817

GAME THEORY

Analysis of Conflict

ROGER B. MYERSON

HARVARD UNIVERSITY PRESS
Cambridge, Massachusetts
London, England 1991

Game Theory

Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

This book is printed on acid-free paper, and its binding materials have been chosen for strength and durability.

Library of Congress Cataloging-in-Publication Data

Myerson, Roger B.
Game theory : analysis of conflict / Roger B. Myerson.
p. cm.
Includes bibliographical references and index.
ISBN 0-674-34115-5 (alk. paper)
1. Game theory. I. Title.
H61.25.M94 1991
519.3—dc20 90-42901
CIP

For Gina, Daniel, and Rebecca

With the hope that a better understanding of conflict
may help create a safer and more peaceful world

In every chapter, there are some topics of a more advanced or specialized nature that may be omitted without loss of subsequent comprehension. I have not tried to "star" such sections or paragraphs. Instead, I have provided cross-references to enable a reader to skim or pass over sections that seem less interesting and to return to them if they are needed later in other sections of interest. Page references for the important definitions are indicated in the index.

In this introductory text, I have not been able to cover every major topic in the literature on game theory, and I have not attempted to assemble a comprehensive bibliography. I have tried to exercise my best judgment in deciding which topics to emphasize, which to mention briefly, and which to omit; but any such judgment is necessarily subjective and controversial, especially in a field that has been growing and changing as rapidly as game theory. For other perspectives and more references to the vast literature on game theory, the reader may consult some of the other excellent survey articles and books on game theory, which include Aumann (1987b) and Shubik (1982).

A note of acknowledgment must begin with an expression of my debt to Robert Aumann, John Harsanyi, John Nash, Reinhard Selten, and Lloyd Shapley, whose writings and lectures taught and inspired all of us who have followed them into the field of game theory. I have benefited greatly from long conversations with Ehud Kalai and Robert Weber about game theory and, specifically, about what should be covered in a basic textbook on game theory. Discussions with Bengt Holmstrom, Paul Milgrom, and Mark Satterthwaite have also substantially influenced the development of this book. Myrna Wooders, Robert Marshall, Dov Monderer, Gregory Pollock, Leo Simon, Michael Chwe, Gordon Green, Akihiko Matsui, Scott Page, and Eun Soo Park read parts of the manuscript and gave many valuable comments. In writing the book, I have also benefited from the advice and suggestions of Lawrence Ausubel, Raymond Denekere, Itzhak Gilboa, Ehud Lehrer, and other colleagues in the Managerial Economics and Decision Sciences department at Northwestern University. The final manuscript was ably edited by Jodi Simpson, and was proofread by Scott Page, Joseph Riney, Ricard Torres, Guangsug Hahn, Jose Luis Ferreira, Ioannis Tournas, Karl Schlag, Keuk-Ryoul Yoo, Gordon Green, and Robert Lapson. This book and related research have been supported by fellowships from the John Simon Guggenheim Memorial Foundation and the Alfred P. Sloan

Foundation, and by grants from the National Science Foundation and the Dispute Resolution Research Center at Northwestern University. Last but most, I must acknowledge the steady encouragement of my wife, my children, and my parents, all of whom expressed a continual faith in a writing project that seemed to take forever.

Evanston, Illinois
December 1990

Contents

Preface

Game theory has a very general scope, encompassing questions that are basic to all of the social sciences. It can offer insights into any economic, political, or social situation that involves individuals who have different goals or preferences. However, there is a fundamental unity and coherent methodology that underlies the large and growing literature on game theory and its applications. My goal in this book is to convey both the generality and the unity of game theory. I have tried to present some of the most important models, solution concepts, and results of game theory, as well as the methodological principles that have guided game theorists to develop these models and solutions.

This book is written as a general introduction to game theory, intended for both classroom use and self-study. It is based on courses that I have taught at Northwestern University, the University of Chicago, and the University of Paris–Dauphine. I have included here, however, somewhat more cooperative game theory than I can actually cover in a first course. I have tried to set an appropriate balance between noncooperative and cooperative game theory, recognizing the fundamental primacy of noncooperative game theory but also the essential and complementary role of the cooperative approach.

The mathematical prerequisite for this book is some prior exposure to elementary calculus, linear algebra, and probability, at the basic undergraduate level. It is not as important to know the theorems that may be covered in such mathematics courses as it is to be familiar with the basic ideas and notation of sets, vectors, functions, and limits. Where more advanced mathematics is used, I have given a short, self-contained explanation of the mathematical ideas.

1

Decision-Theoretic Foundations

1.1 Game Theory, Rationality, and Intelligence

Game theory can be defined as the study of mathematical models of conflict and cooperation between intelligent rational decision-makers. Game theory provides general mathematical techniques for analyzing situations in which two or more individuals make decisions that will influence one another's welfare. As such, game theory offers insights of fundamental importance for scholars in all branches of the social sciences, as well as for practical decision-makers. The situations that game theorists study are not merely recreational activities, as the term "game" might unfortunately suggest. "Conflict analysis" or "interactive decision theory" might be more descriptively accurate names for the subject, but the name "game theory" seems to be here to stay.

Modern game theory may be said to begin with the work of Zermelo (1913), Borel (1921), von Neumann (1928), and the great seminal book of von Neumann and Morgenstern (1944). Much of the early work on game theory was done during World War II at Princeton, in the same intellectual community where many leaders of theoretical physics were also working (see Morgenstern, 1976). Viewed from a broader perspective of intellectual history, this propinquity does not seem coincidental. Much of the appeal and promise of game theory is derived from its position in the mathematical foundations of the social sciences. In this century, great advances in the most fundamental and theoretical branches of the physical sciences have created a nuclear dilemma that threatens the survival of our civilization. People seem to have learned more about how to design physical systems for exploiting radioactive materials than about how to create social systems for moderating human behavior in conflict. Thus, it may be natural to hope that advances in the most fundamental and theoretical branches of the social sciences might be able to provide the understanding that we need to match our great advances in the physical sciences. This hope is one of the motivations that has led many mathematicians and social scientists to work in game theory during the past 50 years. Real proof of the power of game theory has come in recent years from a prolific development of important applications, especially in economics.

Game theorists try to understand conflict and cooperation by studying quantitative models and hypothetical examples. These examples may be unrealistically simple in many respects, but this simplicity may make the fundamental issues of conflict and cooperation easier to see in these examples than in the vastly more complicated situations of real life. Of course, this is the method of analysis in any field of inquiry: to pose one's questions in the context of a simplified model in which many of the less important details of reality are ignored. Thus, even if one is never involved in a situation in which people's positions are as clearly defined as those studied by game theorists, one can still come to understand real competitive situations better by studying these hypothetical examples.

In the language of game theory, a *game* refers to any social situation involving two or more individuals. The individuals involved in a game may be called the *players*. As stated in the definition above, there are two basic assumptions that game theorists generally make about players: they are rational and they are intelligent. Each of these adjectives is used here in a technical sense that requires some explanation.

A decision-maker is *rational* if he makes decisions consistently in pursuit of his own objectives. In game theory, building on the fundamental results of decision theory, we assume that each player's objective is to maximize the expected value of his own payoff, which is measured in some *utility* scale. The idea that a rational decision-maker should make decisions that will maximize his expected utility payoff goes back at least to Bernoulli (1738), but the modern justification of this idea is due to von Neumann and Morgenstern (1947). Using remarkably weak assumptions about how a rational decision-maker should behave, they showed that for any rational decision-maker there must exist some way of assigning utility numbers to the various possible outcomes that he cares about, such that he would always choose the option that maximizes

his expected utility. We call this result the *expected-utility maximization theorem*.

It should be emphasized here that the logical axioms that justify the expected-utility maximization theorem are weak consistency assumptions. In derivations of this theorem, the key assumption is generally a *sure-thing* or *substitution* axiom that may be informally paraphrased as follows: "If a decision-maker would prefer option 1 over option 2 when event A occurs, and he would prefer option 1 over option 2 when event A does not occur, then he should prefer option 1 over option 2 even before he learns whether event A will occur or not." Such an assumption, together with a few technical regularity conditions, is sufficient to guarantee that there exists some utility scale such that the decision-maker always prefers the options that give the highest expected utility value.

Consistent maximizing behavior can also be derived from models of evolutionary selection. In a universe where increasing disorder is a physical law, complex organisms (including human beings and, more broadly speaking, social organizations) can persist only if they behave in a way that tends to increase their probability of surviving and reproducing themselves. Thus, an evolutionary-selection argument suggests that individuals may tend to maximize the expected value of some measure of general survival and reproductive fitness or success (see Maynard Smith, 1982).

In general, maximizing expected utility payoff is not necessarily the same as maximizing expected monetary payoff, because utility values are not necessarily measured in dollars and cents. A risk-averse individual may get more incremental utility from an extra dollar when he is poor than he would get from the same dollar were he rich. This observation suggests that, for many decision-makers, utility may be a nonlinear function of monetary worth. For example, one model that is commonly used in decision analysis stipulates that a decision-maker's utility payoff from getting x dollars would be $u(x) = 1 - e^{-cx}$, for some number c that represents his *index of risk aversion* (see Pratt, 1964). More generally, the utility payoff of an individual may depend on many variables besides his own monetary worth (including even the monetary worths of other people for whom he feels some sympathy or antipathy).

When there is uncertainty, expected utilities can be defined and computed only if all relevant uncertain events can be assigned probabilities, which quantitatively measure the likelihood of each event. Ramsey (1926) and Savage (1954) showed that, even where objective probabilities cannot be assigned to some events, a rational decision-maker should be able to assess all the subjective probability numbers that are needed to compute these expected values.

In situations involving two or more decision-makers, however, a special difficulty arises in the assessment of subjective probabilities. For example, suppose that one of the factors that is unknown to some given individual 1 is the action to be chosen by some other individual 2. To assess the probability of each of individual 2's possible choices, individual 1 needs to understand 2's decision-making behavior, so 1 may try to imagine himself in 2's position. In this thought experiment, 1 may realize that 2 is trying to rationally solve a decision problem of her own and that, to do so, she must assess the probabilities of each of 1's possible choices. Indeed, 1 may realize that 2 is probably trying to imagine herself in 1's position, to figure out what 1 will do. So the rational solution to each individual's decision problem depends on the solution to the other individual's problem. Neither problem can be solved without understanding the solution to the other. Thus, when rational decision-makers interact, their decision problems must be analyzed together, like a system of equations. Such analysis is the subject of game theory.

When we analyze a game, as game theorists or social scientists, we say that a player in the game is *intelligent* if he knows everything that we know about the game and he can make any inferences about the situation that we can make. In game theory, we generally assume that players are intelligent in this sense. Thus, if we develop a theory that describes the behavior of intelligent players in some game and we believe that this theory is correct, then we must assume that each player in the game will also understand this theory and its predictions.

For an example of a theory that assumes rationality but not intelligence, consider price theory in economics. In the general equilibrium model of price theory, it is assumed that every individual is a rational utility-maximizing decision-maker, but it is not assumed that individuals understand the whole structure of the economic model that the price theorist is studying. In price-theoretic models, individuals only perceive and respond to some intermediating price signals, and each individual is supposed to believe that he can trade arbitrary amounts at these prices, even though there may not be anyone in the economy actually willing to make such trades with him.

Of course, the assumption that all individuals are perfectly rational and intelligent may never be satisfied in any real-life situation. On the other hand, we should be suspicious of theories and predictions that are not consistent with this assumption. If a theory predicts that some individuals will be systematically fooled or led into making costly mistakes, then this theory will tend to lose its validity when these individuals learn (from experience or from a published version of the theory itself) to better understand the situation. The importance of game theory in the social sciences is largely derived from this fact.

1.2 Basic Concepts of Decision Theory

The logical roots of game theory are in Bayesian decision theory. Indeed, game theory can be viewed as an extension of decision theory (to the case of two or more decision-makers), or as its essential logical fulfillment. Thus, to understand the fundamental ideas of game theory, one should begin by studying decision theory. The rest of this chapter is devoted to an introduction to the basic ideas of Bayesian decision theory, beginning with a general derivation of the expected utility maximization theorem and related results.

At some point, anyone who is interested in the mathematical social sciences should ask the question, Why should I expect that any simple quantitative model can give a reasonable description of people's behavior? The fundamental results of decision theory directly address this question, by showing that any decision-maker who satisfies certain intuitive axioms should always behave so as to maximize the mathematical expected value of some utility function, with respect to some subjective probability distribution. That is, any rational decision-maker's behavior should be describable by a *utility function*, which gives a quantitative characterization of his preferences for outcomes or prizes, and a *subjective probability distribution*, which characterizes his beliefs about all relevant unknown factors. Furthermore, when new information becomes available to such a decision-maker, his subjective probabilities should be revised in accordance with Bayes's formula.

There is a vast literature on axiomatic derivations of the subjective probability, expected-utility maximization, and Bayes's formula, beginning with Ramsey (1926), von Neumann and Morgenstern (1947), and Savage (1954). Other notable derivations of these results have been offered by Herstein and Milnor (1953), Luce and Raiffa (1957), Anscombe and Aumann (1963), and Pratt, Raiffa, and Schlaiffer (1964); for a general overview, see Fishburn (1968). The axioms used here are mainly borrowed from these earlier papers in the literature, and no attempt is made to achieve a logically minimal set of axioms. (In fact, a number of axioms presented in Section 1.3 are clearly redundant.)

Decisions under uncertainty are commonly described by one of two models: a *probability model* or a *state-variable model*. In each case, we speak of the decision-maker as choosing among *lotteries*, but the two models differ in how a lottery is defined. In a probability model, lotteries are probability distributions over a set of prizes. In a state-variable model, lotteries are functions from a set of possible states into a set of prizes. Each of these models is most appropriate for a specific class of applications.

A probability model is appropriate for describing gambles in which the prizes will depend on events that have obvious objective probabilities; we refer to such events as *objective unknowns*. These gambles are the "roulette lotteries" of Anscombe and Aumann (1963) or the "risks" of Knight (1921). For example, gambles that depend on the toss of a fair coin, the spin of a roulette wheel, or the blind draw of a ball out of an urn containing a known population of identically sized but differently colored balls all could be adequately described in a probability model. An important assumption being used here is that two objective unknowns with the same probability are completely equivalent for decision-making purposes. For example, if we describe a lottery by saying that it "offers a prize of $100 or $0, each with probability ½," we are assuming that it does not matter whether the prize is determined by tossing a fair coin or by drawing a ball from an urn that contains 50 white and 50 black balls.

On the other hand, many events do not have obvious probabilities; the result of a future sports event or the future course of the stock market are good examples. We refer to such events as *subjective unknowns*. Gambles that depend on subjective unknowns correspond to the "horse lotteries" of Anscombe and Aumann (1963) or the "uncertainties" of Knight (1921). They are more readily described in a state-variable model, because these models allow us to describe how the prize will be determined by the unpredictable events, without our having to specify any probabilities for these events.

Here we define our lotteries to include both the probability and the state-variable models as special cases. That is, we study lotteries in which

the prize may depend on both objective unknowns (which may be directly described by probabilities) and subjective unknowns (which must be described by a state variable). (In the terminology of Fishburn, 1970, we are allowing extraneous probabilities in our model.)

Let us now develop some basic notation. For any finite set Z, we let $\Delta(Z)$ denote the set of probability distributions over the set Z. That is,

(1.1) $$\Delta(Z) = \{q{:}Z \rightarrow \mathbf{R} \mid \sum_{y \in Z} q(y) = 1 \text{ and } q(z) \geq 0, \quad \forall z \in Z\}.$$

(Following common set notation, "|" in set braces may be read as "such that.")

Let X denote the set of possible *prizes* that the decision-maker could ultimately get. Let Ω denote the set of possible *states*, one of which will be the *true state of the world*. To simplify the mathematics, we assume that X and Ω are both finite sets. We define a *lottery* to be any function f that specifies a nonnegative real number $f(x|t)$, for every prize x in X and every state t in Ω, such that $\Sigma_{x \in X} f(x|t) = 1$ for every t in Ω. Let L denote the set of all such lotteries. That is,

$$L = \{f{:}\Omega \rightarrow \Delta(X)\}.$$

For any state t in Ω and any lottery f in L, $f(\cdot|t)$ denotes the probability distribution over X designated by f in state t. That is,

$$f(\cdot|t) = (f(x|t))_{x \in X} \in \Delta(X).$$

Each number $f(x|t)$ here is to be interpreted as the objective conditional probability of getting prize x in lottery f if t is the true state of the world. (Following common probability notation, "|" in parentheses may be interpreted here to mean "given.") For this interpretation to make sense, the state must be defined broadly enough to summarize all subjective unknowns that might influence the prize to be received. Then, once a state has been specified, only objective probabilities will remain, and an objective probability distribution over the possible prizes can be calculated for any well-defined gamble. So our formal definition of a lottery allows us to represent any gamble in which the prize may depend on both objective and subjective unknowns.

A *prize* in our sense could be any commodity bundle or resource allocation. We are assuming that the prizes in X have been defined so that they are mutually exclusive and exhaust the possible consequences of the decision-maker's decisions. Furthermore, we assume that each prize in X represents a complete specification of all aspects that the decision-maker cares about in the situation resulting from his decisions. Thus, the decision-maker should be able to assess a preference ordering over the set of lotteries, given any information that he might have about the state of the world.

The information that the decision-maker might have about the true state of the world can be described by an *event*, which is a nonempty subset of Ω. We let Ξ denote the set of all such events, so that

$$\Xi = \{S \mid S \subseteq \Omega \text{ and } S \neq \emptyset\}.$$

For any two lotteries f and g in L and any event S in Ξ, we write $f \gtrsim_S g$ iff the lottery f would be at least as desirable as g, in the opinion of the decision-maker, if he learned that the true state of the world was in the set S. (Here *iff* means "if and only if.") That is, $f \gtrsim_S g$ iff the decision-maker would be willing to choose the lottery f when he has to choose between f and g and he knows only that the event S has occurred. Given this relation ($\gtrsim_S$), we define relations ($\succ_S$) and ($\sim_S$) so that

$$f \sim_S g \text{ iff } f \gtrsim_S g \text{ and } g \gtrsim_S f;$$
$$f \succ_S g \text{ iff } f \gtrsim_S g \text{ and } g \not\gtrsim_S f.$$

That is, $f \sim_S g$ means that the decision-maker would be indifferent between f and g, if he had to choose between them after learning S; and $f \succ_S g$ means that he would strictly prefer f over g in this situation.

We may write $\gtrsim$, $\succ$, and $\sim$ for $\gtrsim_\Omega$, $\succ_\Omega$, and $\sim_\Omega$, respectively. That is, when no conditioning event is mentioned, it should be assumed that we are referring to prior preferences before any states in Ω are ruled out by observations.

Notice the assumption here that the decision-maker would have well-defined preferences over lotteries conditionally on any possible event in Ξ. In some expositions of decision theory, a decision-maker's conditional preferences are derived (using Bayes's formula) from the prior preferences that he would assess before making any observations; but such derivations cannot generate rankings of lotteries conditionally on events that have prior probability 0. In game-theoretic contexts, this omission is not as innocuous as it may seem. Kreps and Wilson (1982) have shown that the characterization of a rational decision-maker's beliefs and preferences after he observes a zero-probability event may be crucial in the analysis of a game.

For any number α such that $0 \leq \alpha \leq 1$, and for any two lotteries f and g in L, $\alpha f + (1 - \alpha)g$ denotes the lottery in L such that

$$(\alpha f + (1 - \alpha)g)(x|t) = \alpha f(x|t) + (1 - \alpha)g(x|t), \quad \forall x \in X, \quad \forall t \in \Omega.$$

To interpret this definition, suppose that a ball is going to be drawn from an urn in which α is the proportion of black balls and $1 - \alpha$ is the proportion of white balls. Suppose that if the ball is black then the decision-maker will get to play lottery f and if the ball is white then the decision-maker will get to play lottery g. Then the decision-maker's ultimate probability of getting prize x if t is the true state is $\alpha f(x|t) + (1 - \alpha)g(x|t)$. Thus, $\alpha f + (1 - \alpha)g$ represents the compound lottery that is built up from f and g by this random lottery-selection process.

For any prize x, we let $[x]$ denote the lottery that always gives prize x for sure. That is, for every state t,

(1.2) $\quad [x](y|t) = 1 \text{ if } y = x, \quad [x](y|t) = 0 \text{ if } y \neq x.$

Thus, $\alpha[x] + (1 - \alpha)[y]$ denotes the lottery that gives either prize x or prize y, with probabilities α and $1 - \alpha$, respectively.

1.3 Axioms

Basic properties that a rational decision-maker's preferences may be expected to satisfy can be presented as a list of axioms. Unless otherwise stated, these axioms are to hold for all lotteries e, f, g, and h in L, for all events S and T in Ξ, and for all numbers α and β between 0 and 1.

Axioms 1.1A and 1.1B assert that preferences should always form a complete transitive order over the set of lotteries.

AXIOM 1.1A (COMPLETENESS). *$f \gtrsim_S g$ or $g \gtrsim_S f$.*

AXIOM 1.1B (TRANSITIVITY). *If $f \gtrsim_S g$ and $g \gtrsim_S h$ then $f \gtrsim_S h$.*

It is straightforward to check that Axiom 1.1B implies a number of other transitivity results, such as if $f \sim_S g$ and $g \sim_S h$ then $f \sim_S h$; and if $f \succ_S g$ and $g \gtrsim_S h$ then $f \succ_S h$.

Axiom 1.2 asserts that only the possible states are relevant to the decision-maker, so, given an event S, he would be indifferent between two lotteries that differ only in states outside S.

AXIOM 1.2 (RELEVANCE). *If $f(\cdot|t) = g(\cdot|t) \quad \forall t \in S$, then $f \sim_S g$.*

Axiom 1.3 asserts that a higher probability of getting a better lottery is always better.

AXIOM 1.3 (MONOTONICITY). *If $f \succ_S h$ and $0 \le \beta < \alpha \le 1$, then $\alpha f + (1 - \alpha)h \succ_S \beta f + (1 - \beta)h$.*

Building on Axiom 1.3, Axiom 1.4 asserts that $\gamma f + (1 - \gamma)h$ gets better in a continuous manner as γ increases, so any lottery that is ranked between f and h is just as good as some randomization between f and h.

AXIOM 1.4 (CONTINUITY). *If $f \gtrsim_S g$ and $g \gtrsim_S h$, then there exists some number γ such that $0 \le \gamma \le 1$ and $g \sim_S \gamma f + (1 - \gamma)h$.*

The substitution axioms (also known as independence or sure-thing axioms) are probably the most important in our system, in the sense that they generate strong restrictions on what the decision-maker's preferences must look like even without the other axioms. They should also be very intuitive axioms. They express the idea that, if the decision-maker must choose between two alternatives and if there are two mutually exclusive events, one of which must occur, such that in each event he would prefer the first alternative, then he must prefer the first alternative before he learns which event occurs. (Otherwise, he would be expressing a preference that he would be sure to want to reverse after learning which of these events was true!) In Axioms 1.5A and 1.5B, these events are objective randomizations in a random lottery-selection process, as discussed in the preceding section. In Axioms 1.6A and 1.6B, these events are subjective unknowns, subsets of Ω.

AXIOM 1.5A (OBJECTIVE SUBSTITUTION). *If $e \gtrsim_S f$ and $g \gtrsim_S h$ and $0 \le \alpha \le 1$, then $\alpha e + (1 - \alpha)g \gtrsim_S \alpha f + (1 - \alpha)h$.*

AXIOM 1.5B (STRICT OBJECTIVE SUBSTITUTION). *If $e \succ_S f$ and $g \gtrsim_S h$ and $0 < \alpha \le 1$, then $\alpha e + (1 - \alpha)g \succ_S \alpha f + (1 - \alpha)h$.*

AXIOM 1.6A (SUBJECTIVE SUBSTITUTION). *If $f \gtrsim_S g$ and $f \gtrsim_T g$ and $S \cap T = \emptyset$, then $f \gtrsim_{S \cup T} g$.*

AXIOM 1.6B (STRICT SUBJECTIVE SUBSTITUTION). *If $f \succ_S g$ and $f \succ_T g$ and $S \cap T = \emptyset$, then $f \succ_{S \cup T} g$.*

To fully appreciate the importance of the substitution axioms, we may find it helpful to consider the difficulties that arise in decision theory when we try to drop them. For a simple example, suppose an individual would prefer x over y, but he would also prefer $.5[y]+.5[z]$ over $.5[x] + .5[z]$, in violation of substitution. Suppose that w is some other prize that he would consider better than $.5[x] + .5[z]$ and worse than $.5[y] + .5[z]$. That is,

$$x \succ y \text{ but } .5[y] + .5[z] \succ [w] \succ .5[x] + .5[z].$$

Now consider the following situation. The decision-maker must first decide whether to take prize w or not. If he does not take prize w, then a coin will be tossed. If it comes up Heads, then he will get prize z; and if it comes up Tails, then he will get a choice between prizes x and y.

What should this decision-maker do? He has three possible decision-making strategies: (1) take w, (2) refuse w and take x if Tails, (3) refuse w and take y if Tails. If he follows the first strategy, then he gets the lottery $[w]$; if he follows the second, then he gets the lottery $.5[x] + .5[z]$; and if he follows the third, then he gets the lottery $.5[y] + .5[z]$. Because he likes $.5[y] + .5[z]$ best among these lotteries, the third strategy would be best for him, so it may seem that he should refuse w. However, if he refuses w and the coin comes up Tails, then his preferences stipulate that he should choose x instead of y. So if he refuses w, then he will actually end up with z if Heads or x if Tails. But this lottery $.5[x] + .5[z]$ is worse than w. So we get the contradictory conclusion that he should have taken w in the first place.

Thus, if we are to talk about "rational" decision-making without substitution axioms, then we must specify whether rational decision-makers are able to commit themselves to follow strategies that they would subsequently want to change (in which case "rational" behavior would lead to $.5[y] + .5[z]$ in this example). If they cannot make such commitments, then we must also specify whether they can foresee their future inconstancy (in which case the outcome of this example should be $[w]$) or not (in which case the outcome of this example should be $.5[x]+.5[z]$). If none of these assumptions seem reasonable, then to avoid this dilemma we must accept substitution axioms as a part of our definition of rationality.

Axiom 1.7 asserts that the decision-maker is never indifferent between all prizes. This axiom is just a regularity condition, to make sure that there is something of interest that could happen in each state.

AXIOM 1.7 (INTEREST). *For every state t in Ω, there exist prizes y and z in X such that* $[y] \succ_{\{t\}} [z]$.

Axiom 1.8 is optional in our analysis, in the sense that we can state a version of our main result with or without this axiom. It asserts that the decision-maker has the same preference ordering over objective gambles in all states of the world. If this axiom fails, it is because the same prize might be valued differently in different states.

AXIOM 1.8 (STATE NEUTRALITY). *For any two states r and t in Ω, if $f(\cdot|r) = f(\cdot|t)$ and $g(\cdot|r) = g(\cdot|t)$ and $f \gtrsim_{\{r\}} g$, then $f \gtrsim_{\{t\}} g$.*

1.4 The Expected-Utility Maximization Theorem

A *conditional-probability function* on Ω is any function $p:\Xi \to \Delta(\Omega)$ that specifies nonnegative conditional probabilities $p(t|S)$ for every state t in Ω and every event S, such that

$$p(t|S) = 0 \text{ if } t \notin S, \text{ and } \sum_{r \in S} p(r|S) = 1.$$

Given any such conditional-probability function, we may write

$$p(R|S) = \sum_{r \in R} p(r|S), \quad \forall R \subseteq \Omega, \quad \forall S \in \Xi.$$

A *utility function* can be any function from $X \times \Omega$ into the real numbers **R**. A utility function $u:X \times \Omega \to \mathbf{R}$ is *state independent* iff it does not actually depend on the state, so there exists some function $U:X \to \mathbf{R}$ such that $u(x,t) = U(x)$ for all x and t.

Given any such conditional-probability function p and any utility function u and given any lottery f in L and any event S in Ξ, we let $E_p(u(f)|S)$ denote the expected utility value of the prize determined by f, when $p(\cdot|S)$ is the probability distribution for the true state of the world. That is,

$$E_p(u(f)|S) = \sum_{t \in S} p(t|S) \sum_{x \in X} u(x,t) f(x|t).$$

THEOREM 1.1. *Axioms 1.1AB, 1.2, 1.3, 1.4, 1.5AB, 1.6AB, and 1.7 are jointly satisfied if and only if there exists a utility function $u:X \times \Omega \to \mathbf{R}$ and a conditional-probability function $p:\Xi \to \Delta(\Omega)$ such that*

(1.3) $\max_{x \in X} u(x,t) = 1$ *and* $\min_{x \in X} u(x,t) = 0, \quad \forall t \in \Omega;$

(1.4) $p(R|T) = p(R|S)p(S|T), \quad \forall R, \ \forall S,$ *and* $\forall T$ *such that* $R \subseteq S \subseteq T \subseteq \Omega$ *and* $S \neq \emptyset;$

(1.5) $f \gtrsim_S g$ *if and only if* $E_p(u(f)|S) \geq E_p(u(g)|S),$ $\forall f,g \in L, \quad \forall S \in \Xi.$

Furthermore, given these Axioms 1.1AB–1.7, Axiom 1.8 is also satisfied if and only if conditions (1.3)–(1.5) here can be satisfied with a state-independent utility function.

In this theorem, condition (1.3) is a normalization condition, asserting that we can choose our utility functions to range between 0 and 1 in every state. (Recall that X and Ω are assumed to be finite.) Condition (1.4) is a version of Bayes's formula, which establishes how conditional probabilities assessed in one event must be related to conditional probabilities assessed in another. The most important part of the theorem is condition (1.5), however, which asserts that the decision-maker always prefers lotteries with higher expected utility. By condition (1.5), once we have assessed u and p, we can predict the decision-maker's optimal choice in any decision-making situation. He will choose the lottery with the highest expected utility among those available to him, using his subjective probabilities conditioned on whatever event in Ω he has observed. Notice that, with X and Ω finite, there are only finitely many utility and probability numbers to assess. Thus, the decision-maker's preferences over all of the infinitely many lotteries in L can be completely characterized by finitely many numbers.

To apply this result in practice, we need a procedure for assessing the utilities $u(x,t)$ and the probabilities $p(t|S)$, for all x, t, and S. As Raiffa (1968) has emphasized, such procedures do exist, and they form the basis of practical decision analysis. To define one such assessment procedure, and to prove Theorem 1.1, we begin by defining some special lotteries, using the assumption that the decision-maker's preferences satisfy Axioms 1.1AB–1.7.

Let a_1 be a lottery that gives the decision-maker one of the best prizes in every state; and let a_0 be a lottery that gives him one of the worst prizes in every state. That is, for every state t, $a_1(y|t) = 1 = a_0(z|t)$ for some prizes y and z such that, for every x in X, $y \gtrsim_{\{t\}} x \gtrsim_{\{t\}} z$. Such best and worst prizes can be found in every state because the preference relation ($\gtrsim_{\{t\}}$) forms a transitive ordering over the finite set X.

For any event S in Ξ, let b_S denote the lottery such that

$$b_S(\cdot|t) = a_1(\cdot|t) \text{ if } t \in S,$$

$$b_S(\cdot|t) = a_0(\cdot|t) \text{ if } t \notin S.$$

That is, b_S is a "bet on S" that gives the best possible prize if S occurs and gives the worst possible prize otherwise.

For any prize x and any state t, let $c_{x,t}$ be the lottery such that

$$c_{x,t}(\cdot|r) = [x](\cdot|r) \text{ if } r = t,$$

$$c_{x,t}(\cdot|r) = a_0(\cdot|r) \text{ if } r \notin t.$$

That is, $c_{x,t}$ is the lottery that always gives the worst prize, except in state t, when it gives prize x.

We can now define a procedure to assess the utilities and probabilities that satisfy the theorem, given preferences that satisfy the axioms. For each x and t, first ask the decision-maker, "For what number β would you be indifferent between $[x]$ and $\beta a_1 + (1 - \beta)a_0$, if you knew that t was the true state of the world?" By the continuity axiom, such a number must exist. Then let $u(x,t)$ equal the number that he specifies, such that

$$[x] \sim_{\{t\}} u(x,t)a_1 + (1 - u(x,t))a_0.$$

For each t and S, ask the decision-maker, "For what number γ would you be indifferent between $b_{\{t\}}$ and $\gamma a_1 + (1 - \gamma)a_0$ if you knew that the true state was in S?" Again, such a number must exist, by the continuity axiom. (The subjective substitution axiom guarantees that $a_1 \gtrsim_S b_{\{t\}} \gtrsim_S a_0$.) Then let $p(t|S)$ equal the number that he specifies, such that

$$b_{\{t\}} \sim_S p(t|S)a_1 + (1 - p(t|S))a_0.$$

In the proof of Theorem 1.1, we show that defining u and p in this way does satisfy the conditions of the theorem. Thus, finitely many questions suffice to assess the probabilities and utilities that completely characterize the decision-maker's preferences.

Proof of Theorem 1.1. Let p and u be as constructed above. First, we derive condition (1.5) from the axioms. The relevance axiom and the definition of $u(x,t)$ implies that, for every state r,

$$c_{x,t} \sim_{\{r\}} u(x,t)b_{\{t\}} + (1 - u(x,t))a_0.$$

Then subjective substitution implies that, for every event S,

$$c_{x,t} \sim_S u(x,t)b_{\{t\}} + (1 - u(x,t))a_0.$$

Axioms 1.5A and 1.5B together imply that $f \gtrsim_S g$ if and only if

$$\left(\frac{1}{|\Omega|}\right) f + \left(1 - \frac{1}{|\Omega|}\right) a_0 \gtrsim_S \left(\frac{1}{|\Omega|}\right) g + \left(1 - \frac{1}{|\Omega|}\right) a_0.$$

(Here, $|\Omega|$ denotes the number of states in the set Ω.) Notice that

$$\left(\frac{1}{|\Omega|}\right) f + \left(1 - \frac{1}{|\Omega|}\right) a_0 = \left(\frac{1}{|\Omega|}\right) \sum_{t\in\Omega} \sum_{x\in X} f(x|t)c_{x,t}.$$

But, by repeated application of the objective substitution axiom,

$$\left(\frac{1}{|\Omega|}\right) \sum_{t\in\Omega} \sum_{x\in X} f(x|t)c_{x,t}$$

$$\sim_S \left(\frac{1}{|\Omega|}\right) \sum_{t\in\Omega} \sum_{x\in X} f(x|t)\big(u(x,t)b_{\{t\}} + (1 - u(x,t))a_0\big)$$

$$\sim_S \left(\frac{1}{|\Omega|}\right) \sum_{t\in\Omega} \sum_{x\in X} f(x|t)\big(u(x,t)\big(p(t|S)a_1 + (1 - p(t|S))a_0\big) + (1 - u(x,t))a_0\big)$$

$$= \left(\frac{1}{|\Omega|}\right) \sum_{t\in\Omega} \sum_{x\in X} f(x|t)u(x,t)p(t|S)a_1 + \left(1 - \sum_{t\in\Omega} \sum_{x\in X} f(x|t)u(x,t)p(t|S)/|\Omega|\right)a_0$$

$$= \big(E_p(u(f)|S)/|\Omega|\big)a_1 + \big(1 - \big(E_p(u(f)|S)/|\Omega|\big)\big)a_0.$$

Similarly,

$$(1/|\Omega|)g + (1 - (1/|\Omega|))a_0 \sim_S \big(E_p(u(g)|S)/|\Omega|\big)a_1 + \big(1 - \big(E_p(u(g)|S)/|\Omega|\big)\big)a_0.$$

Thus, by transitivity, $f \gtrsim_S g$ if and only if

$$\left(E_p(u(f)|S)/|\Omega|\right)a_1 + \left(1 - (E_p(u(f)|S)/|\Omega|)\right)a_0$$
$$\gtrsim_S \left(E_p(u(g)|S)/|\Omega|\right)a_1 + \left(1 - (E_p(u(g)|S)/|\Omega|)\right)a_0.$$

But by monotonicity, this final relation holds if and only if

$$E_p(u(f)|S) \geq E_p(u(g)|S),$$

because interest and strict subjective substitution guarantee that $a_1 \succ_S a_0$. Thus, condition (1.5) is satisfied.

Next, we derive condition (1.4) from the axioms. For any events R and S,

$$\left(\frac{1}{|R|}\right) b_R + \left(1 - \frac{1}{|R|}\right) a_0 = \left(\frac{1}{|R|}\right) \sum_{r \in R} b_{\{r\}}$$
$$\sim_S \left(\frac{1}{|R|}\right) \sum_{r \in R} \left(p(r|S)a_1 + (1 - p(r|S))a_0\right)$$
$$= \left(\frac{1}{|R|}\right) \left(p(R|S)a_1 + (1 - p(R|S))a_0\right) + \left(1 - \frac{1}{|R|}\right) a_0,$$

by objective substitution. ($|R|$ is the number of states in the set R.) Then, using Axioms 1.5A and 1.5B, we get

$$b_R \sim_S p(R|S)a_1 + (1 - p(R|S))a_0.$$

By the relevance axiom, $b_S \sim_S a_1$ and, for any r not in S, $b_{\{r\}} \sim_S a_0$. So the above formula implies (using monotonicity and interest) that $p(r|S) = 0$ if $r \notin S$, and $p(S|S) = 1$. Thus, p is a conditional-probability function, as defined above.

Now, suppose that $R \subseteq S \subseteq T$. Using $b_S \sim_S a_1$ again, we get

$$b_R \sim_S p(R|S)b_S + (1 - p(R|S))a_0.$$

Furthermore, because b_R, b_S, and a_0 all give the same worst prize outside S, relevance also implies

$$b_R \sim_{T\setminus S} p(R|S)b_S + (1 - p(R|S))a_0.$$

(Here $T\setminus S = \{t | t \in T, t \notin S\}$.) So, by subjective and objective substitution,

$$b_R \sim_T p(R|S)b_S + (1 - p(R|S))a_0$$
$$\sim_T p(R|S)\left(p(S|T)a_1 + (1 - p(S|T))a_0\right) + (1 - p(R|S))a_0$$
$$= p(R|S)p(S|T)a_1 + (1 - p(R|S)p(S|T))a_0.$$

But $b_R \sim_T p(R|T)a_1 + (1 - p(R|T))a_0$. Also, $a_1 \succ_T a_0$, so monotonicity implies that $p(R|T) = p(R|S)p(S|T)$. Thus, Bayes's formula (1.4) follows from the axioms.

If y is the best prize and z is the worst prize in state t, then $[y] \sim_{\{t\}} a_1$ and $[z] \sim_{\{t\}} a_0$, so that $u(y,t) = 1$ and $u(z,t) = 0$ by monotonicity. So the range condition (1.3) is also satisfied by the utility function that we have constructed.

If state neutrality is also given, then the decision-maker will give us the same answer when we assess $u(x,t)$ as when we assess $u(x,r)$ for any other state r (because $[x] \sim_{\{t\}} \beta a_1 + (1 - \beta)a_0$ implies $[x] \sim_{\{r\}} \beta a_1 + (1 - \beta)a_0$, and monotonicity and interest guarantee that his answer is unique). So Axiom 1.8 implies that u is state-independent.

To complete the proof of the theorem, it remains to show that the existence of functions u and p that satisfy conditions (1.3)–(1.5) in the theorem is sufficient to imply all the axioms (using state independence only for Axiom 1.8). If we use the basic mathematical properties of the expected-utility formula, verification of the axioms is straightforward. To illustrate, we show the proof of one axiom, subjective substitution, and leave the rest as an exercise for the reader.

Suppose that $f \gtrsim_S g$ and $f \gtrsim_T g$ and $S \cap T = \emptyset$. By (1.5), $E_p(u(f)|S) \geq E_p(u(g)|S)$ and $E_p(u(f)|T) \geq E_p(u(g)|T)$. But Bayes's formula (1.4) implies that

$$E_p(u(f)|S \cup T) = \sum_{t \in S \cup T} \sum_{x \in X} p(t|S \cup T)f(x|t)u(x,t)$$
$$= \sum_{t \in S} \sum_{x \in X} p(t|S)p(S|S \cup T)f(x|t)u(x,t)$$
$$+ \sum_{t \in T} \sum_{x \in X} p(t|T)p(T|S \cup T)f(x|t)u(x,t)$$
$$= p(S|S \cup T)E_p(u(f)|S) + p(T|S \cup T)E_p(u(f)|S)$$

and

$$E_p(u(g)|S \cup T) = p(S|S \cup T)E_p(u(g)|S) + p(T|S \cup T)E_p(u(g)|S).$$

So $E_p(u(f)|S \cup T) \geq E_p(u(g)|S \cup T)$ and $f \gtrsim_{S \cup T} g$. ■

1.5 Equivalent Representations

When we drop the range condition (1.3), there can be more than one pair of utility and conditional-probability functions that represent the same decision-maker's preferences, in the sense of condition (1.5). Such equivalent representations are completely indistinguishable in terms of their decision-theoretic properties, so we should be suspicious of any theory of economic behavior that requires distinguishing between such equivalent representations. Thus, it may be theoretically important to be able to recognize such equivalent representations.

Given any subjective event S, when we say that a utility function v and a conditional-probability function q *represent* the preference ordering $\gtrsim_S$, we mean that, for every pair of lotteries f and g, $E_q(v(f)|S) \geq E_q(v(g)|S)$ if and only if $f \gtrsim_S g$.

THEOREM 1.2. *Let S in Ξ be any given subjective event. Suppose that the decision-maker's preferences satisfy Axioms 1.1AB through 1.7, and let u and p be utility and conditional-probability functions satisfying (1.3)–(1.5) in Theorem 1.1. Then v and q represent the preference ordering $\gtrsim_S$ if and only if there exists a positive number A and a function $B{:}S \rightarrow \mathbf{R}$ such that*

$$q(t|S)v(x,t) = Ap(t|S)u(x,t) + B(t), \quad \forall t \in S, \quad \forall x \in X.$$

Proof. Suppose first that A and $B(\cdot)$ exist as described in the theorem. Then, for any lottery f,

$$\begin{aligned} E_q(v(f)|S) &= \sum_{t\in S}\sum_{x\in X} f(x|t)q(t|S)v(x,t) \\ &= \sum_{t\in S}\sum_{x\in X} f(x|t)(Ap(t|S)u(x,t) + B(t)) \\ &= A\sum_{t\in S}\sum_{x\in X} f(x|t)p(t|S)u(x,t) + \sum_{t\in S} B(t)\sum_{x\in X} f(x|t) \\ &= AE_p\,(u(f)|S) + \sum_{t\in S} B(t), \end{aligned}$$

because $\Sigma_{x\in X}\, f(x|t) = 1$. So expected v-utility with respect to q is an increasing linear function of expected u-utility with respect to p, because $A > 0$. Thus, $E_q(v(f)|S) \geq E_q(v(g)|S)$ if and only if $E_p(u(f)|S) \geq E_p(u(g)|S)$, and so v and q together represent the same preference ordering over lotteries as u and p.

Conversely, suppose now that v and q represent the same preference ordering as u and p. Pick any prize x and state t, and let

$$\lambda = \frac{E_q(v(c_{x,t})|S) - E_q(v(a_0)|S)}{E_q(v(a_1)|S) - E_q(v(a_0)|S)}.$$

Then, by the linearity of the expected-value operator,

$$\begin{aligned} E_q\big(v(\lambda a_1 + (1-\lambda)a_0)|S\big) &= E_q(v(a_0)|S) + \lambda\big(E_q(v(a_1)|S) - E_q(v(a_0)|S)\big) \\ &= E_q(v(c_{x,t})|S), \end{aligned}$$

so $c_{x,t} \sim_S \lambda a_1 + (1-\lambda)a_0$. In the proof of Theorem 1.1, we constructed u and p so that

$$\begin{aligned} c_{x,t} &\sim_S u(x,t)b_{\{t\}} + (1 - u(x,t))a_0 \\ &\sim_S u(x,t)\big(p(t|S)a_1 + (1 - p(t|S))a_0\big) + (1 - u(x,t))a_0 \\ &\sim_S p(t|S)u(x,t)a_1 + (1 - p(t|S)u(x,t))a_0. \end{aligned}$$

The monotonicity axiom guarantees that only one randomization between a_1 and a_0 can be just as good as $c_{x,t}$, so

$$\lambda = p(t|S)u(x,t).$$

But $c_{x,t}$ differs from a_0 only in state t, where it gives prize x instead of the worst prize, so

$$E_q(v(c_{x,t})|S) - E_q(v(a_0)|S) = q(t|S)\left(v(x,t) - \min_{z\in X} v(z,t)\right).$$

Thus, going back to the definition of λ, we get

$$p(t|S)u(x,t) = \frac{q(t|S)(v(x,t) - \min_{z\in X} v(z,t))}{E_q(v(a_1)|S) - E_q(v(a_0)|S)}.$$

Now let

$$A = E_q(v(a_1)|S) - E_q(v(a_0)|S),$$

and let

$$B(t) = q(t|S)\min_{z\in X} v(z,t).$$