We note that the equilibrium conditions (A) and (B) are satisfied for (x,p) if and only if $Z(p)=0$ and $x_i=f_i(p,p\cdot e)$. So if we

can find a solution of $Z(p)=0$ by Section 1, we will have shown the existence of an economic equilibrium in the setting of a pure exchange economy.

Walras Law for Z [(1.4)] is verified directly; if $p \in \text{int } R^{\ell}_{+}$,

$$p \cdot Z(p) = p \cdot D(p) - p \cdot S(p) = \sum p \cdot f_i(p, p \cdot e_i) - p \cdot \sum e_i = 0.$$

Homogeneity, that $Z(\lambda p) = Z(p)$ for $\lambda > 0$ is checked as easily.

To apply the existence theorem, Theorem 1.6, we take $\mathcal{D}$ to be $\text{int } R^{\ell}_{+}$. It remains only to verify the boundary condition (2.5′), that if p tends to a point in the boundary of $R^{\ell}_{+} - 0$, the $\sum Z^i(p) \to \infty$. But that is a consequence of Proposition 2.4, using the fact that Z is bounded below. Thus we have shown the existence of $p^* \in \mathcal{D}$ with $Z(p^*) \leqslant 0$. But by the proposition preceding Theorem 1.4, it must be that $Z(p^*) = 0$ since Walras Law is satisfied. This proves Theorem 2.5.

We give another setting for a pure exchange economy where we use only continuous preferences.

For this consider a preference relation on the full R^{ℓ}_{+} as commodity space (rather than its interior P) represented by a continuous utility function $u: R^{\ell}_{+} \to R$. We replace conditions (2.1) to (2.4) simply by:

$$u: R^{\ell}_{+} \to R \text{ is continuous,} \tag{2.1′}$$

and

$$u(\lambda x + (1-\lambda)x') > c \quad \text{if} \quad u(x) \geqslant c, \quad u(x') \geqslant c \quad \text{and} \quad 0 < \lambda < 1. \tag{2.2′}$$

The latter is a strict convexity condition on the preference relation.

Suppose that to each trader, in addition to a preference of the above type, is associated an endowment e_i in P. Thus each agent has a positive amount of each commodity.

Theorem 2.6

Given a utility $u_i: R^{\ell}_{+} \to R$ for agents $i = 1, \ldots, m$ satisfying (2.1′), (2.2′) above and endowments $e_i \in P$, $i = 1, \ldots, m$, there is a ("free disposal") equilibrium (x^*, p^*). Thus:

(a) $\sum x_i^* \leqslant \sum e_i$, and

(b) x_i^* maximizes u_i on the budget set $\{x_i \in R^{\ell}_{+} \mid p^* \cdot x_i \leqslant p^* \cdot e_i\}$ at x_i^* for each i.

Proof

Before constructing a demand, we cut off commodity space near ∞ to avoid problems with unboundedness. We are able to get away with this because of the

feasibility condition. More precisely choose $c > \|\sum e_i\|$ and let D_c be the ball of radius c or $D_c = \{p \in R^\ell \mid \|p\| \leqslant c\}$. Define an associated *false demand* function $\hat{f}_i : (R^\ell_+ - 0) \times R_+ \to X_c$, $X_c = D_c \cap R^\ell_+$, by taking $\hat{f}_i$ at (p, w) to be the maximum of u_i on $\hat{B}_{p,w} = \{x \in X_c \mid p \cdot x \leqslant w\}$. Then since $\hat{B}_{p,w}$ is compact, convex, and non-empty, by the strict convexity property of u_i, $\hat{f}_i(p, w)$ is well-defined.

Proposition 2.7

The false demand $\hat{f}_i : (R^\ell_+ - 0) \times R_+ \to X_c$ is continuous, $\hat{f}_i(\lambda p, \lambda w) = \hat{f}_i(p, w)$ for $\lambda > 0$, and $p \cdot \hat{f}_i(p, w) \leqslant w$. Also if $\|\hat{f}_i(p, w)\| < c$, then the maximum, $f_i(p, w)$, of u_i on $B_{p,w} = \{x \in R^\ell_+ \mid p \cdot x \leqslant w\}$ exists (the true demand!) and $f_i(p, w) = \hat{f}_i(p, w)$.

Proof

This is straightforward except perhaps for the last. Let $\hat{x}_i = \hat{f}_i(p, w)$ with $\|\hat{x}_i\| < c$ and consider $x_i \in B_{p,w}$ with $u_i(x_i) \geqslant u_i(\hat{x}_i)$. Let S be the segment between $\hat{x}_i$ and x_i in R^ℓ_+. For any $x'_i \neq \hat{x}_i$ on $S \cap X_c$, $u(x'_i) > u_i(\hat{x}_i)$ by strict convexity (2.2′), contradicting the choice of $\hat{x}_i$. This proves Proposition 2.7.

Next define $\hat{D}(p) = \sum \hat{f}_i(p, p \cdot e_i)$, $S(p) = \sum e_i$, and $\hat{Z} : R^\ell_+ - 0 \to R^\ell$ by $\hat{Z} = \hat{D} - S$. Then $\hat{Z}$ satisfies the weak Walras Law, so by Theorem 1.4, there exists p with $\hat{Z}(p) = 0$. Thus if $\hat{f}_i(p, p \cdot e_i) = \hat{x}_i$, $\sum \hat{x}_i = \sum e_i$ and $\|\hat{x}_i\| < c$. Therefore by Proposition 2.7, $\hat{x}_i = x_i = f_i(p, p \cdot e_i)$ and $(x_1, \ldots, x_m, p)$ is an equilibrium; Theorem 2.6 is proved.

Suppose $u_i : R^\ell_+ \to R$ satisfies:

No Satiation Condition: $u_i : R^\ell_+ \to R$ has no maximum.

Then we claim that the commodity vector $x_i = f_i(p, w)$ at the end of the proof of Theorem 2.6 satisfies $p \cdot f_i(p, w) = w$ (rather than inequality). Otherwise choose x_i^* in R^ℓ_+ outside $B_{p,w}$ with $u_i(x_i^*) \geqslant u_i(f_i(p, w))$ by the No Satiation Condition. By strict convexity, as in the proof of Proposition 2.7, we get a contradiction. Thus in this case we have that for the excess demand $Z(p) = \sum f_i(p, p \cdot e_i) - S(p)$, the usual Walras Law is satisfied at equilibrium and we obtain a more satisfactory interpretation of the free disposal equilibrium (see the proposition preceding Theorem 1.4.).

The question of relaxing strict convexity in Theorem 2.6, as well as questions of production, we defer to Appendix A.

3. Pareto optimality

Towards the problems of Pareto optimality in equilibrium theory and the "fundamental theorem of welfare economics", we consider abstract optimization problems in this section.

Our setting is an open set W in R^n (W could be a smooth manifold or submanifold in what follows) together with C^2 functions $u_i: W \to R$, $i=1,\ldots,m$. One might think of W as the space of states of society and the members of that society have preferences represented by the u_i. A point $x \in W$ is called *Pareto optimal* (or just optimal) if there is no $y \in W$ with $u_i(y) \geqslant u_i(x)$ all i and strict inequality for some i. Such a y could be called *Pareto superior* to x. If $m=1$, an optimum is the same thing as an ordinary maximum. The point $x \in W$ is a *local optimum* if there is a neighborhood N of x and x is an optimum for $u_1,\ldots,u_m$ restricted to N. A point $x \in W$ is a *strict optimum* if whenever $y \in W$ satisfies $u_i(y) \geqslant u_i(x)$, all i, then $y=x$ (like a strict maximum). Finally a *local strict optimum* is defined similarly. Note that these definitions apply generally, e.g. to non-open W in R^n. The goal of this section is to give calculus conditions for local optima. The following theorem is proved in Smale (1975) and Wan (1975); we follow the Smale paper especially, which one can see for more history.

Theorem 3.1

Let $u_1,\ldots,u_m: W \to R$ be C^2 functions where W is an open set in R^n. If $x \in W$ is a local optimum, then there exist $\lambda_1,\cdots,\lambda_m \geqslant 0$, not all zero and

$$\sum \lambda_i Du_i(x)=0. \tag{3.1}$$

Further suppose $\lambda_1,\ldots,\lambda_m$, x are as above and

$$\begin{aligned}&\sum \lambda_i D^2u_i(x) \text{ is negative definite on the space}\\ &\{v \in R^n \mid \lambda_i Du_i(x)(v)=0,\ i=1,\ldots,m\}.\end{aligned} \tag{3.2}$$

Then x is a local strict optimum.

Here $Du_i(x)$ is the derivative of u_i at x as a real valued linear function on R^n, and $D^2u_i(x)$ is the second derivative as a quadratic form on R^n [one could think of $D^2u_i(x)$ as the square matrix of second partial derivatives]. $\sum \lambda_i D^2u_i(x)$ is then also a quadratic form.

Note that if one takes $m=1$ and $n=1$, the theorem becomes the basic beginning calculus theorem on maxima. For $m=1$, and n arbitrary, the theorem might be in an advanced calculus course. It has been pointed out to me by several people that one can reduce the proof of Theorem 3.1 to this case of $m=1$. However the direct proof we will give has some advantages with the geometry and symmetry in the u_i's. In the following Im stands for image.

Proof of Theorem 3.1

Let $Pos=\{v \in R^m \mid v=(v_1,\ldots,v_m),\ v_i>0\}$ and $\overline{Pos}$ its closure. Then the first condition of the theorem may be stated as there is $\lambda \in \overline{Pos}-0$ with $\lambda \cdot Du(x)=0$

(dot product). Here $u=(u_1,\ldots,u_m)$ maps W into R^m. Let x be a local optimum and suppose $\mathrm{Im}\,Du(x)\cap Pos\neq\phi$. Then choose $v\in R^n$ with $Du(x)(v)\in Pos$, and $\alpha(t)$ a curve through x in W with $\alpha(0)=x$ and the $\alpha'(0)=v$. Clearly for small values of t, $u_i(\alpha(t))>u_i(\alpha(0))=u_i(x)$ so that x is no local optimum. Thus we know that $\mathrm{Im}\,Du(x)\cap Pos=\phi$.

From this it follows from an exercise in linear algebra that there is some $\lambda\in Pos-0$ with λ orthogonal to $\mathrm{Im}\,Du(x)$. Thus $\lambda\cdot Du(x)=0$, and the first part of the theorem is proved.

Suppose that the theorem (second part) is true in case $\lambda_i>0$, all i, and consider the general case. Let the indices be such that $\lambda_1,\ldots,\lambda_k>0$, $\lambda_{k+1}=\cdots=\lambda_m=0$. Then conditions (3.1) and (3.2) are the same for optimizing $u_1,\ldots,u_m$ at x and optimizing $u_1,\ldots,u_k$ at x. So (3.1) and (3.2) are satisfied for $u_1,\ldots,u_k$ also; and since by assumption the theorem is true in this case, x is a strict local optimum for the $u_1,\ldots,u_k$. But then it is also a strict local optimum for $u_1,\ldots,u_m$. From this it is sufficient to prove the theorem in the case all the λ_i are strictly positive.

We may suppose that x is the origin of R^n and $u(x)=0$ in $\boldsymbol{R}^m$, so that the symbol x will remain free to denote any point in W. Then the condition that $0\in W$ is a local strict optimum is that there is some neighborhood N of 0 in W with $(u(N)-0)\cap\overline{Pos}=\phi$. We will show that under the conditions of Theorem 3.1, indeed there is such an N.

Denote by K or $\mathrm{Ker}\,Du(0)$ the kernel of $Du(0)$ as a linear subspace of R^n and by $K^\perp$ its orthogonal complement.

Lemma 3.2

There exist $r,\delta>0$ with the property that when $\|x\|<r$, $x=(x_1,x_2)$, $x_1\in K$, $x_2\in K^\perp$ and $\delta\|x_1\|\geqslant\|x_2\|$ then $\lambda\cdot u(x)<0$ if $x\neq 0$.

Proof

Let $H=\sum\lambda_i D^2u_i(0)$. By (3.2) there is some $\sigma>0$ so that $H(x,x)\leqslant-\sigma\|x\|^2$ for $x\in K$. For $x\in R^n$, $x=(x_1,x_2)$, $x_1\in K$, $x_2\in K^\perp$, we may write $H(x,x)=H(x_1,x_1)+2H(x_1,x_2)+H(x_2,x_2)$. Since $|H(x_1,x_2)|\leqslant C\|x_1\|\|x_2\|$, $|H(x_2,x_2)|\leqslant C_1\|x_2\|^2$, we choose $\eta,\delta>0$ so that if $\delta\|x_1\|\geqslant\|x_2\|$ then $H(x,x)\leqslant-\eta\|x\|^2$. Write by Taylor's theorem for $\|x\|<r$, $u(x)=Du(0)(x)+D^2u(0)(x,x)+R_3(x)$ where $\|\lambda\cdot R_3(x)\|<\eta/2\|x\|^2$. Taking the dot product with λ yields the lemma.

Now write $J=\mathrm{Im}\,Du(0)$ and write u in R^m as $u=(u_a,u_b)$, $u_a\in J$, $u_b\in J^\perp$.

Lemma 3.3

Given $\alpha>0$ and $\delta>0$ there is $s>0$ so that if $\|x\|<s$, $x=(x_1,x_2)$, $x_1\in K$, $x_2\in K^\perp$ with $\|x_2\|\geqslant\delta\|x_1\|$, then $\|u_b(x)\|\leqslant\alpha\|u_a(x)\|$.

Proof

The restriction

$$D u(0)_{K^{\perp}} : K^{\perp} \to \operatorname{Im} D u(0)$$

is a linear isomorphism so there are positive constants c_1, c with

$$\|Du(0)(x)\| = \|Du(0)(x_2)\| \geqslant c_1 \|x_2\| \quad \text{all} \quad x = (x_1, x_2),$$

$$\geqslant c\|x\| \quad \text{if} \quad \|x_2\| \geqslant \delta \|x_1\|.$$

By the Taylor's series

$$u_a(x) + u_b(x) = u(x) = Du(0)(x) + R(x),$$

so that given $\beta > 0$, we may assume $\|R(x)\| \leqslant \beta \|x\|$ for $\|x\| <$ some number s. With $R = (R_a, R_b)$ we have

$$\|u_a(x)\| = \|Du(0)(x) + R_a(x)\| \geqslant (c - \beta)\|x\|,$$

and

$$\|u_b(x)\| = \|R_b(x)\| \leqslant \beta \|x\|,$$

say with β small enough and $\beta/(c-\beta) < \alpha$. Then $\|u_b(x)\| \leqslant \alpha \|u_b(x)\|$, finishing the proof of the lemma.

To finish the proof of Theorem 3.1, choose α of Lemma 3.3 so that if $\|u_a(x)\| \leqslant \alpha \|u_a(x)\|$ then $u(x) \notin \overline{Pos} - 0$. This can be done since $\operatorname{Im} Du(0) \cap \overline{Pos} = 0$, all the λ_is being strictly positive. Choose a disk around 0 of radius r_0, $r_0 < r$ of Lemma 3.2 and $r_0 < s$ of Lemma 3.3. Let the δ of Lemma 3.3. be given by Lemma 3.2. Now from the two lemmas we have that $u(x) \notin \overline{Pos}$ if $x \neq 0$, $\|x\| < r_0$, proving x to be a local strict optimum and Theorem 3.1.

We pass now to an extension of Theorem 3.1 to the setting of constrained optimization. Thus let C^2 functions $u_1, \ldots, u_m$ be defined on an open set $W \subset R^{\ell}$, subject to constraints given by conditions of the form $g_{\beta}(x) \geqslant 0$, $\beta = 1, \ldots, k$, with $g : W \to R$ of class C^2. One may express the problem by defining $W_0 = \{x \in W \mid g_{\beta}(x) \geqslant 0, \beta = 1, \ldots, k\}$ and seeking conditions for optima of the restrictions $u_1, \ldots, u_m$ to W_0.

Theorem 3.4

Suppose $x \in W_0$ is a local optimum for the functions $u_1, \ldots, u_m$ on W_0, W_0 as above. Then there exist non-negative numbers λ_i, μ_β, not all zero such that

$$\sum_{i=1}^{m} \lambda_i Du_i(x) + \sum \mu_\beta Dg_\beta(x) = 0, \tag{3.1'}$$

where

$$\mu_\beta = 0 \quad \text{if} \quad g_\beta(x) \neq 0.$$

Furthermore suppose $x \in W_0$, $\lambda_0 \geqslant 0$, $\mu_\beta \geqslant 0$ with not all the λ_i, μ_β zero, are given so that (3.1') is true. If the bilinear symmetric form

$$\sum_{i=1}^{m} \lambda_i D^2 u_i(x) + \sum \mu_\alpha D^2 g_\alpha(x) \tag{3.2'}$$

is negative definite on the linear space

$$\{v \in R^\ell \mid v \cdot \lambda_i \operatorname{grad} u_i(x) = 0, \text{ all } i, \text{ and}$$
$$v \cdot \mu_\alpha \operatorname{grad} g_\alpha(x) = 0, \text{ all } \alpha\}$$

then x is a local strict optimum for $u_1, \ldots, u_m$ restricted to W_0.

For the first part let us suppose that $g_\beta(x) = 0$ (by renumbering if necessary) precisely for all $\beta = 1, \ldots, k$, and define $\phi : W \to R^{m+k}$ by $\phi = (u_1, \ldots, u_m, g_1, \ldots, g_k)$. Then we claim that $\operatorname{Im} D\phi(x) \cap Pos = \phi$. Otherwise let $D\phi(x)(v) \in Pos$ and let $\alpha(t)$ be a curve in W satisfying $\alpha(0) = x$, $\alpha'(0) = v$. For small enough ε, $\alpha(\varepsilon)$ is in W_0 and a Pareto improvement over $\alpha(0) = x$. So x could not be locally optimal. So $\operatorname{Im} D\phi(x) \cap Pos = \phi$ and there is a vector $(\lambda_1, \ldots, \lambda_m, \mu_1, \ldots, \mu_k) \in \overline{Pos} - 0$ normal to $\operatorname{Im} D\phi(x)$, as in Theorem 3.1. This proves the first part of Theorem 3.4.

For the proof of the last part we first note, with $\phi : W \to R^{m+k}$ as above, that if $x \in W_0$ is a local strict optimum for ϕ on W, then it is also a local strict optimum for $u_1, \ldots, u_m$ on W_0. This follows from the definitions. But the hypotheses on x in the second part of Theorem 3.4 imply that x is a local strict optimum of ϕ as a consequence of Theorem 3.1. Thus Theorem 3.4 is proved.

We end this section with some final remarks:

(1) Note Theorem 3.1 is the special case of Theorem 3.4 when $k = 0$.
(2) Suppose the g_α satisfies the *Non-Degeneracy Condition* at $x \in W_0$. The set

$Dg_\beta(x)$ for β with $g_\beta(x)=0$ is linearly independent. If this condition is satisfied then in (1) at least one of the λ_i is not zero.

(3) If in Theorem 3.4 $m=1$, the first part is related to the Kuhn–Tucker theorem, and if the Non-Degeneracy Condition is met, one has $\lambda_1=1$.

Theorem 3.4 is in Smale (1974–76, V) and Wan (1975). See also Simon (forthcoming) for further information on this.

4. Fundamental theorem of welfare economics

We return to a pure exchange economy as in Section 2, with traders preferences represented by C^2 utility functions $u_i: P\to R$, $P=\operatorname{int} R^\ell_+$, $i=1,\ldots,m$, satisfying the differentiable convexity, monotonicity and strong boundary conditions (2.2), (2.3), and (2.4). Also as in Section 2, the maps $g_i: P\to S^{\ell-1}_+$ defined by $g_i(x)=\operatorname{grad} u_i(x)/\|\operatorname{grad} u_i(x)\|$ will be used in our approach. While we do not presume that each agent is given an endowment, it will be supposed that the total resources r of the economy are a fixed vector in P.

Thus the set W of attainable allocations or states has the form

$$W=\left\{x\in(P)^m \mid x=(x_1,\ldots,x_m),\ x_i\in P,\ \sum x_i=r\right\}.$$

The individual utility $u_i: P\to R$ of the ith agent induces a map $v_i: W\to R$, $v_i(x)=u_i(x_i)$. After Section 3 it is natural to ask, what the optimal states in W for the functions v_i, $i=1,\ldots,m$, are. The answer is in:

Theorem 4.1

The following three conditions on an allocation $x\in W$ (relative to the induced utilities $v_i: W\to R$) are equivalent:

(1) x is a local Pareto optimum.
(2) x is a strict Pareto optimum.
(3) $g_i(x_i)$ is a vector in $S^{\ell-1}_+$, independent of i.

Let θ be the set of $x\in W$ satisfying one of these conditions. Then θ is a submanifold of W of dimension $m-1$.

In this theorem as in this whole section, we are following Smale (1974–76).

Proof

Note (2) implies (1). We will show that (1) implies (3). For this we do not use any conditions on $u_i: P\to R$ except that the u_i are C^1.

Thus suppose that $x \in W$ is a local optimum. We apply the first part of Theorem 3.1 to obtain $\lambda_1, \ldots, \lambda_m \geqslant 0$, not all zero, such that $\sum \lambda_i D v_i(x) = 0$ or $\sum \lambda_i D u_i(x_i) = 0$. We may suppose that $\lambda_1 \neq 0$ by a change of notation. Apply the sum to the vector $\bar{x} \in (R^\ell)^m$ with $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_m)$, $\sum \bar{x}_i = 0$ (a tangent vector to W). If $\bar{x} = (\bar{x}_1, 0, \ldots, 0, -\bar{x}_1, 0, \ldots, 0)$ with $-\bar{x}_1$ in the kth place we have $\sum \lambda_i D u_i(x_i)(\bar{x}_i) = \lambda_1 D u_1(x_1)(\bar{x}_1) - \lambda_k D u_k(x_k)(\bar{x}_1) = 0$ for all $\bar{x}_1 \in R^\ell$. Thus $\lambda_k D u_k(x_k)$ is not zero all k and equal to $\lambda_1 D u_1(x_1)$. This yields condition (3).

For the equivalence of the three conditions, it remains to prove that if x satisfies (3) then (2), x is a strict optimum. So let x satisfy (3) and let $y \in W$ with $v_i(y) \geqslant v_i(x)$, all i, or equivalently, $u_i(y_i) \geqslant u_i(x_i)$, all i.

We use now:

Lemma 4.2

Let $u: P \to R$ satisfy differentiable convexity (2.3). If $y \in P$, $u(y) \geqslant u(x)$ and $y \neq x$, then $Du(x)(y-x) > 0$. Thus also in this case, $y \cdot g(x) > x \cdot g(x)$.

Proof

For $t \geqslant 0$ and $t \leqslant 1$, Proposition 2.1 (strict convexity) implies that $u(t(y-x)+x) \geqslant u(x)$, and so $(d/dt)u(t(y-x)+x)|_{t=0} \geqslant 0$. Therefore by the chain rule $Du(x)(y-x) \geqslant 0$. On the other hand by Taylor's series if $Du(x)(y-x) = 0$, $u(x+t(y-x)) = u(x) + D^2u(x)((t(y-x))^2) + R_3$ which yields by differentiable convexity [(2.3′)] $u(x+t(y-x)) < u(x)$ for small t. This lies in contradiction with the convexity. The lemma is proved.

By the lemma, for each i, $y_i \cdot g_i(x_i) \geqslant x_i \cdot g_i(x_i)$ with inequality in case $y_i \neq x_i$. Then let $p = g_i(x_i)$ using (2.8), so $\sum p \cdot y_i \geqslant \sum p \cdot x_i$ with inequality if $y_i \neq x_i$ any i. But since $y \in W$, $\sum y_i = r = \sum x_i$ and $\sum p \cdot y_i = \sum p \cdot x_i$. Thus $y_i = x_i$, each i, $y = x$ and x is a strict optimum.

For Theorem 4.1 it remains to prove that θ is an $(m-1)$ dimensional submanifold. For this we use the inverse function theorem in the form of the transversality theorem of Thom which goes as follows:

Let W, V be submanifolds of some Cartesian space (or abstract manifolds) and let Δ be a submanifold of V. Thus given $y \in \Delta$, there is a diffeomorphism h (differentiable map with a differentiable inverse) of a neighborhood U of Y in V onto a neighborhood N of 0 in R^k, $k = \dim V$, and $h(\Delta \cap U) = N \cap C$ where C is a coordinate subspace of R^k. Then $\alpha: W \to V$ is *transversal* to Δ if whenever $x \in W$ with $\alpha(x) = y \in \Delta$, $T_y(V) = \operatorname{Im} D\alpha(x) + T_y(\Delta)$. In other words, the image of the derivative $D\alpha(x): T_x(W) \to T_y(V)$ together with tangent vectors to Δ at y spans the tangent space of V at Y. Also one can think of $D\alpha(x)$ mapping surjectively onto the complement of the tangent space of Δ in $T_y(V)$.

Then the inverse function theorem implies:

Transversality Theorem

Let $\alpha: W \to V$ be transversal to the closed submanifold Δ of V. Then $\alpha^{-1}(\Delta)$ is a submanifold of W with either $\alpha^{-1}(\Delta)$ empty or $\dim W - \dim \alpha^{-1}(\Delta) = \dim V - \dim \Delta$ (codimension is preserved).

Here, the dimension is shortened to dim. References with details are Abraham and Robbin (1967) and Golubitsky and Guilemin (1973).

For the proof let $\alpha(x) = y \in \Delta$ and apply the usual inverse function theorem to the composition $\pi \circ h \circ \alpha: W \to C^{\perp}$ with h as above, $C^{\perp}$ is the orthogonal complement of C above and $\pi: R^k \to C^{\perp}$ is the projection.

Now take the W of the Transversality Theorem as the W in Theorem 4.1 and let V be the Cartesian product of m spheres, $V = (S^{\ell-1})^m$ and Δ to be the diagonal in V,

$$\Delta = \left\{ y \in (S^{\ell-1})^m \mid y = (y_1, \ldots, y_m), y_i \in S^{\ell-1}, y_1 = y_2 = \cdots = y_m \right\}.$$

Define $g: W \to (S^{\ell-1})^m$ by $g(x)$ having ith coordinate given by $g_i(x_i)$ where $g_i: P \to S^{\ell-1}$ is the normalized gradient of the utility of the ith trader. By definition [first part of Theorem 4.1, condition (3)], $g^{-1}(\Delta) = \theta$. We will show that g is transversal to Δ as follows:

Let $K_x = \operatorname{Ker} Du(x)$ where $u: W \to R^m$ is the map with the ith coordinate of $u(x)$ given by $u_i(x_i)$. Then

$$K_x = \left\{ \bar{x} \in (R^{\ell})^m \mid \bar{x}_i \in R^{\ell}, \sum \bar{x}_i = 0, \bar{x}_i \cdot g_i(x_i) = 0 \right\}.$$

Let L_x for $x \in \theta$ be the set of $\bar{x} \in T_x(W)$ with $Dg(x)(\bar{x}) \in T_{g(x)}(\Delta)$ or

$$L_x = \left\{ \bar{x} \in (R^{\ell})^m \mid \sum \bar{x}_i = 0, Dg_i(x_i)(\bar{x}_i) \text{ is independent of } i \right\}.$$

[Eventually we will see that $L_x = T_x(\theta)$ is the tangent space to θ at x.]

Lemma 4.3

$L_x \cap K_x = 0$ for all $x \in \theta$. Moreover $\dim K_x = m\ell - \ell - m + 1$.

Proof

Let $p = g_i(x_i)$ and $\gamma_i: p^{\perp} \to p^{\perp}$ be the restriction of $Dg_i(x_i)$ to $p^{\perp}$. Then γ_i is symmetric with negative eigenvalues [see condition (2.3)]. Also $\sum \gamma_i^{-1}$ is an isomorphism since γ_i^{-1} is symmetric with negative eigenvalues and the sum of

negative definite symmetric linear maps is negative definite (from linear algebra, or look at the corresponding bilinear symmetric forms).

Let $\bar{x} \in L_x \cap K_x$ and $Dg_i(x_i)(\bar{x}_i) = \bar{p}$. Then $\gamma_i^{-1}(\bar{p}) = \bar{x}_i$ since $\bar{x}_i \cdot g_i(x_i) = 0$ and $\sum \bar{x}_i = \sum \gamma_i^{-1}(p) = 0$ so $\bar{p} = 0$. Thus also $\bar{x}_i = 0$ each i, proving the first part of the lemma. The dimension of K_x is easily counted.

To finish the proof of Theorem 4.1, let us count more dimensions. It is easy to see that $\dim W = m\ell - \ell$, $\dim(S^{\ell-1}) = m\ell - m$, $\dim \Delta = \ell - 1$. From these dimensions and the lemma, $Dg(x)$ restricted to K_x maps K_x injectively into the complement of $T_y(\Delta)$ in $T_y((S^{\ell-1})^m)$, $y = (p, \ldots, p)$. This proves that g is transversal to Δ and therefore by the transversality theorem, $g^{-1}(\Delta)$ is empty or a submanifold of dimension $m - 1$. However, it cannot be empty by Theorem 2.5. using any endowments e_i which sums to r. This finishes the proof of Theorem 4.1.

Remark

By the definitions, $L_x = T_x(\theta)$ and so $\dim L_x = m - 1$, and so

$$T_x(W) = T_x(\theta) \oplus K_x \quad \text{(direct sum)}.$$

We give some consequences of Theorem 4.1:

Corollary 4.4

Let W be the space of attainable states of a pure exchange economy with fixed total resources r as above. Consider the map $u: W \to R^m$ defined by: $u(x)$ has ith coordinate $u_i(x_i)$, $i = 1, \ldots, m$, where $u_i: P \to R$ is the utility of agent i. Let θ be the submanifold of Pareto optimal points. Then u/θ, the restriction of u to θ is an imbedding of θ into R^m.

Here an *imbedding* means that the derivative is injective as a linear map from $T_x(\theta) \to R^m$, and the map is injective.

In fact, the corollary is an immediate consequence of the remark that $\operatorname{Ker} Du(x) \cap T_x(\theta) = 0$.

Then since $u(\theta)$ has codimension 1 in R^m, one may define the *Gauss map* $G: \theta \to S^{m-1}$ by letting $G(x)$ be the unit normal to $u(\theta)$ at $u(x)$, oriented so that it lies in R_+^ℓ. By definition $G(x)$ is perpendicular to the image $Du/\theta(x)$ or $G(x) \cdot Du(x)(\bar{x}) = 0$ for all $\bar{x} \in T_x(\theta)$. Since $T_x(\theta) \cap \operatorname{Ker} Du(x) = 0$, this is the same as $G(x) \cdot Du(x)(\bar{x}) = 0$ for all $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_m)$ with $\sum \bar{x}_i = 0$. Thus if we take $\lambda = (\lambda_1, \ldots, \lambda_m) = \lambda_x$ as in Theorem 4.1 and normalized as well, so that $\|\lambda_x\| = 1$, then $\lambda_x = G(x)$. In a certain way the Gauss map G is the curvature of the imbedded manifold $u(\theta)$, so that the λ of Theorem 4.1 may be thought of as a curvature. Note that the previous discussion, in contrast to the rest of this

article, depends on the utility representations u_i, not just the underlying preference.

Remark

In connection with Corollary 4.4, it is worth noting that if $x \notin \theta$, then it can be shown that $Du(x): T_x(W) \to R^m$ is surjective. If $x \in \theta$, then the image (see above) of $Du(x): T_x(W) \to R^m$ has dimension $m-1$ and it can be shown that the map u is a *fold* at x in the sense of singularities of maps. See Smale (1974–76); this aspect of the subject is developed in work of de Melo, Saari, Simon, Titus, and Wan [see Simon (forthcoming) for some references].

Corollary 4.5

Given $e \in W$, there is some x in θ so that $e - x \in K_x$. Furthermore there is a neighborhood $N(\theta)$ of θ in W so that for each $e \in N(\theta)$, there is a *unique* x in θ with $e - x \in K_x$. For an endowment vector e in $N(\theta)$, $e = (e_1, \ldots, e_m)$ there is a corresponding unique Walras equilibrium, (x, p), with $x \in \theta$, $p = g_i(x_i)$, all i, and the budget condition $p \cdot e_i = p \cdot x_i$, all i.

For the proof note that for every $x \in W$ the attainability condition of equilibrium is satisfied. If $x \in \theta$, then the satisfaction condition defining $p = g_i(x_i)$ for some i (hence all i) is also satisfied. Finally the budget condition $p \cdot e_i = p \cdot x_i$ all i may be restated as $g_i(x_i) \cdot (e_i - x_i)$, all i, or simply as $e - x \in K_x (= \operatorname{Ker} Du(x))$. Then the first sentence of Corollary 4.5 just re-expresses the existence Theorem 2.5. The uniqueness theorem, second or third sentence of the corollary, follows from the tubular neighborhood theorem of differential topology [see Golubitsky and Guilemin (1973, ch. 2, sect. 7)]. While we are following Smale (1974–76, VI), this is also close to work of Balasko (1975).

Towards the final corollary of Theorem 4.1 we give the concept of welfare equilibrium. We say that a state $(x, p) \in W \times S_+^{\ell-1}$ is a *welfare equilibrium* if x_i is a (in this case the) maximum of u_i on the budget set $B_{p, p \cdot x_i} = \{x \in P \mid p \cdot x = p \cdot x_i\}$. The subset of welfare equilibria in $W \times S_+^{\ell-1}$ will be called Λ. From this definition it follows that (x, p), $x = (x_1, \ldots, x_m)$, $x_i \in P$, $p \in S_+^{\ell-1}$ is in Λ provided (1_E), (2_E) hold:

(1_E) $\sum x_i = r$.

(2_E) $g_i(x_i) = p$, each $i = 1, \ldots, m$ (from the maximization condition on u_i).

If one has the further data of individual initial endowments, $e_i \in P$, $i = 1, \ldots, m$, summing to r, then a third condition (3_E), with (1_E) and (2_E), defines the equilibria of Section 2 or the Walras equilibria:

(3_E) $p \cdot e_i = p \cdot x_i$, $i = 1, \ldots, m$.

The welfare equilibria are called "equilibria relative to a price system" in

Debreu (1959). They play a central role in theorems of welfare economics as well as non-tatonment dynamics. It is important to distinguish these two kinds of related concepts of equilibria. When there is a danger of confusion, we use the words *Walras equilibria* with emphasis on the budget condition (3_E).

A very sharp, though perhaps not general, version of the *fundamental theorem* of *welfare economics* is the following:

Corollary 4.6

All as above, θ, Λ are $(m-1)$-dimensional submanifolds, closed as subsets of $W, W \times S_+^{\ell-1}$, respectively, and the map $\beta: \Lambda \to W$ defined by $(x, p) \to x$ is a diffeomorphism of Λ onto $\theta \subset W$.

We recall that a *diffeomorphism* is a differentiable map with differentiable inverse so that it is bijective (one to one and onto).

The usual form [compare Debreu (1959), Arrow–Hahn (1971)] states that $\Lambda \to \theta$ is well-defined and surjective, i.e., every optimal allocation is supported by a price system and the allocation part of a welfare equilibrium is optimal.

The proof of Corollary 4.6 goes as follows: Define an imbedding $\alpha: W \to W \times S_+^{\ell-1}$ by $\alpha(x) = (x, g_1(x_1))$. Then $\alpha(\theta) = \Lambda$ using Theorem 4.1; α/θ and β/Λ are inverse to each other with α/θ an imbedding of the submanifold θ. Then Λ is a submanifold and the corollary follows.

We now indicate how some of this goes without assuming any properties on the utilities $u_i: P \to R$ besides differentiability, i.e., C^2. Let θ_s be the subset of the space W of attainable allocations which consists of local strict optima. Emphasizing no hypotheses on the u_i, we still have:

Proposition 4.7

If $x \in W$ is a local optimum for the utility induced functions on W, then

(a) there exists $\lambda_i \geqslant 0$ not all 0 with $\sum \lambda_i Du_i(x_i) = 0$ (which implies that $g_i(x_i)$ is independent of i).

Further let x satisfy (a) and also

(b) $\sum \lambda_i D^2 u_i(x_i)((\bar{x}_i)^2)$ is negative whenever $\sum \bar{x}_i = 0$, $\bar{x}_i \cdot g_i(x_i) = 0$, all i, and $\bar{x}_i \neq 0$, some i.

Then $x \in \theta_s$.

For the proof note that the first part is done (Theorem 4.1). The last part just goes by applying the second part of Theorem 3.1; the situation is similar to the proof of Theorem 4.1.

The condition (b) is considerably weaker than differentiable convexity at x_i, each i. In general one may hope to circumvent convexity hypotheses by using the second-order conditions (as in Theorem 3.1). On the other hand, x may be a strict optimum with no supporting price equilibrium. In that case there is only an "extended price equilibrium" [see e.g. Smale (1974–76, III)].

We now consider the situation of Theorem 4.1 for commodity space with boundary. Up to now in this section the analysis has been interior. Thus suppose that trader i, for $i=1,\ldots,m$, has a C^2 utility representation $u_i: R^\ell_+ \to R$ of his/her preference (so u_i is defined on the full R^ℓ_+, not just the interior). The conditions of differentiable monotonicity and differentiable convexity of Section 2 will be assumed for the rest of this section. We suppose that each $u_i: R^\ell_+ \to R$ is the restriction of a C^2 function defined on some open set of R^ℓ containing R^ℓ_+. Then u_i off R^ℓ_+ will never be used. In this way the derivatives $Du_i(x), D^2u_i(x)$ still make sense for $x \in \partial R^\ell_+$ and so the conditions (2.2) and (2.3) make sense on the boundary as well.

Fix a vector $r \in \operatorname{int} R^\ell_+$ of total resources and let $W_0 = \{x \in (R^\ell_+)^m \mid \sum x_i = r\}$. Then W_0 is the space of attainable states of our pure exchange economy. Let W be a neighborhood of W_0 in $\{x \in (R^\ell)^m \mid \sum x_i = r\}$ on which the functions $v_i: W \to R$, can be defined by $v_i(x) = u_i(x_i)$, $i=1,\ldots,m$. Let $g_i^k: W \to R$ be given by $g_i^k(x) = x_i^k$. Then we are in the situation of optimizing several functions subject to constraints, or Theorem 3.4. These g_i^k are constraints as above and bear no relation to the normalized gradients of utility functions. The problem of optima in W_0 relative to the $v_i: W_0 \to R$ is equivalent to optimizing the $v_i: W \to R$ subject to $g_i^k(x) \geqslant 0$.

Theorem 4.8

For $i=1,\ldots,m$, let $u_i: R^\ell_+ \to R$ satisfy

$$\frac{\operatorname{grad} u_i(x_i)}{\|\operatorname{grad} u_i(x_i)\|} = g_i(x_i) \in S^{\ell-1}_+, \quad \text{each } x_i, \tag{4.1}$$

and

$$D^2u_i(x_i) \text{ on } g_i(x_i)^\perp \quad \text{is negative definite.} \tag{4.2}$$

Suppose $W_0 = \{x \in (R_+)^m \mid \sum x_i = r\}$ with $v_i: W_0 \to R$ defined by $v_i(x) = u_i(x_i)$. If $x \in W_0$ is a local optimum for the v_i:

(a) there exists $p \in S^{\ell-1}_+$ and $\lambda_1,\ldots,\lambda_m \geqslant 0$, not all 0, with $p \geqslant \lambda_i Du_i(x_i)$ each i, where one has equality in the kth coordinate if $x_i^k \neq 0$.

Conversely let $p, x_1,\ldots,x_m, \lambda_1,\ldots,\lambda_m$ be as in (a) with $p \cdot x_i \neq 0$ each i. Then x is a strict optimum.

For the proof let $g_i^j: W \to R$ be defined as above so that $g_i^j(x) = x_i^j$ are constraints for v_i on W. Then the derivatives satisfy $Dg_i^j(x)(\bar{x}) = \bar{x}_i^j$ where $\bar{x} \in (R^\ell)^m$ with $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_m)$ and $\sum \bar{x}_i = 0$. Also $\bar{x}_i = (\bar{x}_i^1, \ldots, \bar{x}_i^\ell)$. If x in W_0 is a local optimum for the v_i, then Theorem 3.4 applies to yield the existence of $\lambda_i \geqslant 0$, $\mu_i^j \geqslant 0$, $i = 1, \ldots, m$, $j = 1, \ldots, \ell$, not all zero with $\mu_i^j = 0$ if $x_i^j \neq 0$ and

$$\sum \lambda_i Du_i(x_i)(\bar{x}_i) + \sum \mu_i^j \bar{x}_i^j = 0, \quad \text{all } \bar{x}_i \text{ as above.}$$

Take $\bar{x}_i^j = 1$, $\bar{x}^j = -1$, all other components of $\bar{x}$ zero to obtain

$$\lambda_i Du_i(x_i)^j + \mu_i^j = \lambda_k Du_k(x_k)^j + \mu_k^j,$$

where $Du_i(x_i)^j$ denotes the jth coordinate of $Du_i(x_i)$.

Alternately we see that $q = \lambda_i Du_i(x_i) + \mu_i$ is independent of i where $\mu_i = (\mu_i^1, \ldots, \mu_i^\ell)$, $\mu_i \geqslant 0$ and $\mu_i \cdot x_i = 0$. Note that $q \neq 0$, for otherwise all the λ_i and μ_i would be zero [recall $Du_i(x_i) \neq 0$]. Let $p = q/\|q\|$ and multiply through $q = \lambda_i Du_i(x_i) + \mu_i$ by $1/\|q\|$. By renaming the λ_i, μ_i we have now

$$p = \lambda_i Du_i(x_i) + \mu_i, \qquad \mu_i \geqslant 0, \quad \lambda_i \geqslant 0, \quad \mu_i \cdot x_i = 0.$$

This yields the first part of Theorem 4.8.

For the converse let $y \in W_0$, $u_i(y_i) \geqslant u_i(x_i)$, $i = 1, \ldots, m$, $x_i, y_i \in R_+^\ell$. We must show that $y_i = x_i$ for each i. By the first lemma in the proof of Theorem 4.1, $Du_i(x_i)(y_i - x_i) \geqslant 0$ with equality only if $y_i = x_i$. By our main condition above $p \cdot x_i = \lambda_i Du_i(x_i)(x_i)$ and so $\lambda_i \neq 0$ since $p \cdot x_i \neq 0$. Then by this same condition $p \cdot (y_i - x_i) \geqslant \mu_i \cdot y_i$ or $p \cdot y_i \geqslant p \cdot x_i$, with equality only if $y_i = x_i$, each i. On the other hand $\sum y_i = \sum x_i = r$; putting this together indeed yields $y_i = x_i$ each i. This finishes the proof.

Remark

Note that if u_i satisfies the stronger monotonicity condition, that $Du_i(x_i) \in \operatorname{int} S_+^{\ell-1}$, then $p \cdot x_i \neq 0$ in Theorem 4.8 can be omitted.

Say that (x, p) is a *welfare equilibrium* (as before), or $(x, p) \in \Lambda \subset W_0 \times S_+^{\ell-1}$ if x_i is a maximum of u_i on the budget set $B_{p, p \cdot x_i} = \{x \in R_+^\ell \mid p \cdot x \leqslant p \cdot x_i\}$, each i. Thus for $(x, p) \in \Lambda$, $\sum x_i = r$, since $x \in W_0$.

Proposition 4.9

If $(x, p) \in \Lambda$, then there exist numbers $\lambda_i \geqslant 0$, $i = 1, \ldots, m$, and $\mu_i \in R^\ell$, $\mu_i \geqslant 0$ with $x_i \cdot \mu_i = 0$ and $p = \lambda_i \cdot Du_i(x_i) + \mu_i$. Conversely, given $(x, p) \in W_0 \times S_+^{\ell-1}$, with $p \cdot x_i \neq 0$, all i, and λ_i, μ_i as above with $p = \lambda_i \cdot Du_i(x_i) + \mu_i$, then $(x, p) \in \Lambda$.

Proof

Since x_i is a maximum of u_i on $B_{p,p\cdot x_i}$, for each i, there exist $\lambda_i \geqslant 0$, $\mu_i \in R^\ell_+$, $\sigma_i \geqslant 0$ not all zero, with

$$\lambda_i Du_i(x_i)(\bar{x}_i) + \sum \mu_i^j Dg_i^j(x_i)(\bar{x}_i) - \sigma_i p \cdot \bar{x}_i = 0, \quad \text{all} \quad \bar{x}_i \in R^\ell,$$

or

$$\sigma_i p = \lambda_i Du_i(x_i) + \mu_i, \qquad \mu_i \cdot x_i = 0.$$

If the σ_i were 0, then so would be λ_i, μ_i. Thus we may rescale by dividing by σ_i to obtain $p = \lambda_i Du_i(x_i) + \mu_i$, $\mu_i \cdot x_i = 0$. This proves the first part. For the second let $y_i \in B_{p,p\cdot x_i}$ with $u(y_i) > u(x_i)$. Then by Lemma 4.2 in the proof of Theorem 4.1, $Du_i(x_i)(y_i - x_i) > 0$, and $p \cdot y_i \geqslant y_i \cdot \lambda_i Du_i(x_i) > p \cdot x_i$, $\lambda_i \neq 0$, as in an earlier argument. Then $y_i \in B_{p,p\cdot x_i}$, contrary to hypothesis. Thus $(x, p) \in \Lambda$. This proves the proposition.

For the rest of this section, let us assume for simplicity the strong monotonicity hypothesis, that $Du_i(x_i) \in \operatorname{int} S^{\ell-1}_+$. The projection map $W_0 \times S^{\ell-1}_+ \to W_0$, $(x, p) \to x$, induces a map $\alpha: \Lambda \to \theta$, from welfare equilibria to Pareto optima. By the proposition above and Theorem 4.8, α is well-defined and it is surjective. While these results have an extensive literature under the topic of "fundamental theorems of welfare economics", the question of uniqueness of a supporting price system seems not so standard. Is α injective?

The answer is affirmative under the further mild hypothesis of "no isolated communities" [Smale (1974–76, V)]. For $x \in W_0$, an *isolated community* is a non-empty proper subset $S \subseteq \{1, \ldots, m\}$ with the property that wherever $i \in S$ and $x_i^j \neq 0$, then $x_k^j = 0$ for all $k \notin S$.

Theorem 4.10

If x is an optimum in W_0 with no isolated communities, then there is a *unique* supporting price system.

Here we are supposing W_0 is the space of attainable states; the utility functions $u_i: R^\ell_+ \to R$ are C^2 with $Du_i(x_i) \in \operatorname{int} S^{\ell-1}_+$ and $D^2 u_i(x_i) < 0$ on $\operatorname{Ker} Du_i(x_i)$.

Lemma 4.11

Suppose $x \in W_0$ has no isolated communities and $i, q \in \{1, \ldots, m\}$ are two agents. Then there is a sequence $i_1, \ldots, i_n$ of agents with $i_1 = i$, $i_n = q$, and a sequence of goods $j_1, \ldots, j_n$ such that $x_{i_k}^{j_k} \neq 0$, all k and for any k, either $j_{k+1} = j_k$ or $i_{k+1} = i_k$.

Proof

Otherwise take any agent, say agent number 1 for convenience, and consider all above such sequences $(i_1,\ldots,i_n)$, $(j_1,\ldots,j_n)$ with $i_1=1$. Let S be the subset of $\{1,\ldots,m\}$ of all possible i_n reached in this way. If S is proper, then it is an isolated community. This proves the lemma.

To prove Theorem 4.10, first obtain p,λ_i,μ_i as in Theorem 4.8, with $p=\lambda_i Du_i(x_i)+\mu_i$, $\lambda_i\geqslant 0$, $\mu_i\in R^\ell_+$ and $\mu_i\cdot x_i=0$. The problem has to do with the ambiguity of the λ_i,μ_i. Suppose by renumbering, that agent 1 has some of the first good so $x_1^1\neq 0$. Normalize p by taking $p^1=1$ (and not $\|p\|=1$). Then $1=p^1=\lambda_1 Du_1(x_1)^1$ since $\mu_1^1=0$, and λ_1 is thus determined. Let q be any other agent; choose a sequence $(i_1,\ldots,i_n)$, $i_1=1$, $i_n=q$, $(j_1,\ldots,j_n)$ as in the lemma. We claim that λ_{i_k} is determined for each i_k. Suppose inductively that $\lambda_{i_{k-1}}$ is determined, and $i_k\neq i_{k-1}$. Then $j_k=j_{k-1}$, both agents i_k and i_{k-1} have some of good j_k. Therefore $p^{j_k}=\lambda_{i_{k-1}}Du_{i_{k-1}}(x_{i_{k-1}})^{j_k}$ determines p^{j_k} and $p^{j_k}=\lambda_{i_k}Du_{i_k}(x_{i_k})^{j_k}$ determines λ_{i_k}. Here we used the fact that the corresponding μ_i^j's are 0. Once all the λ_i's are determined uniquely, let k be any good. Choose i so that $x_i^k\neq 0$. Then $p^k=\lambda_i Du_i(x_i)^k$ determines p^k. This proves Theorem 4.10.

5. Finiteness and stability of equilibria

The first goal is to give a proof that the pure exchange economy described in the first part of Section 2 has only a finite number of Walras equilibria, at least for almost all endowment allocations. At the same time we show that these equilibria are stable (better "robust") in the sense that they persist under perturbations of the endowment allocation. These results are due to Debreu (1970). Our approach to this result is to define an "equilibrium manifold" without passing to the demand functions. The hypotheses, framework (pure exchange economy), and notation will be the same as in the first part of Section 2.

Thus define the *equilibrium "manifold"* Σ as follows: The space $(P)^m\times(P)^m$ consists of (e,x), $e=(e_1,\ldots,e_m)$, $x=(x_1,\ldots,x_m)$ with e_i, $x_i\in P$. Here e will be thought of as an *endowment allocation* parameterizing an economy. Then Σ will be the subset of $(P)^m\times(P)^m$ of (e,x) satisfying:

$$\sum e_i=\sum x_i \qquad \text{(a total resource or attainability condition)}, \tag{5.1}$$

$$g_i(x_i) \text{ is independent of } i \qquad \text{(the first-order condition;} \quad g_i(x_i)=\operatorname{grad} u_i(x_i)/\|\operatorname{grad} u_i(x_i)\|\text{)}, \tag{5.2}$$

$$p\cdot(e_i-x_i)=0 \qquad \text{(budget condition)}. \tag{5.3}$$

Thus if e is fixed, $(e, x)\in\Sigma$, then (x, p) where $p=g_i(x_i)$, is a Walras equilibrium and conversely [see Section 2, (A), (B_1), (B_2)].

Theorem 5.1

Σ is a submanifold of $(P)^m\times(P)^m$ of dimension $m\ell$.

Proof

Define a map

$$\phi:(P)^m\times(P)^m\to R^\ell\times R^{m-1}\times(S^{\ell-1})^m,$$

by sending

$$(e,x)\to\Big(\sum e_i-\sum x_i, p\cdot(e_1-x_1),\ldots, p\cdot(e_{m-1}-x_{m-1}), g_1(x_1),\ldots, g_m(x_m)\Big).$$

Then from the definition of Σ we may write $\Sigma=\phi^{-1}(0\times0\times\Delta)$ where $\Delta=\{(p,\ldots,p)\in(S^{\ell-1})^m\}$, and we have used the fact that conditions $\sum e_i=\sum x_i$ and $p\cdot(e_i-x_i)=0$, $i=1,\ldots,m-1$, imply $p(e_m-x_m)=0$.

As in Section 4, Theorem 5.1 would be a consequence of ϕ being transversal to $0\times0\times\Delta$, using a simple counting of equations. Following the line of proof of Theorem 4.1; if $\phi(e,x)\in\Delta$ define

$$L_{e,x}=\big\{(\bar e,\bar x)\in(R^\ell)^m\times(R^\ell)^m\,|\,\mathrm{D}\phi(e,x)(\bar e,\bar x)\in0\times0\times T(\Delta)\big\},$$

or, equivalently, from differentiating (1), (2) and (3),

$$L_{e,x}=\Big\{(\bar e,\bar x)\in(R^\ell)^m\times(R^\ell)^m\,|\,\sum\bar e_i=\sum\bar x_i, \mathrm{D}g_i(x_i)(\bar x_i)=\bar p\in p^\perp,$$
$$\bar p\cdot(e_i-x_i)+p\cdot(\bar e_i-\bar x_i)=0\Big\}.$$

Here we take $p=g_1(x_1)$ and $\bar p=\mathrm{D}g_1(x_1)(\bar x_1)$.

Now we define a second linear subspace $K_{e,x}$ of $(R^\ell)^m\times(R^\ell)^m$ by

$$K_{e,x}=(\bar e,\bar x)|\,\sum\bar e_i=0,\ \bar x_i\cdot p=0,\ i\leqslant m-1,\ \pi_p\bar e_i=0,\ i\leqslant m-1\Big\}.$$

Here $\pi_p: R^\ell\to P^\perp$ is the orthogonal projection so that $\bar e_i=\pi_p\bar e_i+p\cdot\bar e_i$, each i. This space $K_{e,x}$ is motivated only by the proof of Theorem 5.1. Clearly $\dim K_{e,x}=m\ell$, and one also can see that $\dim R^\ell\times R^{m-1}\times(S^{\ell-1})^m-\dim\Delta=m\ell$. Thus if $L_{e,x}\cap K_{e,x}=0$, we have that ϕ is transversal to $0\times0\times\Delta$, just as the situation was in Section 4.

Lemma 5.2

$L_{e,x} \cap K_{e,x} = 0$.

For the lemma let $(\bar{e}, \bar{x})$ belong to the intersection. As in Section 4, $\gamma_i : p^\perp \to P^\perp$ denotes the restriction of $Dg_i(x_i)$. Then $\sum \bar{x}_i = 0$ since $\sum \bar{e}_i = 0$ and $\sum \bar{e}_i = \sum \bar{x}_i$. So $\bar{x}_i \cdot p = 0$, all i, and $\gamma_i^{-1}(\bar{p}) = \bar{x}_i$, each i. Also $\sum \gamma_i^{-1}(\bar{p}) = \sum \bar{x}_i = 0$ and $\bar{p} = 0$, therefore $\bar{x}_i = 0$. Finally one sees that $\bar{e}_i = 0$ proving the lemma and hence the theorem.

We emphasize that we are taking $u_i : P \to R$, $i = 1, \ldots, m$, as in the first part of Section 2.

Theorem 5.3

There is a closed set $F \subset (P)^m$ of measure 0 so that if $e \notin F$ then there exist a finite (positive) number of Walras equilibria relative to the endowment $e = (e_1, \ldots, e_m)$. This finite set varies continuously in e as long as e does not meet F. Let $\pi : (P)^m \times (P)^m \to (P)^m$ be the projection defined by $\pi(e, x) = e$. Let $\pi_0 : \Sigma \to (P)^m$ be the restriction of π.

Lemma 5.4

The map $\pi_0 : \Sigma \to (P)^m$ is closed. The image of a closed set is closed.

Proof

Consider a sequence $(e^{(j)}, x^{(j)})$ in $(P)^m \times (P)^m$, $j = 1, 2, 3, \ldots$, so that $e^{(j)}$ converges to $e \in (P)^m$. Then by the equilibrium conditions defining Σ, and the boundary condition on u_i, the $x^{(j)}$ have a subsequence converging to some $x \in (P)^m$. This is enough to show that π_0 is closed.

Let $C \subset \Sigma$ be the closed set of critical points of π_0 and $F = \pi_0(C)$. Then F is closed by Lemma and has measure 0 by Sard's theorem. Theorem now is a consequence of the inverse function theorem applied to the map π_0.

A study of comparative statics of equilibria can now be done using these theorems.

While the above approach comes from Smale (1974–76) a closely related way of proving Debreu's theorem is in Balasko (1975).

Appendix A. Existence of economic equilibrium with production

We prove the theorem of Arrow–Debreu on the existence of economic equilibrium with production as treated in Debreu (1959). The reason we include the proof is to show that calculus can indeed be the starting point of equilibrium theory with proofs at least as short and natural as those emphasizing Kakutani's

fixed point theorem. On the other hand, our approach has much in common with that of Debreu; we owe much to his exposition as well as to conversations with him.

Here the treatment is brief. One can see Debreu (1959) for economic interpretations. The proof here is based on Theorem 1.5, and it is somewhat similar to the proofs in Section 2.

An *economy* consists, first, of a production side. We suppose ℓ commodities including labor. To each of n producers, $j=1,\ldots,n$, is associated a "technology" $Y_j \subset R^\ell$ with the conditions:

(T) (a) $0 \in Y_j$, each j (possibility of no production).

Let $Y = \sum Y_j$,

(b) $Y \cap (-Y) = \{0\}$ (an irreversibility condition).

(c) Y is closed and convex.

(d) $Y - R^\ell_+ \subset Y$ (free disposal).

It can be shown that (d) is a consequence of $Y \supset -R^\ell_+$ in the presence of (c); see Debreu (1959). Here Y_j may be thought of as the set of productions that are available to firm j. We suppose that the firm is driven by profit maximization. Thus if a price system p is operative, the production $y \in Y$, is sought so that the profit $p \cdot y$ is a maximum.

Pass now to the consumer side of the economy. To each of m consumers, $i=1,\ldots,m$, is associated a "consumption set" $X_i \subset R^\ell$ and a utility function $u_i: X_i \to R$ which represents his/her preference. The following is assumed:

(C) (a) X_i is a closed convex set.

(b) X_i is bounded below.

That is, there exist $d_1,\ldots,d_n \in R^\ell$ with $X_i \subset \{x \in R^\ell \mid x \geqslant d_i\}$ or $X_i \geqslant d_i$. (Here $x \geqslant d_i$ means that each component of x is $\geqslant$ the corresponding component of d_i.)

(c) u_i satisfies the convexity condition: if $x, x' \in X_i$ with $u_i(x) > u_i(x')$, then $u_i(tx + (1-t)x') > u(x')$ for each $t \in (0,1)$.

(d) u_i has no maximum (no satiation condition)

Remark

One could have used directly a preference relation here, as in Debreu (1959), rather than utility function. No generality is gained as one can see in Debreu's paper.

Furthermore, to each consumer is associated an endowment $e_i \in X_i$ with e_i having all coordinates strictly larger than some element of X_i. As in Debreu (1959), this is an unhappy hypothesis. Finally (private ownership economy) let θ_{ij}, be the share of agent i in firm j. Then it is assumed that $0 \leqslant \theta_{ij} \leqslant 1$ and $\sum_{i=1}^{m} \theta_{ij} = 1$. If a price system p prevails, then the wealth of agent i is given by $w_i = p \cdot e_i + \sum_j \theta_{ij} p \cdot y_j$.

An *equilibrium* for an economy above is a "state" (x, y, p) with $x \in \prod_{i=1}^{m} X_i$, $y \in \prod_{j=1}^{n} Y_j$, $p \in S_+^{\ell-1}$ which satisfies

A) Attainability, or $\sum x_i = \sum y_j + \sum e_i$.

B) Each consumer maximizes satisfaction or:
x_i is a maximum of u_i on the budget set

$$B = \left\{ \bar{x} \in X_i \,|\, p \cdot \bar{x} \leqslant p \cdot e_i + \sum_j \theta_{ij} p \cdot y_j \right\}.$$

C) Each producer maximizes profit or:
y_j is a maximum of Π_p on Y_j, where

$$\Pi_p : Y_j \to R \text{ is } \Pi_p(\bar{y}) = p \cdot \bar{y}.$$

Arrow–Debreu Theorem

For an economy above there is always an equilibrium.

We first give a proof under additional restrictions; then we extend that proof to the General Arrow–Debreu Theorem.

Theorem A.1

Suppose that the economy described above satisfies the further conditions:

(1) Each Y_j is closed and strictly convex.
(2) Each u_i has the strict convexity property of Section 2 or, more precisely, if $u_i(x) \geqslant c$, $u_i(x') \geqslant c$ and $0 < t < 1$, then $u(tx + (1-t)x') > c$.

Then there is an equilibrium.

Toward proving Theorem A.1 we use the following basic lemma for which Bowen gave me this analytic version of my more geometric account:

Lemma A.2 (basic estimate)

Let Y be a closed convex subset of R^ℓ with $Y \cap (-Y) = \{0\}$ and $Y \supset -R_+^\ell$. Then given $b \in R^\ell$ and $n > 0$ there is a constant c so that if $y_1, \ldots, y_n \in Y$ and $\sum y_j \geqslant b$ then $\| y_j \| < c$ each j.

For the proof let $K=\{y\in Y \mid \|y\|=1\}$. We prove three assertions:

Assertion 1

The origin 0 of R^ℓ is not in the convex hull of K.

If $\alpha_1 x_1+\cdots+\alpha_r x_r=0$ with $0<\alpha_i<1$, $\alpha_1+\cdots+\alpha_r=1$, $x_i\in K$, then

$$-\alpha_1 x_1=\alpha_1\cdot 0+\alpha_2 x_2+\cdots+\alpha_r x_r\in Y,$$

and $\alpha_1 x_1$ clearly is in Y. Thus $\alpha_1 x_1\in Y\cap(-Y)$ reaching a contradiction.

Assertion 2

There is a $q=(q_1,\ldots,q_\ell)\in R^\ell$, each $q_i>0$, such that $q\cdot x<0$ for every $x\in K$.

As K is compact, so is its convex hull. By Assertion 1 there is a q in R^ℓ with $q\cdot K<0$. If e_i is a coordinate basis vector then $-e_i\in K$ and $-q_i=q\cdot(-e_i)<0$.

Assertion 3

There are constants $\varepsilon>0$, $\beta>0$, so that if $x\in Y$ then $q\cdot x\leqslant\beta+\varepsilon-\varepsilon\|x\|$.

Let $-\varepsilon=\max\{q\cdot x \mid x\in K\}$ and $\beta=\max\{q\cdot x \mid \|x\|\leqslant 1\}$. The inequality is clear if $\|x\|\leqslant 1$. For $\|x\|>1$, $x\in Y$, and one has $x/\|x\|\in K$ since Y is convex and contains 0. Then $-\varepsilon\geqslant q\cdot x/\|x\|$ or $q\cdot x\leqslant-\varepsilon\|x\|$.

We finish the proof of Lemma 1 as follows: Suppose $\sum y_j\geqslant b$ with $y_j\in Y$. Then $q\cdot b\leqslant\sum q\cdot y_j\leqslant n(\beta+\varepsilon)-\varepsilon\sum\|y_j\|$, so $\sum\|y_j\|\leqslant(n(\beta+\varepsilon)-q\cdot b)/\varepsilon$.

An analogous lemma for the consumption side is:

Lemma A.3

Given $c_1\in R^\ell$, there is $a>0$ such that if $x_i\in X_i$, $X_i\geqslant d_i$ [as in (C) above] for $i=1,\ldots,m$, and $\sum x_i\leqslant c_1$, then $\|x_i\|\leqslant a$, each i.

We omit the very easy proof.

Now let $b=\sum d_i-\sum e_i$ and choose c as in Lemma A.2, so that if $\sum y_j\geqslant b$, then $\|y_j\|<c$, each j. Let $\hat{Y}_j=Y_j\cap D_c$ where $D_r=\{y\in D^\ell \mid \|y\|\leqslant r\}$. For $p\in R^\ell_+-0$, let $\hat{S}_j(p)=$ the maximum of $\Pi_p: Y_j\to R$ where $\Pi_p(y)=p\cdot y$. Then $\hat{S}_j$ is the "false supply function" of firm j.

Lemma A.4

$\hat{S}_j: R^\ell_+-0\to\hat{Y}_j$ is well-defined, continuous, $\hat{S}_j(\lambda p)=\hat{S}_j(p)$ for $\lambda>0$, and if $\|\hat{S}_j(p)\|<c$, then $\hat{S}_j(p)$ is the maximum of Π_p on Y_j (the true supply).

This is clear from the definitions, recalling that we are in the situation of Theorem A.1, so that $\hat{Y}_j$ is strictly convex.

Remark

If Y_j is merely assumed closed and convex (not necessarily strictly convex), one still has $\hat{S}_j$ defined as a correspondence; i.e., $\hat{S}_j: R^\ell_+ - 0 \to S(\hat{Y}_j)$ is a map with values, convex subsets of $\hat{Y}_j$. It is homogeneous and when restricted to $S^{\ell-1}_+$ has a compact graph

$$\Gamma = \{(p, y) \in S^{\ell-1}_+ \times \hat{Y}_j \mid y \in \hat{S}_j(p)\}.$$

Furthermore in this case if $y \in \hat{S}_j(p)$ has norm $\|y\| < c$, then y is a maximum of Π_p on Y_j. Note that $\Pi_p(y)$ is independent of $y \in \hat{S}_j(p)$.

Define $\hat{w}_i: R^\ell_+ - 0 \to R$, the "false income" of consumer i, by $\hat{w}_i(p) = p \cdot e_i + \sum_j \theta_{ij} p \cdot \hat{S}_j(p)$. Then $\hat{w}_i$ is continuous. Let b, c, e, be as above and choose $c_1 \in R^\ell$ such that $\sum y_j + e \leqslant c_1$ if $\|y_j\| < c$ each j. Choose a by Lemma A.3 and let $\hat{X}_i = X_i \cap D_a$.

Define a "false demand" $\hat{D}_i: R^\ell_+ - 0 \to \hat{X}_i$ for each i by $\hat{D}_i(p) =$ the maximum of u_i on $\hat{B}_p = \{x \in \hat{X}_i \mid p \cdot x \leqslant \hat{w}_i(p)\}$ (compare Proposition 2.7).

Lemma A.5

The false demand $\hat{D}_i: R^\ell_+ - 0 \to \hat{X}_i$ is well-defined, continuous, $\hat{D}_i(\lambda p) = D_i(p)$ for $\lambda > 0$, and $p \cdot \hat{D}_i(p) = w_i(p)$. Also if $\|\hat{D}_i(p)\| < a$ then $\hat{D}_i(p)$ is the maximum of u_i on the budget set $B_p = \{x \in X_i \mid p \cdot x \leqslant \hat{w}_i(p)\}$ and $p \cdot \hat{D}_i(p) = \hat{w}_i(p)$.

The proof uses the same arguments as that at the end of Section 2, uses the No Satiation Condition, and the convexity of X_i. The continuity uses the fact that e_i dominates some element of X_i (the basic hypothesis on e_i). We leave the detailed proof, which is not difficult, to the reader.

Remark

In case u_i satisfies the convexity condition (c) of (C) rather than strict convexity of Theorem A.1, then $\hat{D}_i$ is defined as a correspondence with values, convex subsets of X_i. It is homogeneous, and the restriction $\hat{D}_i: S^{\ell-1}_+ \to X_i$ has a compact graph. Also if $x \in D_i$ satisfies $\|x\| < a$, then x is a maximum of u_i on $\{\bar{x} \in X_i \mid p \cdot \bar{x} \leqslant \hat{w}_i(p)\}$ and $p \cdot x = \hat{w}_i(p)$.

Now define these aggregate functions from $R^\ell_+ - 0$ to R^ℓ: $\hat{S} = \sum \hat{S}_j + \sum e_i$, $\hat{D} = \sum \hat{D}_i$, and $\hat{Z} = \hat{D} - \hat{S}$. From Lemmas A.4 and A.5, $\hat{Z}$ satisfies homogeneity and weak Walras, so Theorem 1.5 applies to produce $p^* \in S^{\ell-1}_+$ with $\hat{Z}(p^*) \leqslant 0$. Let $y_j^* = S_j^*(p^*)$, $x_j^* = \hat{D}_j(p^*)$, so then $\sum x_i^* \leqslant \sum y_j^* + \sum e_i$. Since each $x_i^* \in \hat{X}_i \subset X_i$, this implies $b \leqslant \sum y_j^*$ (definition of b). Thus $\|y_j^*\| < c$ (Lemma A.2), and by

Lemma A.4, y_j^* is the maximum of Π_p on Y_j. By the choices of c_1 and a, via Lemma A.3, $\|x_i^*\| < a$, each i. By Lemma A.5, x_i^* is the maximum of u_i on $\{\bar{x} \in X_i | p^* \cdot \bar{x} \leqslant \hat{w}_i(p^*)\}$, with $\hat{w}_i(p^*) = p^* \cdot e_i + \sum_j \theta_{ij} p^* \cdot y_j^*$.

We may choose $z \in R_+^\ell$ so that $\sum x_i^* = \sum y_j^* + \sum e_i - z$. Apply p^* to this to see (using Lemma A.5 again) $p^* \cdot z = 0$. Then $\sum y_j^* - z$ is in $Y = \sum Y_j$ by (T) so we have $y_j \in Y_j$ with $\sum y_j = \sum y_j^* - z$. Then $p \cdot \sum y_j = p \cdot \sum y_j^*$, which implies that y_j also (as well as y_j^*) maximizes Π_p on Y_j, and (x_i^*, y_j, p^*) is an equilibrium, proving Theorem A.1. Note in fact $y_j = y_j^*$ by the strict convexity, but our argument covers the more general case of Theorem A.6.

We next weaken the convexity hypotheses of Theorem A.1 by using the approximation theorem of Appendix B:

Theorem A.6

Theorem A.1 remains true if each Y_j is closed and convex (rather than strictly convex), and instead of the strict convexity hypothesis on each u_i, we only assume (C) as in the Arrow–Debreu Theorem.

Proof

Proceed as in the proof of Theorem A.1. As in the remark after Lemma A.4, we can consider $\hat{S}_j : S_+^{\ell-1} \to \hat{Y}_j$, $j = 1, \ldots, n$, as correspondences.

Suppose $\varepsilon > 0$ is given. Apply the theorem of Appendix B to obtain continuous functions $\hat{S}_{j\varepsilon} : S_+^{\ell-1} \to \hat{Y}_j$ for each $j = 1, \ldots, n$, with $\Gamma_{\hat{S}_{j\varepsilon}} \subset B_\varepsilon(\Gamma_{\hat{S}_j})$. Next note that $\hat{w}_i : S_+^{\ell-1} \to R$, defined by $\hat{w}_i(p) = p \cdot e_i + \sum_j \theta_{ij} p \cdot \hat{S}_j(p)$, is a well-defined continuous function, even with $\hat{S}_j$ a correspondence. As in the remark after Lemma A.5, we can consider $\hat{D}_i : S_+^{\ell-1} \to \hat{X}_i$ defined as a correspondence. Apply the theorem of Appendix B to obtain functions $\hat{D}_{i\varepsilon} : S_+^{\ell-1} \to \hat{X}_i$ such that $\Gamma_{\hat{D}_{i\varepsilon}} \subset B_\varepsilon(\Gamma_{\hat{D}_i})$ and $|p \cdot \hat{D}_{i\varepsilon}(p) - \hat{w}_i(p)| < \varepsilon$, all $p \in S_+^{\ell-1}$.

Define $Z_\varepsilon : S_+^{\ell-1} \to R^\ell$ by $Z_\varepsilon(p) = \sum \hat{D}_{i\varepsilon}(p) - \sum \hat{S}_{j\varepsilon}(p) - \sum e_i$, and $\hat{Z}_\varepsilon(p) = Z_\varepsilon(p) - (p \cdot Z_\varepsilon(p))p$. Then $p \cdot \hat{Z}_\varepsilon(p) = 0$ and $p \cdot Z_\varepsilon(p) \to 0$ as $\varepsilon \to 0$. Apply Theorem 1.5 to obtain p_ε such that $\hat{Z}_\varepsilon(p_\varepsilon) = 0$.

Let $y_{j\varepsilon} = \hat{S}_{j\varepsilon}(p_\varepsilon)$, $x_{i\varepsilon} = \hat{D}_{i\varepsilon}(p_\varepsilon)$. Now take a sequence of ε_k tending to 0. By taking subsequences we obtain $y_{j\varepsilon_k} \to y_j$, $x_{i\varepsilon_k} \to x_i$, $p_{\varepsilon_k} \to p$ to obtain an equilibrium. This finishes the proof of Theorem A.6 as in Theorem A.1.

Now we give the proof of the General Arrow–Debreu Theorem. We need:

Lemma A.7

Let $\widehat{Z}$ denote the convex hull of a subset Z of Euclidean space. Then

$$\sum \widehat{Y_j} = \widehat{\sum Y_j}.$$

Proof

Since $\sum\widehat{Y_j}$ is convex it contains $\widehat{\sum Y_j}$. We will show $\widehat{A}+\widehat{B}\subset\widehat{A+B}$. Let $a_i\geqslant 0$ with $\sum a_i=1$. Then $\widehat{A}+y\subset\widehat{A+y}$ since $\sum a_i x_i+y=\sum a_i(x_i+y)$. Therefore $\widehat{A}+B\subset\widehat{A+B}$. Finally $\widehat{B}+\widehat{A}\subset\widehat{B+\widehat{A}}\subset\widehat{A+B}$, showing indeed that $\widehat{A}+\widehat{B}\subset\widehat{A+B}$. By induction the proof of Lemma A.7 is finished.

With the hypotheses and notation of the beginning of Appendix A, let Y_j^* be the closure of the convex hull of Y_j. Recalling $Y=\sum Y_j$, we have:

Lemma A.8

$\sum Y_j^*=Y$.

Proof

Since $Y_j\subset Y_j^*$, $\sum Y_j^*\supset\sum Y_j$. On the other hand, since the sum of the closure of sets is contained in the closure of the sum, it follows from Lemma A.7 that $\sum Y_j^*\subset Y$ (recall Y is closed and convex). This proves Lemma A.8.

Apply Theorem A.6 to obtain an equilibrium (x_i^*, y_j^*, p) for the economy above with Y_j^* replacing Y_j. Now $\sum y_j^*\in Y$ (Lemma A.8) and so $\sum y_j^*=\sum y_j=y$ with $y_j\in Y_j$.

Furthermore $p\cdot y_j=p\cdot y_j^*$. This is so since y_j^* is a maximum of Π_p on Y_j^* and therefore y is a maximum of Π_p on Y. This implies (since $y=\sum y_j$) that $\Pi_p(y_j)$ is at least as much as $\Pi_p(y_j^*)$ and hence equal. The rest follows and the Arrow–Debreu Theorem is proved.

Appendix B. A theorem on the approximation of multi-valued mappings

We prove the following theorem of Cellina (1969), using extensively an unpublished exposition of W. Hildenbrand:

Theorem B.1

Let K be a compact set (say in some Euclidean space), T a compact convex set of R^ℓ, and $\varphi: K\to S(T)$ a correspondence with values convex subsets of T such that the graph $\Gamma_\varphi=\{(x, y)\in K\times T\,|\,y\in\varphi(x)\}$ is compact. Then given $\varepsilon>0$ there is a continuous function $f: K\to T$ such that $\Gamma_f\subset B_{2\varepsilon}(\Gamma_\varphi)$.

Here Γ_f is the graph of f in $K\times T$ and $B_{2\varepsilon}$ is the open set of all points of $K\times T$ within 2ε of Γ_φ.

For the proof define $\varphi^\delta: K\to S(T)$ by $\varphi^\delta(x)=$ convex hull of $\bigcup_{y\in B_\delta(x)}\varphi(y)$.

Lemma B.2

Let $\varepsilon > 0$ be given. Then there is a $\delta > 0$ such that $\Gamma_{\varphi^\delta} \subset B_\varepsilon(\Gamma_\varphi)$.

Proof

If the lemma were false, one could take $\delta = 1/n$ and obtain a sequence (x_n, y_n) in $K \times T$, with $(x_n, y_n) \notin B_\varepsilon(\Gamma_\varphi)$, all n, and $y_n = \sum \lambda^i{}_n y^i_n$, $\sum \lambda^i{}_n = 1$, $\lambda^i{}_n > 0$, $y^i_n \in \varphi(z^i_n)$, $d(z^i_n, x_n) \leqslant 1/n$. By taking subsequences, we get $x_n \to x$, $y^i_n \to y^i$, $\lambda^i{}_n \to \lambda_n$, $z^i_n \to z^j = x$. So $y = \sum \lambda^i y^i$, $\lambda^i \geqslant 0$, $\sum \lambda^i = 1$ and (x, y^i) is in the closure of Γ_φ. Since $\varphi(x)$ is convex, (x, y) is in the closure of Γ_φ, contradicting $(x_n, y_n) \notin B_\varepsilon(\Gamma)$. The lemma is proved.

Next let δ be as in the lemma and

$$U_y = \{x \in K \mid y \in B(\varphi^\delta(x))\} \quad \text{for each} \quad y \in T,$$

and then choose $U_{y_1}, \ldots, U_{y_k}$ a finite covering of K. Let β_i be a corresponding partition of unity so $\beta_i : K \to [0, 1]$, $i = 1, \ldots, k$, are continuous functions, $\beta_i(x) = 0$ exactly if $x \notin U_i$ and $\sum \beta_i \equiv 1$. For example, one could take

$$\beta_i(x) = \frac{\alpha_i(x)}{\sum_{j=1}^{k} \alpha_j(x)} \quad \text{where} \quad \alpha_j(x) = \inf_{x' \notin U_j} d(x, x').$$

Define $f(x) = \sum \beta_i(x) y_i$. Then f is clearly a continuous function, $f : K \to R^\ell$, such that for $x \in K$, $f(x)$ is a convex combination of those points y_i such that $x \in U_{y_i}$ or $y_i \in B_\varepsilon(\varphi^\delta(x))$.

Since an ε-neighborhood of convex sets is convex, $B_\varepsilon(\varphi^\delta(x))$ is convex and $f(x)$ is in it. Therefore $(x, f(x)) \in B_\varepsilon(\Gamma_{\varphi^\delta})$ and by the lemma $(x, f(x)) \in B_{2\varepsilon}(\Gamma_\varphi)$ proving the approximation theorem.

References

Abraham, R. and J. Robbin (1967), Transversal mappings and flows. New York: Benjamin.

Arrow, K. and F. Hahn (1971), General competitive analysis. San Francisco, CA: Holden-Day.

Balasko, Y. (1975), "Some results on uniqueness and on stability of equilibrium in general equilibrium theory", Journal of Mathematical Economics, 2:95–118.

Cellina, A. (1969), "A theorem on the approximation of compact multi-valued mappings", Rendiconti Academia Nazionale Lincei, 47:fasc.6.

Debreu, G. (1959), Theory of value. New York: Wiley.

Debreu, G. (1970), "Economics with a finite set of equilibria", Econometrica, 38:387–392.

Debreu, G. (1972), "Smooth preferences", Econometrica, 40:603–616.

Golubitsky, M. and V. Guilemin (1973), Stable mappings and their singularities. New York: Springer.

Lang, (1969), Real analysis. Reading, MA: Addison-Wesley.
Simon, C. (forthcoming), "Scalar and vector maximization: Calculus techniques with economics applications", in: S. Reiter, ed., Studies in mathematical economics, MAA studies in mathematics series.
Smale, S. (1974–76), "Global analysis and economics, IIA–VI", Journal of Mathematical Economics, 1:1–14, 107–117, 119–127, 213–221, 3:1–14.
Smale, S. (1975), "Sufficient conditions for an optimum", in: A. Manning, ed., Dynamical systems—Warwick 1974, Lecture notes in mathematics series no. 468. New York: Springer.
Smale, S. (1976a), "A convergent process of price adjustment and Global Newton methods", Journal of Mathematical Economics, 3:1–14.
Smale, S. (1976b), "Dynamics in general equilibrium theory", American Economic Review, 66:288–294.
Varian, H. (1977), "A remark on boundary restriction in the Global Newton method", Journal of Mathematical Economics, 4:127–130.
Wan, H.-Y. (1975), "On local Pareto optima", Journal of Mathematical Economics, 2:35–42.

LIST OF THEOREMS

INDEX