Rational Choice

Itzhak Gilboa

The MIT Press
Cambridge, Massachusetts
London, England

This book was set in Palatino on 3B2 by Asco Typesetters, Hong Kong.
Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Gilboa, Itzhak.
Rational choice / Itzhak Gilboa.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-262-01400-7 (hardcover : alk. paper)
1. Rational choice theory. 2. Social choice. 3. Decision making. 4. Game theory.
5. Microeconomics. I. Title.
HM495.G55 2010
302′.13—dc22 2009037119

10 9 8 7 6 5 4 3 2 1

A Mathematical Preliminaries

A.1 Notation

$\sum$	summation
	$\sum_{i=1}^{n} a_i = a_1 + a_2 + \cdots + a_n.$
$\prod$	product (multiplication)
	$\prod_{i=1}^{n} a_i = a_1 a_2 \ldots a_n.$
$\forall$	for every
$\exists$	there exists
$\Rightarrow$	implies
$\Leftrightarrow$	if and only if (implies and is implied by)
iff	if and only if
$\cup$	union (of sets)
$\cap$	intersection (of sets)
A^c	the complement of (the set) A
$\subset$	a subset of
$\in$	belongs to, member of
$\notin$	doesn't belong to, not a member of
$\varnothing$	the empty set
$f : A \to B$	a function from the set A to the set B
$A \times B$	the product set

$\mathbb{R}$	the set of real numbers
$\mathbb{R}^n$	the n-dimensional Euclidean space
$[a, b]$	a closed interval $\{x \in \mathbb{R} \mid a \leq x \leq b\}$
(a, b)	an open interval $\{x \in \mathbb{R} \mid a < x < b\}$
$\|\cdot\|$	norm, length of a vector
$\#, \|\cdot\|$	cardinality of a set; if the set is denoted A, then $\#A = \|A\|$ denotes its cardinality.

A.2 Sets

A set is a primitive notion. Sets are often denoted by capital letters, A, B, $C, \ldots$ and indicated by braces $\{\ \}$. Inside these braces are listed the elements of the set. For instance, $A = \{0, 1\}$ refers to the set consisting of 0 and 1. Sets can also be described without listing all elements explicitly inside the braces. For instance,

$$N = \{1, \ldots, n\}$$

denotes the set of all natural numbers from 1 to n. Similarly, we define

$$\mathbb{N} = \{1, 2, 3, \ldots\}$$

and

$$\mathbb{Z} = \{\ldots, -1, 0, 1, \ldots\}$$

to be the set of natural numbers and the set of integer numbers, respectively.

The notation $a \in A$ means that a is a *member* of the set A or that a *belongs to* A. $a \notin A$ is the negation of $a \in A$.

The symbol $\subset$ designates a relation between sets, meaning "is a subset of." Explicitly, $A \subset B$ means that A is a subset of B, that is, for all $x \in A$, it is true that $x \in B$. Thus, $x \in A$ iff $\{x\} \subset A$.

The symbol $\varnothing$ denotes the empty set, the set that has no elements.

Sets are also defined by a certain condition that elements should satisfy. For instance,

$$A = \{n \in \mathbb{N} \mid n > 3\}$$

denotes all the natural numbers greater than 3, that is, $A = \{4, 5, 6, \ldots\}$.

$\mathbb{R}$ denotes the set of real numbers. I don't define them here formally, although they can be defined using the rationals, which are, in turn, defined as the ratios of integer numbers.

When we use mathematics to model reality, we also refer to sets whose elements need not be mathematical objects. For instance,

$A = \{humans\}$,

$B = \{mammals\}$.

Such sets are viewed as sets of some mathematical objects interpreted as humans or mammals, respectively. Thus, when we discuss a set of individuals, alternatives, strategies, or states of the world, we mean a set whose elements are interpreted as individuals, alternatives, and so on.

The basic set operations are as follows.

Union ($\cup$). A binary operation on sets, resulting in a set containing all the elements that are in at least one of the sets. Or, for sets A and B,

$$A \cup B = \{x \mid x \in A \quad \text{or} \quad x \in B\}.$$

Here and elsewhere, *or* is inclusive, that is, "p or q" means "p, or q, or possibly both."

Intersection ($\cap$). A binary operation resulting in elements that are in both sets. That is,

$$A \cap B = \{x \mid x \in A \quad \text{and} \quad x \in B\}$$

Two sets A and B are *disjoint* if they have an empty intersection, that is, if $A \cap B = \varnothing$.

Complement (^{c}). A unary operation containing all elements that are not in the set. To define it, we need a reference set. That is, if S is the entire universe,

$$A^c = \{x \mid x \notin A\}.$$

You may verify that

$(A^c)^c = A$,

$A \cap B \subset A, B \subset A \cup B$,

$A \cap A^c = \varnothing$,

$(A \cup B)^c = A^c \cap B^c$,

$(A \cap B)^c = A^c \cup B^c$,

and

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C),$$

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C).$$

Given two sets A and B, we define their *(Cartesian) product* $A \times B$ to be all the ordered pairs whose first element is from A and whose second element is from B. In formal notation,

$$A \times B = \{(x, y) \mid x \in A \quad \text{and} \quad y \in B\}.$$

Note that (x, y) is an *ordered pair* because the order matters. That is, $(x, y) \neq (y, x)$ unless $x = y$. This is distinct from the set containing x and y, in which the order does not matter. That is, $\{x, y\} = \{y, x\}$.

The notation A^2 means $A \times A$. Thus, it refers to the set of all the ordered pairs each element of which is in A. Similarly, we define

$$A^n = A \times \cdots \times A = \{(x_1, \ldots, x_n) \mid x_i \in A, i \leq n\}.$$

The *power set* of a set A is the set of all subsets of A. It is denoted

$$2^A = P(A) = \{B \mid B \subset A\}.$$

A.3 Relations and Functions

A binary relation is a subset of ordered pairs. Specifically, if R is a *binary relation* from a set A to a set B, we mean that

$$R \subset A \times B.$$

This is an extensional definition. The relation R is defined by a list of all pairs of elements in A and in B such that the former relates to the latter. For instance, consider the relation R, "located in," from the set of buildings A to the set of cities B. Then, if we have

$$R = \left\{ \begin{array}{c} (\text{Empire\_State\_Building,New\_York}), \\ (\text{Louvre,Paris}), \\ (\text{Big\_Ben,London}), \ldots \end{array} \right\}$$

we wish to say that the Empire State Building relates to New York by the relation "located in," that is, it is in New York; the building of the Louvre is in Paris; and so forth.

For a relation $R \subset A \times B$ we can define the inverse relation, $R^{-1} \subset B \times A$ by

$$R^{-1} = \{(y, x) \mid (x, y) \subset R\}.$$

Of particular interest are relations between elements of the same set. For a set A, a binary relation on A is a relation $R \subset A^2 (= A \times A)$. For instance, if A is the set of people, then "child_of" is a relation given by

$$R = \left\{ \begin{array}{c} (\text{Cain,Adam}), \\ (\text{Cain,Eve}), \\ (\text{Abel,Adam}), \ldots \end{array} \right\},$$

and the relation "parent_of" will be

$$R^{-1} = \left\{ \begin{array}{c} (\text{Adam,Cain}), \\ (\text{Eve,Cain}), \\ (\text{Adam,Abel}), \ldots \end{array} \right\}.$$

A *function* f from A to B, denoted

$$f : A \to B,$$

is a binary relation $R \subset A \times B$ such that for every $x \in A$, there exists precisely one $y \in B$ such that $(x, y) \in R$. We then write

$$f(x) = y$$

or

$$f : x \mapsto y.$$

The latter is also used to specify the function by a formula. For instance, we can think of the square function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = x^2$$

or write

$$f : x \mapsto x^2.$$

A function $f : A \to B$ is 1–1 (*one-to-one*) if it never attaches the same $y \in B$ to different $x_1, x_2 \in A$, that is, if

$$f(x_1) = f(x_2) \Rightarrow x_1 = x_2.$$

A function $f : A \to B$ is *onto* if every $y \in B$ has at least one $x \in A$ such that $f(x) = y$.

If $f : A \to B$ is both one-to-one and onto, we can define its inverse

$$f^{-1} : B \to A$$

by

$$f^{-1}(y) = x \Leftrightarrow y = f(x).$$

Observe that the notation f^{-1} is consistent with the notation R^{-1} for relations. Recalling that a function is a relation, one can always define f^{-1} as

$$f^{-1} = \{(y, x) \mid y = f(x)\},$$

and if f is one-to-one and onto, this relation is indeed a function, and it coincides with the inverse function of f.

We often also use the notation

$$f^{-1}(x) = \{y \in B \mid y = f(x)\}.$$

With this notation, to say that f is one-to-one is equivalent to saying that $f^{-1}(x)$ has at most one element for every x. To say that it is onto is equivalent to saying that $f^{-1}(x)$ is nonempty. And if f is both one-to-one and onto (a *bijection*), $f^{-1}(x)$ has exactly one element for each x. Then, according to the set notation, for a particular y,

$$f^{-1}(x) = \{y\},$$

and according to the inverse function notation,

$$f^{-1}(x) = y.$$

Using f^{-1} both for the element y and for the set containing only y seems problematic when one is just starting to deal with formal models, but it becomes more common as one advances. This is called an abuse of notation, and it is often acceptable as long as readers know what is meant by it.

Interesting properties of binary relations on a set $R \subset A^2$ include the following.

R is *reflexive* if $\forall x \in A$, xRx, that is, every x relates to itself. For instance, the relations $=$ and $\geq$ on $\mathbb{R}$ are reflexive, but $>$ isn't.

R is *symmetric* if $\forall x, y \in A$, xRy implies yRx, that is, if x relates y, then the converse also holds. For instance, the relation $=$ (on $\mathbb{R}$) is symmetric, but $\geq$ and $>$ aren't. Notice that $>$ does not allow any pair x, y to have both $x > y$ and $y > x$, that is,

$$> \cap >^{-1} = > \cap < = \varnothing,$$

whereas $\geq$ does because if $x = y$, it is true that $x \geq y$ and $y \geq x$. But $\geq$ is not symmetric because it is not always the case that xRy implies yRx.

R is *transitive* if $\forall x, y, z \in A$, xRy and yRz imply xRz, that is, if x relates to z through y, then x relates to z also directly. For example, $=$, $\geq$, and $>$ on $\mathbb{R}$ are all transitive, but the relation "close to" defined by

$$xRy \Leftrightarrow |x - y| < 1$$

is not transitive.

A relation that is reflexive, symmetric, and transitive is called an *equivalence relation*. Equality $=$ is such a relation. Also, "having the same square," that is,

$$xRy \Leftrightarrow x^2 = y^2,$$

is an equivalence relation.

In fact, a relation R on a set A is an equivalence relation if and only if there exist a set B and a function $f : A \to B$ such that

$$xRy \Leftrightarrow f(x) = f(y).$$

A.4 Cardinalities of Sets

The cardinality of a set A, denoted $\#A$ or $|A|$, is a measure of its size. If A is finite, the cardinality is simply the number of elements in A. If A is finite and $|A| = k$, then the number of subsets of A is

$$|P(A)| = 2^k.$$

If we have also $|B| = m$, then

$$|A \times B| = km.$$

Applied to the product of a set with itself,

$$|A^n| = k^n.$$

For infinite sets the measurement of the size, or cardinality, is more complicated. The notation ∞ denotes infinity, but it does not distinguish among infinities. And it turns out that there are meaningful ways in which infinities may differ.

How do we compare the sizes of infinite sets? The basic idea is this. Suppose we are given two sets A and B, and a one-to-one function $f : A \to B$. Then we want to say that B is at least as large as A, that is,

$|B| \geq |A|$.

If the converse also holds, that is, there also exists a one-to-one function $g : B \to A$, then we also have $|A| \geq |B|$, and together these imply that A and B have the same *cardinality*, $|A| = |B|$. (In this case it is also true that there is a one-to-one and onto function from A to B.) Otherwise, we say that the cardinality of B is larger than that of A, $|B| > |A|$.

For example, if

$$A = \{1, 2, \ldots\},$$

$$B = \{2, 3, \ldots\},$$

we find that the function $f : A \to B$ defined by $f(n) = n + 1$ is one-to-one and onto between A and B. Thus, the two sets are just as large. There is something counterintuitive here. A contains all of B plus one element, 1. So it feels like A should be strictly larger than B. But there is no interesting definition of the size of a set that distinguishes between A and B. The reason is that the bijection f suggests that we think of B as identical to A, with a renaming of the elements. With a bijection between two sets, it's hopeless to try to assign them different sizes.

By the same logic, the intervals

$$[0, 1] = \{x \in \mathbb{R} \mid 0 \leq x \leq 1\}$$

and

$$[0, 2] = \{x \in \mathbb{R} \mid 0 \leq x \leq 2\}$$

are of the same cardinality because the function $f(x) = 2x$ is a bijection from the first to the second. This is even more puzzling because these intervals have lengths, and the length of $[0, 2]$ is twice as large as that of $[0, 1]$. Indeed, there are other concepts of size in mathematics that would be able to capture that fact. But cardinality, attempting to count numbers, doesn't.

The cardinality of $(-1, 1)$ is identical to that of the entire real line, $\mathbb{R}$, even though the length of the former is finite and of the latter infinite. (Use the functions tag/arctag to switch between the two sets.)

Combining these arguments, we see that $\mathbb{R}$ has the same cardinality as $[0, 1]$, or $[0, 0.1]$, or $[0, \varepsilon]$ for any $\varepsilon > 0$.

Continuing with the list of counterintuitive comparisons, we find that the naturals $\mathbb{N} = \{1, 2, 3, \ldots\}$ and the integers $\mathbb{Z} = \{\ldots, -1, 0, 1, \ldots\}$ are of the same cardinality even though the integers include all

the naturals, their negatives, and zero. Clearly, we can have a one-to-one function from $\mathbb{N}$ to $\mathbb{Z}$: the identity $(f(n) = n)$. But we can also map $\mathbb{Z}$ to $\mathbb{N}$ in a one-to-one way. For instance, consider the following enumeration of $\mathbb{Z}$:

$$\mathbb{Z} = \{0, 1, -1, 2, -2, 3, -3, \ldots\},$$

that is,

$$\begin{aligned} 0 &\mapsto 1 \\ 1 &\mapsto 2 \\ -1 &\mapsto 3 \\ &\vdots \\ k &\mapsto 2k \\ -k &\mapsto 2k+1 \end{aligned}$$

This function from $\mathbb{Z}$ to $\mathbb{N}$ is one-to-one (and we also made it onto).

Similarly, the set of rational numbers

$$\mathbb{Q} = \left\{ \frac{a}{b} \,\middle|\, a \in Z, b \in \mathbb{N} \right\}$$

is of the same cardinality as $\mathbb{N}$. As previously, it is easy to map $\mathbb{N}$ into $\mathbb{Q}$ in a one-to-one way because $\mathbb{N} \subset \mathbb{Q}$. But the converse is also true. We may list all the rational numbers in a sequence $q_1, q_2, \ldots$ such that any rational will appear in a certain spot in the sequence, and no two rational numbers will claim the same spot. For instance, consider the table

	0	1	−1	2	−2	...
1	q_1	q_2	q_4	q_7	...	
2	q_3	q_5	q_8	...		
3	q_6	q_9	...			
4	q_{10}	...				
...	...					

Note that different representations of the same rational number are counted several times. For instance, $q_1 = q_3 = \cdots = 0$. Hence, define the function from $\mathbb{Q}$ to $\mathbb{N}$ as follows: for $q \in \mathbb{Q}$, let $f(q)$ be the minimal

n such that $q_n = 9$, where q_n is defined by the table. Clearly, every q appears somewhere in the list $q_1, q_2, \ldots$; hence this function is well defined. It is one-to-one because each q_n can equal only one number in $\mathbb{Q}$.

It seems at this point that all infinite sets are, after all, of the same size. But this is not the case. We concluded that the sets

$$\mathbb{N}, \mathbb{Z}, \mathbb{Q}$$

are of the same cardinality, and so are

$$\mathbb{R}, [0, 1], [0, \varepsilon]$$

for any $\varepsilon > 0$. But the cardinality of the first triple is lower than the cardinality of the second.

Clearly, the cardinality of $\mathbb{N}$ cannot exceed that of $\mathbb{R}$, because $\mathbb{N} \subset \mathbb{R}$, and thus the identity function maps $\mathbb{N}$ into $\mathbb{R}$ in a one-to-one manner. The question is, can we have the opposite direction, namely, can we map $\mathbb{R}$ into $\mathbb{N}$ in a one-to-one way, or equivalently, can we count the elements in $\mathbb{R}$? The answer is negative. There are at least three insightful proofs of this fact (not provided here). It suffices to know that there are sets that are not countable, and any interval with a positive length is such a set. Thus, in a well-defined sense, there are as many rational numbers as there are natural numbers, and there are as many numbers in any interval as there are in the entire real line (and, in fact, in any $\mathbb{R}^n$), but any interval (with a positive length) has more points than the natural (or the rational) numbers.

A.5 Calculus

A.5.1 Limits of Sequences

The notion of a limit is intuitive and fundamental. What is the limit of $\frac{1}{n}$ as $n \to \infty$? It is zero. We write this as

$$\lim_{n\to\infty} \frac{1}{n} = 0$$

or

$$\frac{1}{n} \to_{n\to\infty} 0.$$

Formally, we say that a sequence of real numbers $\{a_n\}$ converges to a number b, denoted

$a_n \rightarrow_{n\to\infty} b$

or

$$\lim_{n\to\infty} a_n = b$$

if the following holds: for every $\varepsilon > 0$ there exists N such that

$n \geq N$

implies

$|a_n - b| < \varepsilon$.

Intuitively, $\{a_n\}$ converges to b if it gets closer and closer to b. How close? As close as we wish. We decide how close to b we want the sequence to be, and we can then find a place in the sequence, N, such that all numbers in the sequence from that place on are as close to b as we requested.

If the sequence converges to ∞ (or $-\infty$), we use a similar definition, but we have to redefine the notion of "close to." Being close to ∞ doesn't mean having a difference of no more than ε, but rather, being large. Formally, $a_n \rightarrow_{n\to\infty} \infty$ if, for every M, there exists N such that

$n \geq N \Rightarrow a_n > M$,

and a similar definition is used for convergence to $-\infty$.

A.5.2 Limits of Functions

Again, we intuitively understand what is the limit of a function at a point. For instance, if x is a real-valued variable ($x \in \mathbb{R}$), we can agree that

$$\lim_{x\to\infty} \frac{1}{x} = 0$$

and

$$\lim_{x\to 0} \frac{1}{x^2} = \infty.$$

The formal definition of a limit is the following. The statement

$$\lim_{x\to a} f(x) = b$$

or

$$f(x) \rightarrow_{x \rightarrow a} b$$

means that for every $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$|x - a| < \delta$$

implies

$$|f(x) - b| < \varepsilon.$$

That is, if we know that we want the value of the function to be close to (within ε of) the limit b, we just have to be close enough to (within δ of) the argument a.

The proximity of the argument is defined a little differently when we approach infinity. Being close to ∞ doesn't mean being within δ of it, but being above some value. Explicitly, the statement

$$\lim_{x \rightarrow \infty} f(x) = b$$

or

$$f(x) \rightarrow_{x \rightarrow \infty} b$$

means that for every $\varepsilon > 0$, there exists M such that

$$x > M$$

implies

$$|f(x) - b| < \varepsilon.$$

Similarly, if we wish to say that the function converges to ∞ as x converges to a, we say that for every M, there exists a $\delta > 0$ such that

$$|x - a| < \delta$$

implies

$$f(x) > M.$$

Similar definitions apply to $\lim_{x \rightarrow \infty} f(x) = \infty$ and to the case in which x or $f(x)$ is $-\infty$.

A.5.3 Continuity

A function $f : \mathbb{R} \rightarrow \mathbb{R}$ is continuous at a point a if it equals its own limit, that is, if

$$\lim_{x \to a} f(x) = f(a).$$

The same definition applies to multiple variables. If we have $f : \mathbb{R}^n \to \mathbb{R}$, we say that f is continuous at x if $f(x) \to f(a)$ whenever $x \to a$. Specifically, f is continuous at $a \in \mathbb{R}^n$ if for every $\varepsilon > 0$, there exists $\delta > 0$ such that $\|x - a\| < \delta$ implies $|f(x) - f(a)| < \varepsilon$.

A.5.4 Derivatives

The *derivative* of a real-valued function of a single variable, $f : \mathbb{R} \to \mathbb{R}$, at a point a, is defined as

$$f'(a) = \frac{df}{dx}(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a}.$$

If we draw the graph of the function and let $x \neq a$ be close to a, $\frac{f(x)-f(a)}{x-a}$ is the slope of the string connecting the point on the graph corresponding to a, $(a, f(a))$ and the point corresponding to x, $(x, f(x))$. The derivative of f at the point a is the limit of this slope. Thus, it is the slope of the graph at the point a, or the slope of the tangent to the function.

When we say that a function has a derivative at a point a, we mean that this limit exists. It may not exist if, for instance, the function has a kink at a (for instance, $f(x) = |x - a|$), or if the function is too wild to have a limit even when x approaches a from one side.

The geometric interpretation of the derivative f' is therefore the slope of the function, or its rate of increase, that is, the ratio between the increase (positive or negative) in the value of the function relative to a small change in the variable x. If x measures time, and $f(x)$ measures the distance from a given point, $f'(x)$ is the velocity. If x measures the quantity of a good, and $u(x)$ measures the utility function, then $u'(x)$ measures the marginal utility of the good.

A function that always has a derivative is called *differentiable*. At every point a, we can approximate it by the linear function that is its tangent,

$$g(x) = f(a) + (x - a)f'(a),$$

and for values of x close to a, this approximation will be reasonable. Specifically, by definition of the derivative, the difference between the approximation, $g(x)$, and the function, $f(x)$, will converge to zero faster than x converges to a:

$$\frac{g(x)-f(x)}{x-a}=\frac{f(a)-f(x)-(x-a)f'(a)}{x-a}$$

$$=\frac{f(a)-f(x)}{x-a}-f'(a),$$

where the definition of the derivative means that the latter converges to zero as $x \to a$.

Thus, the zero-order approximation to the function f around a is the constant $f(a)$. The first-order approximation is the linear function $f(a)+(x-a)f'(a)$. Using higher-order derivatives (derivatives of derivatives of ... the derivative), one can get higher-order approximations of f by higher-order polynomials in x.

A.5.5 Partial Derivatives

When we have a function of several variables,

$$f:\mathbb{R}^n \to \mathbb{R}$$

we can consider the rate of the change in the function relative to each of the variables. If we wish to see what is the impact (on f) of changing, say, only the first variable x_1, we can fix the values of the other variables $\bar{x}_2,\ldots,\bar{x}_n$ and define

$$f_{\bar{x}_2,\ldots,\bar{x}_n}(x_1)=f(x_1,\bar{x}_2,\ldots,\bar{x}_n).$$

Focusing on the impact of x_1, we can study the derivative of $f_{\bar{x}_2,\ldots,\bar{x}_n}$. Since the other variables are fixed, we call this a *partial derivative*, denoted

$$\frac{\partial f}{\partial x_1}(x_1,\bar{x}_2,\ldots,\bar{x}_n)=\frac{df_{\bar{x}_2,\ldots,\bar{x}_n}}{dx_1}(x_1).$$

A function $f:\mathbb{R}^n \to \mathbb{R}$ is called differentiable if it can be approximated by a linear function. Specifically, at a point a, define

$$g(x)=f(a)+\sum_{i=1}^{n}(x_i-a_i)\frac{\partial f}{\partial x_1}(x)$$

and require that

$$\frac{|g(x)-f(x)|}{\|x-a\|}$$

converge to 0, as $\|x-a\|$ does.

A.6 Topology

Topology is the study of the abstract notion of convergence. We only need the standard topology here, and the definitions of convergence are given with respect to this topology, as are the definitions that follow. However, it is worthwhile to recall that there can be other topologies and, correspondingly, other notions of convergence.

A set $A \subset \mathbb{R}^n$ is *open* if for every $x \in A$, there exists $\varepsilon > 0$ such that

$$\|x - y\| < \varepsilon \Rightarrow y \in A.$$

That is, around every point in the set A we can draw a small ball, perhaps very small but with a positive radius ε (the more general concept is an open neighborhood) such that the ball will be fully contained in A.

The set

$$(0,1) = \{x \in \mathbb{R} \,|\, 0 < x < 1\}$$

is open (the open interval). Similarly, for $n = 2$, the following sets are open:

$$\{(x,y) \in \mathbb{R}^2 \,|\, x^2 + y^2 < 1\}$$

$$\{(x,y) \in \mathbb{R}^2 \,|\, 3x + 4y < 17\}$$

$$\mathbb{R}^2$$

A set $A \subset \mathbb{R}^n$ is *closed* if for every convergent sequence of points in it, $(x_1, x_2, \ldots)$ with $x_n \in A$ and $x_n \to_{n\to\infty} x^*$, the limit point is also in the set, that is, $x^* \in A$.

The set $[0,1] = \{x \in \mathbb{R} \,|\, 0 \le x \le 1\}$ is closed in $\mathbb{R}$. The following subsets of $\mathbb{R}^2$ are closed (in $\mathbb{R}^2$):

$$\{(x,y) \in \mathbb{R}^2 \,|\, x^2 + y^2 \le 1\}$$

$$\{(x,y) \in \mathbb{R}^2 \,|\, 3x + 4y \le 17\}$$

$$\mathbb{R}^2$$

The set

$$[0,1) = \{x \in \mathbb{R} \,|\, 0 \le x < 1\}$$

is neither open nor closed. It is not open because $0 \in [0,1)$, but no open neighborhood of 0 is (fully) contained in A. It is not closed because the

sequence $x_n = 1 - 1/n$ is a convergent sequence of points in A, whose limit (1) is not in A.

In $\mathbb{R}^n$, the only two sets that are both open and closed are the entire space ($\mathbb{R}^n$ itself) and the empty set. This is true in any space that we call connected.

A.7 Probability

A.7.1 Basic Concepts

Intuitively, an event is a fact that may or may not happen, a proposition that may be true or false. The probability model has a set of states of the world, or possible scenarios, often denoted by Ω or by S. Each state $s \in S$ is assumed to describe all the relevant uncertainty. An *event* is then defined as a subset of states, that is, as a subset $A \subset S$. When S is infinite, we may not wish to discuss all subsets of S. But when S is finite, there is no loss of generality in assuming that every subset is an event that can be referred to.

The set-theoretic operations of complement, union, and intersection correspond to the logical operations of negation, disjunction, and conjunction. For example, if we roll a die and

$$S = \{1, \ldots, 6\},$$

we can think of the events

$$A = \text{“The die comes up on an even number”} = \{2, 4, 6\}$$

$$B = \text{“The die comes up on a number smaller than 4”} = \{1, 2, 3\}$$

and then $A^c = \{1, 3, 5\}$ designates the event “the die comes up on an odd number,” that is, the negation of the proposition that defines A, and $B^c = \{4, 5, 6\}$ is the event described by “the die comes up on a number that is not smaller than 4.” Similarly, $A \cup B = \{1, 2, 3, 4, 6\}$ stands for “the die comes up on a number that is smaller than 4, or even, or both,” and $A \cap B = \{2\}$ is defined by “the die comes up on a number that is both even and smaller than 4.”

Probability is an assignment of numbers to events, which is supposed to measure their plausibility. The formal definition is simpler when S is finite, and we can refer to all subsets of S. That is, the set of events is

$$2^S = \{A \,|\, A \subset S\}.$$

A probability is a function

$$P : 2^S \to \mathbb{R}$$

that satisfies three properties:

1. $P(A) \geq 0$ for every $A \subset S$;
2. Whenever $A, B \subset S$ are disjoint (i.e., $A \cap B = \varnothing$),

$$P(A \cup B) = P(A) + P(B);$$

3. $P(S) = 1$.

The logic behind these conditions is derived from two analogies. First, we can think of a probability of an event as its relative frequency. Relative frequencies are non-negative (property 1), and they are added up when we discuss two disjoint events (property 2). The relative frequency of S, the event that always occurs, is 1 (property 3).

The second analogy, which is particularly useful when an event is not repeated in the same way and relative frequencies cannot be defined, is the general notion of a measure. When we measure the mass of objects or the length of line segments or the volume of bodies, we use numerical functions on subsets (of matter, of space) that satisfy the first two properties. For example, the masses of objects are never negative, and they add up when we take together two objects that had nothing in common. The last property is a matter of normalization, or a choice of the unit of measurement so that the sure event will always have the probability 1.

It is easy to verify that a function P satisfies the *additivity* condition (property 2) if and only if it satisfies, for every $A, B \subset S$,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

These three properties imply that $P(\varnothing) = 0$, so that the impossible event has probability 0.

When S is finite, say, $S = \{1, \ldots, n\}$, we say that $p = (p_1, \ldots, p_n)$ is a probability vector on S if

$$p_i \geq 0, \quad \forall i \leq n,$$

and

$$\sum_{i=1}^{n} p_i = 1.$$

For every probability $P : 2^S \to [0,1]$, there exists a probability vector p such that

$$P(A) = \sum_{i \in A} p_i, \quad \forall A \subset S,$$

and vice versa, every probability vector p defines a probability P by this equation. Thus, the probabilities on all events are in a one-to-one correspondence with the probability vectors on S.

A.7.2 Random Variables

Consider a probability model with a state space S and a probability on it P, or equivalently, a probability vector p on S. In this model a random variable is defined to be a function on S. For example, if X is a random variable that assumes real numbers as values, we can write it as

$$X : S \to \mathbb{R}.$$

The point of this definition is that a state $s \in S$ contains enough information to know anything of importance. If the focus is on a variable X, each state should specify the value that X assumes. Thus, $X(s)$ is a well-defined value, about which there is no uncertainty. Any previous uncertainty is incorporated into the uncertainty about which state s obtains. But given such a state s, no uncertainty remains.

Observe that we can use a random variable X to define events. For instance, "X equals a" is the name of the event

$$\{s \in S \mid X(s) = a\},$$

and "X is no more than a" is

$$\{s \in S \mid X(s) \leq a\},$$

and so forth.

Often, we are interested only in the probability that a random variable will assume certain values, not at which states it does so. If X takes values in some set $\mathcal{X}$, we can then define the *distribution* of a random variable X, as a function $f_X : \mathcal{X} \to [0,1]$ by

$$f_X(x) = P(X = x) = P(\{s \in S \mid X(s) = x\}).$$

For real-valued random variables, there are several additional useful definitions. The cumulative distribution of X, $F_X : \mathbb{R} \to [0,1]$ is

$$F_X(x) = P(X \leq x) = P(\{s \in S \mid X(s) \leq x\}).$$

It is thus a nondecreasing function of x going from 0 (when x is below the minimal value of X) to 1 (when x is greater than or equal to the maximal value of X). This definition can also be used when the state space is infinite and X may assume infinitely many real values.

Trying to summarize the information about a random variable X, there are several central measures. The most widely used is the *expectation*, or the *mean*, which is simply a weighted average of all values of X, where the probabilities serve as weights:

$$EX = \sum_x f_X(x)x$$

and (in a finite state space with generic element i),

$$EX = \sum_{i=1}^{n} p_i X(i).$$

The most common measure of dispersion around the mean is the *variance*, defined by

$$\text{var}(X) = E[(X - EX)^2].$$

It can be verified that

$$\text{var}(X) = E[X^2] - [EX]^2.$$

Since the variance is defined as the expectation of squared deviations from the expectation, its unit of measurement is not intuitive (it is the square of the unit of measurement of X). Therefore, we can often use the *standard deviation*, defined by

$$\sigma_X = \sqrt{\text{var}(X)}.$$

Expectation behaves in a linear way. If X, Y are real-valued random variables, and $\alpha, \beta \in \mathbb{R}$, then

$$E[\alpha X + \beta Y] = \alpha EX + \beta EY.$$

For the variance of sums (or of linear functions in general), we need to take into account the relation between X and Y. The *covariance* of X and Y is defined as

$$\text{cov}(X, Y) = E[(X - EX)(Y - EY)].$$

Intuitively, the covariance tries to measure whether X and Y go up and down together, or whether they tend to go up and down in different directions. If they do go up and down together, whenever X is relatively high (above its mean, EX), Y will be relatively high (above its mean, EY), and $(X - EX)(Y - EY)$ will be positive. And whenever X is below its mean, Y will be also be below its mean, resulting in a positive product $(X - EX)(Y - EY)$. By contrast, if Y tends to be relatively high (above EY) when X is relatively low (below EX), and vice versa, there will be more negative values of $(X - EX)(Y - EY)$. The covariance is an attempt to summarize the values of this variable. If $\text{cov}(X, Y) > 0$, then X and Y are *positively correlated*; if $\text{cov}(X, Y) < 0$, X and Y are *negatively correlated*; and if $\text{cov}(X, Y) = 0$, X and Y are *uncorrelated.*

Equipped with the covariance, we can provide a formula for the variance of a linear combination of random variables:

$$\text{var}[\alpha X + \beta Y] = \alpha^2 \text{ var}(X) + 2\alpha\beta \text{ cov}(X, Y) + \beta^2 \text{ var}(Y).$$

The formulas for expectation and variance also extend to more than two random variables:

$$E\left(\sum_{i=1}^{n} \alpha_i X_i\right) = \sum_{i=1}^{n} \alpha_i EX_i$$

and

$$\text{var}\left(\sum_{i=1}^{n} \alpha_i X_i\right) = \sum_{i=1}^{n} \alpha_i^2 \text{ var}(X_i) + 2\sum_{i=1}^{n} \sum_{j \neq i} \alpha_i \alpha_j \text{ cov}(X_i, X_i).$$

A.7.3 Conditional Probabilities

The unconditional probability of an event A, $P(A)$, is a measure of the plausibility of A occurring a priori, when nothing is known. The conditional probability of A given B, $P(A|B)$, is a measure of the likelihood of A occurring once we know that B has already occurred.

Bayes suggested that this conditional probability be the ratio of the probability of the intersection of the two events to the probability of the event that is known to have occurred. That is, he defined the conditional probability of A given B to be

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

(This definition only applies if $P(B) > 0$.)

The logic of this definition is as follows. Assume that event B has occurred. What do we think about A? For A to occur now, the two events, A and B, have to occur simultaneously. That is, we need their intersection, $A \cap B$, to occur. The probability of this happening was estimated (a priori) to be the numerator $P(A \cap B)$. However, if we just take this expression, probabilities will not sum up to 1. Indeed, the sure event will not have a probability higher than $P(B)$. We have a convention that probabilities sum up to 1. It is a convenient normalization because when we say that an event has probability of, say, .45, we don't have to ask .45 out of how much. We know that the total has been normalized to 1. To stick to this convention, we divide the measure of likelihood of A in the presence of B, $P(A \cap B)$, by the maximal value of this expression (over all A's), which is $P(B)$, and this results in Bayes' formula.

Observe that this formula makes sense in extreme cases. If A is implied by B, that is, if B is a subset of A (whenever B occurs, so does A), then $A \cap B = B$, and we have $P(A \cap B) = P(B)$ and $P(A|B) = 1$; that is, given that B has occurred, A is a certain event. At the other extreme, if A and B are logically incompatible, then their intersection is the empty set, $A \cap B = \varnothing$ and there is no scenario in which both materialize. Then $P(A \cap B) = P(\varnothing) = 0$ and $P(A|B) = 0$; that is, if A and B are incompatible, then the conditional probability of A given B is zero.

If two events are independent, the occurrence of one says nothing about the occurrence of the other. In this case the conditional probability of A given B should be the same as the unconditional probability of A. Indeed, one definition of independence is

$$P(A \cap B) = P(A)P(B),$$

which implies

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A).$$

Rearranging the terms in the definition of conditional probability, for any two events A and B (independent or not),

$$\begin{aligned} P(A \cap B) &= P(B)P(A|B) \\ &= P(A)P(B|A), \end{aligned}$$

that is, the probability of the intersection of two events (the probability of both occurring) can be computed by taking the unconditional

probability of one of them and multiplying it by the conditional probability of the second given the first. Clearly, if the events are independent, and

$$P(A|B) = P(A),$$

$$P(B|A) = P(B),$$

the two equations boil down to

$$P(A \cap B) = P(A)P(B).$$

Note that the formula $P(A \cap B) = P(B)P(A|B)$ applies also if independence does not hold. For example, the probability that a candidate wins a presidency twice in a row is the probability that she wins the first time, multiplied by the conditional probability that she wins the second time given that she has already won the first time.

Let there be two events A and B such that $P(B), P(B^c) > 0$. We note that

$$A = (A \cap B) \cup (A \cap B^c)$$

and

$$(A \cap B) \cap (A \cap B^c) = \varnothing.$$

Hence

$$P(A) = P(A \cap B) + P(A \cap B^c)$$

and, combining the equalities,

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c).$$

Thus, the overall probability of A can be computed as a weighted average, with weights $P(B)$ and $P(B^c) = 1 - P(B)$, of the conditional probability of A given B and the conditional probability of A given B^c.

A.7.4 Independence and i.i.d. Random Variables

Using the concept of independent events, we can also define independence of random variables. Let us start with two random variables X, Y that are defined on the same probability space. For simplicity of notation, assume that they are real-valued:

$$X, Y : S \to \mathbb{R}.$$

Then, given a probability P on S, we can define the joint distribution of X and Y to be the function $f_{X,Y} : \mathbb{R}^2 \to [0,1]$ defined by

$$\begin{aligned} f_{X,Y}(x,y) &= P(X = x, Y = y) \\ &= P(\{s \in S \mid X(s) = x, Y(s) = y\}). \end{aligned}$$

We say that X and Y are *independent random variables* if, for every x, y,

$$f_{X,Y}(x,y) = f_X(x) f_Y(y).$$

In other words, every event that is defined in terms of X has to be independent of any event that is defined in terms of Y. Intuitively, anything we know about X does not change our belief about (the conditional distribution of) Y.

If the state space is not finite, similar definitions apply to cumulative distributions. We can then define independence by the condition

$$\begin{aligned} F_{X,Y}(x,y) &= P(X \le x, Y \le y) \\ &= P(X \le x) P(Y \le y) \\ &= F_X(x) F_Y(y). \end{aligned}$$

All these definitions extend to any finite number of random variables. Thus, if $X_1, \ldots, X_n$ are random variables, their joint distribution and their joint cumulative distributions are, respectively,

$$f_{X_1,\ldots,X_n} : \mathbb{R}^n \to [0,1]$$

and

$$F_{X_1,\ldots,X_n} : \mathbb{R}^n \to [0,1]$$

defined by

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = P(X_1 = x_1, \ldots, X_n = x_n)$$

and

$$F_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = P(X_1 \le x_1, \ldots, X_n \le x_n).$$

Independence of n random variables is similarly defined, by the product rule

$$F_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \prod_{i=1}^{n} P(X_i \le x_i),$$

and it means that nothing that may be learned about any subset of the variables will change the conditional distribution of the remaining ones. If n random variables are independent, then so are any pair of them. The converse, however, is not true. There may be random variables that are pairwise independent but that are not independent as a set. For example, consider $n = 3$, and let X_1 and X_2 have the joint distribution

	0	1
0	0.25	0.25
1	0.25	0.25

with $X_3 = 1$ if $X_1 = X_2$, and $X_3 = 0$ if $X_1 \neq X_2$. Any pair of (X_1, X_2, X_3) are independent, but together the three random variables are not independent. In fact, any two of them fully determine the third.

Two random variables X and Y are *identically distributed* if they have the same distribution, that is, if

$$f_X(a) = f_Y(a)$$

for any value a. Two random variables X and Y are *identical* if they always assume the same value. That is, if, for every state $s \in S$,

$$X(s) = Y(s).$$

Clearly, if X and Y are identical, they are also identically distributed. This is so because, for every a,

$$f_X(a) = P(X = a) = P(Y = a) = f_Y(a),$$

where $P(X = a) = P(Y = a)$ follows from the fact that $X = a$ and $Y = a$ define precisely the same event. That is, since $X(s) = Y(s)$, any $s \in S$ belongs to the event $X = a$ if and only if it belongs to the event $Y = a$.

By contrast, two random variables that are not identical can still be identically distributed. For example, if X can assume the values $\{0, 1\}$ with equal probabilities, and $Y = 1 - X$ (that is, for every s, $Y(s) = 1 - X(s)$), then X and Y are identically distributed, but they are not identical. In fact, they never assume the same value.

The notion of identical distribution is similarly defined for more than two variables. That is, $X_1, \ldots, X_n$ are identically distributed if, for every a,

$$f_{X_1}(a) = P(X_1 = a) = \cdots = P(X_n = a) = f_{X_n}(a).$$

The variables $X_1, \ldots, X_n$ are said to be i.i.d. (identically and independently distributed) if they are identically distributed and independent.

A.7.5 Law(s) of Large Numbers

Consider a sequence of i.i.d. random variables $X_1, \ldots, X_n, \ldots$. Since they all have the same distribution, they all have the same expectation,

$$EX_i = \mu,$$

and the same variance. Assume that this variance is finite

$$\mathrm{var}(X_i) = \sigma^2.$$

When two random variables are independent, they are also uncorrelated (that is, their covariance is zero). Hence the variance of their sum is the sum of their variances.

When we consider the average of the first n random variables,

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

we observe that

$$E(\overline{X}_n) = \frac{1}{n}\sum_{i=1}^{n} EX_i = \mu,$$

and since any two of them are uncorrelated,

$$\mathrm{var}(\overline{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}(X_i) = \frac{\sigma^2}{n},$$

which implies that the more variables we take in the average, the lower will be the variance of the average. This, in turn, means that the average, $\overline{X}_n$, will be, with very high probability, close to its expectation, which is μ.

In fact, more can be said. We may decide how close we want $\overline{X}_n$ to be to μ, and with what probability, and then we can find a large enough N such that, for all n starting from N, $\overline{X}_n$ will be as close to μ as we wish with the probability we specify. Formally, for every $\varepsilon > 0$ and every $\delta > 0$, there exists N such that

$$P(\{s \,|\, |\overline{X}_n - \mu| < \delta \ \forall n \geq N\}) > 1 - \varepsilon.$$

It is also the case that the probability of the event that $\overline{X}_n$ converges to μ is 1:

$$P\left(\left\{s \,\middle|\, \exists \lim_{n\to\infty} \overline{X}_n = \mu\right\}\right) = 1.$$

LLN and Relative Frequencies Suppose a certain trial or experiment is repeated infinitely many times. In each repetition, the event A may or may not occur. The different repetitions/trials/experiments are assumed to be identical in terms of the probability of A occurring in each, and independent. Then we can associate, with experiment i, a random variable

$$X_i = \begin{cases} 1 & A \text{ occurred in experiment } i \\ 0 & A \text{ did not occur in experiment } i \end{cases}$$

The random variables $(X_i)_i$ are independently and identically distributed (i.i.d.) with $E(X_i) = p$, where p is the probability of A occurring in each of the experiments. The relative frequency of A in the first n experiments is the average of these random variables,

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{\#\{i | A \text{ occurred in experiment } i\}}{n}.$$

Hence, the law of large numbers guarantees that the relative frequency of A will converge to its probability p.

Rational Choice

Itzhak Gilboa

The MIT Press
Cambridge, Massachusetts
London, England

This book was set in Palatino on 3B2 by Asco Typesetters, Hong Kong.
Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Gilboa, Itzhak.
Rational choice / Itzhak Gilboa.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-262-01400-7 (hardcover : alk. paper)
1. Rational choice theory. 2. Social choice. 3. Decision making. 4. Game theory.
5. Microeconomics. I. Title.
HM495.G55 2010
302′.13—dc22 2009037119

10 9 8 7 6 5 4 3 2 1

B Formal Models

B.1 Utility Maximization

B.1.1 Definitions

Suppose there is a set of alternatives X. A binary relation $\succsim$ on X is simply a set of ordered pairs of elements from X, that is, $\succsim \subset X \times X$, with the interpretation that for any two alternatives $x, y \in X$,

$$(x, y) \in \succsim,$$

also denoted

$$x \succsim y,$$

means "alternative x is at least as good as alternative y in the eyes of the decision maker" or "given the choice between x and y, the decision maker may choose x."

It is useful to define two binary relations associated with $\succsim$, which are often called the symmetric and the asymmetric parts of $\succsim$. Specifically, let us introduce the following definitions. First, we define the inverse of the relation $\succsim$:

$$\precsim = \succsim^{-1} = \{(y, x) \mid (x, y) \in \succsim\},$$

that is, $y \precsim x$ if and only if $x \succsim y$ (for any x, y). The symbol $\precsim$ was selected to make $y \precsim x$ and $x \succsim y$ similar, but it is a new symbol and requires a new definition.

Do not confound the relation $\succsim$ between alternatives and the relation $\geq$ between their utility values. Later, when we have a representation of $\succsim$ by a utility function, we will be able to do precisely that—to think of

$$x \succsim y$$

as equivalent to

$$u(x) \geq u(y),$$

but this equivalence is the representation we seek, and until we prove that such a function u exists, we should be careful not to confuse $\succsim$ with $\geq$.[1]

Next define the symmetric part of $\succsim$ to be the relation $\sim \subset X \times X$ defined by

$$\sim = \succsim \cap \precsim,$$

that is, for every two alternatives $x, y \in X$,

$$x \sim y \Leftrightarrow [(x \succsim y) \quad \text{and} \quad (y \succsim x)].$$

Intuitively, $x \succsim y$ means "alternative x is at least as good as alternative y in the eyes of the decision maker," and $x \sim y$ means "the decision maker finds alternatives x and y equivalent" or "the decision maker is indifferent between alternatives x and y."

The asymmetric part of $\succsim$ is the relation $\succ \subset X \times X$ defined by

$$\succ = \succsim \setminus \precsim,$$

that is, for every two alternatives $x, y \in X$,

$$x \succ y \Leftrightarrow [(x \succsim y) \quad \text{and not} \quad (y \succsim x)].$$

Intuitively, $x \succ y$ means "the decision maker finds alternative x strictly better than alternative y."

B.1.2 Axioms

The main axioms that we impose on $\succsim$ are as follows.

Completeness For every $x, y \in X$, $x \succsim y$ or $y \succsim x$.

(Recall that *or* in mathematical language is inclusive unless otherwise stated. That is, "A or B" should be read as "A or B or possibly both").

The completeness axiom states that the decision maker can make up her mind between any two alternatives. This means that at each and every possible instance of choice between x and y something will be chosen. But it also means, implicitly, that we expect some regularity in

1. Observe that the sequence of symbols $x \geq y$ need not make sense at all because the elements of X need not be numbers or vectors or any other mathematical entities.

these choices: x is always chosen (and then we would say that $x \succ y$), or y is always chosen ($y \succ x$), or sometimes x is chosen and sometimes y. But this latter case would be modeled as equivalence ($x \sim y$), and the implicit assumption is that the choice between x and y would be completely random. If, for instance, the decision maker chooses x on even dates and y on odd dates, it would seem inappropriate to say that she is indifferent between the two options. In fact, we may find that the language is too restricted to represent the decision maker's preferences. The decision maker may seek variety and always choose the option that has not been chosen on the previous day. In this case, one would like to say that preferences are history- or context-dependent and that it is, in fact, a modeling error to consider preferences over x and y themselves (rather than, say, on sequences of x's and y's). More generally, when we accept the completeness axiom we do not assume only that at each given instance of choice one of the alternatives will end up being chosen. We also assume that it is meaningful to define preferences over the alternatives, and that these alternatives are informative enough to tell us anything that might be relevant for the decision under discussion.

Transitivity For every $x, y, z \in X$, if ($x \succsim y$ and $y \succsim z$), then $y \succsim z$.

Transitivity has a rather obvious meaning, and it almost seems like part of the definition of preferences. Yet, it is easy to imagine cyclical preferences. Moreover, such preferences may well occur in group decision making, for instance, if the group is using a majority vote. This is the famous Condorcet paradox (see section 6.2 of the main text). Assume that there are three alternatives, $X = \{x, y, z\}$ and that one-third of society prefers

$$x \succ y \succ z,$$

one-third

$$y \succ z \succ x,$$

and the last third

$$z \succ x \succ y.$$

It is easy to see that when every two pairs of alternatives come up for a separate majority vote, there is a two-thirds majority for $x \succ y$, a two-thirds majority for $y \succ z$, but also a two-thirds majority for $z \succ x$.

In other words, a majority vote may violate transitivity and even generate a cycle of strict preferences: $x \succ y \succ z \succ x$.

Once we realize that this can happen in a majority vote in a society, we can imagine how this can happen inside the mind of a single individual as well. Suppose Daniel has to choose among three cars, and he ranks them according to three criteria, such as comfort, speed, and price. He finds it hard to quantify and trade off these criteria, so he decides to adopt the simple rule that if one alternative is better than another according to most criteria, then it should be preferred. In this case Daniel can be thought of as if he were the aggregation of three decision makers—one who cares only about comfort, one who cares only about speed, and one who cares only about price—where his decision rule as a "society" is to follow a majority vote. Then Daniel would find that his preferences are not transitive. But if this happens, we expect him to be confused about the choice and to dislike the situation of indecision. Thus, even if the transitivity axiom does not always hold, it is generally accepted as a desirable goal.

B.1.3 Result

We are interested in a representation of a binary relation by a numerical function. Let us first define this concept more precisely.

A function $u : X \to \mathbb{R}$ is said to *represent* $\succsim$ if, for every $x, y \in X$,

$$x \succsim y \Leftrightarrow u(x) \geq u(y). \tag{B.1}$$

Proposition 1 Let X be finite. Let $\succsim$ be a binary relation on X, i.e., $\succsim \subset X \times X$. The following are equivalent: (i) $\succsim$ is complete and transitive; (ii) there exists a function $u : X \to \mathbb{R}$ that represents $\succsim$.

B.1.4 Generalization to a Continuous Space

In many situations of interest, the set of alternatives is not finite. If we consider a consumer who has preferences over the amount of wine she consumes, the amount of time she spends in the pool, or the amount of money left in her bank account, we are dealing with variables that are continuous and that therefore may assume infinitely many values. Thus, the set X, which may be a set of vectors of such variables, is infinite.

Physicists might say that the amount of wine can only take finitely many values because there are a finite number of particles in a glass of wine (and perhaps also in the world). This is certainly true of the amount of money—it is only measured up to cents. And the accuracy

of measurement is also limited in the case of time, temperature, and so forth. So maybe the world is finite after all, and we don't need to deal with extension of proposition 1?

The fact is that finite models may be very awkward and inconvenient. For example, assume there are supply and demand curves that slope in the right directions[2] but fail to intersect because they are only defined for finitely many prices. (In fact, they are not really curves but only finite collections of points in $\mathbb{R}^2$.) It would be silly to conclude that the market will never be at equilibrium simply because there is no precise price at which supply and demand are equal. You might recall a similar discussion in statistics. The very first time you were introduced to continuous random variables, you might have wondered who really needs them in a finite world. But then you find out that many assumptions and conclusions are greatly simplified by the assumption of continuity.

In short, we would like to have a similar theorem, guaranteeing utility representation of a binary relation, also in the case that the set of alternatives is infinite. There are several ways to obtain such a theorem. The one presented here also guarantees that the utility function be continuous. To make this a meaningful statement, we have to have a notion of convergence in the set X, a topology. But in order to avoid complications, let us simply assume that X is a subset of $\mathbb{R}^n$ for some $n \geq 1$ and think of convergence as it is usually defined in $\mathbb{R}^n$.

It is not always the case that a complete and transitive relation on $\mathbb{R}^n$ can be represented by a numerical function. (A famous counterexample was provided by Gerard Debreu.[3]) An additional condition that we may impose is that the relation $\succsim$ be continuous. What is meant by this is that if $x \succ y$, then all the points that are very close to x are also strictly better than y, and vice versa, all the points that are very close to y are also strictly worse than x.

Continuity For every $x \in X$, the sets $\{y \in X \mid x \succ y\}$ and $\{y \in X \mid y \succ x\}$ are open in X.

(Recall that a set is open if, for every point in it, there is a whole neighborhood contained in the set.)

2. The supply curve, which indicates the quantity supplied as a function of price, is increasing. The demand curve, which specifies the quantity demanded as a function of price, is decreasing.

3. G. Debreu, *The Theory of Value: An Axiomatic Analysis of Economic Equilibrium* (New Haven: Yale University Press, 1959), ch. 2, prob. 6.

To see why this axiom captures the notion of continuity, we may think of a function $f : \mathbb{R} \to \mathbb{R}$ and a point $x \in \mathbb{R}$ for which $f(x) > 0$. If f is continuous, then there is a neighborhood of x for which f is positive. If we replace "positive" by "strictly better than y" for a fixed y, we see the similarity between these two notions of continuity.

Alternatively, we can think of continuity as requiring that for every $x \in X$, the sets

$$\{y \in X \,|\, x \succsim y\}$$

and

$$\{y \in X \,|\, y \succsim x\}$$

be closed in X. That is, if we consider a convergent sequence $(y_n)_{n\geq 1}$, $y_n \to y$, such that $y_n \succsim x$ for all n, then also $y \succsim x$, and if $x \succsim y_n$ for all n, then also $x \succsim y$. In other words, if we have a weak preference all along the sequence (either from below or from above), we should have the same weak preference at the limit. This condition is what we were after.

Theorem 2 (Debreu) Let $\succsim$ be a binary relation on X, that is, $\succsim \subset X \times X$. The following are equivalent: (i) $\succsim$ is complete, transitive, and continuous; (ii) there exists a continuous function $u : X \to \mathbb{R}$ that represents $\succsim$.

B.2 Convexity

As a preparation for the discussion of constrained optimization, it is useful to have some definitions of convex sets, convex and concave functions, and so on.

B.2.1 Convex Sets

A set $A \subset \mathbb{R}^n$ is *convex* if, for every $x, y \in A$ and every $\lambda \in [0, 1]$, $\lambda x + (1 - \lambda)y \in A$. That is, whenever two points are in the set, the line segment connecting them is also in the set. If we imagine the set A as a room, convexity means that any two people in the room can see each other.

B.2.2 Convex and Concave Functions

A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if its graph is never above the strings that connect points on it. As an example, we may think of

$$f(x) = x^2$$

for $n = 1$. If we draw the graph of this function and take any two points on the graph, when we connect them by a segment (the string), the graph of the function will be below it (or at least not above the segment). The same will be true for

$$f(x_1, x_2) = x_1^2 + x_2^2$$

if $n = 2$.

Formally, $f : \mathbb{R}^n \to \mathbb{R}$ is *convex* if for every $x, y \in A$ and every $\lambda \in [0, 1]$,

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y),$$

and it is strictly convex if this inequality is strict whenever $x \neq y$ and $0 < \lambda < 1$.

To see the geometric interpretation of this condition, imagine that $n = 1$, and observe that $\lambda x + (1 - \lambda)y$ is a point on the interval connecting x and y. Similarly, $\lambda f(x) + (1 - \lambda)f(y)$ is a point on the interval connecting $f(x)$ and $f(y)$. Moreover, if we connect the two points

$$(x, f(x)), (y, f(y)) \in \mathbb{R}^2$$

by a segment (which is a string of the function f), we get precisely the points

$$\{(\lambda x + (1 - \lambda)y, \lambda f(x) + (1 - \lambda)f(y)) \mid \lambda \in [0, 1]\}.$$

For $\lambda = 1$ the point is $(x, f(x))$; for $\lambda = 0$ it is $(y, f(y))$; for $\lambda = 0.5$, the point has a first coordinate that is the arithmetic average of x and y, and a second coordinate that is the average of their f values. Generally, for every λ, the first coordinate is the $(\lambda, (1 - \lambda))$ average between x and y, and the second coordinate is corresponding average of their f values.

Convexity of the function demands that for every $\lambda \in [0, 1]$, the value of the function at the $(\lambda, (1 - \lambda))$ average between x and y, that is, $f(\lambda x + (1 - \lambda)y)$, will not exceed the height of the string (connecting $(x, f(x))$ and $(y, f(y))$) at the same point.

Next assume that $n = 2$ and repeat the argument to show that this geometric interpretation is valid in general.

A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if the following set is convex[4]

$$\{(x, z) \in \mathbb{R}^{n+1} \mid z \geq f(x)\}.$$

If $n = 1$ and f is twice differentiable, convexity of f is equivalent to the condition that $f'' \geq 0$, that is, that the first derivative, f', is nondecreasing. When $n > 1$, there are similar conditions, expressed in terms of the matrix of second derivatives, which are equivalent to convexity.

Concave functions are defined in the same way, with the converse inequality. All that is true of convex functions is true of concave functions, with the opposite inequality. In fact, we could define f to be concave if $-f$ is convex. But we will spell it out.

A function $f : \mathbb{R}^n \to \mathbb{R}$ is *concave* if for every $x, y \in A$ and every $\lambda \in [0, 1]$,

$$f(\lambda x + (1 - \lambda)y) \geq \lambda f(x) + (1 - \lambda) f(y),$$

and it is strictly concave if this inequality is strict whenever $x \neq y$ and $0 < \lambda < 1$.

Thus, f is concave if the graph of the function is never below the strings that connect points on it. Equivalently, $f : \mathbb{R}^n \to \mathbb{R}$ is concave if and only if the following set is convex

$$\{(x, z) \in \mathbb{R}^{n+1} \mid z \leq f(x)\}.$$

(This set is still required to be convex, not concave. In fact, we didn't define the notion of a concave set, and we don't have such a useful definition. The difference between this condition for convex and concave functions is in the direction of the inequality. The resulting set in both cases is required to be a convex set as a subset of $\mathbb{R}^{n+1}$.)

If $n = 1$ and f is twice differentiable, concavity of f is equivalent to the condition that $f'' \leq 0$, that is, that the first derivative, f', is nonincreasing.

An affine function is a shifted linear function. That is, $f : \mathbb{R}^n \to \mathbb{R}$ is *affine* if

4. Observe that the vector (x, z) refers to the concatenation of x, which is a vector of n real numbers, with z, which is another real number—together a vector of $(n + 1)$ real numbers.

$$f(x) = \sum_{i=1}^{n} a_i x_i + c,$$

where $\{a_i\}$ and c are real numbers.

An affine function is both convex and concave (but not strictly so). The converse is also true: a function that is both convex and concave is affine.

If we take a convex function f, we can, at each x, look at the tangent to the graph of f. This would be a line if $n = 1$ and a hyperplane more generally. Formally, for every x there exists an affine function $l_x : \mathbb{R}^n \to \mathbb{R}$ such that

$$l_x(x) = f(x),$$

and for every $y \in \mathbb{R}^n$

$$l_x(y) \leq f(y).$$

If we take all these functions $\{l_x\}_x$, we find that their maximum is f. That is, for every $y \in \mathbb{R}^n$

$$f(y) = \max_x \; l_x(y).$$

Thus, a convex function can be described as the maximum of a collection of affine functions. Conversely, the maximum of affine functions is always convex. Hence, a function is convex if and only if it is the maximum of affine functions.

Similarly, a function is concave if and only if it is the minimum of a collection of affine functions.

B.2.3 Quasi-convex and Quasi-concave Functions

Consider the convex function

$$f(x_1, x_2) = x_1^2 + x_2^2.$$

Suppose that I cut it at a given height, z, and ask which points (x_1, x_2) do not exceed z in their f value. That is, I look at

$$\{x \in \mathbb{R}^2 \mid f(x) \leq z\}.$$

It is easy to see that this set will be convex. This gives rise to the following definition.

A function $f : \mathbb{R}^n \to \mathbb{R}$ is *quasi-convex* if, for every $z \in \mathbb{R}$,

$$\{x \in \mathbb{R}^n \mid f(x) \leq z\}$$

is a convex set.

Observe that this set is a subset of $\mathbb{R}^n$ and that we have a (potentially) different such set for every value of z, whereas in the characterization of convex functions given previously we used the convexity of a single set in $\mathbb{R}^{n+1}$.

The term *quasi* should suggest that every convex function is also quasi-convex. Indeed, if

$$y, w \in \{x \in \mathbb{R}^n \mid f(x) \leq z\},$$

then

$$f(y), f(w) \leq z,$$

and for every $\lambda \in [0, 1]$,

$$\begin{aligned} f(\lambda y + (1-\lambda)w) &\leq \lambda f(y) + (1-\lambda) f(w) \\ &\leq \lambda z + (1-\lambda) z = z, \end{aligned}$$

and this means that

$$\lambda y + (1-\lambda)w \in \{x \in \mathbb{R}^n \mid f(x) \leq z\}.$$

Since this is true for every y, w in $\{x \in \mathbb{R}^n \mid f(x) \leq z\}$, this set is convex for every z, and this is the definition of quasi-convexity of the function f.

Is every quasi-convex function is convex? The answer is negative, (otherwise we wouldn't use a different term for quasi-convexity). Indeed, it suffices to consider $n = 1$ and observe that

$$f(x) = x^3$$

is quasi-convex but not convex. Indeed, when we look at the sets

$$\{x \in \mathbb{R}^n \mid f(x) \leq z\}$$

for various values of $z \in \mathbb{R}$, we simply get the convex sets $(-\infty, \alpha]$ for some α (in fact, for $\alpha = z^{1/3}$). The collection of these sets, when we range over all possible values of z, does not look any different for the original function, x^3, than it would if we looked at the function x or $x^{1/3}$.

Again, everything we can say of quasi-convex functions has a counterpart for quasi-concave ones. A function $f : \mathbb{R}^n \to \mathbb{R}$ is *quasi-concave* if, for every $z \in \mathbb{R}$,

$$\{x \in \mathbb{R}^n \mid f(x) \geq z\}$$

is a convex set.

Imagine now the parabola upside down,

$$f(x_1, x_2) = -x_1^2 - x_2^2,$$

and when we cut it at a certain height, z, and look at the dome above the cut, the projection of this dome on the x_1, x_2 plane is a circle. The fact that it is a convex set follows from the fact that f is quasi-concave.

B.3 Constrained Optimization

B.3.1 Convex Problems

Constrained optimization problems are much easier to deal with when they are convex. Roughly, we want everything to be convex, both on the feasibility and on the desirability side.

Convexity of the feasible set is simple to define. We require that the set F be convex.

What is meant by "convex preferences"? The answer is that we wish the "at least as desirable as" sets to be convex. Explicitly, for every $x \in \mathbb{R}^n$, we may consider the "at least as good as" set

$$\{y \in X \mid y \succsim x\}$$

and require that it be convex. If we have a utility function u that represents $\succsim$, we require that the function be quasi-concave. Indeed, if u is quasi-concave, then for every $\alpha \in \mathbb{R}$, the set

$$\{y \in X \mid u(y) \geq \alpha\}$$

is convex. When we range over all values of α, we obtain all sets of the form $\{y \in X \mid y \succsim x\}$, and thus a quasi-concave u defines convex preferences. Observe that quasi-concavity is the appropriate term when the utility is given only up to an arbitrary (increasing) monotone transformation (utility is only ordinal). Whereas a concave function can be replaced by a monotone transformation that results in a nonconcave function, a quasi-concave function will remain quasi-concave after any increasing transformation.

Convex problems (in which both the feasible set and preferences are convex) have several nice properties. In particular, local optima are also global optima. This means that looking at first-order conditions is often sufficient. If these conditions identify a local maximum, we can rest assured that it is also a global one. Another important feature of convex problems is that for such problems one can devise simple algorithms of small local improvements that converge to the global optimum. This is very useful if we are trying to solve the problem on a computer. But, more important, it also says that real people may behave as if they were solving such problems optimally. If a decision maker makes small improvements when these exist, we may assume that, as time goes by, he converges to the optimal solution. Thus, for large and complex problems, the assumption that people maximize utility subject to their feasible set is much more plausible in convex problems than it is in general.

B.3.2 Example: The Consumer Problem

Let us look at the consumer problem again. The decision variables are $x_1, \ldots, x_n \in \mathbb{R}_+$, where x_i ($x_i \geq 0$) is the amount consumed of good i. The consumer has an income $I \geq 0$, and she faces prices $p_1, \ldots, p_n \in \mathbb{R}_{++}$ (that is, $p_i > 0$ for all $i \leq n$). The problem is therefore

$$\max_{x_1,\ldots,x_n} u(x_1, \ldots, x_n)$$

subject to

$$p_1 x_1 + \cdots + p_n x_n \leq I$$

$$x_i \geq 0$$

B.3.3 Algebraic Approach

Let us further assume that u is strictly monotone in all its arguments, namely, that the consumer prefers more of each good to less. Moreover, we want to assume that u is quasi-concave, so that the "better than" sets are convex. Under these assumptions we may conclude that the optimal solution will be on the budget constraint, namely, will satisfy

$$p_1 x_1 + \cdots + p_n x_n = I,$$

and if a point $x = (x_1, \ldots, x_n)$ is a local maximum, it is also a global one. Hence it makes sense to seek a local maximum, namely, to ask

whether a certain point on the budget constraint happens to maximize utility in a certain neighborhood (of itself on this constraint).

If the utility function is also differentiable, we may use calculus to help identify the optimal solution. Specifically, the first-order condition for this problem can be obtained by differentiating the Lagrangian

$$L(x_1, \ldots, x_n, \lambda) = u(x_1, \ldots, x_n) - \lambda[p_1 x_1 + \cdots + p_n x_n - I]$$

and equating all first (partial) derivatives to zero. This yields

$$\frac{\partial L}{\partial x_i} = \frac{\partial u}{\partial x_i} - \lambda p_i = 0$$

for all $i \leq n$ and

$$\frac{\partial L}{d\lambda} = -[p_1 x_1 + \cdots + p_n x_n - I] = 0.$$

The second equality is simply the budget constraint, whereas the first implies that for all i,

$$\frac{u_i}{p_i} = \text{const} = \lambda,$$

where $u_i = \frac{\partial u}{dx_i}$. Thus, for any two goods i, j, we have

$$\frac{u_i}{p_i} = \frac{u_j}{p_j} \tag{B.2}$$

or

$$\frac{u_i}{u_j} = \frac{p_i}{p_j}. \tag{B.3}$$

B.3.4 Geometric Approach

Each of these equivalent conditions has an intuitive interpretation. Let us start with the second, which can be understood geometrically. We argue that it means that the feasible set and the desirable ("better than") set are tangent to each other. To see this, assume there are only two goods, $i = 1, 2$. Consider the budget constraint

$$p_1 x_1 + p_2 x_2 = I,$$

and observe that its slope, at a point x, can be computed by taking differentials:

$$p_1\,dx_1 + p_2\,dx_2 = 0,$$

which means

$$\frac{dx_1}{dx_2} = -\frac{p_1}{p_2}. \tag{B.4}$$

Next consider the "better than" set, and focus on the tangent to this set at the point x. We get a line (more generally, a hyperplane) that goes through the point x, and satisfies

$$du = u_1\,dx_1 + u_2\,dx_2 = 0,$$

that is, a line with a slope

$$\frac{dx_1}{dx_2} = -\frac{u_1}{u_2}. \tag{B.5}$$

This will also be the slope of the indifference curve (the set of points that are indifferent to x) at x. Clearly, condition (B.3) means that the slope of the budget constraint, (B.4), equals the slope of the indifference curve (B.5).

Why is this a condition for optimality? We may draw several indifference curves and superimpose them on the budget constraint. In general, we can always take this geometric approach to optimization: draw the feasible set, and then compare it to the "better than" sets. If an indifference curve, which is the boundary of a "better than" set, does not intersect the feasible set, it indicates a level of utility that cannot be reached. It is in the category of wishful thinking. A rational decision maker will be expected to give up this level of utility and settle for a lower one.

If, on the other hand, the curve cuts *through* the feasible set, the corresponding level of utility is reachable, but it is not the highest such level. Since the curve is strictly in the interior of the feasible set, there are feasible points on either side of it. Assuming that preferences are monotone, that is, that the decision maker prefers more to less, one side of the curve has a higher utility level than the curve itself. Since it is feasible, the curve we started with cannot be optimal. Here a rational decision maker will be expected to strive for more and look for a higher utility level.

What is the highest utility level that is still feasible? It has to be represented by an indifference curve that is not disjoint with the feasible set yet does not cut through it. In other words, the intersection of the

feasible set and the "better than" set is nonempty but has an empty interior (it has zero volume, or zero area in a two-dimensional problem). If both sets are smooth, they have to be tangent to each other. This tangency condition is precisely what equation (B.3) yields.

B.3.5 Economic Approach

The economic approach is explained in the main text. But it may be worthwhile to repeat it in a slightly more rigorous way. Consider condition (B.2). Again, assume that I already decided on spending most of my budget, and I'm looking at the last dollar, asking whether I should spend it on good i or on good j. If I spend it on i, how much of this good will I get? At price p_i, one dollar would buy $\frac{1}{p_i}$ units of the good. How much additional utility will I get from this quantity? Assuming that one dollar is relatively small and that correspondingly the amount of good i, $\frac{1}{p_i}$, is also relatively small, I can approximate the marginal utility of $\frac{1}{p_i}$ extra units by

$$\frac{1}{p_i} \cdot \frac{\partial u}{\partial x_i} = \frac{u_i}{p_i}.$$

Obviously, the same reasoning would apply to good j. Spending the dollar on j would result in an increase in utility that is approximately $\frac{u_j}{p_j}$. Now, if

$$\frac{u_i}{p_i} > \frac{u_j}{p_j},$$

one extra dollar spent on i will yield a higher marginal utility than the same dollar spent on j. Put differently, we can take one dollar of the amount spent on j and transfer it to the amount spent on i, and be better off, since the utility lost on j, $\frac{u_j}{p_j}$, is more than compensated for by the utility gained on i, $\frac{u_i}{p_i}$.

This argument assumes that we can indeed transfer one dollar from good j to good i. That is, that we are at an interior point. If we consider a boundary point, where we don't spend any money on j in any case, this inequality may be consistent with optimality.

If one dollar is not relatively small, we can repeat this argument with ε dollars, where ε is small enough for the derivatives to provide good approximations. Then we find that ε dollars are translated to quantities $\frac{\varepsilon}{p_i}$ and $\frac{\varepsilon}{p_j}$, if spent on goods i or j, respectively, and that these quantities yield marginal utilities of $\frac{\varepsilon u_i}{p_i}$ and $\frac{\varepsilon u_j}{p_j}$, respectively. Hence, any of the inequalities

$$\frac{u_i}{p_i} > \frac{u_j}{p_j} \quad \text{or} \quad \frac{u_i}{p_i} < \frac{u_j}{p_j}$$

indicates that we are not at an optimal (interior) point. Condition (B.2) is a powerful tool in identifying optimal points. It says that small changes in the budget allocation, in other words, small changes along the boundary of the budget constraint, will not yield an improvement.

B.3.6 Comments

Two important comments are in order. First, the previous arguments are not restricted to a feasible set defined by a simple budget constraint, that is, by a linear inequality. The feasible set may be defined by one or many nonlinear constraints. What is crucial is that it be convex.

Second, condition (B.2) is necessary only if the optimal solution is at a point where the sets involved—the feasible set and the "better than" set—are smooth enough to have a unique tangent (supporting hyperplane, that is, a hyperplane defined by a linear equation that goes through the point in question, and such that the entire set is on one of its sides). An optimal solution may exist at a point where one of the sets has kinks, and in this case slopes and derivatives may not be well defined.

Still, the first-order conditions, namely, the equality of slopes (or ratios of derivatives) are sufficient for optimality in convex problems, that is, problems in which both the feasible set and the "better than" sets are convex. It is therefore a useful technique for finding optimal solutions in many problems. Moreover, it provides us with very powerful insights. In particular, the marginal way of thinking about alternatives, which we saw in the economic interpretation, appears in many problems within and outside of economics.

B.4 vNM's Theorem

B.4.1 Setup

vNM's original formulation involved decision trees in which compound lotteries were explicitly modeled. We use here a more compact formulation, due to Niels-Erik Jensen and Peter Fishburn,[5] which

5. N. E. Jensen, "An Introduction to Bernoullian Utility Theory," pts. I and II, *Swedish Journal of Economics* 69 (1967): 163–183, 229–247; P. C. Fishburn, *Utility Theory for Decision Making* (New York: Wiley, 1970).

implicitly assumes that compound lotteries are simplified according to Bayes' formula. Thus, lotteries are defined by their distributions, and the notion of mixture implicitly supposes that the decision maker is quite sophisticated in terms of his probability calculations.

Let X be a set of alternatives. X need not consist of sums of money or consumption bundles, and it may include outcomes such as death.

The objects of choice are lotteries. We can think of a lottery as a function from the set of outcomes, X, to probabilities. That is, if P is a lottery and x is an outcome, $P(x)$ is the probability of getting x if we choose lottery P. It will be convenient to think of X as potentially infinite, as is the real line, for example. At the same time, we don't need to consider lotteries that may assume infinitely many values. We therefore assume that while X is potentially infinite, each particular lottery P can only assume finitely many values.

The set of all lotteries is therefore

$$L = \left\{ P : X \to [0,1] \;\middle|\; \begin{array}{c} \#\{x|P(x) > 0\} < \infty, \\ \sum_{x \in X} P(x) = 1 \end{array} \right\}.$$

Observe that the expression $\sum_{x \in X} P(x) = 1$ is well defined thanks to the finite support condition that precedes it.

A *mixing operation* is performed on L, defined for every $P, Q \in L$ and every $\alpha \in [0,1]$ as follows: $\alpha P + (1-\alpha)Q \in L$ is given by

$$(\alpha P + (1-\alpha)Q)(x) = \alpha P(x) + (1-\alpha)Q(x)$$

for every $x \in X$. The intuition behind this operation is of conditional probabilities. Assume that I offer you a compound lottery that will give you the lottery P with probability α and the lottery Q with probability $(1-\alpha)$. You can ask what is the probability of obtaining a certain outcome x, and observe that it is indeed α times the conditional probability of x if you get P plus $(1-\alpha)$ times the conditional probability of x if you get Q.

Since the objects of choice are lotteries, the observable choices are modeled by a binary relation on L, $\succsim \subset L \times L$.

B.4.2 The vNM Axioms

The vNM axioms are

Weak order $\succsim$ is complete and transitive.

Continuity For every $P, Q, R \in L$, if $P \succ Q \succ R$, there exist $\alpha, \beta \in (0,1)$ such that $\alpha P + (1-\alpha)R \succ Q \succ \beta P + (1-\beta)R$.

Independence For every $P, Q, R \in L$, and every $\alpha \in (0,1)$, $P \succsim Q$ if and only if $\alpha P + (1-\alpha)R \succsim \alpha Q + (1-\alpha)R$.

The weak order axiom is not very different from the same assumption in chapter 2 of the main text. The other two axioms are new and deserve a short discussion.

B.4.3 Continuity

Continuity may be viewed as a technical condition needed for the mathematical representation and for the proof to work. To understand its meaning, consider the following example, supposedly challenging continuity. Assume that P guarantees one dollar, Q guarantees zero dollars, and R guarantees death. You are likely to prefer one dollar to no dollars, and no dollars to death. That is, you would probably exhibit preferences $P \succ Q \succ R$. The axiom then demands that for a high enough $\alpha < 1$, you will also exhibit the preference

$$\alpha P + (1-\alpha)R \succ Q,$$

namely, that you will be willing to risk your life with probability $(1-\alpha)$ in order to gain one dollar. The point of the example is that you are supposed to say that no matter how small the probability of death $(1-\alpha)$, you will not risk your life for one dollar.

A counterargument to this example (suggested by Howard Raiffa) is that we often do indeed take such risks. For instance, suppose you are about to buy a newspaper, which costs one dollar. But you see that it is freely distributed on the other side of the street. Would you cross the street to get it at no cost? If you answer in the affirmative, you are willing to accept a certain risk, albeit very small, of losing your life (in traffic) in order to save one dollar.

This counterargument can be challenged in several ways. For instance, you may argue that even if you don't cross the street, your life is not guaranteed with probability 1. Indeed, a truck driver who falls asleep may hit you anyway. In this case, we are not comparing death with probability 0 to death with probability $(1-\alpha)$. And, the argument goes, it is possible that if you had true certainty on your side of the street, you would not have crossed the street, thereby violating the axiom.

It appears that framing also matters in this example. I may be about to cross the street in order to get the free copy of the newspaper, but if you stop me and say, "What are you doing? Are you nuts, to risk your

life this way? Think of what could happen! Think of your family!" I might cave in and give up the free paper. It is not obvious which behavior is more relevant, namely, the decision making without the guilt-inducing speech or with it. Presumably, this depends on the application.

In any event, we understand the continuity axiom. Moreover, if we consider applications that do not involve extreme risks such as death, it appears to be a reasonable assumption.

B.4.4 Independence

The independence axiom is related to dynamic consistency. However, it involves several steps. Consider the following four choice situations:

1. You are asked to make a choice between P and Q.
2. Nature will first decide whether, with probability $(1 - \alpha)$, you get R, and then you have no choice to make. Alternatively, with probability α, nature will let you choose between P and Q.
3. The choices are as in (2), but you have to commit to making your choice before you observe Nature's move.
4. You have to choose between two branches. In one, Nature will first decide whether, with probability $(1 - \alpha)$, you get R, or, with probability α, you get P. The second branch is identical, with Q replacing P.

Clearly, (4) is the choice between $\alpha P + (1 - \alpha)R$ and $\alpha Q + (1 - \alpha)R$. To relate the choice in (1) to that in (4), we can use (2) and (3) as intermediary steps. Compare (1) and (2). In (2), *if* you are called upon to act, you are choosing between P and Q. At that point R will be a counterfactual world. Why would it be relevant? Hence, it is argued, you can ignore the possibility that did not happen, R, and make your decision in (2) identical to that in (1).

The distinction between (2) and (3) has to do only with the timing of your decision. Should you make different choices in these scenarios, you would not be dynamically consistent. It is as if you plan in (3) to make a given choice, but when you get the chance to make it, you do (or would like to do) something else in (2). Observe that when you make a choice in (3), you know that this choice is conditional on getting to the decision node. Hence, the additional information you have should not change this conditional choice.

Finally, the alleged equivalence between (3) and (4) relies on changing the order of your move (to which you already committed) and

Nature's move. As such, this is an axiom of reduction of compound lotteries, assuming that the order of the draws does not matter as long as the distributions on outcomes, induced by your choices, are the same.

B.4.5 The Theorem

Finally, the theorem can be stated.

Theorem 3 (vNM) Let there be given a relation $\succsim \subset L \times L$. The following are equivalent: (i) $\succsim$ satisfies weak order, continuity, and independence; (ii) there exists $u : X \to \mathbb{R}$ such that, for every $P, Q \in L$,

$$P \succsim Q \quad \text{iff} \quad \sum_{x \in X} P(x)u(x) \geq \sum_{x \in X} Q(x)u(x).$$

Moreover, in this case u is unique up to a positive linear transformation (plt). That is, $v : X \to \mathbb{R}$ satisfies, for every $P, Q \in L$,

$$P \succsim Q \quad \text{iff} \quad \sum_{x \in X} P(x)v(x) \geq \sum_{x \in X} Q(x)v(x)$$

if and only if there are $a > 0$ and $b \in \mathbb{R}$ such that $v(x) = au(x) + b$ for every $x \in X$.

Thus, we find that the theory of expected utility maximization is not just one arbitrary generalization of expected value maximization. There are quite compelling reasons to maximize expected utility (in a normative application) as well as to believe that this is what people naturally tend to do (in a descriptive application). If we put aside the more technical condition of continuity, we find that expected utility maximization is equivalent to following a weak order that is linear in probabilities; this linearity is basically what the independence axiom says.

B.5 Ignoring Base Probabilities

The disease example discussed in section 5.4.1 of the main text illustrates that people often mistake $P(A|B)$ for $P(B|A)$. In that example, if you had the disease, the test would show it with probability 90 percent; if you didn't, the test might still show a false positive with probability 5 percent. Suppose you took the test and you tested positive. What was the probability of your actually having the disease?

Let D be the event of having the disease and T be the event of testing positive. Then

$$P(T|D) = .90,$$

$$P(T|D^c) = .05.$$

What is $P(D|T)$?

The definition of conditional probability says that

$$P(D|T) = \frac{P(D \cap T)}{P(T)}.$$

Trying to get closer to the given data, we may split the event T into two disjoint events:

$$T = (D \cap T) \cup (D^c \cap T).$$

In other words, one may test positive if one is sick $(D \cap T)$ but also if one is healthy $(D^c \cap T)$, so

$$P(T) = P(D \cap T) + P(D^c \cap T)$$

and

$$P(D|T) = \frac{P(D \cap T)}{P(T)}$$

$$= \frac{P(D \cap T)}{P(D \cap T) + P(D^c \cap T)}.$$

Now we can try to relate each of the probabilities in the denominator to the conditional probabilities we are given. Specifically,

$$P(D \cap T) = P(D)P(T|D) = .90P(D)$$

and

$$P(D^c \cap T) = P(D^c)P(T|D^c) = .05[1 - P(D)]$$

(recalling that the probability of no disease, $P(D^c)$, and the probability of disease have to sum up to 1.) Putting it all together, we get

$$P(D|T) = \frac{P(D \cap T)}{P(T)}$$

$$= \frac{P(D \cap T)}{P(D \cap T) + P(D^c \cap T)}$$

$$= \frac{P(D)P(T|D)}{P(D)P(T|D) + P(D^c)P(T|D^c)}$$

$$= \frac{.90P(D)}{.90P(D) + .05[1 - P(D)]}.$$

This number can be anywhere in $[0, 1]$. Indeed, suppose that we are dealing with a disease that is known to be extinct. Thus, $P(D) = 0$. The accuracy of the test remains the same: $P(T|D) = .90$, and $P(T|D^c) = .05$, but we have other reasons to believe that the a priori probability of having the disease is zero. Hence, whatever the test shows, your posterior probability is still zero. If you test positive, you should attribute it to the inaccuracy of the test (the term $.05[1 - P(D)]$ in the denominator) rather than to having the disease (the term $.90P(D)$). By contrast, if you are in a hospital ward consisting only of previously diagnosed patients, and your prior probability of having the disease is $P(D) = 1$, your posterior probability will be 1 as well (and this will be the case even if you tested negative).

To see why Kahneman and Tversky called this phenomenon "ignoring base probabilities," observe that what relates the conditional probability of A given B to the conditional probability of B given A is the ratio of the unconditional (base) probabilities:

$$P(A|B) = \frac{P(A)}{P(B)}P(B|A),$$

and the confusion of $P(B|A)$ for $P(A|B)$ is tantamount to ignoring the term $P(A)/P(B)$.

B.6 Arrow's Impossibility Theorem

Let $N = \{1, 2, \ldots, n\}$ be the set of individuals, and let X be the set of alternatives. Assume that X is finite, with $|X| \geq 3$. Each individual is assumed to have a preference relation over X. For simplicity, assume that there are no indifferences, so that for each $i \in N$, there is a relation $\succsim_i \subset X \times X$ that is complete, transitive, and antisymmetric (namely, $x \succsim_i y$ and $y \succsim_i x$ imply $x = y$.) Alternatively, we may assume that for each individual $i \in N$ there is a "strictly prefer" relation $\succ_i \subset X \times X$ that is transitive and that satisfies

$$x \neq y \Leftrightarrow [x \succ_i y \text{ or } y \succ_i x].$$

(If $\succsim_i$ is complete, transitive, and antisymmetric, its asymmetric part $\succ_i$ satisfies this condition.)

The list of preference relations $(\succsim_1, \ldots, \succsim_n) = (\succsim_i)_i$ is called a *profile.* It indicates how everyone in society ranks the alternatives. Arrow's theorem does not apply to one particular profile but to a function that is assumed to define a social preference for *any* possible profile of individual preferences. Formally, let

$$R = \{\succsim \subset X \times X \mid \succsim \text{ is complete, transitive, antisymmetric}\}$$

be the set of all possible preference relations. We consider functions that take profiles, or n-tuples of elements in R into R itself. That is, the theorem will be about creatures of the type

$$f : R^n \to R.$$

Note that all profiles, that is, all n-tuples of relations (one for each individual), are considered. This can be viewed as an implicit assumption that is sometimes referred to explicitly as "full domain."

For such functions f we are interested in two axioms:

Unanimity For all $x, y \in X$, if $x \succsim_i y \ \forall i \in N$, then $xf((\succsim_i)_i)y$.

The unanimity axiom says that if everyone prefers x to y, then so should society.

Independence of Irrelevant Alternatives (IIA) For all $x, y \in X$, $(\succsim_i)_i$, $(\succsim_i')_i$, if $x \succsim_i y \Leftrightarrow x \succsim_i' y$, $\forall i \in N$, then $xf((\succsim_i)_i)y \Leftrightarrow xf((\succsim_i')_i)y$.

The IIA axiom says that the social preference between two specific alternatives, x and y, only depends on individual preferences between these two alternatives. That is, suppose we compare two different profiles, $(\succsim_i)_i$, $(\succsim_i')_i$, and find that they are vastly different in many ways, but it so happens that when we restrict attention to the pair $\{x, y\}$, the two profiles look the same: for each and every individual, x is considered to be better than y according to $\succsim_i$ if and only if it is better than y according to $\succsim_i'$. The axiom requires that when we aggregate preferences according to the function f, and consider the aggregation of $(\succsim_i)_i$, that is $f((\succsim_i)_i)$, and the aggregation of $(\succsim_i')_i$, which is denoted $f((\succsim_i')_i)$, we find that these two aggregated relations rank x and y in the same way.

The final definition we need is the following:

A function f is *dictatorial* if there exists $j \in N$ such that for every $(\succsim_i)_i$ and every $x, y \in X$,

$xf((\succsim_i)_i)y \Leftrightarrow x \succsim_j y.$

That is, f is dictatorial if there exists one individual, j, such that, whatever the others think, society simply adopts j's preferences. We can finally state

Theorem 4 (Arrow) f satisfies unanimity and IIA iff it is dictatorial.

Arrow's theorem can be generalized to the case in which the preference relations admit indifferences (that is, are not necessarily antisymmetric). In this case, the unanimity axiom has to be strengthened to apply both to weak and to strict preferences.[6]

B.7 Nash Equilibrium

A *game* is a triple $(N, (S_i)_{i\in N}, (h_i)_i)$, where $N = \{1, \ldots, n\}$ is a set of *players*, S_i is the (nonempty) set of *strategies* of player i, and

$$h_i : S \equiv \prod_{i\in N} S_i \to \mathbb{R}$$

is player i's vNM *utility function.*

A selection of strategies $s = (s_1, \ldots, s_n) \in S$ is a *Nash equilibrium* (in pure strategies) if for every $i \in N$,

$$h(s) \geq h(s_{-i}, t_i), \qquad \forall t_i \in S_i,$$

where $(s_{-i}, t_i) \in S$ is the n-tuple of strategies obtained by replacing s_i by t_i in s. In other words, a selection of strategies is a Nash equilibrium if, given what the others are choosing, each player is choosing a best response.

To model random choice, we extend the strategy set of each player to mixed strategies, that is, to the set of distributions over the set of pure strategies:

$$\Sigma_i = \left\{ \sigma_i : S_i \to [0,1] \;\middle|\; \sum_{s_i \in S_i} \sigma_i(s_i) = 1 \right\}.$$

6. Other formulations of Arrow's involve a choice function, selecting a single alternative $x \in X$ for a profile $(\succsim_i)_i$. In these formulations the IIA axiom is replaced by a monotonicity axiom stating that if x is chosen for a given profile $(\succsim_i)_i$, x will also be chosen in any profile where x is only "better," in terms of pairwise comparisons with all the others. This axiom is similar in its strengths and weaknesses to the IIA in that it requires that direct pairwise comparisons, not concatenations thereof, would hold sufficient information to determine social preferences.

Given a mixed strategy $\sigma_i \in \Sigma_i$ for each $i \in N$, we define i's payoff to be the expected utility

$$H_i(\sigma_1, \ldots \sigma_n) = \sum_{s \in S} \left[\prod_{j \in N} \sigma_j(s_j) \right] h_i(s),$$

and we define a Nash equilibrium in mixed strategies to be a Nash equilibrium of the extended game in which the sets of strategies are $(\Sigma_i)_i$ and the payoff functions—$(H_i)_i$.

Mixed strategies always admit Nash equilibria.

Theorem 5 (Nash) Let $(N, (S_i)_{i \in N}, (h_i)_i)$ be a game in which S_i is finite for each i.[7] Then it has a Nash equilibrium in mixed strategies.

7. Recall that in the formulation here N was also assumed finite.