Game Theory (W4210)
Course Notes

Macartan Humphreys

September 2005

ABSTRACT These notes are written to accompany my class Political Science W4210 in the political science program at Columbia University. They are in progress. Please send comments and corrections to me at: `mh2245@columbia.edu`

Contents

1
Getting Started

1.1 Reading these notes

These notes will be given out in parts to accompany the first seven weeks of class. The notes do **not** replace the readings but should help with the lectures and should summarize some key information in a single place.

The notes will also contain the **exercises** associated with different parts of the course, these are marked in the text as “Exercise #” and are associated with the lectures from a given week. In all cases they are due at the beginning of class on the following week. There are also **problems** marked in the text as “Problem #.” These do not need to be handed in, rather they are typically simple problems that are worth working through as you read through the notes. The numbering in the text follows the week numbers in the syllabus.

1.2 Reading the Readings

1.2.1 Anticipatory reading

The readings for the course are relatively few in number but you are expected to read them very very closely. The recommended approach might be what’s called “anticipatory reading.”

1. First read the first few pages or the conclusion, or skim through enough to find out what the general problem is.
   ! Now, before going further, write down a wish list of the types of propositions / theorems that you would *like* to see answered in the article (really write down the form of the propositions as formally as you can).

2. Read on to see what kind of results are in fact obtained.
   ! Compare these with your wish list: are the results: Stronger? More general? Deeper? Surprising? Disappointing?
   ! Try to satisfy yourself that the results in the text are true: think of examples and try to think of counterexamples.

3. Write down a proposed strategy of proof.

4. Try to prove the propositions yourself.
   ! If you fail, try to prove a weaker version of the propositions.

5. After succeeding or failing, compare your attempts with the proofs in the paper.
   ! What are the advantages/disadvantages of your approach relative to the approach given in the paper? What tricks did the author use that you had not thought of?
   ! Is the author's proof simpler? Is it constructive? If the author skips some step that is not clear to you ("obviously blah blah", "blah blah is trivially true,"...) try to prove the step.
   ! Don't try to read the proof until you understand exactly what the author is trying to prove.

1.2.2 Pictures and Programs

Throughout your reading: Draw pictures; Create examples; Search for counterexamples. I strongly recommend using some mathematical program to graph the various relations (or special cases of the relations) used in the text, to see what shapes they take, how they relate to other quantities in the text, what a particular solution looks like, and so on. I use Mathcad, other possibilities that are good for graphics are Mathematica and Maple. I find Gauss a lot clunkier.[1] `R` is very good for graphics and you can learn the basics very quickly. (http://www.r-project.org/). Mathematical programs can also be used to develop intuition about the nature of relations by searching through parameter spaces to see when given relationships do or do not hold. Mathcad and Mathematica are also good for

[1] To see Mathematica graphics look here : http://gallery.wolfram.com/, for Maple: http://www.mapleapps.com/categories/graphics/gallery/acgraphicsgallery.shtml and for Mathcad: http://www.mathcad.com/library/Gallery.asp

solving problems analytically, in particular they can do calculus and find symbolic solutions to systems of equations. This year Mathcad has made a set of licences available, for the course of the term, to students taking this class; `R` is available for free. We (Bernd) will give some basic training to get you started in `R` and Mathcad during the first couple of weeks of the course and will assign a some exercises in the problem sets that can be done using these programs.

1.2.3 Writing

I recommend, if you do not already know how, that you learn to use Scientific Word (www.mackichan.com) or LaTeX (www.latex-project.org, www.maths.tcd.ie/~dwilkins/LaTeXPrimer) during the course of the term and use these tools to write your papers. There's a bit of a start-up cost but once you have paid this off you will find that you can write mathematics much more quickly and also more nicely. Again there will be a short introduction to writing in LaTeX or SciWord at the start of the course, but mastering it will be up to you.

1.2.4 Dictionary

Game theory is "notationally challenged." Even simple results often use many more signs and symbols than might have seemed necessary. When reading a text it is always useful to make your own dictionary: keep a page to one side where you record the meanings assigned to symbols in the text—pay special attention to the meaning of subscripts, superscripts and decorations.

It may also be necessary to keep a mathematical dictionary handy. One very good on-line resource is http://mathworld.wolfram.com.

Also try http://www.gametheory.net/Dictionary/

1.3 Sources and Resources

1.3.1 Books on the Syllabus

- * Osborne, Martin, and Ariel Rubinstein. 1994. *A Course in Game Theory.* Cambridge: MIT UP. Also available on-line: http://www.netlibrary.com/ebook_info.asp?product_id=11376

- * Velleman, Daniel. 1994. *How to Prove It: A Structured Approach.* Cambridge: Cambridge University Press.

- Muthoo, Abhinay. 1999. *Bargaining Theory with Applications.* Cambridge: Cambridge University Press. Chapter 3. On reserve at the BUSINESS library.

- Rasmusen, Eric. 2001. *Readings in Games and Information.* London: Blackwell. See BUSINESS: QA269 .R42 2001. On reserve at the BUSINESS library.

- Mas-Colell, Andreu, Michael Whinston, and Jerry Green. 1995. *Microeconomic Theory.* Oxford: Oxford University Press. See BUSINESS: HB172 .M6247 1995.

1.3.2 Recommended Books Not on the Syllabus

- Polya, G. 1945, *How to Solve It: A New Aspect of Mathematical Method.* Princeton: Princeton University Press. (a useful and fairly enjoyable read)

- Myerson, Roger. 1991. *Game Theory: Analysis of Conflict.* Cambridge: Harvard University Press (an excellent textbook)

- Sundaram, Rangarajan. 1996. *A First Course in Optimization Theory.* Cambridge: Cambridge University Press. (good for technical explanations)

- Kreps, David. 1990. *A Course in Microeconomic Theory*, Princeton: Princeton University Press. (broad textbook written in an informal style that some love and some don't)

1.3.3 Recommended On-Line Resources

- Al Roth's page http://www.economics.harvard.edu/~aroth/alroth.html

- David Levine's page http://levine.sscnet.ucla.edu/

- Eric Rasmusen's page: http://php.indiana.edu/~erasmuse/GI/index.html

- e-Journals http://www.columbia.edu/cu/lweb/eresources/ejournals/

- Software for writing up game trees:
  http://www.cmu.edu/comlabgames/efg/index.html

- WoPEc etc.: http://netec.wustl.edu/WoPEc/

  http://econwpa.wustl.edu/months/game

1.4 Notation Refresher

1.4.1 Logic Symbols

$\forall$	For all
$\exists$	There exists...
$\exists!$	There exists a unique...
$\neg$	Not
$\sim$	Not
!	Not
$\vee$	Or
$\wedge$	And
$\vert$	Such that, given that
:	Such that, given that
$\times$	The Cartesian product / Cross product: e.g. the set $A \times B$ is the set of all pairs $\langle a, b\rangle$ in which $a \in A$ and $b \in B$.

1.4.2 Necessary and Sufficient Conditions

There are multiple ways of stating that some statement imples or is implied by another statement. Most commonly such statements are referred to as necessary and sufficient conditions. Here is listing of the equivalent statements that are used for these conditions.

Necessity: The following statements are equivalent.

X is a necessary condition for Y
Y is a sufficient condition for X
X is implied by Y
Only if X then Y (or: Y only if X)
$X \leftarrow Y$ (or: $X \Leftarrow Y$)

Example (i does not shower):
Only if it rains (X) does i get wet (Y)

Sufficiency: The following statements are equivalent.

X is a sufficient condition for Y
Y is a necessary condition for X
X implies Y
If X then Y (or: Y if X)
$X \rightarrow Y$ (or: $X \Rightarrow Y$)

Example (i has no umbrella):
If it rains (X) then i gets wet (Y)

Necessity and Sufficiency: The following statements are equivalent.

X is a necessary and sufficient condition for Y
Y is a necessary and sufficient condition for X
X implies and is implied by Y
X iff Y (or:Y iff X or: X if and only if Y)
$X \leftrightarrow Y$ (or: $X \Leftrightarrow Y$)

Example: i does not shower and has no umberella:
i gets wet (Y) if and only if it rains (X)

1.4.3 Set Operations

$\backslash$	(e.g. $A \backslash B$) The residual of set A after B is removed; $A \backslash B = A - B$.
$\in$	Element of
$\subseteq$, $\subset$	Strict Subset, Subset of
$\cup$	Union
$\cap$	Intersection
$'$	(e.g. A') The Complement of A
^{c}	(e.g. A^c) The Complement of A
$\oslash$	The null or empty set

1.4.4 Some (fairly standard) symbols for particular sets

N	the set of all agents, typically labelled such that $N = \{1, 2, ...n\}$
A_i	the set of actions available to agent $i \in N$, with typical element a_i
Σ_i	the set of mixed strategies available to agent $i \in N$, with typical element σ_i
X	the set of possible *outcomes*, with typical element x, y, or z

1.4.5 Convexity and Concavity

Note that the term convex is used differently when applied to sets and functions.

- A "**convex combination**" of a set of points in a weighted average of those points, for example : $c = \lambda a + (1 - \lambda)b$ is a convex combination of a and b for all $\lambda \in [0, 1]$.

- A set is "**convex**" if it has no indents. More formally, if a set is convex then all convex combinations of points in the set are also in the set (any points lying on a line between two points in the set are also in the set). Hence, for example, an orange is convex but a banana is not.

- The "**convex hull**" of a set is the smallest convex set that contains the set. Its what you'd get if you wrapped a blanket tightly around the set.

- A function is "**convex**" if the line joining any two points of the graph of the function lies above the graph of the function (note that if the set of points *above* the graph of a function form a convex set then the function is convex). Formally, f is convex if for all a, b in the domain of f and for all $\lambda \in (0,1)$, $\lambda f(a) + (1-\lambda)f(b) \geq f(\lambda a + (1-\lambda)b)$.

- A correspondence f is "**convex valued**" if for any point x in the domain of f, $f(x)$ is convex.

- A function is "**concave**" if the line joining any two points of the graph of the function lies *below* the graph of the function (note that if the set of points *below* the graph of a function is convex then the function is concave). Formally, f is concave if for all a, b in the domain of f and for all $\lambda \in (0,1)$, $\lambda f(a) + (1-\lambda)f(b) \leq f(\lambda a + (1-\lambda)b)$.

- A function f is "**quasiconcave**" if for any point, x, in the domain of f, the set of points $\{y : f(y) \geq f(x)\}$ is convex.

1.4.6 Preference Relations $\succsim$, $\mathcal{R}$

- An agent's "**preference relation**" $\succsim$ or $\mathcal{R}$ is a binary relation over a set of alternatives; that is, it tells us something about how two alternatives relate to each other, specifically...

- $x \succsim_i y$, $x\mathcal{R}_i y$ mean "x is **weakly preferred** to y by i."

- $x \succ_i y$, $x\mathcal{P}_i y$ mean "x is **strictly preferred** to y by i."

- $x \sim_i y$, $x\mathcal{I}_i y$ mean "i is **indifferent** between x and y."

- These operators can be strung together: e.g. $x \succ_i y \succsim_i z$.

- A preference relation is "**transitive**" if for any triple (x, y, z), $x \succeq y$ and $y \succeq z$ imply $x \succeq z$.

- A preference relation is "**complete**" if and only if for all pairs (x, y) either $x \succeq y$ or $y \succeq x$ or both.

- A preference relation is "**rational**" if it is complete and transitive. (Note that rationality here then simply means that people have well defined preferences at a given point in time over a set of options, it does not say anything about whether people are selfish, or even that people are in any way clever.)

- We can sometimes represent a preference relation $\succsim_i$ or $\mathcal{R}_i$ as a column of elements in X, with the player's label as column header and the elements ordered from most preferred on top to least preferred at the bottom. For example:

$$aP_ibP_ic \;\Leftrightarrow\; \begin{array}{c} i \\ \hline a \\ b \\ c \end{array}$$

FIGURE 1.1. Player i's ordering over $\{a, b, c\}$.

- The subscript on $\succsim$ or $\mathcal{R}$ tells us *whose* preference relation we are talking about. The subscript may refer to an individual or the group. Conventionally, society's preference relation is represented without any subscript.

- A "**profile**"of preference relations $\{\mathcal{R}_i\}$, $(\succsim_i)$ is an n-tuple (an ordered set with n elements) of preference relations, e.g. $(\succsim_i)_{i \in N} = (\succsim_1 , \succsim_2, ..., \succsim_n)$.

2

General Approaches to the Problem of Group Action

This week we will review some big results from the formal study of political and economic interactions: (i) Kenneth Arrow's Impossibility theorem, a rigorous study with very wide application, both political and philosophical that has produced an enormous literature. For problems of collective decision making, this result suggests that with freedom of opinion no prediction is possible. We then turn to examine a result attributed to Ronald Coase (ii) that suggests that with freedom of trade not only are precise predictions possible but that those predictions correspond to "good" outcomes. Finally we consider how the problem is treated in standard models in the non-cooperative game theoretic tradition (iii). These suggest that for some such problems, with freedom of action, precise predictions may be possible but they do not necessarily produce good outcomes. We close with a discussion of Amartya Sen's result on the impossibility of a Paretian liberal, which is technically simple but highlights well the tensions between these different approaches.

2.1 No Predictions Possible?: Arrow's Theorem

Political scientists (and almost everyone else) like to make statements of the form "the US wants to get rid of Saddam Hussein." Such statements assume the existence of some *method* for making claims about the preferences or interests of a collectivity, based, presumably, on the preferences or interests of all the individuals of the collectivity or some subset of them. In order to interpret such statements properly, we need to know what this method is and what its properties are. We might not require this method, whatever it is, to be based on the complete consensus of the relevant group, we may be happy for example interpreting the statement to read "a majority of Americans want to..." But we do need to have *some* sort of method simply in order to know what such statements mean. Arrow's theorem is about working out what such a method might be. The result of his enquiry is his impossibility theorem.

Arrow's Impossibility Theorem is impossible to state elegantly. In short it says that there is no way aggregate individual preferences into a rational social preference without violating some basic normative principles. But much hinges on what those normative principles are. We provide a formal statement of the theorem—and of the normative principles—next.

2.1.1 SWF, Axioms and Statement of the Theorem

First of all we need to specify what we mean by an aggregation rule. We use the idea of a social welfare function.

Definition 1 *A "**Social Welfare Function**" (**SWF**), is a preference aggregation rule, f, that maps from the set of individual preference profiles over a set of options, X (with typical element $(\succsim_i)_{i\in N}$) to the set of rational ("social") preferences over X (with typical element $\succsim$).*

Note that the requirement that the social preferences be rational implies that the SWF produces a **transitive** social ordering, and that it uses only information from ordinal non-comparable utility functions. For any group of individuals there may of course be any number of SWFs, we could for example arbitrarily choose one person and invert his ordering and call this ordering the "social" preference ordering. (This particular function does not correspond to any mechanisms actually used by any polities. But it does produce a rational ordering!) Intuitively that kind of method probably does not capture what we have in mind when we make statements of the form "Group Z prefers x to y." The reason is that we probably have a number of unstated normative assumptions in mind. One of those is probably that the rule be positively, rather than negatively, responsive to the preferences of the citizens. Arrow tries to bring these normative assumptions out into the open and, in particular, identifies the following four axioms that appear unremarkable. They are:

Definition 2 (N) *f is a "**Non-Dictatorial**" function. That is, the social ordering does not (invariably) reflect the ordering of just one person's preferences. Formally we simply require that there exists no person, i, such that $\forall x, y : \{x \succ_i y, y \succ_j x \forall j \neq i\} \longrightarrow x \succ y$.*

Definition 3 (P) *f is "**Weakly Pareto Efficient.**" That is, $\forall x, y : \{x \succ_i y \forall i\} \longrightarrow x \succ y$: hence, if* everyone *prefers x to y then x should be considered socially preferred to y.*

Definition 4 (U) *f has an "**Unrestricted Domain.**" We do not* ex ante *prohibit people from having particular preferences: individuals should be able rank alternatives in any possible strict ordering.*

Definition 5 (IIA) *f satisfies pair-wise "**Independence of Irrelevant Alternatives.**" The social ordering of x and y does not depend on the ranking of any other alternative, z, in the preference profile of any individual: $\succsim|^{\{x,y\}} = f^{x,y}((\succsim_i|^{\{x,y\}})_{i \in N})$.*

We can think of Arrow's exercise as one of narrowing down the class of possible social functions to see which ones satisfy these four conditions. His surprising result is that *no* social welfare function satisfies these conditions. The four conditions together imply intransitivity. Alternatively, transitivity plus any three of these conditions implies a violation of the fourth. A common way of proving the result is to show that together P, U, and IIA imply a violation of N. This is the approach that I follow next; the proof is based primarily on the very clear exposition by Mueller (1989) with minor modifications. Similar proofs can be found in Ordeshook (1986), Vickery (1960) and Shubik (1987).

The proof makes use of the idea of a *decisive set*, this is defined as follows:

Definition 6 *D is "**decisive**" over (x, y) if $x \succ_i y \forall i \in D$ and $y \succ_i x \forall i \notin D$ imply $x \succ y$.*

Note that a decisive set always exists since the Pareto principle implies that the group as a whole is a decisive set.

The proof then proceeds in two stages. The first shows that if any group is decisive over one pair of alternatives then it is decisive over all pairs of alternatives. The second shows that if any group is decisive, then one individual is decisive. Together these imply that the existence of any group that is decisive implies the existence of a dictator.

2.1.2 Formal Statement and Proof

Theorem 7 (Arrow) *Any SWF that satisfies* P*,* U*, and* IIA *violates* N*.*

Proof. *The proof follows in two stages:*

Stage 1: To show that if a group is decisive over one pair, it is decisive over all pairs.

1. *Let* $\mathcal{D}$ *be decisive over* x *and* y*. We aim to show that* $\mathcal{D}$ *is also decisive over any other pair* z *and* w*.*
   [The existence of some such $\mathcal{D}$ *follows from the Pareto principle]*
2. *Assume* $x \succ_i y \succ_i z$ *for all* $i \in \mathcal{D}$ *and* $z \succ_j x \succ_j y$ *for all* $j \notin \mathcal{D}$
   [Unrestricted domain lets us assume any orderings we like!]
3. *Then from* P*:* $x \succ y$

   But since $\mathcal{D}$ *is decisive we have* $y \succ z$
   So from transitivity: $x \succ z$
4. *Considering only relative rankings over* x *and* z *we have* $x \succ_i z$ *for all* $i \in \mathcal{D}$ *and* $z \succ_j x$ *for all* $j \notin \mathcal{D}$ *implies* $x \succ z$*. Hence* $\mathcal{D}$ *decisive over* x *and* y *implies that* $\mathcal{D}$ *is decisive over* z *and* x*.*
5. *By repeating steps 1-4 we can establish that* $\mathcal{D}$ *decisive over* z *and* x *implies that* $\mathcal{D}$ *is decisive over* z *and* w*.*

Stage 2: We next aim to show that if $\mathcal{D}$ *contains more than one player and is decisive over all pairs, then a strict subset of* $\mathcal{D}$ *is decisive over all pairs.*

1. *Assume* $\mathcal{D}$ *has more than one individual and partition* $\mathcal{D}$ *into non-empty groups* $\mathcal{D}^1$ *and* $\mathcal{D}^2$*.*
2. *Assume* $x \succ_i y \succ_i z$ *for all* $i \in \mathcal{D}^1$ *and* $z \succ_j x \succ_j y$ *for all* $j \in \mathcal{D}^2$*; for all other players assume* $y \succ_k z \succ_k x$
   [Note, we again use unrestricted domain]
3. *Since* $\mathcal{D}$ *is decisive and* $x \succ_i y$ *for all* $i \in \mathcal{D}$ *we have* $x \succ y$
4. *Now, either* $z \succ y$ *or* $y \succsim z$*.*
   If $z \succ y$ *then* $\mathcal{D}^2$ *is decisive.*
   But if $y \succsim z$ *then (since* $x \succ y$*) we have* $x \succ z$ *and so* $\mathcal{D}^1$ *is decisive. Either way one subset is decisive over this pair (and hence over all pairs)*
5. *By repeating steps 1-4 we establish that if* $\mathcal{D}$ *is decisive then one player is decisive over all pairs and hence a dictator exists.*

■

2.1.3 Examples of SWFs and Consequences of Axiom Violation

We have then that these four properties can not all be simultaneously satisfied for a SWF. However it is also the case that all of these properties are necessary to get the impossibility result; a SWF can be found that satisfies any three of the four properties. We demonstrate this by example next.

- A Dictatorship violates N but satisfies all the other axioms.
- A host of SWFs violate P but satisfy IIA, N and U. They tend however to have an arbitrary aspect to them. The "inverse dictator" mentioned above was one such example.
- The Borda count violates IIA but satisfies all other axioms. Violating IIA may seem unimportant—it certainly doesn't have the same moral weight as violating U, P or N. It leads however to many uncomfortable paradoxes. Here I illustrate one that may arise, the "inverted order paradox".

  Consider a population of three types that are demographically distributed with preferences as in Figure **??**. Assume now that we use the Borda count to create a social ordering. The Borda count takes each person's ranking, gives a "0" score to the lowest rank, a "1" to the second lowest and so on up to the highest. It then adds up the scores received by all options, compares these scores and creates a global ranking, violating IIA along the way.

Type	I	II	III
Weight	(3)	(2)	(2)
	x	a	b
	c	x	a
	b	c	x
	a	b	c

Inverted Order Paradox

- If preferences are constrained to be "single peaked" then majority rule is transitive and satisfies all the axioms except U. The restriction, is extremely strong and unlikely to hold in practice. With unrestricted domain satisfied however majority rule produces intransitive orderings, the classic case being the *Condorcet Paradox.*

Problem 8 *Use this information to check that the ordering of the ranking of a, b and c is inverted with the removal or addition of option x.*

Problem 9 *Arguably Arrow's result is driven by the poverty of the information he uses. What if the function used as inputs not the individual's* *ordering over options but the individuals ordering over orderings of options, this should provide much richer information. Would it resolve the problem? To answer this, you can use a related result to Arrow's that shows that it is also not possible to find a function satisfying the axioms that takes a profile of preference relations into a single best option (a social choice) rather than into a preference relation (a social ordering).*

2.2 Freedom to Trade and Precise and Pleasant Predictions

Whereas Arrow's theorem is often used to highlight the arbitrary nature of political processes, the Coase theorem is often used to argue that political processes are not "needed" at all. The Coase theorem, though rarely stated formally, appears to imply that markets can reach optimality even in cases where people tend to think states are needed. The field of application is wide ("*That follows from the Coase Theorem*!" is listed by Stigler as one of the top 30 comments at economics conferences). It has been influential in forming development policies and is often invoked for evaluating social policies. Informally, the theorem states that in the absence of "transactions costs" and assuming that property rights are fully assigned, the ability to trade will produce the same Pareto optimal[1] level of public goods production (or negative externality production) no matter how those property rights are allocated. [In weaker versions of the theorem the "the same" clause is dropped]

To see the logic of the strong version of the theorem, consider the following simple "quasilinear" environment in which each player $i \in N = \{1, 2, ...n\}$ can take actions $a_i \in A_i$ where A_i is a non-empty compact and convex subset of $\mathbb{R}^n$ for each i.

Assume that individuals can engage in trade in which they can make monetary transfers, t_{ji}, to each other (where t_{ji} is the net transfer from j to i and so $t_{ji} = -t_{ij}$) in exchange for commitments to take particular actions (a_i). Let t denote a vector containing all such transfers. Assume finally that individuals have no budget constraints.

Let the utility of each $i \in N$ be given by her independent evaluation of the good $u_i(a)$ (where a is the vector of all choices made by all individuals and $u_i : \times_{i \in N} A_i \rightarrow \mathbb{R}^1$ is continuous and strictly concave), plus her money income, $\sum\limits_{j \in N} t_{ji}$. Hence we write $v_i(a,t) = u_i(a) + \sum\limits_{j \in N} t_{ji}$. (Note: This is the quasilinear part—players have linear preferences in income.)

An "**outcome**" in this game is a pair (a, t).

[1] An outcome, x, is "Pareto Optimal" if there exists no other outcome that all players like at least as much as x and that at least one player strictly prefers to x.

Now we state and prove the following:

Claim 10 *(i) There is a unique solution, call it a^*, to the problem of maximizing $f(a) = \sum\limits_{i \in N} u_i(a)$. (ii) Furthermore, outcome (a', t') is (Pareto) efficient if and only if $a' = a^*$.*

Proof. (i) (Existence) With A_i compact and non-empty for all i, $\times_{i \in N} A_i$ is compact; with $u_i(a)$ continuous for all i, $\sum\limits_{i \in N} u_i(a)$ is continuous, hence, from the Weierstrass theorem the function $f : \times_{i \in N} A_i \rightarrow \mathbb{R}^1$ attains a maximum. (Uniqueness) With each $u_i(a)$ strictly concave, we have that $f(a)$ is also strictly concave and hence achieves a unique maximum.

$(ii.1)$ (First we do the *if* part using a Direct Proof) Assume that a^* maximizes $f(a)$. But then a^* maximizes $\sum\limits_{i \in N} u_i(a) = \sum\limits_{i \in N} (u_i(a) + \sum\limits_{j \in N} t_{ji}) = \sum\limits_{i \in N} v_i(a)$. But this implies that (a^*, t^*) achieves the utilitarian optimum and therefore that it is Pareto efficient.[2]

$(ii.2)$ (Next we do the *only if* part and use a Proof by Contradiction) Assume that (a', t') is Pareto efficient but that a' does not maximize $f(a)$. Consider now the rival outcome in which a^* is implemented and transfers, t^* are made that differ from t' only in that: $t^*_{i1} = u_i(a^*) - u_i(a') + t'_{i1}$ for all $i \in N$.[3]

The net gain to player $i \neq 1$ is then:

$$\begin{aligned}
\Delta_i &= u_i(a^*) - u_i(a') + t^*_{1i} - t'_{1i} \\
&= u_i(a^*) - u_i(a') - t^*_{i1} + t'_{i1} \\
&= u_i(a^*) - u_i(a') - u_i(a^*) + u_i(a') - t'_{i1} + t'_{i1} \\
&= 0
\end{aligned}$$

Hence, each player other than Player 1 is indifferent between (a', t') and (a^*, t^*).

[2] Why is the utiltarian optimum necessarily Pareto efficient. Write out a short proof.

[3] The interpretation here is that each player hands her "bonus", $u_i(a^*) - u_i(a')$ that she gains from the implementation of a^*, over to Player 1, along with whatever transfer, t'_{i1}, was previouly being made.

Player 1's net gain is given by:

$$\begin{aligned}
\Delta_1 &= u_1(a^*) - u_1(a') + \sum_{i \in N \backslash 1} (t^*_{i1} - t'_{i1}) \\
&= u_1(a^*) - u_1(a') + \sum_{i \in N \backslash 1} (u_i(a^*) - u_i(a') + t'_{i1} - t'_{i1}) \\
&= u_1(a^*) - u_1(a') + \sum_{i \in N \backslash 1} (u_i(a^*) - u_i(a')) \\
&= \sum_{i \in N} u_i(a^*) - \sum_{i \in N} u_i(a')
\end{aligned}$$

But from the fact that a^* solves $\max\limits_{a} \sum\limits_{i \in N} u_i(a)$ but a' does not, we have that $\sum_{i \in N} u_i(a^*) > \sum_{i \in N} u_i(a')$ and hence $\Delta_1 > 0$. Hence (a', t') cannot be Pareto efficient, a contradiction. ■

Exercise 11 (Coase Theorem) *The claim was stated and proved for cases where players have "quasilinear preferences." Does it hold when they do not? Consider for example a case in which the utility of a is different depending on whether a player is rich or poor and given by $v_i(a, t) = u_i(a) \times \sum\limits_{j \in N} t_{ji}$. Does part (ii) of the claim hold? [Optional: Can you identify classes of utility functions where part (ii) does or does not hold?]*

Remark 12 *A very similar proof can be used to show the same result for weak Pareto optimality (Property [P] above).*

If we are willing to believe that bargaining will always lead to an efficient outcome, and we buy into the other assumptions on the individuals' preferences, then the theorem implies that bargaining will always lead to the same public goods outcome (a^*).

This result has been used to argue that it doesn't matter who is in government as long as people can trade. And it doesn't matter whether the decision to look for work is taken by a woman or by her husband...

2.3 Freedom of Action and Precise but Pessimistic Predictions

Now consider an illustration of how the problem we have just examined is handled in a non-cooperative framework. A criticism of the result we saw above is that we did not specify *how* trade was conducted. By ignoring the question we fail to provide an explanation for how optimality is achieved.

Consider the same problem as described above: each player $i \in N$ can take actions $a_i \in A_i$; individuals can make monetary transfers, t_{ji}, to each other; i's utility is given by $v_i(a,t) = u_i(a) + \sum\limits_{j \in N} t_{ji}$. In this case however players cannot write "contracts," rather they can make informal agreements but once it comes to taking their actions, each player chooses a_i plus a vector of transfers simultaneously. Given this problem we look for a Nash equilibrium in which no player wishes to change her actions, given the set of actions by all other players.

With this structure of no trades and simultaneous action it is clear that any Nash equilibrium must involve no transfers from or to any player. We focus then on the "policy" actions employed by the players. Ignoring transfers, each player i chooses a_i to maximize $u_i(a_1, a_2, ..., a_i, ..., a_n)$. First order conditions are given by $\frac{\partial u_i(a_1,a_2,...,a_i,...,a_n)}{\partial a_i} = 0$. At the equilibrium such conditions must hold for all players, and so, (placing a "$*$" on the player's strategies to denote Nash equilibrium strategies) a set of conditions of the following form must be satisfied at the equilibrium:

$$\begin{bmatrix} \frac{\partial u_1(a^*)}{\partial a_1} \\ \frac{\partial u_2(a^*)}{\partial a_2} \\ \vdots \\ \frac{\partial u_n(a^*)}{\partial a_n} \end{bmatrix} = 0$$

Say such an equilibrium exists, can we say anything about its properties? In many cases we can. For example, assume that for any two players, i and j, $\frac{\partial u_i(a^*)}{\partial a_j} \neq 0$. This implies that the strategy that j chooses to maximize his own welfare does not happen to be the strategy that is also best for him to have chosen from i's perspective. In this case although j cares little for a small change at this point, such a small change in j's actions is measurably good (or perhaps, bad), for i. In this case a trade is possible that improves the welfare of both i and j, but it is not being realized under the Nash equilibrium.

Clearly if, furthermore, it is the case for all $i \neq j$, $\frac{\partial u_i(a^*)}{\partial a_j}$ is positive (or negative), then an increase (or reduction) in j's actions will be Pareto improving, subject to some small compensation for j.

Example 13 *Let* $N = \{1, 2, ...n\}$, $n > 2$, *and for* $i \in N$ *let* $A_i = \mathbb{R}^1$ *and* $u_i(a_i|a_{-i}) = (\sum_{j \in N} a_j)^2 - 2(a_i + 1)^2$.[4]

The Nash equilibrium for this game is found by maximizing $u_i(a_i|a^*_{-i})$ *for each individual, conditional upon the actions of each other individual*

[4] This is a system in which all players benefit enormously from extra contributions from all other players but each player pays a private cost for contributing. In this system the utilitarian objective function, $\sum_{i \in N} u_i(a_i|a_{-i})$, is convex in each a_i and unless bounds are put on the actions people can take there is no limit to how large a contribution is socially desirable.

being themselves best responses (hence the "$*$" in a^*_{-i}). Taking first order conditions for player i we have that a^*_i must satisfy:

$$\frac{\partial u_i(a_i|a^*_{-i})}{\partial a_i} = 2(\sum_{j\in N} a^*_j) - 4(a^*_i + 1) = 0$$

or:

$$\sum_{j\in N} a^*_j - 2a^*_i - 2 = 0 \qquad (2.1)$$

(check the second order conditions!). We have one of first order condition like this for each individual. At a Nash equilibrium they must all hold. To solve for them then we simply have to solve a system of n linear equations. With n undetermined that might not be easy, but there are typically shortcuts for such problems.

Here is one shortcut: If each condition holds, then the sum of the conditions also holds. Hence:

$$\sum_{i\in N}(\sum_{j\in N} a^*_j - 2a^*_i - 2) = \sum_{i\in N} 0$$

and hence

$$\sum_{j\in N} a^*_j = \frac{2n}{n-2}$$

Using Equation 2.1[5] *we can then have that the Nash equilibrium is given by*

$$a^*_i = \frac{2}{n-2} \textit{ for all } i.$$

We have then identified a Nash equilibrium and we are ready to start analyzing it. Key points to note in this case as part of your analysis are that $n > 2$, a_i is decreasing rapidly and that, the total contributions, $\sum_{j\in N} a^*_j$ are also decreasing, approaching a limit of $\sum_{j\in N} a^*_j = 2$ as $n \to \infty$.

Exercise 14 *Given this example, use* R *or Mathcad (or another program), (i) to graph a given players' utility function over $\mathbb{R}^1$ (that is, over one dimension), (ii) to graph an individual's equilibrium action as a function of n, (iii) to graph the total equilibrium contributions as a function of n.*

[5] A second shortcut that can work for this problem, is to assume that the equilibrium is symmetric. In this case using Equation 2.1 and substituing a_i for each a_j we have $na^*_i - 2a^*_i - 2 = 0 \leftrightarrow a^*_i = \frac{2}{n-1}$. If you use this approach you should then substitute these values into the a^*_{-i} vector and confirm that indeed a^*_i is a best response to a^*_{-i}, thereby confirming (rather than assuming) that there is in fact a symmetric Nash equilibrium.

2.4 Sen: On the Impossibility of a Paretian Liberal

The very different results we have seen follow from different approaches to modelling the world, different assumptions about what can be known about preferences and different assumptions about how groups make collective decisions. Although the differences arise from modelling choices there is also a philosophical tension between the positive Paretian result in our discussion of Coase and the adverse results found when we employ Nash's theorem. This tension is made explicit in a result due to Amartya Sen that is known as The Impossibility of a Paretian Liberal. It demonstrates the impossibility of creating any mechanism that produces a transitive ordering that guarantees both a minimal set of rights and Pareto optimality.

Traditionally the result is generated by looking at a two person setting where the two players, A and B are each allowed to be decisive over some outcome pair in X. This is seen as a way of modeling liberal "rights": if for example, A has the right to read a book then A is decisive over the pair $\{$ A reads a book, A does not read a book$\}$.

Sen's story goes like this: two agents, Lewd and Prude are deciding whether to read *Lady Chatterley's Lover*. Lewd thinks that *LC's Lover* is a terrific book and would love to read it. He would get an enormous kick however out of Prude reading it. And he would simply hate the idea of nobody reading it. In contrast, Prude thinks that *LC's Lover* is a terrible rag and would like nobody to read it. If someone were to read it however, he would rather that it be he and not Lewd since Lewd might actually enjoy it. Their somewhat unusual preferences then are given by:

Lewd	Prude
1. Prude reads (p)	1. Nobody reads (n)
2. Lewd reads (l)	2. Prude reads (p)
3. Nobody reads (n)	3. Lewd reads (l)

FIGURE 2.1. Prude and Lewd

Lewd is assumed to be decisive over (l, n) and over this pair he would choose l; Prude is decisive over (p,n) and he chooses n. The two sets of rights then give us the liberal social ordering $l \succ n \succ p$. By transitivity the liberal social ordering yields $l \succ p$. Clearly however for both agents $p \succ l$. If we required Pareto optimality we would then have the cycle $l \succ n \succ p \succ l \succ n \ldots$ Visually the cycle involves Lewd picking up the copy of *Lady Chatterley's Lover*, handing it to Prude, Prude putting it down on the table, Lewd lifting it up again and so on.

Problem 15 *Does the Coase theorem provide a solution to Arrow's problem? Does it contradict Sen's result?*

2.5 Note on readings for next week

There's a relatively heavy reading load for next week but little by way of problem sets. First—do read through the pieces assigned for this week; the Coase and Hardin are both easy reads but open up lots of interesting avenues; the Geanakopolis is more difficult and of lower priority but is a very elegant example of rigorous formal writing. From next week's class on we will turn to looking systematically at strategies for proving propositions. *Read Velleman carefully*! Preempt him and try to solve his examples before he does. When you're finished test yourself using the tables on page 305-306: on a blank piece of paper, write out each goal listed in 1-9 in the table, close the book and try to write down the strategy of proof. Maybe do this before you try to do the exercises from this week's notes. The short "How to Write Mathematics" piece is very useful and a fairly enjoyable read.

3
How to Prove it I: Strategies of Proof

The notes in this section go over some of the material in Velleman only very briefly and certainly do *not* substitute for the Velleman readings. Additional material in the notes includes a small library of useful theorems that help in proving the existence and uniqueness of different kinds of points. We then consider applications of some of the results and in particular walk through a proof for Nash's theorem on the existence of an equilibrium. We end with a discussion of the use of w.l.o.g. in proof writing.

3.1 By Example / by Counterexample

The easiest proofs are by example. In proof writing, examples can serve three related functions. First they can establish the *possibility* of an outcome. For example the claim that for some class of games "Individual maximization may produce Pareto inferior outcomes" can be proved by an example. Relatedly, positive claims can be disproved with examples: find a single counterexample to a theorem and you know that the theorem is wrong. In fact it is very important to be able to falsify a proposition by counterexample. The reason is this: before you prove a proposition, you want to be pretty sure that it is true. You can lose a *lot* of time trying to prove things that turn out to be false. A thorough (failed) search for counterexamples will help you to focus only on propositions that are true and hopefully save a lot of time. Third, counterexamples can be used as the basis for a proof by contradiction, as described below.

Sometimes finding counterexamples can be tricky. Here are two approaches. The first, is based on the principle of unconditional love towards counterexamples: to be accepted, a counterexample does not have to be plausible, it only has to exist. If you falsify a proposition with an implausible counterexample, it's falsified and you are done (the proposition may then have to modified to deal with implausible cases, but very likely you may find that the implausible case leads you down to consider more plausible cases). So one approach is to think imaginatively through unlikely cases that have the features that the proposition says shouldn't exist; do not stick to thinking about likely cases. For example: Claim (to be falsified) "Consider a set of allocations of $1 between person 1 and person 2. and assume that no player gets negative utility from any outcome. Then, if some allocation maximizes the product of all players' utilities it also maximizes the sum of all players' utilities." This sounds reasonable, after all, maximizing the product is the same as maximizing the sum of the logs, and you might think (wrongly) that if utility functions are unique up to a monotone transformation, these two maximization problems should be the same. So let's find a counterexample. In fact let's see if we can find a case where the product of the utilities is always low, but the sum of utilities can be high. We can keep the product low by ensuring that at least one player gets 0 utility for all outcomes. Here are somewhat unusual utility functions that might do the trick. Let's say that each person's utility is equal to their allocation whenever they get strictly more than the other person, but that their utility is 0 if they get the same or less than the other person. Then it's easy to see that the product of the players' utilities is always 0. But the sum if the utilities can be as high as 100 if either player is given the whole $1. So we know the proposition is wrong. What part of it is wrong? You can get some hint of the answer if you plot out the utility possibility set for the counterexample we just constructed.

Exercise 16 *Plot out the utility possibility set for the problem we just constructed (that is, the set of all realizable payoff vectors (pairs, in this case)). Is it symmetric? Is it convex?*

Exercise 17 *Perhaps the problem is convexity. Consider the following candidate proposition: "Consider a set of lotteries of allocations of a $1 between person 1 and person 2. (In this case one lottery might be of the form 'with a 50% probability we give 80 cents to player 1 and 20 cents to player 2 and with a 50% probability we give 20 cents to player 1 and 80 cents to player 2.') Then, if some allocation maximizes the product of all players' utilities it also maximizes the sum of all players' utilities." Is this true? If not, prove it. (Hint: it's not true)*

The second approach is to use a computer to locate a counterexample. In situations where the parameters of the problem are very well defined you can use mathematical software to search for a counterexample. Consider for example the following claim: “Assume Players 1 and 2 have utility functions given by $u_i(x_i) = x_i^\alpha, \alpha \in [0,2]$ and that all allocations are such that $x_i = 1 - x_j$ and $x_i \geq 0$ and $x_j \geq 0$ for $i \neq j$. Then if some allocation maximizes the product of all player’s utilities it also maximizes the sum of all players’ utilities.” This problem is quite easy to get a handle on. You can solve it analytically. But, since the method is important for more complex problems, let’s think about how to set it up so that a computer can solve it. The question ‘Is there a counterexample?’ can be written as an existence question of the form: is there an α such that there exists an x_i that maximizes $(x_i^\alpha \times (1-x_i)^\alpha)$ but does not maximize $(x_i^\alpha + (1-x_i)^\alpha)$. The approach you will use then is to search through the space of utility functions—which in this case is the same as examining all values of α in the range $[0,2]$–and for each acceptable value in the parameter space, find maximizers of $(x_i^\alpha \times (1-x_i)^\alpha)$ and $(x_i^\alpha + (1-x_i)^\alpha)$, and see if the maximizers of $(x_i^\alpha \times (1-x_i)^\alpha)$ are as good at maximizing $(x_i^\alpha + (1-x_i)^\alpha)$ as the maximizers of $(x_i^\alpha + (1-x_i)^\alpha)$ are. The answer to the question will be yes or no for each α, if you get a no, then the proposition is false; if you then plot the yes’s and no’s as a function of α then you may get more insights into the conditions under which the proposition is true.

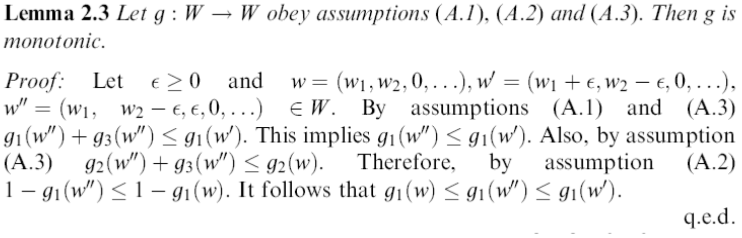
Exercise 18 *Try it.*

3.2 Direct Proof

To establish that $A \rightarrow B$ using a direct proof we assume that A is true and deduce B.

Example 19 *To prove: For x, $y \in \mathbb{R}^1$, $x^2 + y^2 \geq 2xy$.*
Proof: $(x-y)^2 \geq 0 \rightarrow x^2 + y^2 - 2xy \geq 0 \rightarrow x^2 + y^2 \geq 2xy$.

The last example is a classic direct proof. It proceeds by showing how one true implication implies another true implication, implies another true implication, until you imply your proposition. Oftentimes the chain of reasoning is not easy to see. The following is one method that can often help to identify the chain (although there are dangers involved in this!): work backwards. In practice you can often show that your proposition implies something that you know to be true. If you can do this—and if each implication “→ ” in the chain of reasoning you identify can also be reversed to produce an “←” statement—then you can turn the result around and state the reversed chain as a direct proof. The proof above for example was actually constructed by noticing that $x^2 + y^2 \geq 2xy \rightarrow x^2 + y^2 - 2xy \geq 0 \rightarrow (x-y)^2 \geq 0$. The last of these claims we know to be true. This gave a chain. But before we have a proof we also have to ensure that the chain of reasoning works in the opposite direction too, that is that $x^2 + y^2 \geq 2xy \leftrightarrow x^2 + y^2 - 2xy \geq 0 \leftrightarrow (x - y)^2 \geq 0$. If the reasoning works in both directions, as it does in this case, we are done and we simply have to turn our construction around when we write up the proof. This method can be dangerous if you are not careful about checking that each step holds in both directions.[1]



Lemma 2.3 *Let $g : W \to W$ obey assumptions (A.1), (A.2) and (A.3). Then g is monotonic.*

Proof: Let $\epsilon \geq 0$ and $w = (w_1, w_2, 0, \ldots)$, $w' = (w_1 + \epsilon, w_2 - \epsilon, 0, \ldots)$, $w'' = (w_1, \ w_2 - \epsilon, \epsilon, 0, \ldots) \ \in W$. By assumptions (A.1) and (A.3) $g_1(w'') + g_3(w'') \leq g_1(w')$. This implies $g_1(w'') \leq g_1(w')$. Also, by assumption (A.3) $g_2(w'') + g_3(w'') \leq g_2(w)$. Therefore, by assumption (A.2) $1 - g_1(w'') \leq 1 - g_1(w)$. It follows that $g_1(w) \leq g_1(w'') \leq g_1(w')$.

q.e.d.

FIGURE 3.1. A direct proof; from Gerber and Ortuno-Ortin 1998

Exercise 20 *Prove that the square of an even integer is even.*[2]

In many cases, direct proofs involve writing down the properties that we expect some object to have and demonstrating that those properties obtain. For example, often we may want to show that a set, S, is convex. This can be established using a direct proof by choosing two arbitrary elements of S, say a and b and deducing from the properties of S that any point $c = \lambda a + (1 - \lambda)b$ is an element of S for $\lambda \in (0, 1)$.

Exercise 21 *The set $Y = \{y \in \mathbb{R}^n : \sum_{i=1}^n \ln(y_i) \geq \sum_{i=1}^n \ln(y_i^0)\}$ is convex for any $y^0 \in \mathbb{R}^n$. Prove this statement for the case where $n = 2$.*

[Hint] To get you started: you need to choose two elements of Y, y^1 and y^2 (each with: $\sum_{i=1}^2 \ln(y_i^1) \geq \sum_{i=1}^2 \ln(y_i^0)$ and $\sum_{i=1}^2 \ln(y_i^2) \geq \sum_{i=1}^2 \ln(y_i^0)$) and show that for $\lambda \in (0, 1)$: $\sum_{i=1}^2 \ln(\lambda y_i^1 + (1-\lambda) y_i^2) \geq \sum_{i=1}^2 \ln(y_i^0)$. [Note] This is a result that we will use when we turn to study Nash's bargaining solution.

In some instances it is possible to prove something directly by considering the entire class of cases that the proposition covers. While this class of cases

[1] For example, say we want to prove that whenever $(x - y)^2 > 0$, it must be that $x > y$. This is clearly false. However, by noting that $x > y \to x - y > 0 \to (x - y)^2 > 0$ we may erroneously conclude that $(x - y)^2 > 0 \to x - y > 0 \to x > y$. The problem is that $x - y > 0 \to (x - y)^2 > 0$ but $(x - y)^2 > 0 \not\to x - y > 0$.

[2] An integer, a, is even if there exists an integer, b, such that $a = 2b$.

can be large, it can often be divided into simple groups that share a common property. Consider the following direct proof.

Example 22 *To prove (directly) that the number 3 is an odd integer.*[3] *This is equivalent to proving that there is no integer x, such that $2x = 3$. Let's consider every possible integer. We can divide these into all the integers less than or equal to* 1 *and all the integers greater than or equal to* 2*. For every integer $x \leq 1$ we have $2x \leq 2$; but for every integer $x \geq 2$ we have $2x \geq 4$. Hence every even integer must be less than or equal to* 2 *or else greater than or equal to* 4*. Hence* 3 *is not an even integer.*

In such cases, when you can successfully divide the set of all possible cases and prove the proposition for each one, be doubly sure to explain why the set of cases you have chosen is complete.

In many cases however it is enough to establish the result for a single 'arbitrary' element of the class. This often simplifies the problem. By arbitrary we simply mean that the element that we select has got no features of relevance to the proof that is not shared by all other elements in the set. Here is an example.[4]

Example 23 *Proposition: Every odd integer is the difference of two perfect squares. Proof: Choose an arbitrary odd integer, x. Being odd, this integer can be written in the form $x = 2y + 1$. But $2y + 1 = (y + 1)^2 - y^2$. Hence x is the difference of two perfect squares.*

3.3 Proof by Contradiction

To establish proposition P using a proof by contradiction you assume that not-P is true and then deduce a contradiction. On the principle that a true proposition does not imply a false proposition (but a false proposition implies any proposition[5]) we then deduce that not-P is false and hence

[3] An integer is odd if it is not even.

[4] Consider also the followin joke from Solomon W. Golomb (in "The Mathemagician and the Pied Puzzler, A Collection in Tribute to Martin Gardner," E. Berlekamp and T. Rogers (editors), A K Peters 1999).

Theorem. All governments are unjust.

Proof. Consider an arbitrary government. Since it is arbitrary, it is obviously unjust. The assertion being correct for an arbitrary government, it is thus true for all governments.

[5] Use the fact that a false proposition implies any proposition to solve the following problems from the island of knights and knaves where knights always tell truth whereas knaves always lie. (due to Raymond Smullyan):

- A says "If I'm a knight then P." Is A a knight? Is P true?

that P is true. It sounds like a long way of going about proving something but in fact it often opens up many avenues.[6]

Proofs by contradiction are particularly useful for proving negative statements, such as non-existence statements. They are also useful for proving uniqueness statements.[7]

Example 24 *Claim: There do not exist integers* a *and* b *such that* $2a+4b=3$.

A direct proof of this claim would require finding a way to demonstrate that for any pair of integers a *and* b, $2a+4b \neq 3$. *With an infinite number of integers, this seems difficult. A proof by contradiction would proceed as follows: Assume that there exist integers* a, b *such that* $2a+4b=3$. *Now since* $2a+4b=2(a+2b)$ *and* $a+2b$ *is an integer we have that* $2a+4b=3$ *is even. But this is false since 3 is odd. Hence we have a contradiction and so there exist no integers* a, b *such that* $2a+4b=3$.

Problem 25 *(Strict monotonicity of contract curves) Consider a setting in which two players have strictly convex preferences over points in* $\mathbb{R}^n$. *Consider a set of points,* C, *with the property that for any* $x \in C$ *there exists no point* y *such that* y *is weakly preferred to* x *by both players and strictly preferred by at least one. Show that for any two points,* x *and* y *in* C, *one player strictly prefers* x *to* y *while the other player strictly prefers* y *to* x.

Exercise 26 *Prove that if* x^2 *is odd then* x *is odd.*

Exercise 27 *Show that in any* n*-Player normal form game of complete information, all Nash Equilibria survive iterated elimination of strictly dominated strategies.*

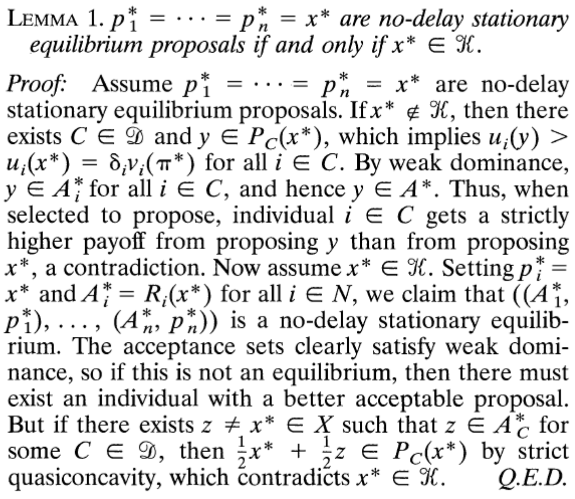
- Someone asks A, “Are you a knight?” A replies: ”If I am a knight then I’ll eat my hat”. Must he eat his hat?

- A says, ”If B is a knight then I am a knave.” What are A and B?

The following moral conundrum is also due to Raymond Smullyan: Suppose that one is morally obliged to perform any act that can save the universe from destruction. Now consider some arbitrary act, A, that is impossible to perform. Then it is the case that if one performs A, the universe *will* be saved, because it’s false that one will perform this impossible act and a false proposition implies any proposition. One is therefore morally obligated to perform A (and every other impossible act).

[6] A closely related method is the *reductio ad absurdum*.The reductio attempts to prove a statement directly and reaches a conclusion that cannot be true.

[7] These are of course related since a statement about uniqueness is a statement about the non existence of a *second* object in some class.



Lemma 1. *$p_1^* = \cdots = p_n^* = x^*$ are no-delay stationary equilibrium proposals if and only if $x^* \in \mathcal{H}$.*

Proof: Assume $p_1^* = \cdots = p_n^* = x^*$ are no-delay stationary equilibrium proposals. If $x^* \notin \mathcal{H}$, then there exists $C \in \mathcal{D}$ and $y \in P_C(x^*)$, which implies $u_i(y) > u_i(x^*) = \delta_i v_i(\pi^*)$ for all $i \in C$. By weak dominance, $y \in A_i^*$ for all $i \in C$, and hence $y \in A^*$. Thus, when selected to propose, individual $i \in C$ gets a strictly higher payoff from proposing y than from proposing x^*, a contradiction. Now assume $x^* \in \mathcal{H}$. Setting $p_i^* = x^*$ and $A_i^* = R_i(x^*)$ for all $i \in N$, we claim that $((A_1^*, p_1^*), \ldots, (A_n^*, p_n^*))$ is a no-delay stationary equilibrium. The acceptance sets clearly satisfy weak dominance, so if this is not an equilibrium, then there must exist an individual with a better acceptable proposal. But if there exists $z \neq x^* \in X$ such that $z \in A_C^*$ for some $C \in \mathcal{D}$, then $\frac{1}{2}x^* + \frac{1}{2}z \in P_C(x^*)$ by strict quasiconcavity, which contradicts $x^* \in \mathcal{H}$. *Q.E.D.*

FIGURE 3.2. Necessary and Sufficient Conditions with Proofs by Contradiction. From Banks and Duggan 2000

3.4 Establishing Monotonicity

For many problems you may want to show that some function or some process is monotone. Establishing monotonicity is often interesting in its own right; but it can also be used to establish that processes do not cycle. In some cases establishing particular types of monotonic relations may be used to establish the possibility of cycles. For continuous functions this can often be done by signing the first derivative. But for more complex processes this might not be possible. Consider for example the set of possible paths that a policy in $\mathbb{R}^n$ may take as it gets moved around by votes and amendments. The path may be very chaotic. As another example consider shifts in memberships of voting blocks, we can consider processes in which individuals move backwards and forwards, but we may be interested in whether a given pattern of blocks can ever resurface twice. This can be done by establishing some strictly monotonic feature of the process.

A method that can be used in many cases is to identify a set of one dimensional metrics that can be written as a function of the process at hand. In the example of shifting policies it may be possible to show that in each movement, the policy moves closer to some point, or that some player's utility always increases, or that the policy is contained n a sphere with shrinking radius. For many complex processes there may be many possible metrics—the size of the smallest group, the variance of some distribution, and so on. If any of these metrics change monotonically with each step in the process then the monotonicity of the metric can be used to establish the monotonicity of the process (and hence, in the case of strict monotonicity, the absence of cycles).

Exercise 28 ***Problem 29*** *Consider the following process. A population $N = \{1, 2, ...n\}$ is divided into m groups, $g_1, g_2, ...g_m$ with membership size of g_k denoted by $|g_k|$. Each individual has an associated weight of α_i. There exists a bargaining process through which each group receives a share of \$1, proportionate to its size, and each individual receives a share of her group's share in proportion to her relative weight. Hence $y_i = \frac{\alpha_i}{\sum_{j \in g_k} \alpha_j} \frac{|g_k|}{N}$. In each period one member and one group is randomly selected. The individual member changes group membership from her own group into the selected group if (a) her payoff is strictly higher in the selected group and (b) the payoff of all members of the group also receive higher payoffs from the switch. Question: does the same allocation of individuals to groups ever occur twice. Prove it.*

3.5 Establishing Uniqueness

Two approaches are proposed to prove uniqueness of a point a with some property. The first is direct: first show that some point a exists with the right properties, then show that any point b that possesses these properties is equivalent to a. The second approach is proof by contradiction: assume that two distinct points exist (satisfying the properties) and then show how this produces a contradiction.

Example 30 *Problem 78 will ask you to prove that if X is convex and compact then "the four Cs" guarantee the existence of a unique "ideal point." The existence part follows from the Weierstrass theorem and the continuity of u (see below). Uniqueness can be established as follows: Assume (contrary to the claim that there is a unique ideal point) that two distinct points x and y both maximize u on X. Then, from the convexity of X, any point z in the convex hull of x and y lies in X. Strict quasiconcavity implies that for any $z \in (x, y)$, $u_i(z) > \min(u_i(x), u_i(y))$. In particular, with $u_i(x) = u_i(y)$ we have $u_i(z) > u_i(x)$. But this contradicts our assumption that x maximizes u on X. This establishes that there cannot be two distinct points in X that maximize u.*

3.6 Turning the Problem on Its Head

When you are stuck, progress can often be made by reformulating the problem. The most common way of doing this is by "proving the contrapositive".

Proving the contrapositive:

- If you are having difficulties proving that $A \rightarrow B$, try proving the equivalent statement that *not* $B \rightarrow$ *not* A.
- Similarly $A \rightarrow$ *not* B is equivalent to $B \rightarrow$ *not* A.

Equivalencies involving quantifiers:

- "There does not exist an x such that $P(x)$" is equivalent to "For all x, not $P(x)$"
- "It is not true that for all x, $P(x)$" is equivalent to "There exists some x such that $P(x)$ is not true"
- (This last equivalency also holds for "bounded quantifiers"): "It is not true that for all $x \in X$, $P(x)$" is equivalent to "There exists some $x \in X$ such that $P(x)$ is not true"

3.7 Style

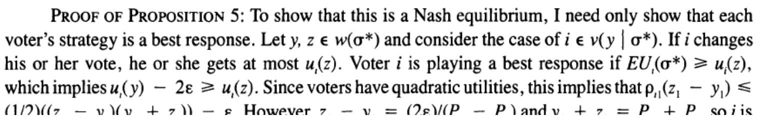
Proofs do not have to read like computer code. Unlike your code you normally want people to read your proofs; first of all to check that they are right, but also because proofs often contain a lot of intuition about your results and if people read them they will likely have a better appreciation of your work. So make them readable.

There is a lot of variation in the style of proof writing in political science but in general it is fine to have proofs written mostly in English. Indeed some object to the use of logical symbols of the form $\forall$ or $\exists$ when the English is just as clear and not much longer. Typically you want a proof to be compact but this means that there should not be any substantive flab, not that there should not be any English. Elegance is a goal but the key to the elegance of a proof is not in density of the prose but the structure of the underlying argument. You can't make up for the inelegance of a clumsy proof by stripping out the English. In fact English can often make up for a lot of math. It is permissible to use phrases such as "To see that the same holds for Case 2, repeat the argument for Case 1 but replacing s' for s''"; such sentences are much clearer and more compact than repeating large segments of math multiple times.

Beyond this, good proof writing is much like all good writing.

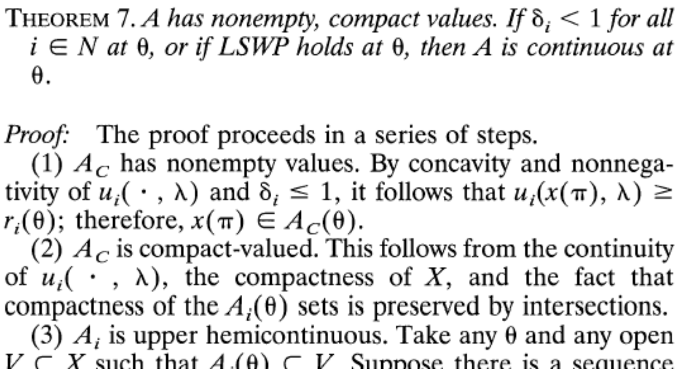
At the **beginning**, it is a good idea to say where you are going: setting out the logic of your proof at the beginning makes the train of your logic is easy to follow (Figure 3.3).

In the **middle**, it is good to use formatting to indicate distinct sections of your proof; for example, different parts (the 'if' part, the 'only if' part), cases and so on (Figure 3.4).



Proof of Proposition 5: To show that this is a Nash equilibrium, I need only show that each voter's strategy is a best response. Let y, $z \in w(\sigma^*)$ and consider the case of $i \in v(y \mid \sigma^*)$. If i changes his or her vote, he or she gets at most $u_i(z)$. Voter i is playing a best response if $EU_i(\sigma^*) \geqslant u_i(z)$, which implies $u_i(y) - 2\varepsilon \geqslant u_i(z)$. Since voters have quadratic utilities, this implies that $\rho_{i1}(z_1 - y_1) \leqslant$ $(1/2)((z - y)(y + z)) - \varepsilon$. However $z - y = (2\varepsilon)/(P - P)$ and $y + z = P + P$ so i is

FIGURE 3.3. Describe the strategy of your proof in the first sentence.Excerpt from Fedderson (1992).



Theorem 7. *A has nonempty, compact values. If $\delta_i < 1$ for all $i \in N$ at θ, or if LSWP holds at θ, then A is continuous at θ.*

Proof: The proof proceeds in a series of steps.

(1) A_C has nonempty values. By concavity and nonnegativity of $u_i(\,\cdot\,, \lambda)$ and $\delta_i \leq 1$, it follows that $u_i(x(\pi), \lambda) \geq r_i(\theta)$; therefore, $x(\pi) \in A_C(\theta)$.

(2) A_C is compact-valued. This follows from the continuity of $u_i(\,\cdot\,, \lambda)$, the compactness of X, and the fact that compactness of the $A_i(\theta)$ sets is preserved by intersections.

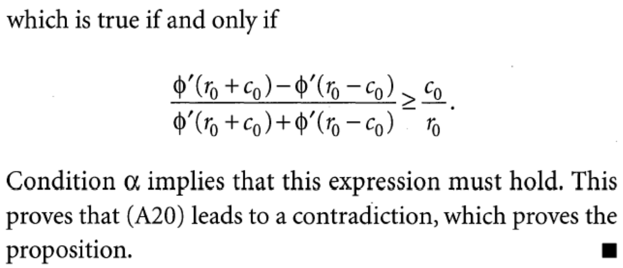
(3) A_i is upper hemicontinuous. Take any θ and any open $V \subset X$ such that $A_i(\theta) \subset V$. Suppose there is a sequence

FIGURE 3.4. Fragment of complex proof in which multiple steps are laid out explicitely. From Banks and Duggan (2000)

At the **end** it is good to say that you've shown what you wanted to show. This can be a little redundant and is not necessary in a short proof but is good form in a longer proof (Figure 3.5).

Finally, even if a lot of math is woven into your text, you are still responsible for ensuring that the whole follows the rules of English grammar and syntax. Every sentence should be a complete English sentence, beginning with a capital letter (in fact almost always beginning in English) and ending with a period. A few rules of thumb follow:

- You cannot start a sentence in math like this "$i \in N^*$ implies condition z..."; instead you need something like: "If $i \in N^*$ then condition z holds..."
- For clarity it is useful to use a few imperatives early in the proof: "Assume that...", "Suppose that...", "Let ..."
- Most sentences that are math heavy begin with leaders like "Note that...", "Recall that...", "Since...", "But...", "Therefore..."
- When in the middle of a sequence of steps, begin a sentence with a gerund of the form "Adding...", "Taking the derivative..." and so on



which is true if and only if

$$\frac{\phi'(r_0+c_0)-\phi'(r_0-c_0)}{\phi'(r_0+c_0)+\phi'(r_0-c_0)} \geq \frac{c_0}{r_0}.$$

Condition α implies that this expression must hold. This proves that (A20) leads to a contradiction, which proves the proposition. ■

FIGURE 3.5. Concluding a proof (Groseclose 2001)

(although this does not excuse you from using a main verb, even if the verb is mathematical).

- When using theorems or lemmas you begin with “Applying...” or “From...” or “By...”

4
How to Prove It II: Useful Theorems

4.1 Existence I: Maximum and Intermediate Values

The first theorem we present is very simple but useful in a large range of contexts. It provides sufficient conditions for the existence of a maximum:

Theorem 31 (Weierstrass Theorem) (or: the Maximum Value Theorem). *Suppose that $A \subset \mathbb{R}^n$ is non-empty and compact and that $f : A \rightarrow \mathbb{R}^n$ is a continuous function on A. Then f attains a maximum and a minimum on A.*

Example 32 *Consider a setting like that we discussed in our treatment of the Coase theorem in which players can make transfers to each other. In particular, assume that net transfers from Player 1 to Player 2 are given by $t_{12} \in A$ where A is a non-empty subset of $\mathbb{R}$. Assume that each player has utility $u_i : A \rightarrow \mathbb{R}$ over these transfers, where u_i is continuous for $i \in \{1, 2\}$. In this context we may be interested in knowing whether there is a solution to the Nash bargaining problem, or whether there is a utilitarian or Rawlsian optimum or a Nash equilibrium. The Weierstrass theorem provides a positive answer to all of these questions, no matter what the u functions look like (as long as they are continuous)* if *A is compact. If A is not-compact, then all bets are off and more information is needed about the u functions in order to be sure of solutions to these problems.*

This result is commonly used for the first part of "existence and uniqueness" claims regarding optimizing behavior. Note that compactness of A and continuity of f are important to make sure that the graph of f is closed, if it were open then there would be risk that the supremum of the graph would not be attainable.

The next result is useful in claiming the existence of some identifiable threshold that can be used to partition actors into sets (e.g. those who like some outcome and those who do not...).

Theorem 33 (Bolzano's Theorem) (or: the Intermediate Value Theorem). *Suppose that $A \subset \mathbb{R}^1$ is non-empty and compact and that $f : A \to \mathbb{R}^1$ is a real continuous function on A, then for any two points a and b in A, f takes every value between $f(a)$ and $f(b)$, for some points between a and b.*

Example 34 *Assume that there is a continuum of individuals that can be ordered in terms of how much they agree with the policies of some party. Assume that the net transfers to each individual in the run-up to an election is a continuous (although not necessarily monotonic) function of the individual's support for the party. Then if there is some individual who stands to make a net gain and some individual who stands to make a net loss, then we know that there is some individual with preferences between these two who will be unaffected.*

4.2 Existence II: Fixed Point Theorems

We now begin with a series of fixed point theorems. These have very wide application but have been used especially for proving the existence of an equilibrium of various forms (including Nash's result). The idea is that if you can use a function or a correspondence, f, to describe the way that a system moves from one state, x, to another, $f(x)$, then you can describe a stable state as a point x^* that is a fixed point of f, that is, with the property that $f(x^*) = x^*$. A large number of fixed point theorems exist, but they each have different constraints on the functions f that are required to guarantee the existence of the fixed point.

Theorem 35 (Brouwer's Fixed Point Theorem) *Suppose that $A \subset \mathbb{R}^n$ is non-empty, compact and convex and that $f : A \to A$ is a continuous function from A into itself. Then there is an $a \in A$ with $a = f(a)$.*

Example 36 *(of a mapping from $\mathbb{R}^2$ to $\mathbb{R}^2$) If one copy of this page is crumpled up and placed on top of an uncrumpled copy, then there is a part of the notes on the crumpled page that lies directly over that same part on the uncrumpled page (even if you place the page upside down).*

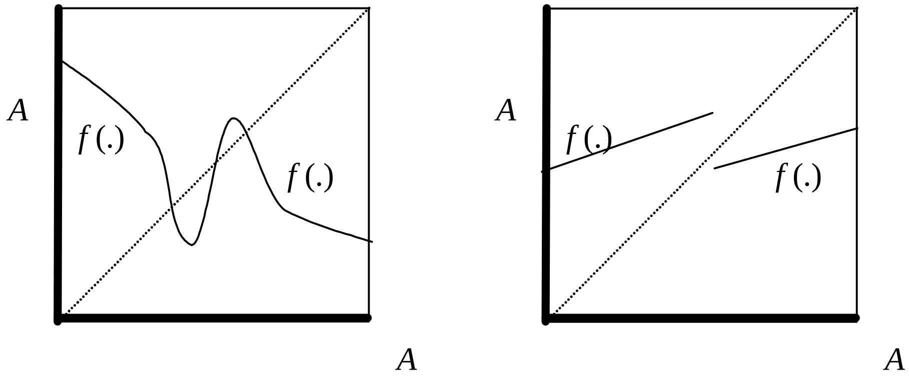


FIGURE 4.1. Brouwer: There are 3 fixed points in the first panel, but none in the second panel (because f is not continuous)

Example 37 *Let individuals have ideal points distributed with positive density over the ideological space* $[0,1]$. *Assume that there is some* continuous *shock to everyone's ideology (that is, individuals may take any new position on* $[0,1]$ *but the shock is such that ideological neighbors remain ideological neighbors). Then at least one person will not change her position.*

The following theorem is useful if you can not rely on your function being continuous but you do know that it is non-decreasing...

Theorem 38 (Tarski's Fixed Point Theorem) *Suppose that* $A = [0,1]^n$ *and that* $f : A \to A$ *is a non-decreasing function. Then there is an* $a \in A$ *with* $a = f(a)$.

The third fixed point theorem that we introduce is good when reaction functions are set-valued rather than point-valued. This is the case for example whenever there is more than one best response to the actions of other players. In such contexts we need a fixed point for *correspondences* rather than for functions. And we need a new notion of continuity:

Definition 39 *let* A *and* B *be closed subsets of* $\mathbb{R}^n$ *and* $\mathbb{R}^k$ *respectively. The correspondence* $f : A \to B$ *is "**upper hemicontinuous**" if it has a closed graph and for any compact subset,* C*, of* A*,* $f(C)$ *is bounded.*

The intuitive definition given by Starr (1997) for upper hemicontinuity (also called upper semicontinuity) is this: if you can sneak up on a value in the graph of f then you can catch it.[1]

[1] His corresponding idea for lower hemicontinuity is: if you can catch a value, then you can sneak up on it.

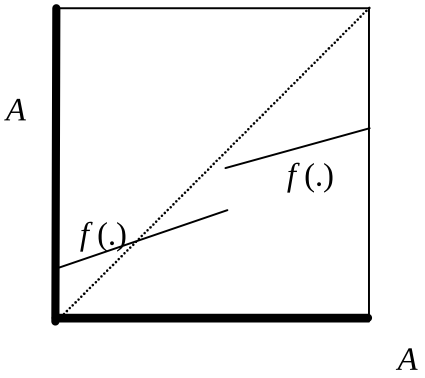
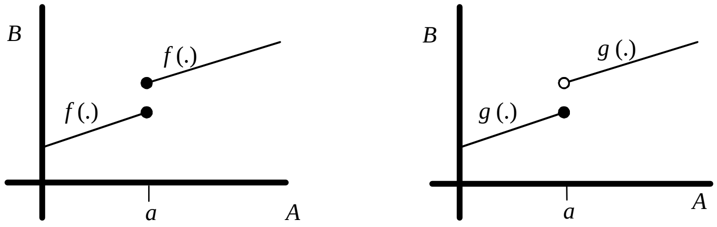


FIGURE 4.2. Tarski: Note that f is not continuous but there is still a fixed point because f is non-decreasing



(i) Here the graph of g is closed (there are two values for f(a)) and bounded and so f is upper hemicontinuous

(ii) Here the graph of g is open (the white circle represents a point that is not in the graph of g) and so g is not upper hemicontinuous

FIGURE 4.3. upper hemicontinuity

Theorem 40 (Kakutani's Fixed Point Theorem) *Suppose that $A \subset \mathbb{R}^n$ is non-empty, compact and convex and that $f : A \to A$ is an upper hemicontinuous correspondence mapping from A into itself with the property that for all a in A we have that $f(a)$ is non-empty and convex. Then there is an $a \in A$ with $a \in f(a)$.*

See Figure 4.4 for an illustration of the theorem.

The following theorem, closely related to Brouwer's fixed point theorem, is useful for mappings from spheres to planes (for example, if the action set is a set of *directions* that players can choose or if the outcome space is spherical, such as is the case arguably for outcomes on a calendar or on the borders of a country.)

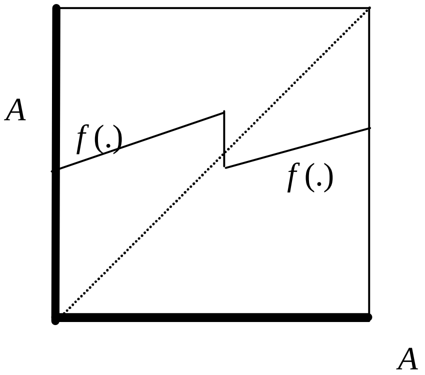


FIGURE 4.4. The figure shows the correspondance f. Note that the graph of f is convex valued and that there is a fixed point.

Theorem 41 (Borsuk-Ulam) *Any continuous function from an n-sphere into $\mathbb{R}^n$ maps some pair of antipodal points to the same point. (Two vectors v and v' on a sphere centered on the origin are "**antipodal**" if $v = -v'$)*

Example 42 *There is always a pair of points on opposite sides of the Earth that have both the same temperature and the same barometric pressure.*

4.3 Application: Existence of Nash Equilibrium

Example 43 (Existence of a Nash equilibrium) *Recall that we said before that a profile of mixed strategies σ^* is a "**Nash equilibrium**" if for each $i \in N$, we have: $\sigma_i^* \in \arg \max\limits_{\sigma_i \in \Delta(A_i)} (u_i(\sigma_i, \sigma_{-i}^*))$. Now we want to show that such an equilibrium exists whenever A is finite. The strategy of proof is simple: we wish to find a fixed point of the correspondence $f(\sigma) = (\arg \max\limits_{\sigma_i \in \Delta(A_i)} (u_i(\sigma_i, \sigma_{-i})))_{i=1,\ldots,n}$. Any such fixed point is a Nash equilibrium. In practice this requires a search to see if this correspondence satisfies the conditions of some fixed point theorem. Once the search is complete the actual proof simply requires showing that all of the conditions are satisfied. In this case we will show that the conditions of Kakutani's fixed point theorem are satisfied. We work through the conditions one by one.*

First, for convenience, let's use Σ_i to denote $\Delta(A_i)$ and Σ to denote $\times_{j \in N} \Delta(A_j)$. Now, define the correspondence $f_i(\sigma) = \arg \max\limits_{\sigma_i \in \Sigma_i} (u_i(\sigma_i, \sigma_{-i}^*))$ where $f_i(\sigma)$ is the set of "best responses" by player i, conditional upon all the strategies played by all the other players (as recorded in the vector σ). Note that $f_i(\sigma)$ maps from Σ into Σ_i. Next, define a correspondence that *is the Cartesian product of the individual reaction correspondence: $f = (f_1, f_2, ..., f_n) : \Sigma \to \Sigma$; hence f is a correspondence that describes the best responses of each player to the strategies of all the other players.*

Step 1. Σ ***is non-empty, compact and convex****. Let's take a second to think about what each Σ_i looks like. If player i has k pure strategies available, then a possible mixed strategy available to him is a set of probabilities associated with each of the k strategies: $(p_1, p_2, ..., p_k)$. This is just a point in $\mathbb{R}^k$; in fact, because all the probabilities sum to 1 it is a point lying on a simplex*[2] *of dimension $k - 1$. And so the set of all such points, Σ_i, is a simplex and in particular it is closed, bounded, non-empty and convex. Similarly, Σ is closed, bounded, non-empty and compact. Hence f maps from a non-empty, compact and convex set into itself.*

Step 2. For every σ***, each*** $f(\sigma)$ ***is (i) non-empty and (ii) convex.*** *We check these conditions for each $f_i(\sigma)$. To check (i) we simply need to make sure that $u_i(\sigma_i, \sigma_{-i})$ attains a maximum for each for each σ_{-i}. For this we use the Weierstrass theorem, making use of the fact that Σ_i is compact and $u_i(\sigma_i, \sigma_{-i})$ is continuous in σ_i (continuity follows from the fact that $u_i(.)$ is linear in σ_i which itself follows from the Expected Utility Theorem). To check (ii) we also use the Expected Utility Theorem. Assume that there exist two distinct values, σ_i' and σ_i'', that maximize $u_i(\sigma_i, \sigma_{-i})$. We wish to confirm that for any $\lambda \in (0, 1)$, $\lambda\sigma_i' + (1 - \lambda)\sigma_i''$ is also a best response. But, from linearity, $u_i(\lambda\sigma_i' + (1 - \lambda)\sigma_i'', \sigma_{-i}) = \lambda u_i(\sigma_i', \sigma_{-i}) + (1 - \lambda)u_i(\sigma_i'', \sigma_{-i})$ and hence $u_i(\lambda\sigma_i' + (1 - \lambda)\sigma_i'', \sigma_{-i}) = u_i(\sigma_i'', \sigma_{-i}) = u_i(\sigma_i', \sigma_{-i})$. This establishes convexity of $f_i(\sigma)$ and hence of $f(\sigma)$.*

Step 3. The correspondence f ***is upper hemicontinuous****. This is perhaps the least obvious part. Boundedness of the graph of f is not a problem since we have that Σ is itself bounded (Σ is compact). So we just need to establish that the graph of f is closed. A good rule of thumb is that if it seems difficult, try a proof by contradiction. Assume then that f is not upper hemicontinuous and in particular that the graph of f is not closed. Consider a sequence $(\sigma^n, \tilde{\sigma}^n)$ where each $\tilde{\sigma}^n$ is an element of $f(\sigma^n)$. Assume now that that the sequence converges to $(\sigma, \tilde{\sigma})$, but that, contrary to upper hemicontinuity $\tilde{\sigma}$ is not an element of $f(\sigma)$. In particular, for some player there exists some rival strategy σ_i' such that $u_i(\sigma_i', \sigma_{-i}) > u_i(\tilde{\sigma}_i, \sigma_{-i})$. But if this is true then for sufficiently small $\varepsilon > 0$ we have $u_i(\sigma_i', \sigma_{-i}) > u_i(\tilde{\sigma}_i, \sigma_{-i}) + 2\varepsilon$ and hence $u_i(\sigma_i', \sigma_{-i}) - \varepsilon > u_i(\tilde{\sigma}_i, \sigma_{-i}) + \varepsilon$ $[*]$. Now, for sufficiently large n (when $(\sigma^n, \tilde{\sigma}^n)$ is arbitrarily close to $(\sigma, \tilde{\sigma})$) we have, from the continuity of u in σ, that $u_i(\tilde{\sigma}_i, \sigma_{-i})$ is arbitrarily close to $u_i(\tilde{\sigma}_i^n, \sigma_{-i}^n)$ and hence (for sufficiently large n), $u_i(\tilde{\sigma}_i, \sigma_{-i}) + \varepsilon > u_i(\tilde{\sigma}_i^n, \sigma_{-i}^n)$ $[**]$. By the same reasoning we have $u_i(\sigma_i', \sigma_{-i}^n) + \varepsilon > u_i(\sigma_i', \sigma_{-i})$ $[**$*

[2] An $n - 1$ dimensional simplex is a shape (a polytope) with n affinely independent vertices. A two dimensional simplex for example is a triangle (this simplex can be formed by taking all the convex combinations of three affiniely independent points in $\mathbb{R}^3$).

$*]$. Drawing together the findings in $[*] - [***]$ we have $u_i(\sigma_i', \sigma_{-i}^n) > u_i(\sigma_i', \sigma_{-i}) - \varepsilon > u_i(\tilde{\sigma}_i, \sigma_{-i}) + \varepsilon > u_i(\tilde{\sigma}_i^n, \sigma_{-i}^n)$ and hence $u_i(\sigma_i', \sigma_{-i}^n) > u_i(\tilde{\sigma}_i^n, \sigma_{-i}^n)$, but this contradicts our assumption that $\tilde{\sigma}_i^n \in f_i(\sigma^n)$. This contradiction establishes upper hemicontinuity.

Together Steps 1-3 provide all the conditions needed to use Kakutani's fixed point theorem to establish that f has a fixed point.

Exercise 44 (Brouwer) *This exercise establishes conditions necessary for the existence of an equilibrium in a two player alternating offers bargaining game over $\mathbb{R}^n$. Assume that players have strictly convex preferences defined over $X \times T$ given by $\succsim_1$and $\succsim_2$ and that for any element z in X, $(z, 0) \succsim_i (z, 1)$ for $i \in \{1, 2\}$.*

Prove that there exists a pair $(\bar{x}, \bar{y}) \in \mathbb{R}^n \times \mathbb{R}^n$ such that:
$\bar{x}$ maximizes $\succsim_1$ subject to $(x, 0) \succsim_2 (\bar{y}, 1)$ and
$\bar{y}$ maximizes $\succsim_2$ subject to $(y, 0) \succsim_1 (\bar{x}, 1)$

4.4 Theorems to Establish Uniqueness

Guaranteeing the *existence* of a fixed point does not tell you much about the fixed point. One important property of fixed points that you might want to know is whether the fixed points are unique. This is particularly important for contexts in which you want to compare properties of equilibria across settings. In general, guaranteeing uniqueness is difficult, however, for some applications, the following result may help.[3]

Theorem 45 (Contraction Mapping Theorem) (Also called "Banach's Contraction Principle"). *Suppose that $A \subset \mathbb{R}^n$ is non-empty and that $f : A \rightarrow A$ has the property that for some measure of distance $d(.,.)$ there exists some constant $k < 1$ such that $d(f(a), f(b)) \leq kd(a, b)$ for all a, b in A, then any fixedd point of f is unique.*

A related and easy-to-use result is that if in a two player game the reaction functions of both players are continuous and have slope less than 1, then there is a unique Nash equilibrium.

[3] For more in this area you can consult work that aims to identify conditions for a reaction function to be a contraction. One approach that has proved fruitful is identifying when the Hessian u_{aa} displays "diagonal domnance." See for example references in: Gerard P. Cachon and Serguei Netessine. 2003. "Game Theory in Supply Chain Analysis." Working Paper. http://opim.wharton.upenn.edu/~cachon/pdf/game_theoryv1.pdf Further results can be found in the literature on "supermodular games."

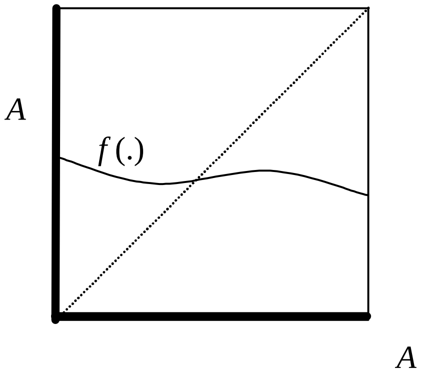


FIGURE 4.5. Example of a Contraction Mapping

4.5 Methods to Make Results More General: Genericity

One way to think about the generality of a proposition is in terms of the **genericity** of the set for which the result is true.

The idea here is that we may want to make statements of the form: "such and such a property never occurs except in *really exceptional* circumstances" or "I have proved this for such and such a situation, but in fact that situation is the type of situation that we would '*essentially always*' expect to observe."

The question is: how do we make statements of this form precise? Here the notion of **genericity** comes in handy.

Definition 46 *Consider some set $\mathcal{S} \subset \mathcal{A}$. $\mathcal{S}$ is "**generic**" in $\mathcal{A}$ if it is open and dense.*

Let's unpack that statement:

- *if it is open*: for any point $z \in \mathcal{S}$, there exists an ε such that every point in $\mathcal{B}(z, \varepsilon)$ lies in $\mathcal{S}$. (that is, $\forall z \in \mathcal{S} \exists \varepsilon > 0 : \mathcal{B}(z, \varepsilon) \subset \mathcal{S}$)
- *if it is dense*: for any point $x \in \mathcal{A} \backslash \mathcal{S}$ and for every $\varepsilon > 0$ there exists some point $y \in \mathcal{S}$ in $\mathcal{B}(x, \varepsilon)$ (that is, $\forall \varepsilon > 0 \forall x \in (\mathcal{A} \backslash \mathcal{S}) \exists y \in \mathcal{B}(x, \varepsilon) \cap \mathcal{S}$)

Example 47 *Consider the set of preference profiles in which three players each have an ideal point in $\mathbb{R}^2$. Let's say that we want to show that generically these three ideal points will not lie on a straight line. To do this we need to define a set, $\mathcal{A}$, whose typical member is a set of three points in $\mathbb{R}^2$ (e.g. $x = \{x_1, x_2, x_3\} \in \mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{R}^2$). We say that an element y of $\mathcal{A}$ is* *in the neighborhood $\mathcal{B}(x, \varepsilon)$ of $x \in \mathcal{A}$ if $|x_i - y_i| \leq \varepsilon$ for $i = 1, 2, 3$. Let $\mathcal{T}$ denote the set of elements of $\mathcal{A}$ whose components all lie in a straight line. The aim is then to show that the complement of $\mathcal{T}$ in $\mathcal{A}$ is open and dense.*

4.6 General functional forms and implicit function theorems

In many cases, to render your results more general, you should avoid choosing particular functional forms. For example, in the case of utility functions, such as $u_i : A_i \times A_j \rightarrow \mathbb{R}^1$, you ideally will want to put as few constraints as possible on the function, and simply impose conditions of the form $\frac{\partial^2 u_1}{\partial x_1{}^2} < 0$ or $\frac{\partial^2 u_1}{\partial x_1 \partial x_2} < 0$. The problem with such a lack of specificity is that it often seems difficult to make substantive statements about players' optimal behavior. In particular it is often impossible to derive explicit expressions describing behavior.

Commonly we might know something of the form: at an equilibrium point some condition might hold; for example $f(x^*, y^*) = 0$, but this on it's own does not give us much of an indication of how x^* relates to y^*. Such a relation might however be of substantive interest. Implicit function theorems allow you to get a handle on this by letting you to *infer* the existence of an explicit function, using other properties you know about the problem. This can be enough to make statements regarding how behavior changes as features of the game change.

Theorem 48 (An Implicit Function Theorem) *Let $f(x, y) : \mathbb{R}^2 \rightarrow \mathbb{R}^1$ denote a continuous function on a ball around some point (x^*, y^*). If $\frac{\partial f(x^*, y^*)}{\partial y} \neq 0$, then there exists a continuous function $y = y(x)$, defined on an interval about the point x^*,such that*
(i) $f(x, y(x)) = f(x^*, y^*)$
(ii) $y = y(x^*) = y^*$
(iii)$\frac{dy(x^*)}{dx} = -\frac{\frac{\partial f(x^*, y^*)}{\partial x}}{\frac{\partial f(x^*, y^*)}{\partial y}}$

Similar theorems exist for more general functions f. Implicit function theorems like this can be used to combine information on the signs that we have for the first and second derivatives of functions with general conditions that we know have to hold when players play optimally in order to make predictions.

To develop your intuitions consider the following, let $f(x, y)$ be given by $f(x, y) = \ln(x) + y$. And assume that at x^*, y^*, $f(x^*, y^*) = 0$. Hence

$\ln(x^*)+y^* = 0$. Clearly it is possible to write an explicit function of the form $y^* = -\ln(x^*)$. If we consider small movements around x^* we could think of this as a function of the form $y^*(x) = -\ln(x)$. For this function we can see readily that $\frac{dy(x^*)}{dx} = -\frac{1}{x}$. We get this from differentiating an explicit function. But even if we had not created an explicit function, we could use the fact that $\frac{dy(x^*)}{dx} = -\frac{\frac{\partial f(x^*,y^*)}{\partial x}}{\frac{\partial f(x^*,y^*)}{\partial y}}$ to deduce that $\frac{dy(x^*)}{dx} = -\frac{\frac{1}{x}}{1} = -\frac{1}{x}$.

Let's see this in operation. In the following example, the x's and y's are the strategies taken by players, and the f functions are their utilities over the actions. we use the theorem to see how one player's optimal strategy changes as a function of the other player's optimal strategy.

Example 49 *As an example, consider a two stage extensive form public goods game of perfect information like the Coase problem we saw before. Assume that each player $i \in \{1,2\}$ has to choose some x_i from a range $[\underline{x}_i, \overline{x}_i] \subset \mathbb{R}^1$. Assume that each player has a twice continuously differentiable utility function $u_i(x_i, x_j)$ with $\frac{\partial^2 u_i}{\partial x_i \partial x_i} < 0$. Assume further that each player benefits from the contribution of the other, $\frac{\partial u_i(x_i,x_j)}{\partial x_j} > 0$, but only benefits from their own contributions within certain ranges; in particular we assume that $\frac{\partial u_i(\underline{x}_i,x_j)}{\partial x_i} > 0$, and $\frac{\partial u_i(\overline{x}_i,x_j)}{\partial x_i} < 0$ for all feasible x_j. This will guarantee that the solution to the players' maximization problems are interior solutions.*

Note that we have not yet made any assumptions about $\frac{\partial^2 u_i}{\partial x_i \partial x_j}$.

Let's see if we can work out what the impact of the structure of the game and the order of play is on the players' actions and how this might depend on the values of $\frac{\partial^2 u_i}{\partial x_i \partial x_j}$.

So, as always, to solve the game, go to the end of the game tree and consider player 2's choice of x_2, conditional upon whatever value of x_1, say x_1^*, that player 1 chooses. Now, if there is an interior optimum, then given the concavity of Player 2's optimization problem, we have that at the optimum the following first order condition should be satisfied:

$$\frac{\partial u_2(x_2^*, x_1^*)}{\partial x_2} = 0 \tag{4.1}$$

The problem we have is that with these general functional forms we cannot write down explicitly what value of x_2 satisfies this condition. However, if we can assume that there exists a value that satisfies the condition, then we can make use of "implicit function theorem" to make statements about how that value depends on x_1. The "implicit function theorem" tells us that, even if we can't solve for x_2, there does exist a function $x_2(x_1)$, such that $x_2(x_1^*) = x_2^*$ and :

$$\frac{dx_2(x_1^*)}{dx_1} = -\frac{\frac{\partial^2 u_2(x_2^*,x_1^*)}{\partial x_1 \partial x_2}}{\frac{\partial^2 u_2(x_2^*,x_1^*)}{\partial x_2 \partial x_2}} \tag{4.2}$$

This result already tells you about how Player 2's optimal actions are affected by 1's actions: It means that if $\frac{\partial^2 u_2}{\partial x_2 \partial x_2} < 0$ *then, around the optimum, the sign of* $\frac{dx_2(x_1^*)}{dx_1}$ *is the same as the sign of* $\frac{\partial^2 u_2(x_2^*, x_1^*)}{\partial x_1 \partial x_2}$*. This is somewhat intuitive but still useful: it means that if the marginal utility to 2 of taking an action is greater the higher Player 1's action, then, when she optimizes, she will choose higher values of* x_2 *whenever* x_1 *is higher.*

Now, let's go up the game tree to Player 1's optimization problem.

Player 1 wants to maximize $u_1 = u_1(x_1, x_2)$ *but knows that, in equilibrium,* x_2 *will be determined by the function we have identified above, that is, by* $x_2 = x_2(x_1)$*.*

So to work out her optimal strategy she totally differentiates u_1*.*

Total differentiation then gives:

$$du_1 = \frac{\partial u_1(x_1, x_2(x_1))}{\partial x_1} dx_1 + \frac{\partial u_1(x_1, x_2(x_1))}{\partial x_2} dx_2 \tag{4.3}$$

and so:

$$\frac{du_1}{dx_1} = \frac{\partial u_1(x_1, x_2(x_1))}{\partial x_1} + \frac{\partial u_1(x_1, x_2(x_1))}{\partial x_2} \frac{dx_2}{dx_1} \tag{4.4}$$

We have then (substituting from above) that if both players have interior solutions to their maximization problems, we need:

$$\frac{\partial u_1(x_1^*, x_2(x_1^*))}{\partial x_1} + \frac{\partial u_1(x_1^*, x_2(x_1^*))}{\partial x_2} \left[-\frac{\frac{\partial^2 u_2(x_2^*, x_1^*)}{\partial x_1 \partial x_2}}{\frac{\partial^2 u_2(x_2^*, x_1^*)}{\partial x_2 \partial x_2}} \right] = 0 \tag{4.5}$$

Bringing our results together then we have that the equilibrium strategies (x_1^*, x_2^*) *satisfy:*

$$\frac{\partial u_2(x_2^*, x_1^*)}{\partial x_2} = 0 \tag{4.6}$$

and

$$\frac{\partial u_1(x_1^*, x_2(x_1^*))}{\partial x_1} + \frac{\partial u_1(x_1^*, x_2(x_1^*))}{\partial x_2} \left[-\frac{\frac{\partial^2 u_2(x_2^*, x_1^*)}{\partial x_1 \partial x_2}}{\frac{\partial^2 u_2(x_2^*, x_1^*)}{\partial x_2 \partial x_2}} \right] = 0 \tag{4.7}$$

Let's now use this condition to answer some questions:

Question*: Do players contribute more of less in the subgame perfect equilibrium of the extensive form game, relative to the Nash equilibrium of the normal form game? Does the order of play matter?*

Answer*: it depends on the sign of* $\frac{\partial^2 u_2(x_2^*, x_1^*)}{\partial x_1 \partial x_2}$*. And we can work out how.*

The conditions for equilibrium in the normal form game are that the equilibrium strategies, call them (x_1^{n*}, x_2^{n*}), satisfy:

$$\frac{\partial u_1(x_1^{n*}, x_2^{n*})}{\partial x_1} = 0 \tag{4.8}$$

and

$$\frac{\partial u_2(x_2^{n*}, x_1^{n*})}{\partial x_2} = 0 \tag{4.9}$$

Now, with $\frac{\partial^2 u_2(x_2^*, x_1^*)}{\partial x_1 \partial x_2} = 0$, this is clearly the same condition as the condition for the subgame perfect equilibium. And so order of play makes no difference nor does the sequential structure. Not so however when $\frac{\partial^2 u_2(x_2^*, x_1^*)}{\partial x_1 \partial x_2} \neq 0$. To see why consider the following claim:

Claim 50 *Let (x_1^{n*}, x_2^{n*}) denote the equilibrium in the simultaneous game. If $\frac{\partial^2 u_2}{\partial x_1 \partial x_2} = 0$, and $\frac{\partial^2 u_1}{\partial x_1 \partial x_2} > 0$ then $(x_1^*, x_2^*) = (x_1^{n*}, x_2^{n*})$. If $\frac{\partial^2 u_2}{\partial x_1 \partial x_2} > 0$, and $\frac{\partial^2 u_1}{\partial x_1 \partial x_2} = 0$ then $(x_1^*, x_2^*) \gg (x_1^{n*}, x_2^{n*})$.*

Remark 51 *The first part of the claim is what we saw already: if the optimal actions of the second player are independent of the actions of the first player, then the sequential structure adds nothing. So let's look at the second part: the second part tells us both that the subgame perfect Nash will be different to the Nash of the normal form game and that the order of play matters.*

Proof. (Of the second part) If $x_2^* > x_2^{n*}$, then with $\frac{\partial u_2(x_2^*, x_1^*)}{\partial x_2} = \frac{\partial u_2(x_2^{n*}, x_1^{n*})}{\partial x_2} = 0$ and $\frac{\partial^2 u_2(x_2^*, x_1^*)}{\partial x_1 \partial x_2} > 0$ we must have $x_1^* > x_1^{n*}$.[4] Similarly if $x_1^* > x_1^{n*}$ we must have $x_2^* > x_2^{n*}$. Hence either $(x_1^*, x_2^*) \gg (x_1^{n*}, x_2^{n*})$ or $(x_1^*, x_2^*) \leq (x_1^{n*}, x_2^{n*})$. Assume that $(x_1^*, x_2^*) \leq (x_1^{n*}, x_2^{n*})$, then, since $\frac{\partial u_1}{\partial x_1}$ is decreasing in x_1, but $\frac{\partial^2 u_1}{\partial x_1 \partial x_2} = 0$, we have $\frac{\partial u_1(x_1^*, x_2^*)}{\partial x_1} \geq \frac{\partial u_1(x_1^{n*}, x_2^{n*})}{\partial x_1}$. In that case, since $\frac{\partial u_1(x_1^*, x_2^*)}{\partial x_1} + \frac{\partial u_1(x_1^*, x_2^*)}{\partial x_2}\left[-\frac{\frac{\partial^2 u_2(x_2^*, x_1^*)}{\partial x_1 \partial x_2}}{\frac{\partial^2 u_2(x_2^*, x_1^*)}{\partial x_2 \partial x_2}}\right] = \frac{\partial u_1(x_1^{n*}, {x_2}^{n*})}{\partial x_1} = 0$ we must have that $\frac{\partial u_1(x_1^*, x_2^*)}{\partial x_2}\left[-\frac{\frac{\partial^2 u_2(x_2^*, x_1^*)}{\partial x_1 \partial x_2}}{\frac{\partial^2 u_2(x_2^*, x_1^*)}{\partial x_2 \partial x_2}}\right] \leq 0$. However since $\frac{\partial u_1}{\partial x_2} > 0$, $\frac{\partial^2 u_2}{\partial x_1 \partial x_2} > 0$ and $\frac{\partial^2 u_2}{\partial x_2 \partial x_2} < 0$, we have $\frac{\partial u_1(x_1^*, x_2^*)}{\partial x_2}\left[-\frac{\frac{\partial^2 u_2(x_2^*, x_1^*)}{\partial x_1 \partial x_2}}{\frac{\partial^2 u_2(x_2^*, x_1^*)}{\partial x_2 \partial x_2}}\right] > 0$, a contradiction. ■

[4] [Smaller steps] To see this, note that $\frac{\partial u_2(x_1, x_2)}{\partial x_2}$ is strictly decreasing in x_2, but strictly increasing in x_1, so if $x_2^* > x_2^{n*}$ then $\frac{\partial u_2(x_2^*, x_1^*)}{\partial x_2} < \frac{\partial u_2(x_1^*, x_2^{n*})}{\partial x_2}$, but if $x_1^* \leq x_1^{n*}$ then $\frac{\partial u_2(x_1^*, x_2^{n*})}{\partial x_2} \leq \frac{\partial u_2(x_2^{n*}, x_1^{n*})}{\partial x_2}$ and so $\frac{\partial u_2(x_2^*, x_1^*)}{\partial x_2} < \frac{\partial u_2(x_1^*, x_2^{n*})}{\partial x_2} \leq \frac{\partial u_2(x_2^{n*}, x_1^{n*})}{\partial x_2} \rightarrow \frac{\partial u_2(x_2^*, x_1^*)}{\partial x_2} < \frac{\partial u_2(x_2^{n*}, x_1^{n*})}{\partial x_2}$, contradicting $\frac{\partial u_2(x_2^*, x_1^*)}{\partial x_2} = \frac{\partial u_2(x_2^{n*}, x_1^{n*})}{\partial x_2} = 0$.

Remark 52 *The logic of the second part is illustrated in Figure 4.6. In the figure, Player 2's best response function* is *the implicit function. The intuition is that with* $\frac{\partial^2 u_2(x_2^*,x_1^*)}{\partial x_1 \partial x_2} > 0$*, Player 1 will put more in precisely so that this will increase the amount that the second player puts in. The logic is reversed however for* $\frac{\partial^2 u_2(x_2^*,x_1^*)}{\partial x_1 \partial x_2} < 0$*, in this case Player 1 will contribute less in equilibrium than she would were she non-strategic.*[5] *The intuition is that the more I put in, the less that player 2 will put in and so the less I benefit, on the margin, from my contribution. This amplifies the Prisoner's dilemma aspect of a public goods game of this form.*

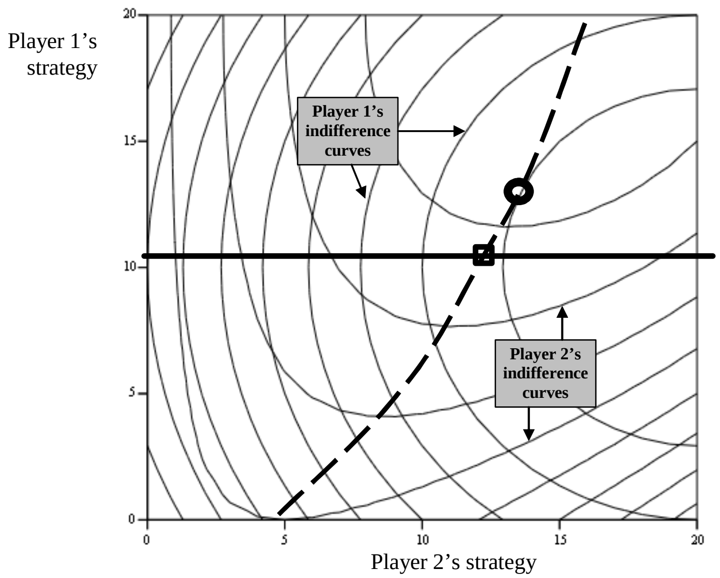


FIGURE 4.6. The dashed line is Player 2's best response function to Player 1; the solid (horizontal) line is 1's best response to Player 2's satrategy. the intersection is a Nash equilibrium (marked with a bold square). But if Player 1 moves first and player 2 responds, the equilibrium will be the point marked with a bold circle.

This example illustrates how, using relatively little information on preferences, we can make substantively interesting statements. In this case about

[5] To see this, let's take the first case. Notice that with $\frac{\partial^2 u_2(x_1^*,x_2^*)}{\partial x_1 \partial x_2} < 0$, the term in square brackets is positive and this makes the marginal contribution of x_1 to player 1's welfare more negative everywhere. For the condition to hold, the more negative is $-\left[\frac{\partial u_1(x_1^*,x_2(x_1^*))}{\partial x_2}\frac{\frac{\partial^2 u_2(x_1^*,x_2^*)}{\partial x_1 \partial x_2}}{\frac{\partial^2 u_2(x_1^*,x_2^*)}{\partial x_2 \partial x_2}}\right]$ the larger must $\frac{\partial u_1(x_1^*,x_2(x_1^*))}{\partial x_1}$ be; but with $\frac{\partial^2 u_2}{\partial x_1 \partial x_1} < 0$, this requires that x_1 be lower.

how the sequential structure of the game affects equilibrium strategies, and about how it matters who, if anyone, goes first.

4.7 Methods to Make Proof-writing Easier (Without Loss of Generality...)

You will often find a line in the middle of a proof that says "assume without loss of generality that..." or simply "assume w.l.o.g. that..." followed by what may appear to be an unjustified simplification of the model or unjustified adjustment of the "givens" of a proposition.

The idea behind the use of w.l.o.g. is that you can make assumptions about the values or behavior of parts of your model that make proving propositions clearer and easier without altering the structure of the problem.

Warning: To invoke w.l.o.g. you need to be sure that the adjustments you make do not alter the structure of the problem. If it is not obvious that the adjustment is without loss of generality you should add an explanation for why it is in fact w.l.o.g.

In practice w.l.o.g. is often invoked inappropriately. Here are a couple of guidelines on when w.l.o.g. can be invoked.

W.l.o.g. can be invoked to normalize utility functions. Or, often, to assume that one quantity is at least as great as another (whenever it is the case that corresponding results follow whenever the opposite inequality holds). From our discussion of Von Neumann Morgenstern utility functions we should feel comfortable normalizing the utility functions of all players such that their utility at some outcome, such as the status quo, sq, is given by some arbitrary number (e.g. 0). Having done this we can further normalize by fixing the utility of any player at some other outcome, such as his ideal point, y to 1, provided that the player prefers y to sq.[6] We can not, w.l.o.g., impose an arbitrary monotonically increasing transformation of the utility function and then proceed to evaluate a player's preferences over lotteries.

In spatial games we often have degrees of freedom in labelling the axes and the orientation of the policy space. Hence for example if there are two players with distinct ideal points in the interior of $\mathbb{R}^2$, we can, w.l.o.g., relabel the space so that one player has an ideal point at $(0, 0)$ and the other's ideal is at $(1, 1)$. We cannot do so, w.l.o.g., if the ideal points are not necessarily distinct or if the labels of the axes have meaning in the context of the model (as they do in models of structure-induced equilibrium). We cannot arbitrarily assume values for the ideal points of 3 players in $\mathbb{R}^1$ (but can we

[6] The point is that with Von Neuman Morgenstern utilities we have two degrees of freedom to play with for each individual—an intercept term and a slope term.

in $\mathbb{R}^2$?). If X is a convex and compact subset of $\mathbb{R}^1$ we can assume, w.l.o.g., that $X = [0,1]$. If X is not necessarily compact or convex, we cannot make this assumption, w.l.o.g. If two players have elliptical indifference curves, we can assume, w.l.o.g., that one of them has circular indifference curves; but we cannot assume, w.l.o.g., that they both do.

If we know that bargaining outcomes are efficient and independent of irrelevant alternatives, then we can restrict our attention in two person bargaining over a multidimensional space to bargaining over a one dimensional space, w.l.o.g., we cannot do so, w.l.o.g., if there are three players or if irrelevant alternatives matter.

Problem 53 *Go to JSTOR and select all political science journals. Search for all articles containing the text "without loss of generality." You should find at least 120 hits. Randomly sample 4-5 hits from this list and go to the "page of first match." For each of these hits try to satisfy yourself whether or not the w.l.o.g. assumption is indeed w.l.o.g.*

Exercise 54 *A very common w.l.o.g. assumption is that in voting over outcomes arrayed on a single dimension in which all players have single peaked preferences, collectivities (such as legislatures) can be represented by the median of those collectivities. Is this assumption w.l.o.g.? In particular is it the case that the preferences of the majority of a legislature over any two options in* $[0,1]$ *will be the same as the preferences of the median of the legislature over those two options?*

4.8 Readings for Next Week

The Varian reading is unavoidable, it's excellent and extremely useful. The Davies piece, it's very short and also fun, and a good model for our classroom discussion in later weeks... It is not however top priority. The Starmer piece is well worth while, it gives a thorough treatment of the representation of preferences by VNM utility functions, as well as some alternatives to this practice. Make sure you understand the logic of the graphs on p. 340-341, these will be used throughout and contain quite a lot of information. This paper is worth reading in part because it shows how open this field is, even if the concerns raised are concerns that formal modelers typically ignore. if you want more teh first two pieces on the recommended readings should be read at some point over the course of the term and before you write up your paper.

5
What to Prove I: First Define your game

It's a good idea to begin by clearly defining your game (or class of games) and there are conventions for this that you can follow. Following these makes reading your paper easier. It also provides you with a check list to make sure that you have in fact made all the assumptions you need in order to get going. And finally this check list can be used as a tool for you to scrutinize the set of assumptions that you have imposed on your model. The items on your list should typically include your assumptions about the game's: Players, Action Sets, Payoffs and Information (PAPI). Exactly what you need will depend on the form of the game:

5.1 Normal Form Games (Games in Strategic Form)

5.1.1 Definition

To define a "**normal form**" (or, equivalently a "**strategic form**") game you need to define three sets:

- A **set of players** N. The set of players is normally a *finite* set, for example: "let $N = \{1, 2, ..., n\}$ denote the set of players...." It can however sometimes be useful to represent a large population with a continuum.[1] Typically the set of players are individuals; they may however be groups or other entities if those entities satisfy whatever behavioral or rationality assumptions your model requires (and Arrow's theorem warns you that assuming they do is not an easy assumption).

- A **set of feasible strategies** (that includes pure as well as mixed strategies) for each player A_i. Again this can be finite ("Let Player i's strategy set be given by $A_i = \{L, R\}$") or infinite ("Let Player i's strategy set be given by $A_i = [0, 1]$"). Common impositions for infinite strategy sets are that they be *closed* and *bounded* and that they be convex.
- Either a **set of preference relations** ($\succsim_i$) or a **set of utility functions** (u_i) that represent the players' attitudes to the outcomes that result from the different choices of actions. Each utility function maps from $A_1 \times A_2 \times ... \times A_n$ into $\mathbb{R}^1$.

These sets provide sufficient information to represent the game. Depending on the representation of preferences, the normal form game, G, is completely described by the triples:

$$G = \langle N, (A_i)_{i\in N}, (\succsim_i)_{i\in N}\rangle \qquad \textit{or} \qquad G = \langle N, (A_i)_{i\in N}, (u_i)_{i\in N}\rangle$$

Note that in the above representation I did not include an "**outcome space**." The reason for this is that strategies determine the outcome, hence preferences over outcomes can be represented as preferences over collections of strategies. If however a representation of outcomes is important for your game then you can define an outcome mapping, say $f : A_1 \times A_2 \times ... \times A_n \rightarrow X$, where X, as before, is the set of feasible outcomes and then represent utility functions as a mapping from X into $\mathbb{R}^1$.[2] if you choose to do this then the game may be represented as:

$$G = \langle N, (A_i)_{i\in N}, X, (\succsim_i)_{i\in N}\rangle$$

...where $(\succsim_i)_{i\in N}$ is defined over elements of X.

[1] In this case, typically the distribution of characteristics of N can be represented by a continuous distribution over some possible set of types. For example: "Assume a contimuum of players with ideal points uniformly distributed over the interval [0,1]." Or more generally: "Let the set of possible types be given by $\Theta \subset \mathbb{R}^n$.And let the set of players be represented by a measure space $(\Theta, \mathcal{B}, f)$ where $\mathcal{B}$ is the σ-algebra of Borel subsets of Θ and f is a finite measure with $\int_\Theta df = 1$."

[2] Doing the two step process does not add much but you may find that it provides an easier way to think about the political situation you are trying to model. Warning: if you do this you will need to define the outcome space so that it includes all results from all possible strategies, and this can include lotteries over more primitive outcomes.

5.1.2 Pure and Mixed Strategies

For the game $G = \langle N, (A_i)_{i \in N}, (\succsim_i)_{i \in N} \rangle$, let us assume that the set A_i corresponds to a discrete set of options for player i, which we term i's set of "**pure strategies.**"

Consider now a larger set of strategies corresponding to the set of all possible lotteries, or "**mixed strategies**" over these action sets[3]; these lotteries assign some probability to each pure strategy, with all probabilities summing to 1. We will call the set of such mixed strategies $\Delta(A_i)$, or more simply, Σ_i, and use σ_i to refer to a typical element of Σ_i. In this case, for any action a_i in the player's action set we let $\sigma_i(a_i)$ denote the probability that the player takes action a_i.

Defining $\Sigma = \times_{i \in N} \Sigma_i$, we can then let σ denote a *profile* of mixed strategies where $\sigma \in \Sigma$. It's easy to check that by this definition any pure strategy is itself a mixed strategy (sometimes called a "degenerate mixed strategy"). Finally, we can define a utility function as a function that maps from Σ into $\mathbb{R}^1$.

5.1.3 Illustrating normal form games

Unless the strategy set is non-finite or very complex or there are many players, you can normally represent a normal form game with a payoff (bi-)matrix. Doing so will make the game more transparent for your readers and will typically make the solution more obvious. In a two player game, if the strategy space for each player is not finite but is "one dimensional" you can present a continuous version of the payoff matrix as a 3-d graph whose two dimensional floor (s_i, s_j) represents the combinations of strategies of the two players and whose vertical axis measures the payoffs to each of the two players that correspond to the strategy combinations of the two players (hence the graph would show two surfaces corresponding to $u_i(s_i, s_j)$ and $u_j(s_i, s_j)$).

Example 55 (Prisoners' Dilemma with continuous action space) *Let the set of players be given by $N = \{1, 2\}$. Let the strategy space for each $i \in N$ by $S_i = [0, 1]$ with typical element s_i. Finally, assume that each player has utility function $u_i = 2s_j - s_i^2$. Try to draw a continuous version of the payoff matrix; and then try to get a computer to draw it.*

[3] To distinguish the game $G = \langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$ from the game $G = \langle N, (\Delta(A_i))_{i \in N}, (u_i)_{i \in N} \rangle$, Osborne and Rubinstein refer (briefly) to the latter as the "*mixed strategy extension*" of the former.

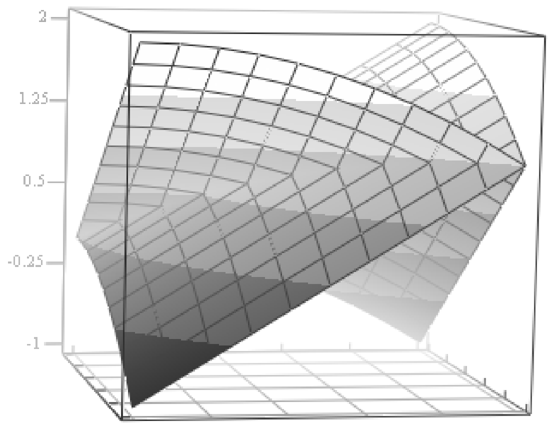


FIGURE 5.1. Representation of a 2 Player Prisoners' Dilemma with Continuous Action Spaces. The graph shows two surfaces, representing player payoffs over continuous action spaces, each player's payoff is increasing in the other person's strategy and decreasing in her own strategy.

5.2 Extensive Form Games

5.2.1 Defining Extensive Form Games of Complete Information

To represent an extensive form game with complete information you need to replace the notion of an action set with the richer notion of a game tree, T. The game tree for extensive form games with complete information consists of a pair $\langle H, P\rangle$ in which:

1. H is a set of sequences or "**histories**" that satisfies a couple of simple nesting properties. To describe these, let us say that a finite sequence h' with k' elements is a "**truncation**" of a possibly infinite sequence h with $k > k'$ if h' consists of the first k' elements of h. Hence $\varnothing$, (a), and (a, b) are truncations of (a, b, c). The nesting properties are then: (1.) if a sequence h is in H, then any truncation of h is also in H. (2.) If all truncations of an infinite sequence h are in H, then h is in H.
   We say that a sequence is "**terminal**" if it is not a truncation of some other history. All infinite sequences are terminal.

1. P, a "**player function**," which indicates which player plays after each non-terminal history in H. Hence $P(h) = 1$ means that Player 1 moves after history h.

Given these definitions we can represent an extended form game of complete information with:

$$G = \langle N, H, P, (\succsim_i)_{i \in N} \rangle \qquad \textit{or} \qquad G = \langle N, H, P, (u_i)_{i \in N} \rangle$$

Note that we have not defined the **action set** as part of the game. It can be defined however from the elements that we have used; in particular the action set of a player i at a history h is given by the set of actions that, when appended to h, are themselves possible histories of the game; formally: $A_i = \{a : (h, a) \in H\}$.

Problem 56 *Consider a two person game of chicken in extensive form in which one player must first decide whether to swerve and the second player decides second. Write out the set H and the function P.*

5.2.2 Defining Extensive Form Games of Incomplete Information

To represent an extended form game with incomplete information we need three extra items. First we add a player, $\mathcal{N}$ (for "nature") to the set of players. Second we assume that $\mathcal{N}$ chooses moves probabilistically rather than strategically. This allows us to represent the possibility that when a player takes action a, she is not sure whether this will produce one history or another (although this may be observed once it has occurred). The second novelty is the introduction of player ignorance about histories *after* they occur—we allow for the possibility that after a sequence h, a given player is not sure whether the sequence h or h' has been followed. These last two additions can be described formally as follows:

- A function c that maps from a history h in which $P(h) = \mathcal{N}$ into a probability measure over $A_{\mathcal{N}}$, with all such probability measures being independent of each other. As an example, the probability that nature chooses $a \in A_{\mathcal{N}}$ after history h (given that $P(h) = \mathcal{N}$) could be written: $c(a|h)$.
- For each player, an information partition $\mathcal{I}_i$ of the set of histories that could lead to player i playing. The different cells of a player i's partition are collections of histories that i can not tell apart. For example say that the set of possible histories leading up to my play is

> $H_i = (h_1, h_2, h_3, h_4, h_5, h_6)$ and that my partition of H_i is $\mathcal{I}_i =[h_1, h_3 \mid h_2 .\mid h_4, h_5, h_6]$, then: should h_1 or h_3 occur, I cannot be sure which of these occurred (although I can tell that none of h_2 or $h_4 - h_6$ occurred); if h_2 occurs I will not confuse it for anything else; but if h_4, h_5 or h_6 occur then I can not be sure of which of them is the real history (but I do know that it is not h_1, h_2 or h_3). The partitions $(\mathcal{I}_i)_{i\in N}$ should have the property that if h and h' are in the same cell of a partition, then $A_i(h) = A_i(h')$—as otherwise if a player knew what actions she had available to her and if different histories resulted in different options, then the player would be able to use this information to work out what history had occurred.

Given these definitions we can represent an extended form game of incomplete information with:

$$G = \langle N, H, P, c, (\mathcal{I}_i)_{i\in N}, (\succsim_i)_{i\in N}\rangle \qquad \textit{or} \qquad G = \langle N, H, P, c, (\mathcal{I}_i)_{i\in N}, (u_i)_{i\in N}\rangle$$

5.2.3 Behavioral Strategies

In our discussion of normal form games we considered the set of all mixed strategies over a player's set of pure strategies. In the study extensive form games however we make use of the idea of a "behavioral strategy": a strategy that specifies what actions a player will take at every point in the game tree (at which she has a choice to make), whether or not that point is in fact reached.[4] A behavioral strategy may involve mixing within an information set, but it is not necessarily equivalent to a mixed strategy.

To clarify, let's begin with a formal definition:

Definition 57 *A "**behavioral strategy**", σ_i for player i in extensive form game Γ is a collection of probability distributions over the options in each of the player's information sets, $\iota \in I$.*

So for example, if in a game a player has to choose between moving Right or Left at two distinct information sets, $I = \{\iota_1, \iota_2\}$, his strategy, σ_i, may contain two probability distributions, each conditional upon his information set: $\sigma_i = \{\sigma_i(\iota_1), \sigma_i(\iota_2)\}$ where $\sigma_i(\iota_j) = \{Prob(a = L|\iota = \iota_j), Prob(a = R|\iota = \iota_j)\}$.

[4] In some cases then a strategy will specify an action $a(h)$ to be taken by player i after some history, h, even if i's strategy precludes the possibility that h will ever be reached. In this sense a strategy differs from a plan of action.

Note that sometimes p_i or π_i is used rather than σ_i to distinguish "behavioral" from "mixed strategies." The difference between the two is somewhat subtle and is discussed in the next remark.

Note also: for all discussions of extensive form games to follow, we will consider only behavioral strategies and define utilities directly over these strategies.

Remark 58 *In an extensive form game a "behavioral strategy" is a collection of probability measures, one for each information set reached by player i; a "mixed strategy"is a single probability measure over the set of all pure strategies. In games of "perfect recall"*[5]*, mixed strategies have exactly one behavioral strategy representation; behavioral strategies have one—but possibly more than one—mixed strategy representation.*

For games of perfect recall the Nash equilibria concepts are equivalent also. But a Nash equilibrium in behavioral strategies is not necessarily the same as a Nash equilibrium in mixed strategies if players do not have perfect recall.

The following nice example is an adaptation of one in Osborne and Rubinstein (Figure 214.1) and should help clarify the difference (if you look at it long enough). It might also convince you that there may be interest in studying one person games. Here it is: consider the game in Figure 5.2. Player i has a unique information set. If he follows pure strategy "L" he receives payoff 0 (he takes the first left). When he follows pure strategy S he also receives payoff of 0 (missing his turn). His payoff from every mixed strategy in this game is derived from mixing between these payoffs from his use of pure strategies and so gives a payoff of 0 (that is, any weighted average of 0 and 0 is 0). This however is not the same as what he could achieve by employing a behavioral strategy that involves mixing over the elements in his information set. In this case if his strategy is to take Left with probability p each time he has a choice to make, he can expect to gain a payoff of $\Pr(1-p)$*: a payoff that is maximized with* $p = .5$ *and that produced an expected payoff of* $.25$.

5.2.4 Illustrating Extensive Form Games

Unless the game tree is either crystal clear or impossibly complex, *draw it out.* The game tree helps in your thinking, it also helps the reader a lot, especially readers during talks who want to get a quick overview of the

[5] In games of perfect recall a player remembers past actions that he has taken, graphically this means that if a player's actions distinguish between two nodes, then those two nodes should not be in the same information set.

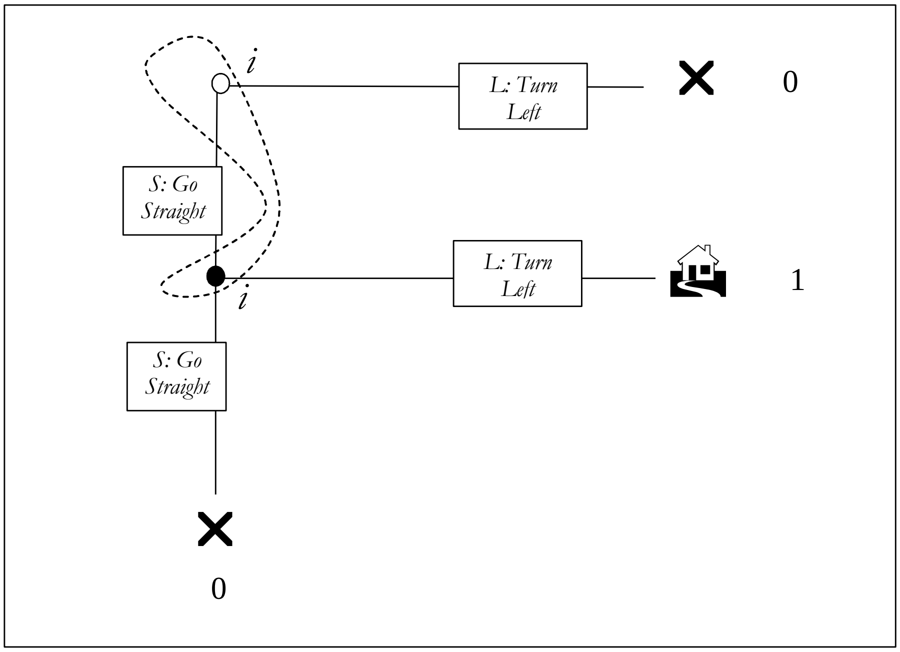


FIGURE 5.2. Memoryless man trying to get home.

game. Once you have a game tree a particular sequence h in H is just a series of connected branches, starting at the beginning of the game tree and moving some way down. A terminal history is any path from the beginning to the end of the tree. The function P can be represented with a label at each node indicating who chooses at that node. it is good practice to:

- have a hollow circle, ○, at the beginning of the game tree
- have a filled circle, at each of the other nodes (including the terminal nodes)
- have time moving in a single direction (e.g. from top to bottom, from left to right[6]
- indicate the name of the player that moves at every information set
- indicate the label of the action taken by the player at every branch leading out of every node

[6] Although in games with complicated information sets it is sometimes clearest to have time moving in multiple directions so that information sets do not have to do tricky maneuvers to avoid each other.

- indicate the payoffs in a vector at the end of the game tree with the payoffs ordered in the order in which the players move. If there are complex move orderings write the names of the players beside the payoffs.
- if players choose from a continuous set, indicate their options with a cone
- for information sets that contain multiple nodes, envelope the nodes inside a dotted oval; if there are just two nodes in an information set you can simply connect them with a dotted line

To illustrate particular equilibria on a game tree:

- If pure strategies: use a thick line to mark the branches specified by the equilibrium both on and off the equilibrium path; the thick line may even contain an arrow.
- If there is mixing at some information sets: indicate the probability with which each branch is played in equilibrium beside the relevant branch
- if the solution concept requires the specification of a player's beliefs regarding which node within an information set she is at: indicate the beliefs (a number between 0 and 1) with a number in square brackets beside the node

5.3 Coalitional Games

5.3.1 Coalitional Games with non-transferable utility

To define a "**coalitional game**" you need, alongside N, X and $(\succsim_i)_{i \in N}$ (defined over X):

- An outcome function f, that assigns elements of X to each non-empty subset of N.

Given this, the game is defined by:

$$G = \langle N, X, f, (\succsim_i)_{i \in N} \rangle$$

Note that in this game description there is no action set describing what each individual does: the idea is that what goes on inside of a coalition is unmodelled, instead it is assumed that coalitions can solve their problems in various ways that players inside the coalition may value differently.

5.3.2 Coalitional Games with Transferable Utility

To define a "**coalitional game with transferable utility**" you only need, alongside the set of players N, a "value function," v, that associates a number with every subset of N. Hence the game may be written:

$$G = \langle N, v \rangle$$

Note that you do not need an action space or a set of player preferences or a game tree. The reason is that in coalitional games you can abstract away from the individual strategies of the players and focus instead on the maximum possible values that can be achieved by coalitions (ignoring how they actually achieve them). Having"transferable payoffs" means that an individual can "give" any amount of "his utility" to another player, much like he can transfer money. In effect by assuming this you can work in utility space and you have already captured all relevant information about utility (since by definition players prefer more utility to less utility).

6

What to Prove II: Making Assumptions, Making Points

6.1 Weak Assumptions, Strong Theories

If you make the right assumptions you can prove anything. If you believe a result is true and want to prove it, one (albeit inelegant), approach is to make as many assumptions as you need to prove it and then chip away at your assumptions.

Whatever approach you use, once you have a version of your result, you then want to make sure that:

1. The assumptions you have are as **weak** as possible

2. Only **plausible** assumptions are driving the results and

3. The **distance** between your assumptions and your conclusions is as great as possible

I consider these in turn.

1. Why do you want *weak* assumption? Let's begin by defining what we mean by *weak* and *strong* (relatively speaking):

Definition 59 *Assumption $\mathcal{A}$ is stronger than Assumption $\mathcal{B}$ if $\mathcal{A}$ implies $\mathcal{B}$ but $\mathcal{B}$ does not imply $\mathcal{A}$. Proposition $\mathcal{A}$ is stronger than Proposition $\mathcal{B}$ if $\mathcal{A}$ implies $\mathcal{B}$ but $\mathcal{B}$ does not imply $\mathcal{A}$.*

Given these definitions you can see that: *the weaker your assumptions, the stronger your theory.* Hence consider two sets of assumptions, a stronger set A_s and a weaker set A_w, $A_s \rightarrow A_w$. (e.g. A_s "John is in Manhattan" A_w "John is in New York"; note that $A_s \rightarrow A_w$ but $A_s \nleftarrow A_w$). Now say that we have the choice between two propositions, P_1: $A_s \rightarrow B$ and P_2 : $A_w \rightarrow B$. Which of P_1 and P_2 is the stronger proposition? Clearly P_2 is stronger since if P_2 is true then we have both $A_s \rightarrow A_w$ and $A_w \rightarrow B$ and hence $A_s \rightarrow B$. But this is just what P_1 says. Hence $P_2 \rightarrow P_1$ Since we do not have $P_1 \rightarrow P_2$ we have that P_2 is stronger than P_1.

Example 60 *(Preferences)*

Proposition 61 *Assume Players 1 and 2 have preferences over points in* $\mathbb{R}^2$ *representable by utility functions* $u_i(x) = -(|p_i - x|)^2$ *for* $i \in (1,2)$ *and where* $p_i \in \mathbb{R}^2$ *for* $i \in (1,2)$*. Then Player 1 and Player 2's contract curve will be a straight line.*

Proposition 62 *Assume Players 1 and 2 have preferences over points in* $\mathbb{R}^n$ *representable by utility functions* $u_i = u_i(|p_i - x|)$ *for* $i \in (1,2)$ *where* $p_i \in \mathbb{R}^n$ *for* $i \in (1,2)$ *and* $u_i' < 0$*. Then Player 1 and Player 2's contract curve will be a straight line.*

In this example, Proposition 61 makes stronger assumptions about preferences than Proposition 62. Note that $u_i(x) = -(|p_i - x|)^2 \longrightarrow u_i = u_i(|p_i - x|)$, $u_i' < 0$; but $u_i(x) = -(|p_i - x|)^2 \nleftarrow u_i = u_i(|p_i - x|)$, $u_i' < 0$. Hence if you are willing to assume quadratic preferences, (the assumption you need to make to employ Proposition 61) then you could employ Proposition 62 to get the result. If on the other hand you are only willing to assume "Euclidean preferences" (the assumption you need to make to employ Proposition 62) then you can't be sure that you can employ Proposition 61. Equivalently, if Proposition 62 is true, then Proposition 61 is true, but not vice-versa.

Problem 63 *Arrow's Theorem states that there is no way, other than through dictatorship, to aggregate a collection of rational preference profiles into a rational preference profile that does not violate Universal Domain, the Pareto Principle or Independence of Irrelevant Alternatives. Recall that "rational" implies complete and transitive. Now, consider a theorem that made the same claim as Arrow's but that did not* ***require*** *individual preferences to be transitive. Would this be a stronger or a weaker theorem? Why?*

Exercise 64 *We defined transitivity of the weak preference relation* $\succcurlyeq$ *as the property that if* $x \succcurlyeq y$ *and* $y \succcurlyeq z$ *then* $x \succcurlyeq z$*. Consider a rival notion of transitivity, call it q-transitivity, as the property that if* $x \succ y$ *and*

$y \succ z$ then $x \succ z$. Is the assumption of q-transitivity a stronger or a weaker assumption that the assumption of transitivity? Prove it. Does Arrow's impossibility result obtain if we replace the requirement of transitivity with the requirement of q-transitivity? Prove it.

2. Only plausible assumptions should drive your results. This does not mean that you should not have implausible assumptions, but rather that your key results should not depend on the implausible assumptions.

Implausible assumptions are typically needed to make models models. The problem is that there is always, always a temptation to add more realism into a model, sometimes by adding extra players, extra strategies, sometimes by adding greater "realism" such as by adding uncertainty or piling up extra subgames. You have to avoid these temptations. Models can *very* quickly become unsolvable. Even if they are solvable they can quickly become uninterpretable. Even if they are solvable and interpretable, their central insights and lessons can quickly become obscured. And an unsolved or uninterpretable or aimless model is not much good to anyone.[1]

This then is a form of the principle of parsimony: a model should be as simple as possible in order to describe the aspect of the problem that you care about—*but no simpler*! The tricky thing is to know if your model is "too simple." The key test is to ask the question:

> Does the exclusion of some piece of realism result in qualitatively different conclusions than if that piece of realism is introduced?

If not, then keep it simple. If so, then the exclusion is not legitimate.[2]

Example 65 *The median voter theorem for electoral politics assumes (a) that candidates are office seekers and (b) that candidates know with certainty the preferences of the electorate. Both assumptions are unreasonable, although how unreasonable they are varies across contexts. Does this matter? Yes; at least in the following way. We can show that relaxing just assumption (a) has no affect on the model's results; neither does relaxing assumption (b). Relaxing both assumptions at the same time may however lead to divergence in the platforms selected by candidates in equilibrium rather than convergence to the median. Hence if we are concerned with political contexts where assumptions (a) and (b) are unreasonable, then the standard median voter model is inappropriate.*

[1] Unless, perhaps you can prove that the problem has no solution...

[2] Hence Occam's version of the principle of parsimony: *if two theories explain a problem equally well*, choose the simpler. The principle does not state that simpler models are (necessarily) preferable to more complex models when the results of the two differ.

FIGURE 6.1. Maximize the distance between your assumptions and your results.

3. Maximize the "**distance**" between your assumptions and your conclusions. The idea is this: when you sell a model, the product people receive is your set of conclusions and the price they pay is your set of assumptions (that is, what they have to swallow in order to get the conclusions). If they're going to feel like they've got a good deal then they need to see a big difference between the price they pay and the product they get. *Never* present a proposition that is "true by assumption."

In short:

- If you think your model is too complex, remove the flab. Remove parts of the model that appear not to be doing any of the work and check whether your results change.
- If your model is simple but you can show that it's not *too* simple: prove your results in the simple case and then show in an appendix that the results are robust to changes in assumptions. How much goes in to the appendix and how much in the main text will depend on your audience.
- If your model is *too* simple: if feasible, write the more complex model such that the simple model is a special case of the complex model. You can then show how the simple model's results are altered as a function of the introduction of complexity.
- If in fact your problem and results are *so* simple that they can be expressed clearly enough in words rather than in equations, then write in words and expand your audience.

Problem 66 *Read Chapter 22 of Rasmussen, "Shooting the Bird's Eye."*

6.2 Solution Concepts

After you have designed a game—specified the actors, the possible strategies, the information available to them, their preferences and so on—you will move into analyzing it. Analysis may describe actions that you might expect individuals to take in response to particular actions by others, normative properties of the players' behavior, the information that players may gather over the course of a game and so on. Traditionally however a centerpiece of your analysis will be your attempt to *solve* the game and then describe the properties of the solution. To do this you need a solution concept.

A "**solution concept**" is a rule for predicting how players will play a game. More formally (loosely following Myerson) it can be described as a mapping ϕ from some set of games, Γ, into some set, B, of descriptions of admissible behaviors. For example, Γ could be the set of all finite normal form games and B could be the set of all randomized strategy profiles.

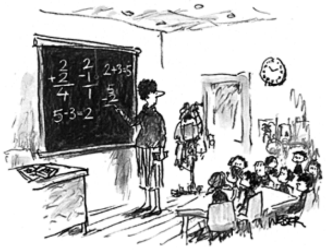
We will see that there is a very large range of solution concepts available to a modeller—even for the set of finite normal form games and randomized strategy profiles. In selecting a solution concept then, modellers first need some sense of what a "good" solution is. What makes a solution concept good is up for grabs, but typically solution concepts should be *accurate* in the sense that they select outcomes that are "reasonable" according to some stated criterion and not identify outcomes that are not reasonable.[3] After accuracy, *precision* is often desired: we can say that one solution concept is more precise than another if the predictions of the former are a subset of the predictions of the latter; ideally the solution concept should predict few rather than many possible outcomes. A third property, particularly of use for empirical work, is *existence*—falsifying a prediction is difficult if the solution concept does not predict anything.[4] A good solution concept should be *insensitive to irrelevant information* and in particular it should not depend on the description of the problem and on the method employed to identify the solution (some reasonable-sounding solution concepts fail this last condition). Finally a desirable property of the solution concept is that it be *strong* in the sense that it only requires weak assumptions

[3] These two requirements of accuracy are quite distinct: Following Myerson we say that a solution, ϕ is a "**lower solution**" if for any element, p, of ϕ there exists an environment where p is an accurate prediction. A lower solution may however fail to identify good predictions. In contrast ϕ is an "**upper solution**" if for any, p, that is *not* an element of ϕ there is no environment where p is an accurate prediction. Upper solutions may include predictions that are never accurat predictions, but excludes some class of inaccurate predictions. Finally, a solution, ϕ is an "**exact solution**" if it is both an upper solution and a lower solution.

[4] As you read through the discussions on the merits and the demerits of the core as a solution concept, a pertinent question to ask yourself is whether you buy into the idea that existence is a desirable property.

about human behavior. I don't know of any theorem that says that it is impossible to have a solution concept that satisfies all these properties, but in any case when choosing among concepts you'll typically have to make trade-offs among these criteria.

In future weeks we work through a menu of possible solution concepts and identify some of the merits and demerits of each, as well as figuring out how to use each one. We begin with normal form games, then we consider extensive form games and cooperative games. We hold back on a detailed discussion of solution concepts for games where imperfect information is an important feature until after we introduce techniques for studying these games.



"Please, Ms. Sweeney, may I ask where you're going with all this?"

FIGURE 6.2. From *The New Yorker*

6.3 I've solved the game, what now?

Here is a small and non-exhaustive list of the kind of things that you may want to do once you have settled on a solution concept.

1. Establish the existence of a solution to your game

2. Establish the uniqueness of the solution

3. Or, identify the number of solutions

4. Describe the types of actions / learning / coalitions that occur at the solution

5. Characterize the efficiency properties at the solution

6. If there are inefficiencies identify the source of the inefficiencies

7. Characterize the distributive properties at the solution (how does it compare to one that a social welfare maximizer might choose?)

8. Compare the outcome to what would occur under different behavioral assumptions

9. Perform "comparative statics" by describing how properties of the solution depend on the number of players, parameters in or properties of their utility functions, their strategy sets, information available to them, the structure of the game tree...

10. Identify testable propositions from your model

7
Representing Preferences: A Menu

7.1 Represenation: Preferences and Utility Functions

In many applications it is useful to work with utility functions rather than with preference relations. A question arises however as to whether, or to what extent, a given utility function "represents" a preference relation.

Formally, we say that a utility function $u_i : X \rightarrow \mathbb{R}$ **represents** the preference relation $\succsim_i$ if for all outcomes x, y in X we have: $u_i(x) \geq u_i(y) \leftrightarrow x \succsim_i y$. Under what conditions is such a representation possible?

The following proposition provides one necessary condition:

Proposition 67 *If a utility function u_i can represent $\succsim_i$ then $\succsim_i$ is rational.*

Exercise 68 *Prove this proposition.*

Unfortunately, this proposition does not tell us whether rationality of $\succsim_i$ is a *sufficient* condition for there to be a utility function that can represent $\succsim_{i.}$ In fact, there are rational preference relations that cannot be represented by utility functions. A well-known one of these is the **lexicographic preference relation**. Such a preference relation takes a form like the following: Let $X = \mathbb{R}^2$, with typical elements (x_1, x_2) or (y_1, y_2). Then let $\succsim_i$ be given by $x \succ_i y \leftrightarrow \{x_1 > y_1 \text{ or } \{x_1 = y_1 \text{ and } x_2 > y_2\}\}$ and $x \sim_i y \leftrightarrow x = y$. An example is someone who always prefers more general theories to less general theories, no matter what the level of mathematical complexity, but, conditional upon a given level of generality prefers mathematically simple to mathematically complex theories.

Sufficient conditions do however exist for the problem of representing preferences with utility functions. Here is one useful one:

Definition 69 ***Continuous Preferences****. We say that a preference relation is "****continuous****" if for any sequence $\{x_n\} \subset X$ that converges to a point x in X, and sequence $\{y_n\} \subset X$ that converges to a point y in X, we have that if $x_n \succsim_i y_n$ for all elements of $\{x_n\}$ and $\{y_n\}$ then $x \succsim_i y$.*

Proposition 70 *If $\succsim_i$ is rational and continuous then there exists a continuous utility function u_i that can represent $\succsim_i$.*

Example 71 *An implication of the last proposition is that the lexicographic preferences that we described above must not be continuous. To verify, note that the sequence $x_n = (1, \frac{1}{n})$ has the limit $x = (1, 0)$, whereas the sequence $y_n = (\frac{n-1}{n}, 1)$ converges to $y = (1, 1)$. In this example we have that for all $x_n \succ_i y_n$, however, $y \succ_i x$.*

Health Warning!: The key concern is that if all you impose is rational preference relations, or perhaps continuous and rational preference relations, then the utility numbers that you assign to a set of outcome have almost *none* of the properties that you normally assume for numbers. All that matters is order. It is not meaningful to add two numbers such as $u(x)$ or $u(y)$ together. Nor is it meaningful to compare differences such as $(u(x) - u(y))$ and $(u(z) - u(w))$.

To see the range of possible representations of $\succsim$ let g be any (strictly) monotonically increasing transformation in the sense that $g(a) \geq g(b) \leftrightarrow a \geq b$ for any a, b. And let u_i represent $\succsim_i$. Then clearly $g(u_i(x)) \geq g(u_i(y)) \leftrightarrow u_i(x) \geq u_i(y) \leftrightarrow x \succsim_i y$. But this implies that $g(u(.))$ (or $g \circ u$) represents $\succsim_i$. Hence we have that u_i represents $\succsim_i$uniquely only up to a strictly increasing monotonic transformation.[1]

7.2 Representation: Von Neumann-Morgenstern Utility Functions

Consider now situations of *risk*: that is where individuals are not sure exactly what outcome will occur but she has some priors about what is

[1] An advantage of this for solving problems is that if you only want to represent rational preferences and you do not need to make assumptions about attitudes to risk then you are free to choose any monotonically increasing function of your utility function. Sometimes this is useful for maximization problems. For example if you find $u(x) = \frac{1}{1+e^{-x^3}}$ difficult to work with, then just use $u(x) = x$. There is nothing at stake here.

or is not likely. To capture this idea for situations in which only a finite number of outcomes are likely we use the idea of a lottery. A lottery is an assignment of probabilities to each of some set of states, for example $L = ((x_1, p_1), (x_2, p_2), (x_3, p_3), ..., (x_n, p_n))$ is a lottery in which each state x_i occurs with probability $p_{i.}$ The question is: how should these lotteries be valued?

The expected utility hypothesis–that utility of a lotteries is valued as the expected utility from the lottery is a cornerstone for analyzing choice in these situations. The hypothesis states that we can represent an individual's preferences over the objects in X and lotteries over objects in X with an expected utility function u that is "linear in probabilities." Informally:

$$u(((x_1, p_1), (x_2, p_2), (x_3, p_3), ..., (x_n, p_n))) = \sum_{i=1}^{n} p_i u(x_1)$$

The Starmer article was intended to put the hegemony of this approach into question. In fact the hypothesis has weak theoretical foundations and very weak empirical support. Nonetheless it is attractive because (1) it provides a tractable way to evaluate risky choices and (b) it can be easily derived from innocuous seeming axioms.

Having stated the hypothesis informally, let's now derive it from a set of axioms (the treatment below follows that found in Luce and Raiffa 1957).

Assumptions:

1. (Rationality) The preference relation $\succsim$ over elements in X is rational. The preference relation over lotteries of elements in X is also rational.

2. (Reduction of Compound Lotteries) A compound lottery (a lottery over lotteries over elements in X) is equivalent to a simple lottery over elements in X with probabilities computed according to the ordinary probability calculus.

3. (Continuity) For any outcome in X a player is indifferent between that outcome and some lottery involving the player's most preferred and least preferred outcomes.

4. (Substitutability) If the player is indifferent between an outcome x and a lottery L then L may be substituted for x in any lottery involving X.

5. (Monotonicity) If a the player prefers x to y then given two lotteries between x and y he prefers the one that assigns the higher probability to x.

Proposition 72 *If Assumptions 1-5 hold then there are numbers u_i associated with each outcome in X such that for any any pair of lotteries $L = ((x_1, p_1), (x_2, p_2), ..., (x_n, p_n))$ and $M = ((x_1, q_1), (x_2, q_2), ..., (x_n, q_n))$: $L \succsim_i M \leftrightarrow \sum_{j=1}^{n} p_j u_j \geq \sum_{j=1}^{n} q_j u_j$*

Proof. Without loss of generality assume that x_1 is i's most preferred option and x_n is player i's least preferred option. Consider the lottery $L = ((x_1, p_1), (x_2, p_2), ..., (x_n, p_n))$ (**note** we can also write this more compactly as $((x_j, p_j))_{j=1,2...,n}$).

From **continuity** we have that for each x_j, we have that player i is indifferent between x_j and some lottery that awards x_1 with probability u_j and x_n with probability $1 - u_j$.

From **substitutability** we have that i is indifferent between a lottery, L, and one in which the simple elements of L are replaced with these lotteries over x_1 and x_n Hence:

$$((x_j, p_j))_{j=1,2...,n} \sim_i (([(x_1, u_j), (x_n, 1 - u_j)], p_j))_{j=1,2...,n}$$

By **reducing this compound lottery** we have:

$$((x_j, p_j))_{j=1,2...,n} \sim_i \left[(x_1, \sum p_j u_j), (x_n, 1 - \sum p_j u_j)\right]$$

From **rationality** we then have:

$$\begin{array}{rcl} ((x_j, p_j))_{j=1,2...,n} & \succsim & {}_i((x_j, q_j))_{j=1,2...,n} \\ & \leftrightarrow & \\ \left[(x_1, \sum p_j u_j), (x_n, 1 - \sum p_j u_j)\right] & \succsim & {}_i \left[(x_1, \sum q_j u_j), (x_n, 1 - \sum q_j u_j)\right] \end{array}$$

Finally from **monotonicity** we have that since $x_1 \succsim_i x_n$:

$$\begin{array}{rcl} \sum p_j u_j & \geq & \sum q_j u_j \\ & \leftrightarrow & \\ \left[(x_1, \sum p_j u_j), (x_n, 1 - \sum p_j u_j)\right] & \succsim & {}_i \left[(x_1, \sum q_j u_j), (x_n, 1 - \sum q_j u_j)\right] \end{array}$$

Combining these gives us that there exist numbers $(u_j)_{j=1,2...,n}$ such that:

$$\begin{array}{rcl} \sum p_j u_j & \geq & \sum q_j u_j \\ & \leftrightarrow & \\ ((x_j, p_j))_{j=1,2...,n} & \succsim & {}_i((x_j, q_j))_{j=1,2...,n} \end{array}$$

And this proves the proposition. ■

In applications, utility functions of this form are assumed directly rather than deduced from axioms. The theoretical justification for them nonetheless relies on the plausibility of the axioms (each of which is worth thinking about). Utility functions with this property are then typically termed **Von Neumann-Morgenstern** utility functions or **Bernoulli utility functions**. We will now see why such utility functions are useful in evaluating a player's attitude to risk.

7.3 Utility over a Single Dimension

7.3.1 Attitudes to Risk

For some problems, a von Neumann-Morgenstern utility function provides a measure of an individual's attitude to risk.

Assume that a player has a continuous utility function u over $X \subset \mathbb{R}$. Consider now a lottery given by $L = ((x_1, p), (x_2, 1-p))$. From the above we have that the individual's utility for this lottery is given by: $u(L) = p.u(x_1) + (1-p).u(x_2)$. And from the continuity of u (and the intermediate value theorem) we have that there must exist some point $x^* \in X$ such that $u(x^*) = p.u(x_1) + (1-p)u(x_2)$ (Prove this!). That is, the individual is indifferent between x^* and a lottery between x_1 and x_2. We call x^* the "**certainty equivalent**" of the lottery L.

The value of the certainty equivalent tells us something about the player's attitude to risk. In particular if $x^* < px_1 + (1-p)x_2$ then in some sense the individual is "risk averse"—they prefer a sure thing that has a value lower than the expectation of the lottery.

More generally we define risk aversion, risk seekingness and risk neutrality as follows::

Definition 73 *An individual with utility function u is* ***risk averse*** *if for any $p \in (0,1)$ and outcomes x_1, $x_2 \in X$ we have $u(px_1 + (1-p)x_2) > pu(x_1) + (1-p)u(x_2)$. An individual with utility function u is* ***risk seeking*** *if for any $p \in (0,1)$ and outcomes x_1, $x_2 \in X$ we have $u(px_1 + (1-p)x_2) < pu(x_1) + (1-p)u(x_2)$. An individual with utility function u is* ***risk neutral*** *if for any $p \in (0,1)$ and outcomes x_1, $x_2 \in X$ we have $u(px_1 + (1-p)x_2) = pu(x_1) + (1-p)u(x_2)$.*

Key result: If you refer back to the definitions of a concave and convex function from the beginning you should be able to convince yourself easily that under the assumption of von Neumann-Morgenstern utility the statements "Player i has a (strictly) concave utility function" and "Player i is risk averse" are equivalent. Similarly the statement "Player i has a (strictly) convex utility function" and "Player i is risk seeking are equivalent." Risk neutrality is equivalent to possessing a linear utility function.

Hence information on the concavity or convexity of a utility function is sufficient information to make qualitative statement regarding the risk attitudes of an individual. How though could one begin to quantify an individual's degree of risk aversion / risk seekingness?

One approach is to calculate the difference, $px_1 + (1-p)x_2 - x^*$, between the expectation of a lottery and the quantity of a sure thing that the individual values equivalently. This difference is termed the "**risk premium**"; if positive it means that an individual would forgo some of the expected value of a lottery in order to receive a sure thing instead.

A useful utility function that we will use at a later stage has particularly interesting properties in terms of its implications for risk premiums. The so-called **constant absolute risk aversion** utility function is given by:

$$u(x) = -e^{-\lambda(x)}$$

Among other useful properties, this utility function is *unique* (up to a positive affine transformation) in producing a risk premium that is independent of wealth in the following sense: for any p, x_1, x_2 and w, the risk premium for the lottery $L(w) = ((x_1+w, p), (x_2+w, 1-p))$ is independent of w.

Problem 74 *Prove that if a player has utility function $u(x) = -e^{-\lambda(x)}$ then her risk premium for the lottery $L(w) = ((x_1 + w, p), (x_2 + w, 1 - p))$ is independent of w.*

7.3.2 Functional Forms

For solving particular problems choosing the constant absolute risk aversion utility function (or some other functional form) can be a useful way to proceed. The aim however should be to make the assumptions weaker once you are convinced that your result is true. To help guide you, here is a listing of possible assumptions you might want to impose on utility functions In the case of utility over outcomes that can be arrayed on a single dimension (roughly ordered from weaker to stronger)

- **Smoothness** (that is, u is differentiable)
- **Single peakedness** (if X is the real line then single peakedness obtains if all players have quasiconcave utility: for any point $x \in X$ the set of points that i prefers to x is convex).
- **Monotonicity** (the first derivative) for example $u'(x) > 0$; or $u'(x) \geq 0$ for $x \geq x^*$ and $u'(x) \leq 0$ for $x \leq x^*$.
- **Attitudes to risk** (the second derivative), for example $u''(x) < 0$.;
- **Symmetry** (a strong assumption): there exists a point x such that for all $y \in X$, $u(x + y) = u(x - y)$.

- **Functions with free parameters**: One common family of functional forms over a single dimension is the CRRA (constant relative risk aversion) utility function, given by $u(x) = \frac{x^{1-a}}{1-a}$, for $a \neq 1$. The name follows from the fact that the coefficient of relative risk aversion (given by $x\frac{u''}{u'}$) is in this case constant: ($x\frac{u''}{u'} = x\frac{-ax^{-a-1}}{x^{-a}} = -a$). A simpler function with similar properties is given by $u(x) = x^{\alpha}$ (this may be increasing, decreasing, concave or convex, but still belongs to a restricted family of functional forms; what is it's coefficient of relative risk aversion?)
- **Fully specified functional forms**. Most restrictive are functional forms with fully specified parameters. e.g. For example the CRRA for $a = 0$ is simply $u(x) = x$; in the limit as α approaches 1, it is given by $u(x) = ln(x)$. Another function that is often easy to work with is the case where $\alpha = .5$, in which case $u(x) = 2\sqrt{x}$.

7.4 Preferences over many Dimensions With No Satiation

In the last section we considered utility functions to represent preferences over single dimensional outcome spaces. Broadly analogous considerations are germane for situations in which outcomes are values across multiple dimensions. Commonly such situations are modelled either for situations in which individuals have "spatial" preferences, discussed below, or for problems with "economist" preferences in which player utilities are increasing, without limit, in multiple dimensions, discussed now.

A particularly useful family of utility functions is given by the *Constant Elasticity of Substitution* family.

The Constant Elasticity of Substitution utility function takes the following form:

$$u(x_1, x_2, x_3, ...) = (\alpha_1 x_1^{\theta} + \alpha_2 x_2^{\theta} + \alpha_3 x_3^{\theta} ...)^{\frac{1}{\theta}} \text{ for } \theta \in [-\infty, 0) \cup (0, 1]$$

or more compactly:

$$u(x) = \left(\sum\nolimits_i \alpha_i x_i^{\theta}\right)^{\frac{1}{\theta}} \text{ for } \theta \in [-\infty, 0) \cup (0, 1] \tag{7.1}$$

Note that the function is not defined for $\theta = 0$; in this case though it is possible to define the function in the limit as θ tends towards 0. Doing so (this requires using l'Hôpital's rule) we find that

$$u(x|\theta = 0) = x_1^{\alpha_1} x_2^{\alpha_2} x_3^{\alpha_3} ... \tag{7.2}$$

This is the case of *unit* elasticity of substitution, also known as the **Cobb-Douglas** function. Applying a monotonic transformation to this utility we can see that the following function is also Cobb-Douglas: $u(x) = \alpha_1 \ln(x_1) + \alpha_2 \ln(x_2) + ...$

The other limiting cases are also interesting:

$$u(x|\theta = 1) = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 ... \tag{7.3}$$

$$u(x|\theta = -\infty) = \min(\alpha_1 x_1, \alpha_2 x_2, \alpha_3 x_3 ...) \tag{7.4}$$

From this we can see that three seemingly distinct approaches to welfare can be captured by the CES function: if we set the weights equal to each other, $\alpha_i = \alpha_j$ and think of the variables $x_1, x_2, x_3, ...$ as representing the utility of individuals $1, 2, 3, ...$ then a **Utilitarian** welfare maximize would try to maximize equation 7.3; a **Rawlsian** would attempt to maximize equation 7.4; while someone implementing **Nash**'s axioms for a bargaining solution (see below) would maximize equation 7.2.

The difference between these three functions (and more generally between different types of CES functions) is the degree to which you are willing to trade off loses on one dimension for gains on another. Utilitarians are perfectly willing to do this. The Nash product also does it but is more conservative, Rawlsians refuse to do it and focus only on the weakest link.

Such a willingness to trade across dimensions is referred to simply as the *elasticity of substitution*, described by Hicks (1932) as "a measure of the ease with which the varying factor can be substituted for others." The elasticity of substitution can be defined for any differentiable utility function as follows:

$$\epsilon_{i,j} = \frac{d\left(\frac{x_i}{x_j}\right)}{d\left(\frac{\frac{\partial u}{\partial x_j}}{\frac{\partial u}{\partial x_i}}\right)} \frac{\frac{\partial u}{\partial x_j}}{\frac{\partial u}{\partial x_i}} \frac{x_j}{x_i}$$

To make sense of this awkward expression, we can note that it can be interpreted as [2] the percentage change in the amount of x_i (relative to x_j) that corresponds to a given percentage change in the curve of the utility function around some point x_i, x_j. For a more economic interpretation, it is useful to think of prices that can be associated with the variables x_i; if the price p_i reflects the marginal utility of x_i, as is the case whenever an individual is purchasing the optimal amount of x for a given budget, then this expression can be written: $\epsilon_{i,j} = \frac{d\left(\frac{x_i}{x_j}\right)}{d\left(\frac{p_j}{p_i}\right)} \frac{p_j}{p_i} \frac{x_j}{x_i} = \frac{d\left(\frac{x_i}{x_j}\right)/\frac{x_i}{x_j}}{d\left(\frac{p_j}{p_i}\right)/\frac{p_j}{p_i}} = \frac{d\ln(\frac{x_i}{x_j})}{d\ln(\frac{p_j}{p_i})} =$

[2] Note: $\frac{d\left(\frac{x_i}{x_j}\right)}{d\left(\frac{\frac{\partial u}{\partial x_j}}{\frac{\partial u}{\partial x_i}}\right)} \frac{\frac{\partial u}{\partial x_j}}{\frac{\partial u}{\partial x_i}} \frac{x_j}{x_i} = \frac{d\left(\frac{x_i}{x_j}\right)}{d\left(\frac{dx_i}{\partial x_j}\right)} \frac{dx_i}{dx_j} \frac{x_j}{x_i} = \frac{d\left(\ln\left(\frac{x_i}{x_j}\right)\right)}{d\left(\ln\left(\frac{dx_i}{\partial x_j}\right)\right)}$

$-\frac{d\ln(\frac{x_i}{x_j})}{d\ln(\frac{p_i}{p_j})}$. Hence we see that the elasticity is the percentage *reduction* in the purchase of x_i, relative to x_j, for a given percentage *rise* in the price of x_i (again, relative to the price of x_j).

To see why the CES is called the CES, note that $\epsilon_{i,j} = \frac{d\left(\frac{x_i}{x_j}\right)}{d\left(\frac{\frac{\partial u}{\partial x_j}}{\frac{\partial u}{\partial x_i}}\right)} \frac{\frac{\partial u}{\partial x_j}}{\frac{\partial u}{\partial x_i}} \frac{x_j}{x_i} = \frac{1}{1-\theta}$, which is independent of all the α_i terms as well as the particular values of the variables x_i.[3]

Note that with unit elasticity (Cobb-Douglas), if the price of a good increases by a small percentage, then the quantity consumed goes down by the same percentage, with the result that the total amount spent on the good does not change. Hence, *as the price of one good changes, the demand for another good is unaffected.* In fact, with Cobb-Douglas, the total amount spent on a good is just a share of the total budget (the share spent on good i is given by $\alpha_i/\sum_j \alpha_j$) and is independent of the prices of the good. This feature often simplifies the mathematics greatly but it may be substantively unreasonable.

7.5 Preferences in the Spatial Theory

In most treatments of the spatial theory of politics, and unlike economic models, players are generally assumed to have some point of "global satiation"—a set of policies that they would like to see enacted were they a policy dictator, alternatively referred to as an "ideal point" or a "bliss point." The existence of an ideal point could reflect fundamental disagreements about what types of policies are good. Or they could reflect the way different individuals make trade-offs in situations where they can not get the best of everything. Given the ideal point, the value of different policy options declines as you move "away" from the ideal point of a player.[4]

[3] To get this result note that: $\frac{\partial u}{\partial x_i} = \alpha_i x_i^{\theta-1}(\alpha_1 x_1^\theta + ... + \alpha_i x_i^\theta + ...)^{\frac{1-\theta}{\theta}}$ and hence: $\frac{\frac{\partial u}{\partial x_j}}{\frac{\partial u}{\partial x_i}} = \frac{\alpha_j}{\alpha_i}\left(\frac{x_i}{x_j}\right)^{1-\theta}$. Then since $\frac{x_i}{x_j} = \left(\left[\frac{\alpha_j}{\alpha_i}\left(\frac{x_i}{x_j}\right)^{1-\theta}\right]\frac{\alpha_i}{\alpha_j}\right)^{\frac{1}{1-\theta}}$, it is easy to see that $\frac{d\left(\frac{x_i}{x_j}\right)}{d\left(\frac{\frac{\partial u}{\partial x_j}}{\frac{\partial u}{\partial x_i}}\right)} = \frac{1}{1-\theta}\frac{\alpha_i}{\alpha_j}\left(\left(\frac{\alpha_j}{\alpha_i}\left(\frac{x_i}{x_j}\right)^{1-\theta}\right)\frac{\alpha_i}{\alpha_j}\right)^{\frac{1}{1-\theta}-1} = \frac{1}{1-\theta}\frac{\alpha_i}{\alpha_j}\left(\frac{x_i}{x_j}\right)^{\theta}$. Hence $\frac{d\left(\frac{x_i}{x_j}\right)}{d\left(\frac{\frac{\partial u}{\partial x_j}}{\frac{\partial u}{\partial x_i}}\right)} \frac{\frac{\partial u}{\partial x_j}}{\frac{\partial u}{\partial x_i}} \frac{x_j}{x_i} = \left(\frac{1}{1-\theta}\frac{\alpha_i}{\alpha_j}\left(\frac{x_i}{x_j}\right)^{\theta}\right)\left(\frac{\alpha_j}{\alpha_i}\left(\frac{x_i}{x_j}\right)^{1-\theta}\right)\frac{x_j}{x_i} = \frac{1}{1-\theta}$.

[4] For example Ordeshook claims: "The idea of spatial preferences, of representing the set of feasible alternatives as a subset of an m-dimensional Euclidean space, of labelling

However there are many different ways in which ideal point information can be introduced; furthermore, some models though strongly "spatial" in feel do not explicitly *require* the notion of an ideal point. Also, there are many ways to operationalize the notion of moving "away" from an ideal—couldn't for example the "awayness" relation be defined in terms of utility losses, however utility is defined, rather than in terms of some arbitrary metric? Such considerations makes it difficult to know what makes a spatial model spatial. While I don't know of any generally accepted definition of what constitutes *the* spatial model, I think the following distinction is useful.

In a "*weakly spatial model*" the set of outcomes, X, can be mathematically represented as a space, for example a vector space, a smooth manifold or, most commonly, a Euclidean space. Agents can be assumed to have (complete and transitive) rational preference orderings—given by the binary relation $\succsim_i$—over all pairs of elements in this set. But it is not required that agents cognitively perceive the relative utilities of outcomes in terms of relative distances in the underlying mathematical space. Agent preferences are characterized by an abstract agent-specific utility function $u : W \to \mathbb{R}^1$ with the property that, for any pair of outcomes y, y', we have $y \succsim_i y' \Leftrightarrow u_i(y) \geq u_i(y')$. In this sense the preferences over many dimensions with no satiation discussed above can be treated formally as weakly spatial models. But weakly spatial models also allow for the possibility that players have ideal points — points of global satiation. For many these are the hallmarks of any spatial model, but it is important to note that while they *may* exist in weakly spatial models, these ideal points do not necessarily have a privileged position and do not have to be specified as a part of the utility function.

In a "*strongly spatial" model*, the set of outcomes is characterized as being located in a space, but human agents' *evaluations* of these same outcomes are also assumed to be "spatial". This implies that each agent locates the set of outcomes, as well as his/her most-preferred outcome (ideal point) x_i, in a cognitive space and uses some distance measure $d_i(.,.)$ defined over the entire space[5], to characterize the relative utility of outcomes. This means that, given preference relation $\succsim_i$, we have

the dimensions 'issues,' of assuming that people (legislators or voters) have an ideal policy on each issue, and of supposing that each person's preference (utility) decreases as we move away from his or her m-dimensional ideal policy, is now commonplace and broadly accepted as a legitimate basis for modelling electorates and parliaments."

[5] A distance measure is a function $d : W \times W \to \mathbb{R}^1$ that has the properties that for points a,b, and c in W:

1. $d(a, b) = 0 \Leftrightarrow a = b$
2. $d(a, b) = d(b, a)$
3. $d(a, b) + d(b, c) \geq d(a, c)$

$y \succsim_i y' \Leftrightarrow d_i(x_i, y) \leq d_i(x_i, y')$. Commonly, these preferences are represented by an agent specific utility function $u : W \rightarrow \mathbb{R}^1$ with the property that for any two points y, y' we have $y \succsim_i y' \Leftrightarrow u_i(y) \geq u_i(y')$ where $u_i(.)$ is itself a composition of a loss function f and the distance function d_i; hence, $u_i(y) = f_i(d_i(x_i, y))$.

The models in McKelvey and Schofield (1986, 1987), generating the chaos results, are *weakly* spatial models. Most spatial models of party competition in the classic Downsian tradition are *strongly* spatial, in our sense, since they assume each voter to evaluate the set of potential outcomes in terms of each potential outcome's relative "closeness" to the voter's ideal point. The class of strongly spatial models extends, for essentially the same reason, to spatial models of probabilistic voting, with or without valence parameters (e.g. Groseclose 2001; Schofield, 2003, 2004). A useful rule of thumb is that models that *require* the notion of an agent "ideal point" are strongly spatial models. This is because the distance between the agent ideal point and each element in the set of potential outcomes typically enters as an argument in the agent's utility schedule.[6]

7.5.1 Strongly Spatial Models

I now consider some of the more common ways of representing preferences in spatial games. The listing that follows is ordered from strongest assumption to weakest assumption. In each case the stronger assumption implies the weaker assumption.

Linear or Quadratic Utility Functions

Perhaps the most restrictive of utility functions used in the spatial model is given by $u(x|p) = -(|x - p^i|)^2$ or sometimes simply $u(x|p) = -|x - p^i|$. While reasonable for a first cut and highly tractable as well as popular, these utility functions will put severe limitations on the generality of your results.

Euclidean Preferences

The class of Euclidean preferences corresponds to the family of utility functions for which the utility of a point x, is a decreasing function of the distance between x and some player ideal point p^i. Hence $f_i(|x - p^i|) \geq f_i(|y - p^i|) \leftrightarrow x \succsim_i y$ for some strictly decreasing function f_i. A number of functions that do the trick are graphically represented in Figure 7.1.

[6] For more on this see Humphreys and Laver. 2005. "Spatial Models, Cognitive Metrics and Majority Voting Equilibria"

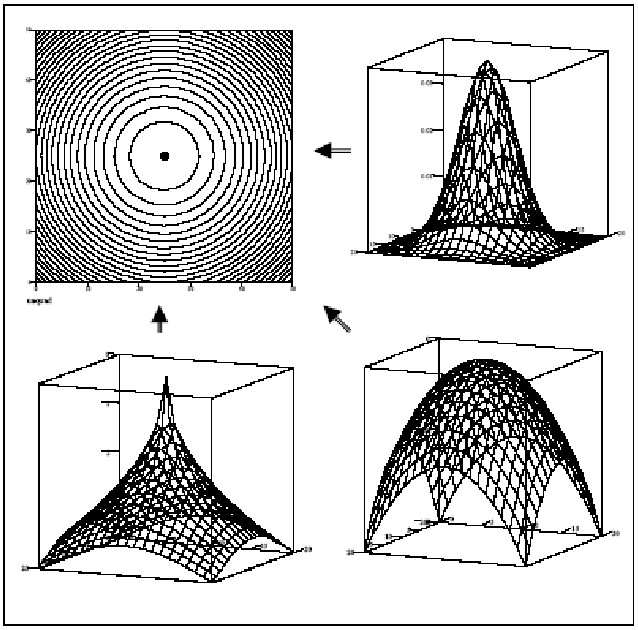


FIGURE 7.1. Three Representations of Euclidean Preferences

The figure highlights the point that many different functional forms may yield Euclidean preferences, and that the assumption of Euclidean preferences says nothing about the curvature of the utility function. Herein some other useful facts regarding Euclidean preferences:

- Indifference curves are circles (in $\mathbb{R}^2$) or spheres.

- Contract curves are straight lines.

- More generally, the Pareto Set of a set of players is the convex hull of the ideal points of the group's ideal points.

- If there is an odd number of players then you can identify "median hyperplanes"—planes with the property that half the player ideal points are on or to one side of the plane and half the ideal points are on or to the other side. If a point, x, is not on a median hyperplane then there is another point, y, (that *is* on the plane) that a majority prefers to x.

Problem 75 *If n players have Euclidean preferences, then the contract curves between each pair form a straight line. Is this also true if n players have other-regarding preferences in a distributive game, representable by a*

Cobb-Douglas utility function over an $n-1$ dimensional simplex with non-zero exponents? (That is, let utilities be of the form $u_i(x_1, x_2, ..., x_i, ..., x_n) = x_1^{\alpha_1^i}, x_2^{\alpha_2^i}, ..., x_i^{\alpha_i^i}, ..., x_n^{\alpha_n^i}$ where the exponent α_j^i captures the weight that player i places on the allocation of x to player j; and where each x_i is positive, and $\sum x_i = 1$.) Prove one way or the other.

The notion of a "median hyperplane" is especially useful in studying the geometry of majority rule. By "median hyperplane" we mean a plane with the property that for each side of the plane there lies a majority of ideal points on or to that side of it; in other words, if $M(v, a) = \{x|x.v = a\}$ is a median plane, there is a majority of points in the (closed) upper half space $\overline{M^+} = \{x|x.v \geq a\}$, and a majority of points in the (closed) lower half space $\overline{M^-} = \{x|x.v \leq a\}$.

Any majority rule equilibrium (core point) must lie on every median hyperplane. To see this, and to get a sense of how to work with such objects consider the following claim:

Claim 76 *Assume all players have Euclidean preferences over points in $\mathbb{R}^n$. Then if a point, x, does not lie on a median line (hyperplane) then some point on the line (hyperplane) will be preferred by some majority of players.*

Proof. Consider a median hyperplane $M(v, a)$ and some point x with $x \in M^-$. Now consider a rival point $x^* = x + (a - x.v)v$. Notice that with v an element of the unit sphere (that is $v.v = 1$) x^* is in M (note also that since $x \in M^-$, $x.v < a$ and $(a - x.v > 0)$). Also, for any point p:

$$\begin{aligned} |p - x^*|^2 &= (p - x^*).(p - x^*) \\ &= (p - x).(p - x) + (a - x.v)(a - x.v)v.v - 2(a - x.v)(p - x).v \\ &= |p - x|^2 + (a - x.v)((a - p.v) + (x.v - p.v)) \end{aligned}$$

But for $p \in \overline{M^+}$, $(a - x.v) > 0$, $(a - p.v) \leq 0$, and so $(x.v - p.v) < 0$. This implies $(a - x.v)((a - p.v) + (x.v - p.v)) < 0$ and hence $|p - x^*| < |p - x|$. Hence for any player i with ideal point $p \in \overline{M^+}$, $x^* \succ_i x$. Since there is a majority of players from group G^j with ideals in $\overline{M^+}$ we have $x^* \succ_{G^j} x$. ■

Generalized Euclidean Preferences

A somewhat more general class of utility functions for spatial games is generated by a metric that may place different weights on different dimensions and that may allow for non-separability across dimensions. A standard functional form used is the family of utility functions for which the utility of a point x, is a decreasing function of $(x - p^i)^T Z(x - p^i)$.where $(x - p^i)^T$ denotes the transpose of the vector $(x - p^i)$ and Z is any symmetric $n \times n$ positive definite matrix. Hence $f_i((x - p^i)^T Z(x - p^i)) \geq f_i((y - p^i)^T Z(y - p^i)) \leftrightarrow x \succsim_i y$ for some strictly decreasing function f_i. Euclidean preferences are given by the special case where Z is the identity matrix.

Again a number of functions with different curvatures will work. In the case of generalized Euclidean preferences however, indifference curves are ellipses and contract curves curve. The representation can be useful for studying the effects of differences in the relative salience of different dimensions to different players.

General strongly spatial utilities

Again somewhat more general assume that a player has an ideal point p^i and that preferences $\succsim_i$ can be represented by a utility function $u(x) = f_i(||x-p^i||)$ with $f_i(||x-p^i||) \geq f_i(||y-p^i||) \leftrightarrow x \succsim_i y$ for some strictly decreasing function f_i and where $||.||$ is a *norm*. Most commonly we think of the norm as being a vector norm[7] or Minkowski metric denoted by $||x||_\theta$ for $\theta = 1, 2, 3, ...$ where:

$$||x||_\theta = \left(\sum\nolimits_i |x_i|^\theta\right)^{\frac{1}{\theta}} \text{ for } \theta \in \{1, 2, 3, ...\} \tag{7.5}$$

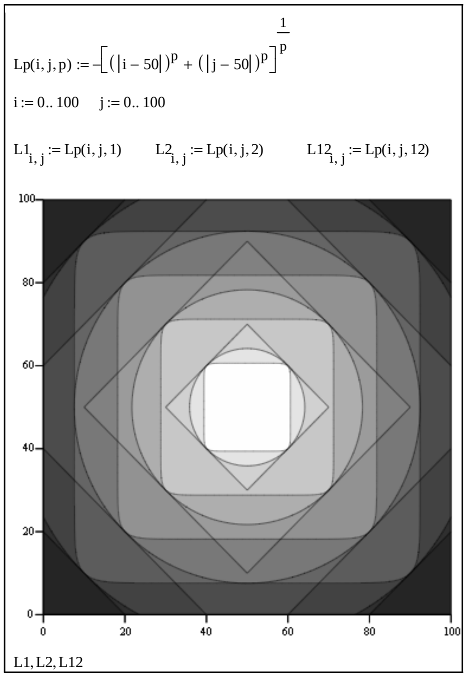
Commonly a $||x||_\theta$ is referred to as a L_θ norm (or commonly L_p norm). Note that using the $||x||_1$ metric, the distance between two points, a and b is simply $||a-b||_1 = \sum_i |a_i - b_i|$, that is, the sum of the distances on each dimension. This is in fact the city-block norm or Manhattan norm (based on the idea that to walk from a to b in a city you can't cut diagonally but have to add up the distances walked going north to south and the distances walked going east to west). Preferences based on this metric are sometimes called City-block preferences. The $||x||_2$ metric, or L_2 norm, is simply the Euclidean metric and so preferences based on this distance function are simply Euclidean preferences. For higher values of θ there is an increasing unwillingness to trade off across dimensions. For $||x||_\infty$, sometimes called the Tschebyscheff (Chebyshev) norm, or L_∞, the distance of a vector is given by the the length of the longest component of the vector. [8] For an illustration of these norms see figure 7.2. For any of these weights can be readily attached to any dimension and various rotations of the space can be introduced.

[7] A vector norm associates a number with each vector, subject to:

- $||a|| \geq 0$ and $||a|| = 0$ iff $a = \mathbf{0}$.
- $||ka|| = ||k|| \times ||a||$ for any scalar k (in particular this implies that for norms for which $||-1|| = ||1||$, $||a|| = ||-a||$.
- $||a|| + ||b|| \leq ||a+b||$

[8] There is then an evident analogy between the family of preferences based on Minkowski norms and the family of CES preferences given in Equation 7.1. For $\theta = 1$, in both cases tradeoffs along each dimension take place along a line; there is perfect substitution. At the opposite extreme we have $||x||_\infty = \max(|x_i|)$, just as for CES utility we saw $u(x|\theta = -\infty) = \min(\alpha_i x_i)$.

A slightly less general family of strongly spatial utilities requires the "triangle inequality" to hold strictly: hence for any a, $b \neq 0$ we require that $||a|| + ||b|| = ||a + b||$ only if there exist non negative λ_1, λ_2 not both 0 for which $\lambda_1 a = \lambda_2 b$. This excludes the L_1 and L_∞ norms (see for example McKelvey and Wendell, "Voting Equilibria in Multidimensional Choice Spaces").



$\mathrm{Lp}(i, j, p) := \left[(|i - 50|)^p + (|j - 50|)^p\right]^{\frac{1}{p}}$

$i := 0..100 \quad j := 0..100$

$\mathrm{L1}_{i,j} := \mathrm{Lp}(i, j, 1) \quad \mathrm{L2}_{i,j} := \mathrm{Lp}(i, j, 2) \quad \mathrm{L12}_{i,j} := \mathrm{Lp}(i, j, 12)$

FIGURE 7.2. Minkowski metrics: Utility based on L_1, L_2 and L_{12} norms.

7.5.2 Weakly Spatial Models

Weakly spatial models do not require information about the metrics agents use to measure the space, nor do they require the existence of an ideal point. All that is required is that, given the way that the modeler has chosen to represent the space, we can describe the agent's preferences over points in the space.

Again there can be great variation in the generality of the representation of player preferences. Here I consider two degrees of generality: Linear preferences and convex preferences.

Linear Preferences

Following Caplin and Nalebuff (1997) we say that players have "**linear preferences**" if each player has a "type" p^i where $p^i \in \mathbb{R}^n$ and the player's preferences can be represented by a utility function over points in x that takes the form:

$$u(x, p^i) = \sum_{j=1}^{n} p^i_j f_j(x) + g(x) \tag{7.6}$$

in which each function $f_1, f_2, ..., f_n$ and g map from X to $\mathbb{R}^1$.

While "linear" sounds like a narrow class of preferences, the freedom to choose the g and f_j functions actually allows this class of functions to incorporate a wide variety of possibilities. Note importantly that the g and f_j functions are not player specific; the player specific part is captured by p^i.

The class of linear preferences includes some strongly spatial preferences such as Euclidean preferences as well as some "economic preferences" such as the CES and Cobb-Douglas preferences described above. To see that Euclidean preferences fall within this family note that if preferences can be represented by any utility function they can be represented by the function $u(x, p^i) = -|x - p^i|^2$ and also by the function $u(x, p^i) = -|x - p^i|^2 + p^i.p^i$ (where we simply add a player specific constant (that does not depend on x) onto a player's utility). In this case, multiplying out we have $u(x, p^i) = -x.x + p^i.x$. Defining $f_i(x) = x$ and $g(x) = -x.x$ we have $u(x, p^i) = \sum_{j=1}^{n} p^i_j f_j(x) + g(x)$.

Problem 77 *Convince yourself that CES and Cobb-Douglas utility functions satisfy the form in Equation 7.6.*

Convex Preferences

More generally (including all the linear cases and more) assume that preferences are convex. A player has convex preferences if the set of points that he prefers to any given point x is a convex set. Often convex preferences are represented simply by quasiconcave utility functions (quasiconcave utility functions imply convex preferences by definition). Formally: for any x, $y \in X$, $x \neq y$, $y \succsim_i x$ implies $\lambda x + (1 - \lambda)y \succ_i x$ for any $\lambda \in (0, 1)$.

Strict convexity, along with completeness, consistency (transitivity) and continuity are collectively referred to as the "four Cs" . Over a convex and compact outcome space, these are enough to generate a unique ideal point.

Problem 78 *If the set of alternatives is convex and compact then "the four Cs" guarantee the existence of a unique "ideal point." Prove it.*

7.6 Intertemporal Preferences

Representing a player's preference over periods of time presents special problems. The key problem is that we may need to find an expression for how an individual at a single time point evaluates outcomes that have effects over multiple time periods. The problem goes to the philosophy of identity. If an individual only exists at a given point in time should she care about consumption in future periods or only with the present? That is, can we expect her to be "altruistic" with respect to her future selves? And if so, how can we write an expression for her altruistic evaluation of the welfare of all her future selves...

7.6.1 The Discounted Utility Model

By far the most common approach to the problem is to employ the **discounted utility** (DU) model.

The DU model attempts to disaggregate a person's utility over a collection of payoffs, delivered over time, into two parts: a time independent instantaneous utility function and a discount function, multiplied together in each period of time and added up over time, that is $U((x_1, t_1), (x_2, t_2)..., (x_s, t_s)) = \sum_{i=1}^{s} u(x_i)D(t_i)$.

In practice the discount function is given by the function $D(t_i) = \frac{1}{(1+r)^t} = \delta^t$, in which r is the "**discount rate**" (corresponding, for economists at least, to the interest rate) and $\delta = \frac{1}{1+r}$ is the "**discount factor**" (used more often by political scientists and corresponding, under different interpretations to the share of the pie left over after a given period or to the probability of survival into the next period). Virtually all models of intertemporal choice in political science employ the DU model. Unfortunately however the model has very shaky theoretical and empirical support.[9]

To get a sense of the restrictions imposed by the DU model, note that the model implies:

- **Utility Independence**: Hence preferences over the ordering or pattern of instantaneous utility packets are excluded except insofar as they affect the discounted summation. In fact, studies show that people seem to prefer increasing sequences to decreasing sequences.

- **Consumption independence**: Complementarity of goods over time is excluded. Hence the amount I eat for lunch, or what I eat, should not affect my preferences for what I eat for dinner.

[9] For more on this read: Frederick, S., G. Loewenstein and T. O'Donoghue (2002), "Time Discounting: A Critical Review," *Journal of Economic Literature*, 40, 351-401; www.iies.su.se/nobel/papers/intertemporal_review_final_2.PDF

- **Time invariant instantaneous utility functions**

- **Time invariant discount functions**: hence if I prefer a dollar today to 2 dollars tomorrow, that means that given a choice between receiving a dollar in one year and two dollars in a year and a day, I'd go for a dollar in a year over 2 in a year and a day. If that's not true for you then you exhibit "intertemporal preference reversals"—a characteristic commonly revealed in the study of humans and pigeons.

- **Invariance of time preference across all forms of consumption**. In fact though gains appear to be discounted more than losses small objects appear to be discounted more than large objects.

- **Present-bias**. The DU model typically assumes that $\delta < 1$. But why?

7.6.2 Representing time preferences in a prize-time space.

In some instances, notably bargaining environments, we are concerned not with aggregating utilities over consumption streams but in evaluating the utility of a good received at a given period in time. In this context we represent the preference of individuals over pairs in "prize-time space"—that is, the product $X \times T$ of a time and product space with typical element (x, t), representing the delivery of prize x in time t.

It can be shown that if the following conditions hold, then an individual's preferences may be represented with an exponential discount rate δ and utility function u such that $(x, t) \succsim (y, s) \Leftrightarrow \delta^t u(x) \geq \delta^s u(y)$:

1. Time is valuable: Let r, s and t be elements of T with $s > t$ and let D denote some object in $X \times T$ that is least preferred by i: that is $D \in \{(x, t) \in X \times T | \nexists (y, r) \in X \times T : (x, t) \succ_i (y, r)\}$. Then: for all $x \in X$, $(x, t) \succsim_i (x, s)$, and $(x, t) \succ_i (x, s)$ if $(x, t) \succ_i D$.

2. Preferences are stationary: $(x, t) \succsim_i (y, t + 1) \leftrightarrow (x, 0) \succsim_i (y, 1)$ and $(x, t) \succsim_i (y, t) \leftrightarrow (x, 0) \succsim_i (y, 0)$.

3. Preferences are continuous. A preference relation is "**continuous**" if for any sequence $\{x_n\} \subset X$ that converges to a point x in X, and sequence $\{y_n\} \subset X$ that converges to a point y in X, we have that if $(x_n, t) \succsim_i (y_n, s)$ for all elements of $\{x_n\}$ and $\{y_n\}$ then $(x, t) \succsim_i (y, s))$. Note this is the same as above but with added references to elements of T.

(see Osborne and Rubinstein **7.2** and Fishburn, P. and A. Rubinstein. 1982. "Time Preference," International Economic Review, 23, 677-694; Online at JSTOR).[10]

Importantly however, the fact that an individual's preferences may be represented by a utility function and discount factor does not mean that we can endow the discount factor with an absolute meaning. In particular we can not make claims of the form: "Player i with a discount factor of $\delta_i = .5$ is more patient than player j with a discount rate of $\delta_j = .25$." Instead, to interpret the implications of the discount factor for time preferences you first have to specify a utility function to represent a player's preferences.

Indeed, if a player's preferences, $\succsim$, can be represented by: $\delta^t u(x)$ in the sense that $(x,t) \succsim (y,s) \Leftrightarrow \delta^t u(x) \geq \delta^s u(y)$ then they can also be represented by $\tilde{\delta}^t \tilde{u}(x)$ where $\tilde{u}(x) = [u(x)]^{\frac{\ln(\tilde{\delta})}{\ln \delta}}$, and the redefined discount rate $\tilde{\delta}^t$ can be anything between 0 and 1. To see this, use the fact that $a^{\frac{\ln b}{\ln a}} = b$ to note that:

$$\begin{aligned}
\delta^t u(x_t) &\geq \delta^s u(y_s) \\
&\leftrightarrow \delta^{t\frac{\ln(\tilde{\delta})}{\ln \delta}} [u(x_t)]^{\frac{\ln(\tilde{\delta})}{\ln \delta}} \geq \delta^{s\frac{\ln(\tilde{\delta})}{\ln \delta}} [u(y_s)]^{\frac{\ln(\tilde{\delta})}{\ln \delta}} \\
&\leftrightarrow \delta^{\frac{\ln(\tilde{\delta}^t)}{\ln \delta}} [u(x_t)]^{\frac{\ln(\tilde{\delta})}{\ln \delta}} \geq \delta^{\frac{\ln(\tilde{\delta}^s)}{\ln \delta}} [u(y_s)]^{\frac{\ln(\tilde{\delta})}{\ln \delta}} \\
&\leftrightarrow \tilde{\delta}^t [u(x_t)]^{\frac{\ln(\tilde{\delta})}{\ln \delta}} \geq \tilde{\delta}^s [u(y_s)]^{\frac{\ln(\tilde{\delta})}{\ln \delta}}
\end{aligned}$$

Hence if $(x,t) \succsim (y,s) \Leftrightarrow \delta^t u(x) \geq \delta^s u(y)$ is what we mean by the "representation of preferences" then we are free to choose any discount rate we like in order to represent a given individual's preferences.

Note however that although we have $\delta^t u(x_t) \geq \delta^s u(y_s) \leftrightarrow \tilde{\delta}^t [u(x_t)]^{\frac{\ln(\tilde{\delta})}{\ln \delta}} \geq \tilde{\delta}^s [u(y_s)]^{\frac{\ln(\tilde{\delta})}{\ln \delta}}$ it does *not* follow that $\sum_{s=0}^{t} \delta^s u(x_s) \geq \sum_{s=0}^{r} \delta^s u(y_s)$ if and only if $\sum_{s=0}^{t} \tilde{\delta}^s [u(x_s)]^{\frac{\ln(\tilde{\delta})}{\ln \delta}} \geq \sum_{s=0}^{r} \tilde{\delta}^s [u(x_s)]^{\frac{\ln(\tilde{\delta})}{\ln \delta}}$.[11]

[10] For really recent work on this, see: Efe A. Oky and Yusufcan Masatlioglu, 2003. A General Theory of Time Preferences: http://home.nyu.edu/~eo1/Papers-PDF/Timepreference1.pdf

[11] The problem is that $\sum_{s=0}^{t} \delta^s u(x_s)$ represents a valuation of a t-tuple of timed outcomes. For a 2-period counterexample consider the case where $u(x) = x, \delta = .5$; $x_0 = 1$, $x_1 = 1$, $y_0 = 0$, $y_1 = 3$. In this case: $\sum_{s=0}^{1} \delta^s u(x_s) = 1.5 = \sum_{s=0}^{1} \delta^s u(y_s)$. But if we consider $\tilde{\delta} = .25$ and $\tilde{u}(x) = [x]^{\frac{\ln(.25)}{\ln .5}} = x^2$ we have $\sum_{s=0}^{1} \tilde{\delta}^s \tilde{u}(x_s) = .25 < 2.25 = \sum_{s=0}^{1} \tilde{\delta}^s \tilde{u}(x_s)$.

7.6.3 Alternatives to the DU Model

Alternatives now include:

- **Hyperbolic discounting**. The idea behind models with **hyperbolic discounting** is that discount factors are likely to rise over time. A simple way to model this is to assume that instead of $D(t) = \frac{1}{(1+r)^t}$ we have $D(t) = \begin{cases} 1 & \text{if } t=0 \\ \frac{\beta}{(1+r)^t} & \text{if } t>0 \end{cases}$. Such functional forms can lead to interesting predictions of intertemporal inconsistencies in an individual's choices and are used to explain over-consumption and procrastination.... (Read Jon Elster on this!)
- In models of "**Habit Formation**" preferences in time t depend on actions in previous times. A simple functional form that you can try playing with is: $u(c_t|c_{t-1}, c_{t-2}...) = u(c_t, \sum_{s=1}^{\infty} c_{t-s})$
- Other possibilities include "**Anticipatory utility**" and "**Reference-dependent utility**"

7.7 Readings for Next Week

We have then the tools we need to represent preferences over particular outcomes as well as preferences over uncertain outcomes. Our discussion treated certainty as exogenous however whereas in fact in many games a player's information may depend on how the game is played and in particular on the actions and the inferred information of other players.We turn to these issues next. The main reading, besides these notes, is Geanakoplos' excellent discussion of Common Knowledge in which he introduces in a non-technical manner a number of core results in "interactive epistemology." If you wish to go further in this direction, Geanakoplos refers to more formal treatments in the text and the Branderburger pieces on the syllabus are very interesting. The second reading by Roger Myerson is also relatively straightforward, but is good at communicating how the approach taken by Harsanyi to model games of incomplete information (and subsequently adopted by everyone else) was a major and difficult step forward in the development of game theory. For more, read the Harsanyi piece itself, given on the syllabus.

8
Information

We begin with a recap on the use of Bayes' rule in game theory, along with a discussion of how it is used when type or actions spaces are discrete or continuous. We then touch on a rich and fast growing branch of game theory – interactive epistemology – that draws on Bayes' rule and related work but provides a series of surprising results that are of general interest but also help in modelling games of incomplete information.

8.1 Bayes' Rule

8.1.1 Bayes' Rule with Discrete Type and Action Spaces

In our discussion of von Neumann-Morgenstern utilities[1] we showed how to model a player's evaluation of uncertain outcomes given his beliefs about which outcomes were more or less likely. There we had that given a set of possible outcomes, $X = (x_1, x_2, ..., x_s)$ and probabilities associated with each one, $p = (p_1, p_2, ..., p_s)$ we could write the player's expected utility as:

$$U(p) = \sum\nolimits_{j=1}^{s} p_j u(x_j)$$

[1] In all that follows this week we assume that players have von Neumann Morgenstern utilities. An interesting topic might be to think about how these results change if we cannot make use of the Expected Utility Theorem.

Now we are going to deal with where those probabilities come from. We need to do this because, in practice, they may depend on aspects of the game and in particular we need to know how players form their beliefs about what these probabilities are during the course of play: in other words, we need to know how players *update.*

The classic place to start to model updating is Bayes' Rule. The rule is relatively simple. Unfortunately it has been shown in experiments and in survey work that people do not in fact follow Bayes' rule closely and are terrible at answering questions that involve Bayesian updating. So, because it doesn't come naturally to us, we got to learn it. (Even as we recognize that, in part *because* it doesn't come naturally to us it might not be terribly appropriate for positive theory).

The easy equation to remember is:

$$\Pr(a,s) = \Pr(a|s)\Pr(s) = \Pr(s|a)\Pr(a)$$

For example let $\Pr(a,s)$ denote the probability that when you pull out a card from a deck it turns out to be the Ace of Spades; that is, $\Pr(a,s)$ is the probability that it will be both an Ace, "a" and a spade "s." And of course in this case $\Pr(a,s) = \frac{1}{52}$.

However, $\Pr(a,s) = \frac{1}{52}$, is also the answer to the product of the answers to the two questions:

1. What's the chance, $\Pr(s)$, of choosing a spade? ($\frac{1}{4}$). And:

2. Given that a spade is chosen, what's the chance, $\Pr(a|s)$, of choosing an Ace? ($\frac{1}{13}$).

Asking these two questions is one convoluted way to ask the original question. Nevertheless it's good to know that multiplying the answers from these two questions gives the same answer for the probability of choosing the Ace of Spades:$\frac{1}{13}\frac{1}{4} = \Pr(a|s)\Pr(s) = \Pr(a,s) = \frac{1}{52}$. So far so sensible. Similarly we have $\Pr(s|a)\Pr(a) = \Pr(a,s)$ and so $\Pr(a|s)\Pr(s) = \Pr(s|a)\Pr(a)$. Assuming $\Pr(s) > 0$, one manipulation of this equality then gives:

$$\Pr(a|s) = \frac{\Pr(s|a)\Pr(a)}{\Pr(s)}$$

To calculate $\Pr(s)$ we can consider any partition of the sample space into an exhaustive and mutually exclusive set of events $(a, b, c...)$. We then have $\Pr(s) = \Pr(s|a)\Pr(a) + \Pr(s|b)\Pr(b) + \Pr(s|c)\Pr(c) + ...$ And hence we have:

$$\Pr(a|s) = \frac{\Pr(s|a)\Pr(a)}{\Pr(s)} = \frac{\Pr(s|a)\Pr(a)}{\Pr(s|a)\Pr(a) + \Pr(s|b)\Pr(b) + ...}$$

And this is Bayes' rule. Before using the rule to describe updating let's get used to using it for simple inference problems. Consider the following examples.

Example 79 *Assume that 2% of politicians are honest and 1% of political scientists are honest (H). Assume that there is an equal number of political scientists and politicians: say a person is chosen at random and found to be honest. What are the chances that he is a politician? This is found using:*

$$\begin{aligned} \Pr(\textit{pol'n}|H) &= \frac{\Pr(H|\textit{pol'n})\Pr(\textit{pol'n})}{\Pr(H|\textit{pol'n})\Pr(\textit{pol'n}) + \Pr(H|\textit{poli scientist})\Pr(\textit{polit scientist})} \\ &= \frac{\frac{2}{100}\frac{1}{2}}{\left[\frac{2}{100}\frac{1}{2}+\frac{1}{100}\frac{1}{2}\right]} \\ &= \frac{2}{3}. \end{aligned}$$

Example 80 *What if there are twice as many political scientists as there are politicians? We then have:*

$$\Pr(\textit{politician}|\textit{honest}) = \frac{\frac{2}{100}\frac{1}{3}}{\left[\frac{2}{100}\frac{1}{3}+\frac{1}{100}\frac{2}{3}\right]} = \frac{1}{2}$$

Hence even though a politician is twice as likely to be honest as a politician, the chances that a random honest person is a politician is just .5.

Exercise 81 *(From Gintis) The poet and the saint each tell the truth one third of the time. The poet says of the saint "she just told the truth." What's the probability that the saint told the truth?*

Exercise 82 *(From Gintis) The Greens and the Blacks are playing bridge. After a deal, an on-looker, Mr Brown, asks Mr Black "Do you have an ace in your hand?" He nods yes (truthfully!). After the next deal Mr Brown asks Mr Black "Do you have the Ace of spades in your hand?"He nodes yes again (again, truthfully!). In each case, what are the probabilities that Mr Black has a second Ace.*

For describing rational updating, Bayes' little equation is surprisingly useful. We will see that a classic usage is for situations in which Player 1 is uncertain about the type of the person she is playing with, Player 2, but has beliefs about how each different type of player might behave.

- Let the set of possible types be given for Player 2 by: $\Theta_2 = \{\theta_1, \theta_2, ..., \theta_n\}$.
- Assume that 2 can take an action $a \in A$
- Assume that Player 1 has estimates of what a given type is likely to do: in particular she has an estimate of $\Pr(a|\theta)$ for every $a \in A$, $\theta \in \Theta_2$.

- Assume also that Player 1 has some "priors" on Player 2's type. She then has an estimate of $\Pr(\theta)$ for each $\theta \in \Theta_2$.

The question then is, if Player 1 observes a particular action, a, made by Player 2 can she then get a better estimate for $\Pr(\theta_2)$ than what she had before?

Surely. Using Bayes rule, she has for each type θ' :

$$\Pr(\theta'|a) = \frac{\Pr(a|\theta')\Pr(\theta')}{\sum\limits_{\theta\in\Theta_2}\Pr(a|\theta)\Pr(\theta)}$$

What if she then observed a second, independent, action a', will she learn something new? Yes:

$$\begin{aligned}\Pr(\theta'|a', a) &= \frac{\Pr(a'|\theta')\left(\frac{\Pr(a|\theta')\Pr(\theta')}{\sum\limits_{\theta\in\Theta_2}\Pr(a|\theta)\Pr(\theta)}\right)}{\sum\limits_{\theta\in\Theta_j}\Pr(a'|\theta')\left(\frac{\Pr(a|\theta')\Pr(\theta')}{\sum\limits_{\theta\in\Theta_2}\Pr(a|\theta)\Pr(\theta)}\right)} \\ &= \frac{\Pr(a'|\theta')\Pr(a|\theta')\Pr(\theta')}{\sum\limits_{\theta\in\Theta_j}\Pr(a'|\theta)\Pr(a|\theta)\Pr(\theta)}\end{aligned}$$

And so on. After observing t such independent actions, Player 1 has a new estimate:

$$\Pr(\theta'|a_1, ..., a_t) = \frac{\prod_{k=1}^{t}\Pr(a_k|\theta')\Pr(\theta')}{\sum\limits_{\theta\in\Theta_j}\prod_{k=1}^{t}\Pr(a_k|\theta)\Pr(\theta)}$$

A related usage is this: When applied to working out which node "x" you are at, given that you are in a particular information set, ι and given a strategy profile σ. The appropriate formula is:

$$\Pr(x|\iota, \sigma) = \frac{\Pr(\iota|x, \sigma)\Pr(x|\sigma)}{\Pr(\iota|\sigma)}$$

However, if $x \in \iota$, then by the definition of information sets, we have $\Pr(\iota|x, \sigma) = 1$, and so:

$$\Pr(x|\iota, \sigma) = \frac{\Pr(x|\sigma)}{\Pr(\iota|\sigma)} = \frac{\Pr(x|\sigma)}{\sum_{x'\in\iota}\Pr(\iota|\sigma)}$$

Finally, worth noting (and of great importance in what follows in later weeks): $\Pr(x|\iota, \sigma)$ is **undefined** if $\Pr(\iota|\sigma) = 0$. In words: *if we believe that no one could do a particular thing, but that thing is nevertheless done, then we have* no clue *who did it.*

Some of the implications of Bayes' rule are surprising and many do not come easily to humans. In many cases the frequency method can be employed with a computer to use Bayes' rule "empirically" by simply taking large numbers of samples and recording frequncies of cases in which events obtain or do not obtain given different conditions. Try modelling the following either directly or using a computer:

Exercise 83 *A game is played in which one of two players chooses a dice from a set of three dice. After the first player chooses, the second player chooses a dice from the remaining two that are left. Then the two role their dice simultaneously. The player that turns up the highest number wins. In this game however, although the dice are fair in the sense that any side is equally likely to come up, the* numbers *on the face of the three dice differ. Dice 1 has the following six numbers on its face: {5, 7, 8, 9, 10, 18}; dice 2 has {15, 16, 17, 2, 3, 4} and dice 3 has {1, 6, 11, 12, 13, 14}. Find player 2's optimal choice of dice given any choice of dice by Player 1. Which dice should player 1 choose? The problem is just one of finding the probability of winning given any two dice that are chosen. The result is however a little confusing.*

8.1.2 Bayes' Rule with Continuous Type and Action Spaces

When type and/or action spaces are continuous we typically need to work with distributions rather than with probabilities. The principle and the rule however, is essentially the same. Now however we work out the *posterior distribution* $f(\theta|m)$ of a type, θ given some signal, m. The trick then is to find $f(\theta|m)$. The version of Bayes' rule that we use for density functions is given by:

$$f(\theta|m) = \frac{g(m|\theta)f(\theta)}{g(m)} = \frac{g(m|\theta)f(\theta)}{\int g(m|\theta)f(\theta)d\theta}$$

One common application is the need to calculate the expectation of θ when you observe a signal which is drawn from some distribution that itself is a function of θ. Consider for example the following problem:

- θ is distributed over X, with density function $f(\theta)$
- For any true value of θ, some signal (message) $m \in X$ is generated according to the density function $g(m|\theta)$
- You need to calculate your best guess for θ given that you observe a signal m.

To solve the problem you need to work out the distribution of θ given m, call this *posterior distribution* $f(\theta|m)$. Once you know $f(\theta|m)$, you can work out $\mathsf{E}(\theta|m)$ using $\mathsf{E}(\theta|m){=}\int\theta f(\theta|m)d\theta$. Using Bayes' rule we then have, simply, $\mathsf{E}(\theta|m){=}\displaystyle\int\theta\frac{g(m|\theta)f(\theta)}{\int g(m|\theta)f(\theta)d\theta}d\theta$.

Example 84 *To develop your intuition, work through the following example. Say that θ is distributed uniformly over $[0,1]$ (hence $f(\theta)=1$). The expected value of θ is $\mathsf{E}\theta=\int_0^1\theta f(\theta)d\theta = \int_0^1\theta d\theta = \frac{\theta^2}{2}\Big|_0^1 = .5$. If θ is realized, the signal m, is drawn from the distribution with density function, $g(m|\theta)=4\theta m+2(1-m-\theta)$. (This density function puts more weight on lower declarations if indeed lower θ's are observed, and higher weight if higher θ's are observed). Let's now work out $\mathsf{E}(\theta|m)$.*

Using Bayes' rule we have:

$$f(\theta|m)=\frac{g(m|\theta)f(\theta)}{\int g(m|\theta)f(\theta)d\theta}=\frac{4\theta m+2(1-m-\theta)}{\int_0^1 4\theta m+2(1-m-\theta)d\theta}=4\theta m{+}2(1{-}m{-}\theta)$$

In this simple very symmetric case then, the density function $f(\theta|m)$ in fact equals $g(m|\theta)$. (This is not generally true.[2])

We can now work out $\mathsf{E}(\theta|m)$, and this is given by:

$$\mathsf{E}(\theta|m){=}\int_0^1[4\theta m+2(1-m-\theta)]\theta d\theta=\frac{1+m}{3}$$

Note that if a signal of $m=.5$ is observed, then your best guess for θ is simply $\theta=.5$; if however, $m=1$, then your best guess for θ is as high as $\frac{2}{3}$.

8.2 Interactive Epistemology

8.2.1 Information Partitions

Recall in our description of extensive form games with incomplete information, we used the notion of an information partition. We said for example that if a player had partition $\mathcal{I}_i = [h_1, h_3 \mid h_2.\mid h_4, h_5, h_6]$, then should h_1 or h_3 occur, the player knows that one of these occurred but not which one of these occurred; if h_2 occurs she will not confuse it for anything else; if h_4, h_5 or h_6 occurs then she knows one of these occurred but not which one.

[2] To see an asymmetric case, imagine that $f(\theta)=2\theta$ but $g(m|\theta)$ is as before. In this case your prior is such that you are more likely to see a high θ in the first place. We have: $\mathsf{E}\theta=\int_0^1\theta f(\theta)d\theta=\frac{2\theta^3}{3}\Big|_0^1=\frac{2}{3}$. But, $f(\theta|m)=\frac{g(m|\theta)f(\theta)}{\int_0^1 g(m|\theta)f(\theta)d\theta}=\frac{(4\theta m+2(1-m-\theta))\theta}{\frac{1+m}{3}}$. And so $\mathsf{E}(\theta|m){=}\int_0^1\frac{(4\theta m+2(1-m-\theta))\theta}{\frac{1+m}{3}}\theta d\theta=\frac{m+.5}{m+1}$.

We now generalize this notion of a partition somewhat. More generally, let Ω denote the set of all possible states of the world (previously we used H). A typical element of Ω, such as ω is a state of the world.

Let a partition of Ω by player i be denoted P_i (previously we used $\mathcal{I}_i$). Now it is useful to think of P_i as a function that maps from the states of the world to the cells of P_i. For example if P_i divides the set of integers into the evens and the odds, then $P_i(1)$ is the same as $P_i(3)$ and $P_i(5)$; each corresponds to the set of odd numbers. The set of even numbers is given for example by $P_i(2)$, or by $P_i(4)$ and so on. In the example of indistinguishable histories used above $P_i(h_1) = \{h_1, h_3\}$.

We can also usefully apply P_i to *sets* of elements in Ω. Now for event E (i.e. set of states of the world E), let $P_i(E)$ denote the set of all states of the world that player i thinks might be possible if the true state were an element of E. Hence: $P_i(E) = \cup_{\omega \in E} P_i(\omega)$.

We now have what we need to say that somebody "*knows*" something: i **knows** E at ω if $P_i(\omega) \subset E$. This means is that if the true state is ω, i may be uncertain what the true state of the world is, but if all of the things that he believes are possible imply E, then i knows E. (note that we are relying on the fact that is possible also to make statements about the truth of *groups* of elements in Ω if we are able to make statements about an individual element in Ω: if E is a set of states of the world, if the true state of the world is ω and if $\omega \in E$, then not only is ω true, but E is true as well[3]).

For some of the discussion that follows, it is useful to employ a "knowledge function" of the following form:

$$K_i(E) = \{\omega \in \Omega | P_i(\omega) \subseteq E\}$$

Hence for a given event E, the knowledge function returns all elements of Ω at which i knows E. In the example above $K_i(h_1) = \varnothing$, $K_i(h_2) = h_2$, $K_i(\{h_1, h_2\}) = h_2$, $K_i(\{h_1, h_2, h_3\}) = \{h_1, h_2, h_3\}$. Note that $K_i(E)$ is itself an "event" (that can occur at many different states of the world); it is precisely the event "i know E." Since it is an event it can also be treated as an argument in a knowledge function. Evidently if $\omega \in K_i(E)$ then i knows E at ω.

Problem 85 (see O&R) *What do the following statements mean? (i) For all E, $K_i(E) \subseteq E$ (ii) For all E, $K_i(E) \subseteq K_i(K_i(E))$ (iii) For all E, $\Omega \backslash K_i(E) \subseteq K_i(\Omega \backslash K_i(E))$*

[3] For example ω could be the state in which oil costs \$70 a barrel. Another state of the world, ω' might be one in which oil costs \$60. And define $E = \{\omega, \omega'\}$. In this case we might label the event E with "states in which oil is \$60 or \$70 a barrel." For this example the statement $\{$for $\omega \in E$, ω is true implies E is true$\}$, is obvous once given in English: if oil is \$70 a barrel, then oil is \$60 or \$70 a barrel.

To recap: $P_i(E)$ returns all states that i thinks possible if the truth were in E, $K_i(E)$ returns all states that, were they in fact the true state, i would know E.

Relatedly we can define a function to describe the event "everyone in N knows." Everyone knows E is the set:

$$K_N(E) = \{\omega \in \Omega | \bigcup_{i \in N} P_i(\omega) \subseteq E\}$$

Evidently the set of states in $\bigcup_{i \in N} P_i(\omega)$, is larger than the set of states in $P_i(\omega)$ or some particular i, and hence the condition $\bigcup_{i \in N} P_i(\omega) \subseteq E$ is harder to satisfy. Clearly then the set $K_N(E)$ is a subset of $K_i(E)$ for each i (or, more trivially, if everyone knows E, then each person knows E).

The event everybody knows that everybody knows E can be written $K_N(K_N(E))$, or more compactly, $K_N^2(E)$. We define $K_N^r(E)$ for any such sequence of statements.

If for every $\omega \subset E$, $P_i(\omega) \subset E$, then E is "*self-evident*" (to i)—that is, *whenever* E occurs, i knows it. Using the knowledge function: E is self evident to i if $E = K(E)$. Clearly $P_i(E) = E$ implies that E is self evident. (Note that the idea of being self evident here is a little different to normal usage, rather than being a property of the thing itself, self-evidence is a function of an individual's information; that might be self-evident to you but not to me...)

This framework starts getting really interesting when we start asking what do players know about what other players know. Let's motivate this with an example.

Example 86 *Suppose that player 1's partition divides prices into two cells, those between 0 and \$33 and those between \$34 and \$100. Suppose that player 2's partition divides prices into two cells, those between 0 and \$66 and those between \$67 and \$100. Say that a price is low if it is \$75 or below. Hence:* $L = \{\omega \in \Omega | \omega \leq 75) = [0, 75]$. *Assume furthermore that each player knows the other's partition.*

Now, say the price is in fact \$25. Then both Players know the price is low. But do they know that they know that? Formally, we know that the players know the price is low because $P_1(\omega^*) = [0, 33] \subset L$ *and* $P_2(\omega^*) = [0, 66] \subset L$. *Applying the* P *operators over sets we have* $P_2(P_1(\omega^*)) = [0, 66] \subset L$: *in words: 1 knows that 2 knows that the price is low. In contrast,* $P_1(P_2(\omega^*)) = \Omega \not\subset L$. *and so 2 does* not *know that 1 knows that the price is low (since plausibly the true price is between* 33 *and* 66*). Furthermore,* $P_1(P_2(P_1(\omega^*))) = P_2(P_1(P_2(\omega^*))) = \Omega$, *and so neither knows that the other knows that the other knows that the price is low.*

As the example shows, if players know the other players' partitions, even though they do not necessarily know what the other knows, we can still, given any state of the world, work out what each player knows and what they know about what the other players know, and so on...

For a second example, consider the player partitions shown in Figure 8.1:

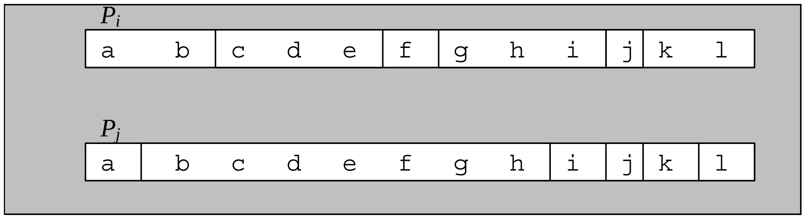

P_i
a b | c d e | f | g h i | j | k l
P_j
a | b c d e f g h | i | j | k | l

FIGURE 8.1. Example of partitions of players i and j two partitions over events $a - l$.

There are states of the world a - l. Some are distinguishable to some players, others to others. For example, state of the world f is self evident to player i. But j cannot tell f apart from outcomes b, c, d, or e. Player j knows that event a occurs whenever it occurs, but i can't distinguish between a and b.

Given these partitions, Figure 8.2 shows some examples of iterated applications of the P_i, P_j and the K_i, K_j functions.

Note that with repeated iteration of the P_i operators the resulting sets get larger until some point where they remain constant (but possibly containing all of Ω); with repeated iteration of the K_i operators the resulting sets get smaller until some point where they remain constant (but possibly empty).

Problem 87 *Find* $P_i(P_j(P_i(P_j(a))))$, $P_j(P_i(P_j(P_i(f))))$, $K_j(K_i(K_j(K_i(\{a,b,c,d,e,f,g,h,i\}))))$, *and* $K_j(K_i(K_j(K_i(P_i(P_j(P_i(P_j(l))))))))$. *Provide an interpretation for each of these.*

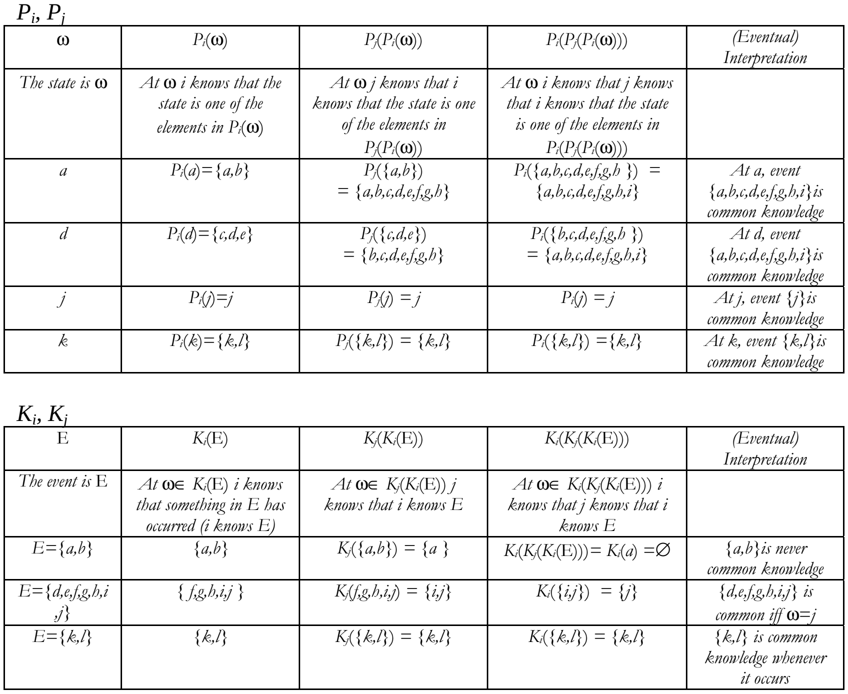
The final example, discussed in Figure 8.3, is based on a well known riddle and shows how these concepts can be used to study a tricky puzzle.

8.2.2 Common Knowledge

The example also shows how to apply the partition operator iteratively in order to identify what states agents deem possible given a chain of reasoning. We use this iterated application to define a notion of *reachability*:

Definition 88 *State* ω' *is "**reachable**" from state* ω *if there exists a sequence of agents* $i, j, ..., k$ *for whom* $\omega' \in P_k(...(P_j(P_i(\omega))))$

The interpretation of $\omega' \in P_k(...(P_j(P_i(\omega))))$ in English is at state ω, i thinks that j *may* think that....k *may* think that ω' is *possible*. If for some



P_i, P_j

ω	$P_i(\omega)$	$P_j(P_i(\omega))$	$P_i(P_j(P_i(\omega)))$	*(Eventual) Interpretation*
The state is ω	*At* ω *i knows that the state is one of the elements in* $P_i(\omega)$	*At* ω *j knows that i knows that the state is one of the elements in* $P_j(P_i(\omega))$	*At* ω *i knows that j knows that i knows that the state is one of the elements in* $P_i(P_j(P_i(\omega)))$	
a	$P_i(a)=\{a,b\}$	$P_j(\{a,b\})$ $= \{a,b,c,d,e,f,g,h\}$	$P_i(\{a,b,c,d,e,f,g,h\}) =$ $\{a,b,c,d,e,f,g,h,i\}$	*At a, event* $\{a,b,c,d,e,f,g,h,i\}$ *is common knowledge*
d	$P_i(d)=\{c,d,e\}$	$P_j(\{c,d,e\})$ $= \{b,c,d,e,f,g,h\}$	$P_i(\{b,c,d,e,f,g,h\})$ $= \{a,b,c,d,e,f,g,h,i\}$	*At d, event* $\{a,b,c,d,e,f,g,h,i\}$ *is common knowledge*
j	$P_i(j)=j$	$P_j(j) = j$	$P_i(j) = j$	*At j, event* $\{j\}$ *is common knowledge*
k	$P_i(k)=\{k,l\}$	$P_j(\{k,l\}) = \{k,l\}$	$P_i(\{k,l\}) =\{k,l\}$	*At k, event* $\{k,l\}$ *is common knowledge*

K_i, K_j

E	$K_i(E)$	$K_j(K_i(E))$	$K_i(K_j(K_i(E)))$	*(Eventual) Interpretation*
The event is E	*At* $\omega\in K_i(E)$ *i knows that something in* E *has occurred (i knows* E*)*	*At* $\omega\in K_j(K_i(E))$ *j knows that i knows* E	*At* $\omega\in K_i(K_j(K_i(E)))$ *i knows that j knows that i knows* E	
$E=\{a,b\}$	$\{a,b\}$	$K_j(\{a,b\}) = \{a\}$	$K_i(K_j(K_i(E)))= K_i(a) =\varnothing$	$\{a,b\}$ *is never common knowledge*
$E=\{d,e,f,g,h,i,j\}$	$\{f,g,h,i,j\}$	$K_j(f,g,h,i,j) = \{i,j\}$	$K_i(\{i,j\}) = \{j\}$	$\{d,e,f,g,h,i,j\}$ *is common iff* $\omega=j$
$E=\{k,l\}$	$\{k,l\}$	$K_j(\{k,l\}) = \{k,l\}$	$K_i(\{k,l\}) = \{k,l\}$	$\{k,l\}$ *is common knowledge whenever it occurs*

FIGURE 8.2. Usage of the P_i and K_i operators.

event, $\omega \in E$, every state that is reachable from ω is in E, then no chain of reasoning will lead any one to doubt that E is true at state ω. In such situations we say that E is *common knowledge*. The formal definition is as follows:

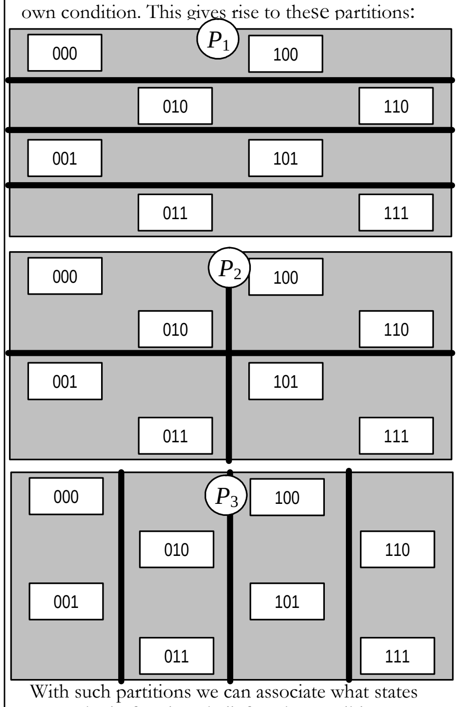
Definition 89 (Lewis 1969) *Event E is "**common knowledge**" at state ω if for every $n \in \{1, 2, ...\}$ and every sequence $(i_1, i_2, ...i_n)$ we have* $P_{i_n}(...(P_{i_2}(P_{i_1}(\omega)))) \subset E$.

- *(Variation 1; Using the idea of reachability) E is common knowledge at ω if every state that is reachable at ω is in E.*
- *(Variation 2; Using the knowledge operator) E is common knowledge at state ω if for every n and every sequence $(i_1, i_2, ...i_n)$ we have $\omega \in K_{i_n}(...(K_{i_2}(K_{i_1}(E))))$.*
- *(Variation 3; Using the "everybody knows" operator) E is common knowledge at state ω if $\omega \in K_N^{\infty}(E)$.*

Problem 90 *Prove that these four definitions are equivalent.*

Egg on your face

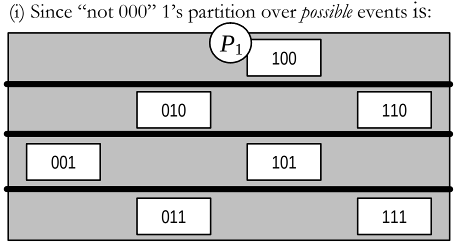
In this example a state, given in the form, "*abc*" denotes whether each of players 1, 2 or 3 has egg on her face; 000 indicates that none do, 101 indicates that 1 and 3 do but 2 does not and so on. Each player can observe the condition of other players but not their own condition. This gives rise to these partitions:



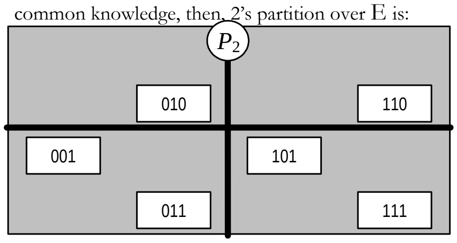
With such partitions we can associate what states must obtain for given beliefs to be possible. For example: Assume that all know that somebody has egg on their face. Given this, the event "1 does not know that she has egg on her face" is then: {010,110,001,101,011,111}. The event "2 knows that 1 does not know that she has egg on her face" is: {001,101,011,111}. Then the event "3 knows that 2 knows that 1 does not know that she has egg on her face" is {001}. Hence 3 knows that 2 knows that 1 does not know that she has egg on her face only if 3 is the only person with egg on her face. Intuitively, because 3 knows in this state that she (3) has egg on her face, she knows that 2 knows that 1 knows that 3 has egg on her face and hence that 2 knows that 1 knows that it is not the case that both 2 and 3 do not have egg on their face, hence 1 is unsure whether she (1) has egg on her face. If in fact either 1 or 2 had egg on their faces, then 3 would not know that she had egg on her face in which case plausibly she would not in which case plausibly 2 might think that only 1 had egg on her face in which case 1would know that she had egg on her face.

Now assume that it is made common knowledge that "not 000." Each woman is asked in turn whether she has egg on her face; the first two say "no" the third says "yes." Why is this?

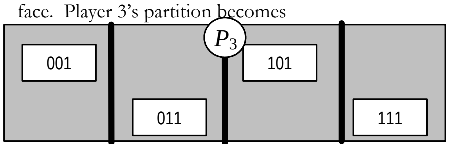
(i) Since "not 000" 1's partition over *possible* events is:



(ii) if 1 announces that the event "1 does not know whether she has egg on her face" E= {010,110,001,101,011,111} is true and this is common knowledge, then, 2's partition over E is:



(iii) Given this partition, the event "2 does not know that she has egg on her face" is given by {001,101,011,111}. Once announced, it now becomes common knowledge that 3 has egg on her face. Player 3's partition becomes



Hence, 3 can tell the exact state of the world. But since at this stage it is common knowledge that she has egg on her face, announcing that she has egg on her face adds no new information and the partitions of 1 and 2, shown below, do not allow them to work out their state or to provide new knowledge.

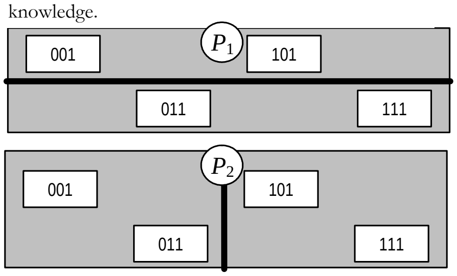


FIGURE 8.3. Learning from the Ignorance of others

These definitions, though intuitive, require that the analyst checks, in principle, an infinite number of statements in order to ensure that an event is common knowledge. The following theorem, due to Aumann provides a tool for checking for common knowledge in a finite number of steps:

Theorem 91 (Aumann's definition of Common Knowledge) *Let $M(\omega)$ denote the smallest set containing ω that is simultaneously self evident to all players. Then E is common knowledge at ω if and only if $M(\omega) \subset E$.*

Remark 92 *Note, although Theorem 91 reads like a definition, if we already have Lewis' definition (or related definitions) then Theorem 91 can be read as the identification of necessary and sufficient conditions for an event to be common knowledge.*

I find that the following equivalent statement lends itself more easily to identifying when an event is common knowledge: let M denote the finest common coarsening (the "meet") of the players' partitions, then, to be common knowledge at ω, an event, E, must be a superset of $M(\omega)$.

For intuition refer back to Figure 8.1. The finest common coarsening of these partitions, denoted by M, is shown in Figure 8.4.[4]

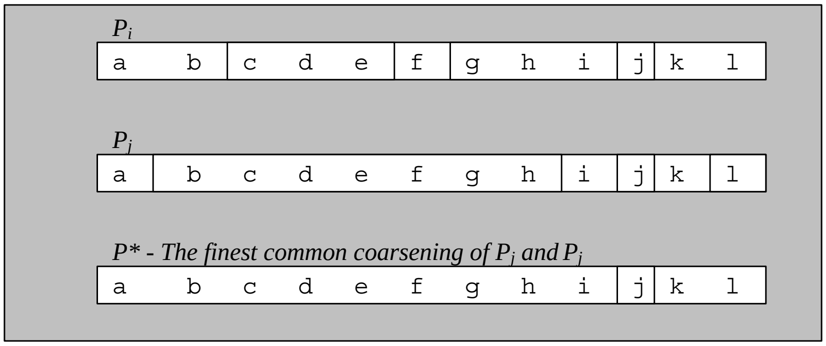

P_i
a b | c d e | f | g h i | j | k l
P_j
a | b c d e f g h | i | j | k | l
$P*$ - The finest common coarsening of P_i and P_j
a b c d e f g h i | j | k l

FIGURE 8.4. Coarsenings: An example of the finest common coarsening (the "meet") of two partitions.

What does $M(\omega)$ look like for different states ω? Begin with $\omega = a$. We can see that if a occurs, all players know that the state is either a or b. In particular $P_i(a) = \{a, b\}$ and $P_j(a) = a$. But is the event "ω is either a or

[4] To see that M is a common coarsening note simply that is is a coarsening of each of P_i and P_j. to see that it is the finest such common coarsening consider a partitioning M' that differs from M only in that some two elements that lie within a single cell of M lie within separate cells of M'. It is easy to check that any such division must separate two elements that both lie in a single cell of P_i or P_j in which case M is not a coarsening of one of P_i or P_j.

b" common knowledge? We can see from the above definition that it is not. Looking at M we see that $M(a) = \{a, b, c, d, e, f, g, h, i\}$, clearly $\{a, b\}$ is not a superset of $M(a)$ and hence is not common knowledge. The reason, intuitively is that, although at a, j knows that $\omega = a$ and also knows that i knows that $\omega \in \{a, b\}$, i does not know that j knows that $\omega \in \{a, b\}$, since j believes it possible that $\omega = b$ in which case it is possible that i believes $\omega \in \{b, c, d, e, f, g, h\}$.

Nonetheless supersets of $M(a)$ do exist, and hence there are events that are common knowledge at $\omega = a$. In particular $M(a)$ is itself common knowledge; so are other combinations of the cells in M involving $M(a)$, such as $M(a) \cup \{j\}$, and so too are other unions that involve any other states beyond those contained in $M(a)$,such as $\{a, b, c, d, e, f, g, h, i, l\}$, even if these are not themselves unions of cells of M.

By a similar logic we see that $M(j) = j$, and hence j is itself common knowledge when it occurs, as is the event $M(a) \cup \{j\}$ when j occurs.

Using the language of reachability we have that a state ω' is reachable, if and only if it is an element of $M(\omega)$.

Proof of Theorem 91. The proof we use shows that Variation 3 in definition 89 implies and is implied by the condition in Theorem 91.

- *Sufficiency.* We first show that if an event is a superset of $M(\omega)$ then it is common knowledge.

- Step 1. **The event $M(\omega)$ is common knowledge for every event ω' in $M(\omega)$.** Using our "everyone knows" operator we establish in Step 1a that $K_N(M(\omega)) = M(\omega)$, in step 1b we show that this implies that $K_N^\infty(M(\omega)) = M(\omega)$.

- Step 1a. $K_N(M(\omega)) = M(\omega)$. This follows from the definition of K_N and the fact that $M(\omega)$ is a coarsening of each P_i. To see this, recall that $K_N(M(\omega)) = \{\omega | \bigcup_{i \in N} P_i(\omega) \subseteq M(\omega)\}$. Assume first that some $\omega \in M(\omega)$ is not in $K_N(M(\omega))$, but this implies that for this ω, $P_i(\omega) \not\subseteq M(\omega)$, contradicting the fact that $M(\omega)$ is a coarsening of $P_i(\omega)$. Assume next that some ω' is in $K_N(M(\omega))$ but not in $M(\omega)$. But if ω' is in $K_N(M(\omega))$ then $\bigcup_{i \in N} P_i(\omega') \subseteq M(\omega)$, since $\omega' \in P_i(\omega')$ for each i, we have $\omega' \in \bigcup_{i \in N} P_i(\omega')$ and hence $\omega' \in M(\omega)$, a contradiction Hence $K_N(M(\omega)) = M(\omega)$. But since for any ω each set $P_i(\omega)$ is in $M(\omega)$, the union of these sets is in $M(\omega)$.

- Step 1b. Since $K_N(M(\omega)) = M(\omega)$, it follows that $K_N^2(M(\omega)) = K_N(M(\omega)) = M(\omega)$ and by induction that $K_N^\infty(M(\omega)) = M(\omega)$.

- Step 2. **Every superset of $M(\omega)$ is common knowledge at ω.** This step follows trivially from the fact that for superset E of $M(\omega)$,

$K_N^\infty(M(\omega)) \subseteq K_N^\infty(E)$. Since $M(\omega)$ is common knowledge, so too is E.

- Steps 1 and 2 establish sufficiency.
- *Necessity.* Now we want to show that if E is common knowledge at ω, $M(\omega) \subseteq E$. To see this, suppose that there exists some ω' in $M(\omega)$ but not in E. Since ω' is in $M(\omega)$ there exists a sequence $k = 0, 1, ..., m$ with associated states $\omega_0, \omega_0, ...\omega_m$ with $\omega_0 = \omega$ and $\omega_m = \omega'$ such that $\omega_k \in P_{i(k)}(\omega_{k-1})$ (this operationalizes the idea of reachability). Now, at information set $P_{i(m)}(\omega_{m-1})$, ω_{m-1} and ω_m are in the same partition of player $i(m)$ but since ω_m is not in E, $i(m)$ does not know E at ω_{m-1}.[5] Working backwards on k, we have that E cannot be common knowledge: that is, at ω_{m-1}, $i(m)$ does not know E; hence at ω_{m-2}, it is not true that $i(m-1)$ knows that $i(m)$ knows E;.at ω_{m-3}, it is not true that $i(m-2)$ knows that $i(m-1)$ knows that $i(m)$ knows E... .; continuing like this we have that at ω_0, it is not true that $i(1)$ knows that $i(2)$ knows that... $i(m)$ knows E. Hence E is not common knowledge.

■

8.2.3 Agreeing to disagree

Now let us introduce a notion of subjective probabilities over states of the world. Letting $p_i(E)$ denote a player's prior belief that event E will occur and $q_i(E)$ the player's posterior belief.

Aumann's agreement theorem is as follows.

Theorem 93 (Agreeing to disagree) *If q_i is common knowledge and if all players have common priors, then $q_i = q_j$ for all i and j.*

As a preliminary to the proof consider the following immediate but useful facts:

- If for two elements $p_i(\omega') \neq 0$ but $p_i(\omega'') = 0$, then $\omega'' \notin P_i(\omega)$. In other words, a zero probability event does not appear in the same partition as a positive probability event. Why?
- If for any $\omega' \in P_i(\omega)$, $p_i(\omega') > 0$, then $q_i(\omega'')$ is well defined for all ω'' in $P_i(\omega)$. This follows easily from an application of Bayes' rule: $\Pr(\omega''|P(\omega)) = \frac{\Pr(P(\omega)|\omega'')\times\Pr(\omega'')}{\Pr(P(\omega))} = \frac{\Pr(\omega'')}{\Pr(P(\omega))}$. The denominator is well defined if for some $\omega' \in P_i(\omega)$, $p_i(\omega') > 0$

[5] To work through the logic graphically refer to figure 8.4, let $E = \{a, b, c, d, e, f, g, h\}$ and identify a path from a to i.

- Similarly, using Bayes' rule, a player with partition P_i in state of the world ω (and assuming $p(P_i(\omega)) \neq 0$) assigns posterior probability to event E by:

$$q_i = q_i(E \cap P_i(\omega)|P_i(\omega)) = \frac{p_i(P_i(\omega)|E \cap P_i(\omega))p_i(E \cap P_i(\omega))}{p_i(P_i(\omega))} = \frac{p_i(E \cap P_i(\omega))}{p_i(P_i(\omega))}$$

Example 94 *Consider the roll of a dice and an information partitioning for player i given by: $P_i = (\{1,2,3\},\{4,5\},\{6\})$. Let E denote the set of cases in which the state of the world is even. The subjective probability of E, given that a 1 is observed is clearly $q_i(\omega \in 2|\omega \in \{1,2,3\}) = \frac{\Pr(\omega \in 2)}{\Pr(\omega \in \{1,2,3\})}$*

- The subjective probabilities of any event E at state of the world ω depend on $P(\omega)$, and hence is the same for any other state of the world ω' in $P(\omega)$. You can't condition on something you don't know. In other words, the subjective probability is constant across elements in a player's information partition $P_i(\omega)$.

The key move in Aumann's proof relies on the fact that with common knowledge, these probabilities are also *constant across elements on a cell of the meet of the players' partitions*: $M(\omega)$. This is the least obvious step in the proof that follows so let's look at it more closely. The intuition is the following: say that at state of the world ω, player i holds the posterior probability $q_i(E|\omega) = q^*$. Clearly he also holds this for any event ω' in $P_i(\omega)$. Now, player j does not know the true state ω and hence does not know the set of events $P_i(\omega)$ that i is conditioning upon when calculating $q_i(E)$. But if it is the case that at ω, j *knows* that $q_i = q^*$ then she knows this not just at ω but also for all elements in $P_j(\omega)$. Similarly for all other states ω in $P_i(\omega)$ and hence for all states in $P_j(P_i(\omega))$. Since i knows that j knows that $q_i(E|\omega) = q^*$ we then have it must be that $q_i(E|\omega') = q^*$ for all ω' in $P_i(P_j(P_i(\omega)))$; and so on. Continuing in this manner we simply have $q_i(E|\omega' \in M(\omega)) = q_i(E|\omega) = q^*$.

Armed with these elements we move to the proof.

Proof of Theorem 93. Now since players have common priors over the elements of Ω, using Bayes' rule, as above, a player with partition P_i in state of the world ω (and assuming $p(P_i(\omega)) \neq 0$) has posterior:

$$q_i = \frac{p(E \cap P_i(\omega))}{p(P_i(\omega))}$$

Note that q_i has subscript i even though p does not because p is applied over sets P_i subscripted by i. And hence: $q_i p(P_i(\omega')) = p(E \cap P_i(\omega'))$.

However, since q_i is common knowledge, we have that this posterior is constant across all elements ω' in $M(\omega)$. Hence: $q_i p(P_i(\omega')) = p(E \cap P_i(\omega'))$, for all ω' in $M(\omega)$.

Summing up over the disjoint sets $P_{i1}, P_{i2}..$ of $M(\omega)$ we have

$$q_i\left(p(P_{i1}) + p(P_{i2}) + ...\right) = p(E \cap P) + p(E \cap P_{i2} + ...)$$

$$q_i\left(p(M(\omega'))\right) = p(E \cap M(\omega')))$$

And hence $q_i = \frac{p(E \cap M(\omega')))}{p(M(\omega'))}$

We then have that q_i does not depend on any term that is subscripted by i but only on properties of the meet of the players' partitionings, which is constant across players. Repeating the reasoning for player j we have $q_i = q_j$. ■

Remark 95 (Application to trade) *: A series of no-trade or no-specultion results follow fairly quickly from the previous result. The basic idea is that in zero sum games with risk averse agents, if in equilibrium both parties simultaneously agree on a trade, then it is common knowledge that each believes that the trade favors them. But since they cannot agree to disagree, each must assign the same value of the trade to each of them, so if i thinks the trade is good for i but bad for j, so must j, in which case j will not trade.*

Remark 96 (Application to conflict) *Consider the implication of the theorem for models of inter state conflict. The theorem states that for any model in which two countries, i and j assign subjective probabilities to the chances that i wins a war, if the countries have common priors (although quite possibly different information on which they can base their posterior probabilities), then if p_i and p_j are common knowledge, $p_i = p_j$.*

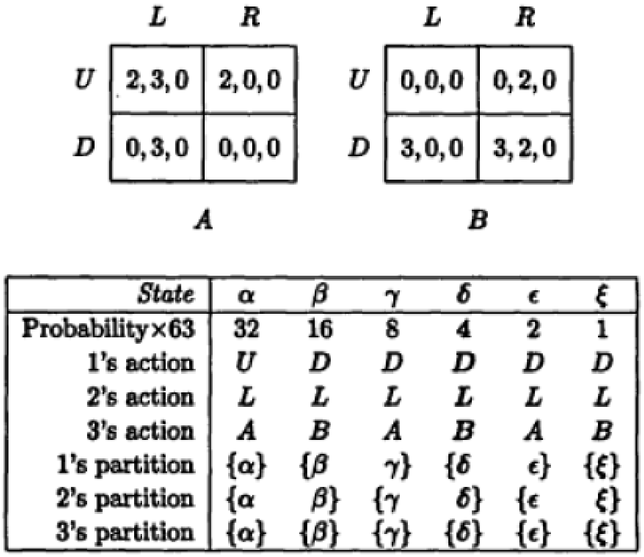
8.2.4 Rationalizability and Common Knowledge of Rationality

While these results, and related ones, help in modelling and in describing the informational structure of games, and their implications, they also provide guidance regarding *what solution concepts are defensible.* Here we ask the question: "how much" common knowledge do you need in order to defend a particular solution concept?

For the analysis we treat actions as being a function of the state of the world, and hence write action of player i in the form of a function $a_i = a_i(\omega)$. This may seem odd since ω may be unknown to i, but this concern is easily handled by the condition that $a_i(\omega) = a_i(\omega')$ for all $\omega' \in P_i(\omega)$. We also introduce a belief function, also a function of the state, that describes the belief that i has about the actions of other players: $\mu_i(\omega)$; $\mu_i(\omega)$ is a probability distribution over the set of profiles of strategies of all other players.

Given these elements, we ask: what do we need to know about a_i and μ_i to justify the use of the Nash solution concept?

It is often said that common knowledge of rationality is enough to justify Nash equilibrium as a solution concept. In fact this is not the case; nor even are a series of stronger restrictions, at least in games with more than two players. The following counterexample is provided by Osborne and Rubinstein



	L	R
U	2,3,0	2,0,0
D	0,3,0	0,0,0

A

	L	R
U	0,0,0	0,2,0
D	3,0,0	3,2,0

B

State	α	β	γ	δ	ϵ	ξ
Probability×63	32	16	8	4	2	1
1's action	U	D	D	D	D	D
2's action	L	L	L	L	L	L
3's action	A	B	A	B	A	B
1's partition	{α}	{β	γ}	{δ	ϵ}	{ξ}
2's partition	{α	β}	{γ	δ}	{ϵ	ξ}
3's partition	{α}	{β}	{γ}	{δ}	{ϵ}	{ξ}

FIGURE 8.5. one of the two rows, player 2 chooses one of the two columns,

.

Problem 97 *Consider Figure 8.5. Assume (i) that all players have beliefs that are consistent with their knowledge; that is, they assign positive probability only to actions profiles that are a function of states of the world that i believes possible. (ii) (a strong assumption) all players know the beliefs of all other players–that is for every state that i deems possible, j has the same beliefs about the actions of all other players. (iii) all players know all players to be rational, that is, that given a player's equilibrium beliefs, his equilibrium action is a best response to his beliefs (and this is known by all about all players). Convince yourself* [1] *Strategy profile* $\{D, L, B\}$ *is not a Nash equilibrium. and* [2]*that if the state of the world is* δ*, the following profile* $\{D, L, B\}$ *is consistent with (i), (ii) and (iii).*

What of rationalizability as a solution concept? It turns out that we can justify rationalizability. Any action that satisfied quite weak conditions on the common knowledge of rationality is rationalizable.

Define a profile of rationalizable strategies as follows (another, although equivalent, definition is given next week).

Definition 98 *An action a_i is "**rationalizable**" in a normal form game if for each $i \in N$, a_i is a member of a set Z_i where each $Z_i \subseteq A_i$ and every action $a_i' \in Z_i$ is a best response to some belief μ_i whose support is a function of a subset of Z_{-i}.*

Now we have the following claim (this given for two players, but extends beyond two players):

Claim 99 *Suppose $|N| = 2$ and that at state ω there is a self evident event E containing ω and that it is common knowledge that (i) each player is rational in the sense that $a_i(\omega')$ is a best response to $\mu_i(\omega')$ for all ω' in E and (ii) each player's belief is consistent with his knowledge in the sense that for all ω' in E, $\mu_i(\omega')$ places positive probability only on actions that are functions of states that i deems possible. Then a is rationalizable.*

Proof. For each $i \in N$ define $Z_i = \{a_i(\omega')|\omega' \in E\}$; that is Z_i contains all actions that i could take, given that he is rational, under any state that is reachable from ω.

For each $i \in N$ and ω' in E define $S_i(\omega')$ as that set of actions that i believes j might take if the state were ω'. $S_i(\omega') = \{a_j(\omega'')|\omega'' \in P_i(\omega')\}$.

For any ω' in E, $a_i(\omega')$ is a best response to beliefs that have a support on a subset of $S_i(\omega')$ (this is just the consistency condition (ii) in the statement of the claim).

Since E is self evident $P_i(\omega') \subseteq E$.

Hence $S_i(\omega') \subseteq Z_j$.

In other words, each $a_i \in Z_i$ is rationalizable because its support is a subset of the set of actions that are rationalizable for j. ■

9

Solution Concepts for Normal Form Games

9.1 Iterated Elimination of Strictly Dominated Strategies

There are some solution concepts that make less demands on a player's abilities. One relies on the elimination of dominated strategies. We use the following definition:

Definition 100 *In the normal form game $\mathcal{G} = \langle N, (A_i), (u_i)\rangle$, strategy $a_i' \in A_i$ is "**strictly dominated**" if there exists an $a_i'' \in A_i$ such that $u_i(a_i', a_{-i}) < u_i(a_i'', a_{-i})$ for all feasible a_{-i}.*

Given this definition, one simple solution concept is the set of strategy combinations that survive a "**process of iterated deletion of strictly dominated strategies**." The idea is this: eliminate all strictly dominated strategies from each player's strategy set, then eliminate any further strategies that are strictly dominated once the first set of strategies have been removed. Keep repeating this process until no more strategies can be removed. The remaining strategy combinations are possible predictions under this solution concept.

Problem 101 *The idea of iterated deletion of strictly dominated strategies is quite simply, albeit loosely, stated in words. A more formal statement would have to express the "keep repeating this process" idea precisely. One possibility to express this kind of idea is to define an infinite series of sets recursively and look for the intersection of all these sets. The formal* *definition will also have to be explicit about the order of deletion. As an exercise in definition-writing, use this approach to define the set of strategies (pure as well as mixed) that survive iterated deletion of strictly dominated strategies formally. (For one possibility see Definition 2.1 in Fudenberg and Tirole, for another see Definition 60.2 in Osborne and Rubinstein)*

In simple games this approach makes weak demands on players' strategic abilities—they can use a simple algorithm to work out how to play a game. Furthermore you can construct simple learning algorithms that lead players that interact repeatedly to end up playing the strategies that are selected by the elimination of strictly dominated strategies. And it does not even (always) require "**common knowledge**" about the rationality of the players (although infinite sequences of eliminations will require common knowledge).[1] Also, the set of strategies that survives the process does not depend on the details of how elimination is done. These are all good properties of a solution concept. An unfortunate aspect however is that the method may be very indeterminate in its predictions and predict a large number of outcomes in games where other solution concepts (such as Nash Equilibrium) can make unique predictions.

9.2 Iterated Elimination of Weakly Dominated Strategies

We could define an analogous solution concept based on the elimination of weakly dominated strategies (similar to the definition given above, we say that strategy $a_i' \in A_i$ is "**weakly dominated**" if there exists an $a_i'' \in A_i$ such that $u_i(a_i', a_{-i}) \leq u_i(a_i'', a_{-i})$ for all feasible a_{-i} and where the inequality is strict for at least one a_{-i}). Even though intuitively this approach has many of the same merits as the elimination of strictly dominated strategies, it turns out that the approach has lots of problems. Most importantly, what predictions the method produces depends not just on the game but on the order in which you delete strategies. The problem is that a combination of strategies that survives the iterated elimination of weakly dominated strategies may itself be weakly dominated, so whether it gets selected or not depends on whether strategies that eliminate it get eliminated before it or not. Furthermore, Nash equilibria may be eliminated by this process.

[1] We do not cover in this course the rich literature on the study of different types of knowledge. Crudely however, if players i and j have **common knowledge** about x that means that both i and j know x, they both know that the other knows x, they also both know that they both know that the other knows x and so on. For more on knowledge read Chapter 5 of Osborne and Rubinstein.

Problem 102 *Consider the following game: each of two players has to guess a number between 1 and 5; if the two players match in their guesses, or are no more than one out, they win $1 each, otherwise they get nothing. What strategies survive iterated elimination of weak strategies (i) if the order of deletion is: "delete one of player 1's weakly dominated strategies, then one of player 2's strategies, then one of player 1's remaining strategies and so on... and (ii) if the order is "delete all of player 1's weakly dominated strategies, then all of player 2's, then all of player 1's remaining weakly dominated strategies and so on...? Note how in (i) the answer depends on which strategy you begin by deleting. Are the predictions from this method more or less precise than predictions based on Nash equilibrium reasoning? How would you play this game?*

		Player II				
		1	2	3	4	5
Player I	1	(1,1)	(1,1)	(0,0)	(0,0)	(0,0)
	2	(1,1)	(1,1)	(1,1)	(0,0)	(0,0)
	3	(0,0)	(1,1)	(1,1)	(1,1)	(0,0)
	4	(0,0)	(0,0)	(1,1)	(1,1)	(1,1)
	5	(0,0)	(0,0)	(0,0)	(1,1)	(1,1)

FIGURE 9.1. Weakly Dominated Strategies

Problem 103 *In an intriguing class of games, including some Battle of the Sexes games, in which there are multiple Pareto efficient pure strategy Nash equilibria, elimination of weakly dominated strategies can be used to select one of the two equilibria if one of the players, say player 1, is given an option to take some third action that affects his own payoffs adversely. In these games Player 1 need never resort to this third, masochistic, action, but the threat of being able to do so is enough to lead to the selection of her preferred outcome from the Battle of the Sexes game through iterated deletion of weakly dominated strategies. Construct an example of such a game.*

Exercise 104 *Do Exercise 63.1 from Osborne and Rubinstein.*

9.3 Rationalizable Strategies

Common knowledge of each other's rationality can justify the iterated elimination of dominated strategies: it can also be used to motivate a more precise solution concept: the set of rationalizable strategies.[2] We say that a strategy σ_i is a "**best response**" to a set of strategies σ_{-i} if $u_i(\sigma_i, \sigma_{-i}) \geq u_i(\sigma'_i, \sigma_{-i})$ for all σ'_i in ΔA_i.We say that a strategy σ_i is a "**never a best response**" if there is no σ_{-i} such that σ is a best response.

The set of "**rationalizable strategies**" is the set of strategies that survives the iterated elimination of strategies that are never a best response. Most importantly, any element of the set (and only elements of the set) can be justified by a "chain of reasoning" of the form: I will play σ_i because I believe that player j will play σ_j where σ_j is a best response by j if she thinks that I am going to play σ'_i (where σ'_i is not necessarily equal to σ_i) and so on. In other words: *this is as good as you can get if all you want to impose on player behavior is common knowledge of rationality.*

Nice properties of the set of rationalizable strategies include

1. It exists (Bernheim 1984)
2. like the iterated elimination of dominated strategies, the order of elimination does not matter
3. It can be identified in a finite number of iterations (in a finite game)
4. the set is a subset of the set of strategies that survive the iterated elimination of dominated strategies

Exercise 105 *Prove Property (4): the set of rationalizable strategies is a subset (although not necessarily a strict subset) of the set of strategies that survive the iterated elimination of dominated strategies. Also prove that Nash equilibrium strategies are a subset of rationalizable strategies.*

Problem 106 *Consider the following game (Figure 9.2, taken from Vega-Redondo, 2003). Identify the Nash equilibrium. Show that strategies (A,B,C,W,X,Y) are all rationalizable.*

9.4 Nash Equilibrium

Nash equilibrium is perhaps the best known solution concept for normal form games. Recall that we defined the normal form game as a 3-tuple, $G = \langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$. The set of strategies for each player, A_i, could be discrete actions (such as Left, Right) or elements from continuous sets—such as a number between 0 and 1, or randomizations over more primitive strategies (more on this below). Let a denote a typical profile of strategies where $a \in A_1 \times A_2 \times ... \times A_n$ We then have the following definition:

[2] Although in two player games the two solution concepts are equivalent.

				II	
		A	B	C	D
I	W	7 0	5 2	0 7	1 0
	X	2 5	3 3	2 5	1 0
	Y	0 7	5 2	7 0	1 0
	Z	0 0	-2 0	0 0	-1 10

FIGURE 9.2. Identifying Rationalizable Strategies

Definition 107 *A profile of strategies* a^**, is a "****Nash equilibrium****" if for each* $i \in N$*, we have:* $a_i^* \in \arg\max_{a_i \in A_i} (u_i(a_i, a_{-i}^*))$

Note that the equilibrium is a *profile* of *strategies*: it is not a particular outcome, or an action by any individual. To describe an equilibrium you have to say what each player's strategy is.

Nash equilibrium is a compelling solution concept primarily because anything that is *not* a Nash equilibrium requires that some individual take an action even though some other action would serve her interests better, whatever those interests are.

But the concept has problems; common worries with the concept include the following:

- **Existence**: a Nash equilibrium may not exist. Two examples are the following. (I) Consider a two-person game in which each player can declare heads or harps. Player 1's utility is maximized if both players declare the same thing; player 2's is maximized if they each declare different things. (II) Consider a two person game in which each person can declare any integer, the person who declares the highest integer wins.

- **Precision**: For other games the Nash solution concept can provide so many different solutions that it provides little guidance. Consider a game in which two players have to divide up a dollar between them. Each individual has to declare some division. If the two divisions coincide, then that division is implemented, if not then no division is implemented and the game ends. Here there is an infinity of Nash equilibria and the solution concept gives no determinate prediction about how the dollar will be divided.

- **Accuracy**: In the context of extensive form games there are many outcomes that may be Nash equilibria but that are rendered implausible by the structure of the game. Consider the divide the dollar game mentioned above in which each player cares only about her own share of the dollar and in which Player 1 first declares a division and Player 2, after hearing Player 1's declaration then makes her own declaration. It is a Nash equilibrium for Player 1 to suggest 1 cent for herself and 99 cents for Player 2. But in practice we would expect a rational Player 1 to suggest something like 99 cents for herself and 1 cent for player 2...

- **Plausibility**: In some contexts it's hard to know what exactly the Nash equilibrium is describing. For situations where there are multiple Nash equilibria and players have never encountered each other before, there are only weak grounds for expecting that the set of strategies selected by players will correspond to a Nash equilibrium profile unless players have some reasons to form expectations of the strategies of others around some particular profile.

There are ways to refine the solution concept in order to address most of these issues. First however we consider how the use of a generalization, equilibrium in "mixed strategies" can solve some part of the existence problem. First though, some engineering.

9.4.1 Locating Nash Equilibria in Games with Continuous Action Spaces

The method for locating a Nash equilibrium a^* when players have continuous action spaces and differentiable utility functions is to differentiate $u_i(a_i|a_{-i})$ with respect to a_i, keeping a_{-i} fixed. Taking the first order conditions then allows us to write a^* as a (possibly implicit) function of a_{-i}. Doing this for all players gives us a system of n equations, with each equation giving a relationship between one element of a^* and all the other elements. Then *any* vector a^* that solves this system of equations is a Nash equilibrium. For an example, refer to our discussion of the Coase theorem above.

9.5 Nash Equilibrium with Mixed Strategies

In Section 5.1.2 we defined the notion of mixed strategies. we did not however say how player's evaluate a profile of mixed strategies. We are now in a position to do that.

Making use of the Expected Utility Theorem from before we have that if players have von Neumann-Morgenstern utilities, we can estimate the *expected* utility that a player gets from some randomized strategy profile.

Let a typical profile of pure strategies be written $a \in A_1 \times A_2 \times ... \times A_n$ and recall that players have utility functions that are defined over such profiles. Then, assuming independent randomization, a player's expected utility over a mixed strategy profile, σ, is given by:

$$u_i(\sigma) = \sum_{a \in A_1 \times ... \times A_n} p(a)u(a) = \sum_{a \in A_1 \times ... \times A_n} \prod_{i \in N} \sigma_i(a_i)u(a)$$

where $p(a) = \prod_{i \in N} \sigma_i(a_i)$ is the probability of observing a given the mixed strategies of all the players. The definition analogous to that given above for a Nash equilibrium for this class of games is given then by:

Definition 108 *A profile of mixed strategies σ^* is a "**Nash equilibrium**" if for each $i \in N$, we have: $\sigma_i^* \in \arg \max_{\sigma_i \in \Delta(A_i)} (u_i(\sigma_i, \sigma_{-i}^*))$*

An advantage of the mixed strategy solution is that it **exists** for a large class of games (Nash's theorem).

9.5.1 Locating Mixed Strategy Nash Equilibria

An important property of a mixed strategy is the following. Assume that σ_i^* is part of an equilibrium profile σ^* in which player i places positive weight on some subset, A_i^*, of elements of A_i. Then player i is indifferent between σ_i^* and any randomized strategy involving elements of A_i^*, including each of the pure strategies corresponding to playing elements of A_i^* with certainty.

The reason for this is simply that if playing some pure strategy produces utility higher than σ_i^*, then σ_i^* can not be a best response, whereas if playing some pure strategy $a_i \in A_i^*$ produces a worse outcome than σ_i^* then a rival strategy $\sigma_i^{*\prime}$ in which less weight is placed on strategy a_i produces a higher payoff than σ_i^* and so, again, σ_i^* can not be a best response (see Osborne and Rubinstein Lemma 33.2 for a formal proof).

This has got good and bad implications. The good implication is that it makes identifying the mixed strategy Nash equilibrium easy: if a player i mixes over a set of actions A_i^*, then the *other* players' mixing must be such that i is indifferent between choosing each element in A_i^*.

Here is a simple example. Consider the battle of the sexes game in normal form given by:

At the mixed strategy Nash equilibrium in which both players mix over L and R, and letting $\sigma_{II}^*(L)$ denote the probability that II plays L and letting $\sigma_{II}^*(R) = 1 - \sigma_{II}^*(L)$ denote the probability that II plays R, we must have have:

		I	
		L	R
II	L	2,1	0,0
	R	0,0	1,2

FIGURE 9.3. Battle of the Sexes

$$\begin{array}{rcl}\text{Expected utility to I from } L \text{ (given } \sigma_{II}^*) & = & \sigma_{II}^*(L).1+(1-\sigma_{II}^*(L)).0 \\ & = & \\ \text{Expected utility to I from } R \text{ (given } \sigma_{II}^*) & = & \sigma_{II}^*(L).0+(1-\sigma_{II}^*(L)).2.\end{array}$$

Hence $\sigma_{II}^*(L).1=(1-\sigma_{II}^*(L)).2,$ and so $\sigma_{II}^*(L)=\frac{2}{3}$. Similarly $\sigma_I^*(L)=\frac{1}{3}$. It is worth constructing a set of examples and solving for the mixed strategy Nash equilibrium to get the hang of using this technique. Try one with more than 2 players. To deepen your understanding, try to solve for a mixed strategy equilibrium of a Prisoners' Dilemma. Having done this, now try the following n person problem:

Exercise 109 *Let $N=\{1,2,...n\}$ and let each person have a strategy set given by $A_i=\{L,R\}$; assume that players are permitted to randomize over elements of A_i. Let utilities be given by:*

$$u_i(a)=\begin{cases} 0 & \textit{if all players play } L \\ 1 & \textit{if at least one player, including } i\textit{, plays } R \\ 2 & \textit{if at least one player plays } R \textit{ but } i \textit{ plays } L\end{cases}$$

Find a mixed strategy Nash equilibrium of this game. How does it depend on n? What kind of situation does this game describe?

Problem 110 *For n odd solve for the symmetric mixed strategy Nash equilibria of the game in which*

$$u_i(a)=\begin{cases} 0 & \textit{if all players play } L \\ 1 & \textit{if at least } \frac{n}{2}+1 \textit{ players, including } i\textit{, plays } R \\ 2 & \textit{if at least } \frac{n}{2}+1 \textit{ players plays } R \textit{ and player } i \textit{ plays } L\end{cases}$$

9.5.2 Mixing Over Continuous Pure Strategy Sets

So far we have mixed considered strategies of the form σ_i where σ_i is a vector of length equal to the cardinality of the player's pure strategy set.

In many instances however, we consider games in which the player's pure strategy set is not finite—for example if the players strategy set is the set $A_i = [0, 1]$. In these instances it may still be the case that a player mixes using probabilities for some finite set of points in A_i; but in some cases this may not be enough.

To get an intuition for why this might not be enough consider a penalty shootout in the world cup, in which the striker can aim at any point in goal (for example she chooses a point in $A = [0, 1]^2$). The goalie can position herself anywhere in A. She succeeds in catching the ball if she chooses the same spot as the striker, but her probability of catching the ball is a diminishing and convex function of how far she is from the struck spot. If the goalie chooses a finite set of spots over which she randomizes, then the striker can identify the spots in-between the goalie's selected spots where a shot is most likely to be successful, and it is over these spots that she will randomize. But in that case, the goalie should also randomize over these spots, not the original set of spots. The logic implies that no finite set of spots will not the trick.

One way out is to consider a strategy as a probability distribution over the strategy set. In this case the player's strategy is a function, f_i, defined over A, with the property that $f(a) \geq 0$ for all $a \in A$ and $\int_A f(a)da = 1$. Given this reformulation of the strategy, the logic we discussed above for identifying mixed strategies is the same. In particular, for any two strategies, a and a', in the support of f_i, it must be the case that the player's expected utility from playing these strategies is the same. So, for example, if there are two players playing with strategies f_i, f_j, with strategy space A serving as the support for each player's strategy, then it must be that for each player:

$$\int_A f_j(a)u_i(a', a)da = \int_A f_j(a)u_i(a'', a)da \text{ for all } a', a'' \in A \tag{9.1}$$

In the more general case where each player has a strategy set A_i but only randomizes over some subset, A_i^* of A_i, the corresponding condition is then that for each player:

$$\int_{A_j^*} f_j(a)u_i(a', a)da = \int_{A_j^*} f_j(a)u_i(a'', a)da \text{ for all } a', a'' \in A_i^*$$

and

$$\text{for all } a^* \in A_i^* : \{a' \in A_i | \int_{A_j^*} f_j(a)u_i(a', a)da > \int_{A_j^*} f_j(a)u_i(a^*, a)da\} = \varnothing$$

The trick then is to find, or to characterize, functions, f, such that these conditions are true.

Example 111 *Consider a game with two players, $N = \{1, 2\}$, each with strategy set $A_i = [0, 1]$, and utility function $u_i = \begin{cases} 1 - a_i & \text{if } a_i > a_j \\ -a_i & \text{otherwise} \end{cases}$.*

We could think of this as a lobbying game, in which each player can pay a cost a_i but they win only if their payment is strictly higher than the other player's payment. There is clearly no pure strategy Nash equilibrium since if any player plays $a_i < 1$, the other's best response is to play ever-so-slightly more than a_i,in which case player i should in turn increase her offer, ever-so-slightly more. If however, a player plays $a_i = 1$, then the other player's best response is to play $a_j = 0$; in which case the first player should drop his move.

However a mixed strategy does exist in which players mix over the entire range, using density function f such that if player i uses f, then player j, is indifferent between all strategies in A, and hence is also willing to play f. Now lets see how we can locate f.

Using Equation 9.1, we need that $\int_A f_j(a)u_i(a', a)da$ does not depend on a'. Given the utility functions in this game, we need a function such that:

$$\int_0^{a'} f_j(a)(1 - a')da + \int_{a'}^{1} f_j(a)(-a')da \qquad (9.2)$$

does not depend on a'. The first part of this expression is the probability that player i "wins" using strategy a' (multiplied by the payoff from winning when a' is used); the second part is the expected payoff from losing.

There are a few ways of working out characteristics of f, using the fact that the value of expression 9.2 does not depend on a'. Most straightforward in this case is to note that $\int_0^{a'} f_j(a)(1 - a')da + \int_{a'}^1 f_j(a)(-a')da = \int_0^{a'} f_j(a)da - a'$. But this is exactly equal to 0 if f is such that $\int_0^{a'} f_j(a)da -$ a'; or, defining the distribution function F in the normal way, if $F(a')=a'$ for all a' in A This is true for the uniform density function $f(a) = 1$ defined over $[0, 1]$. If you do not spot that, another approach is to note that since expression 9.2 should not depend on a', then its first derivative with respect to a', is 0. The first derivative of $\int_0^{a'} f_j(a)da - a'$ is simply $f_j(a') - 1$. If this is equal to 0, we must have $f_j(a') = 1$, for all a', and hence $f_j(a) = 1$.

Exercise 112 *Create a game in which players have continuous action spaces but in which no pure strategy equilibrium exists and identify a mixed strategy Nash equilibrium in the game.*

Problem 113 *Can you find a game in which players have continuous action spaces but in which no pure strategy equilibrium exists and in which only one player mixes as a part of some Nash equilibrium?*

For an example where such strategies are used see Grossman and Helpman's *Special Interest Politics*, Section 9.3.1. For a ruch generalization of the logic in the example given above ina model of candidate behavior under rival voting rules, see Myerson, Roger. 1993. "Incentives to Cultivate Favored Minorities under Alternative Electoral Systems." American Political Science Review, 87. In more general cases, solving for f_i is often more difficult than in this example and sometimes requires solving differential equations.

9.5.3 Are Mixed Strategy Equilibria Believable?

In the last sections we saw that one property of mixed strategy Nash equilibria is that if a player places positive weight on more than one strategy, then her Nash equilibrium strategy is not a unique best response to the strategies of the other players.

Indeed there exists easier-to-implement-options (the pure strategies) that each player likes just as much as her equilibrium strategy but that do not themselves form part of an equilibrium. So why would a player chose to mix? In some games, in equilibrium, the level of mixing for a player i may not be determined by features of the game that affect i's payoffs *at all*; rather i's mixing may be driven by a concern of no interest to her: the requirement that, in equilibrium, *other* players be indifferent over some set of pure strategies. Such concerns then lead to some skepticism over the solution concept.

Before leaving this solution concept let's note one smart response to this concern due to Harsanyi (1973; see also Fudenberg and Tirole 6.7.2). Harsanyi shows that any mixed strategy equilibrium "almost always" corresponds to the limit of a pure strategy equilibrium in a sequence of slightly perturbed games in which players have uncertainty over each others' payoffs. In this case we can think of strategies as approximating unique best responses by individuals in an uncertain world.

You should be able to convince yourself of the logic of this approach by working through the next problem.

Problem 114 *Consider the following penalty shootout game:*

where ε is some arbitrarily small constant and C_G is a random variable that describes the goalie's type (in particular, a slight edge on his ability to catch the ball when he jumps left and the striker shoots left); given a population of Goalie types, C_G is positive half the time and negative half the time; similarly C_S describes the striker's type and is positive half the time and negative half the time. Each player knows his type (C_i) but only knows the probabilities associated with the other player's type. How would you expect each player to behave. How would the play of the game look to an observer as ε tends to 0?

		Goalie	
		L	R
Striker	L	$0,1+\varepsilon c_G$	$1+\varepsilon c_S,0$
	R	1,0	0,1

FIGURE 9.4. Penalty Shootout

9.6 Correlated Equilibria

So far we have assumed that although a player's equilibrium strategy may depend on the strategies of other players (it should be a best response to them), the strategies are are statistically independent in the sense that $Prob(a_1, a_2) = Prob(a_1).Prob(a_2)$–that is, the randomizations themselves are independent. For example consider the game in which $N = \{1, 2\}$, and for each i, $A_i = \{1, -1\}$, and $u_i : A_i \times A_j \to \mathbb{R}^1 = a_i \times a_j$. This is a simple coordination game. Using the expected utility theorem and the independence assumption, we can define the expected utility functions as

$$\begin{aligned} U_i(\sigma_i, \sigma_j) &: \Sigma_i \times \Sigma_j \to \mathbb{R}^1 \\ &= \sigma_i\sigma_j + (1-\sigma_i)(1-\sigma_j) - (\sigma_j(1-\sigma_i) + \sigma_i(1-\sigma_j)) \\ &= 1 - 2\sigma_i(1-\sigma_j) - 2\sigma_j(1-\sigma_i). \end{aligned}$$

A mixed strategy Nash equilibrium is given by $\sigma_i = \sigma_j = .5$, giving expected utility of 0. This is Pareto inferior to the equilibria in which $\sigma_i = \sigma_j = 0$ and $\sigma_i = \sigma_j = 1$. The reason that they do badly under the mixed strategy equilibrium is because their independent randomizations out some positive probability on the un-coordinated outcomes. They could do better if their randomizations were not independent. As an extreme case, imagine that in order to choose their strategy they each look at their watches and play "1" if there are an odd number of seconds displayed and "-1" if there is an even number displayed If the players take their decisions at the same time –and have well synchronized watches—their randomization is more likely to result in coordination. in effect, they are both using some public randomization device.

The study of correlated makes use of the possibility of such public randomization mechanisms to avoid Pareto dominated outcomes. In the case of multiple Nash equilibria, using a public signal can readily be employed to select among the equilibria. Since the strategies considered are themselves equilibria, it is an equilibria for players to implement the outcome determined by the public signal. Hence, equilibria are selected by a coin toss.

The notion gets more bite however in cases where the expected value of some distribution over Nash equilibria is itself dominated by some other outcome that is not itself a Nash equilibrium. In this case, correlated signals can still be used to weight players strategies towards outcomes that are not themselves the result of a pure strategy Nash equilibrium. In this case we do not require that a public signal in the sense that all players observe the *same* signal; rather, it is enough if all players receives some private signal, but that these signals are correlated.[3]

Formally we define a correlated equilibrium (adapted from Vega-Redondo, 2003) as follows:

Definition 115 *A "**correlated equilibrium**" is a probability density* $p : \times_{i \in N} \Sigma_i \rightarrow [0,1]$ *such that for each* $i \in N$*, and for all mappings* $\tilde{\sigma}_i : \Sigma_i \rightarrow \Sigma_i$ *we have:* $\sum\limits_{\sigma \in \Sigma} p(\sigma) u_i(\sigma) \geq \sum\limits_{\sigma \in \Sigma} p(\sigma) u_i(\tilde{\sigma}_i(\sigma_i), \sigma_{-i})$*.*

This means that if the public randomization device (privately) recommends playing strategy σ_i to player, i, then player i's, expected payoff from playing this strategy–where her expectations are formed by considering the probabilities of each *profile* being played–is better than what she would get by deviating (playing some $\tilde{\sigma}_i(\sigma_i)$ instead of σ_i).

Exercise 116 (Aumann) *Consider the following symmetric game (Figure 9.5). There are two pure strategy Nash equilibria, but symmetric randomization between them leads to an expected payoff less than would be obtained from* (a, a)*. Is there a device that could allow players to do better, using a correlated equilibrium, than they do under this symmetric randomization across Nash equilibria?*

9.7 Focal Equilibria and Pareto Dominance

Consider the game represented in Figure 9.6.

In this game there are 3 pure strategy as well as mixed strategy equilibria (find some mixed strategy equilibria!). But the Nash equilibrium solution does not provide any guidance on which of these outcomes is more likely to occur. In experimental settings however when this type of game is played, the outcome (M, M) is selected as being more likely than the other outcomes. Observations of this form have led to the notion of "focal equilibria,"

[3] For example, a dice may be rolled and Player 1 may be recommended to take action a whenever the number that came up is greater than 3, and to take action b otherwise; while player 2 may be recommended to take an action a whenever the number that came up is a multiple of 2, and to take action b otherwise. In this case, if Player 2 is told to take action a, he knows that the outcome was either 2, 4 or 6 and so he puts a 2/3 probability on player 1 also having been given the recommendation to do action a. If however player 1 receives the signal to do a, she believes that the dice must have come up 4, 5 or 6, so she also believes that there is a 2/3 probability on player 2 also having been given the recommendation to do action a.

		II	
		a	*b*
I	*a*	4 4	5 1
	b	1 5	0 0

FIGURE 9.5. Correlated Equilibria

		I		
		L	M	R
II	L	100, 100	0, 0	0, 0
	M	0, 0	99, 99	0, 0
	R	0, 0	0, 0	100, 100

FIGURE 9.6. Focal Point Equilibrium

a solution concept associated with Thomas Schelling (1960), in which an outcome is not simply a Nash outcome but is one that distinguishes itself from all others and around which players may form their expectations about the strategies that other players may employ. If there is a unique such equilibrium we call it the focal equilibrium.

Warning: Use this idea with caution! The notion of *which* outcomes are more "distinguished" is deeply problematic as different outcomes may be uniquely distinguished along different dimensions of variation, yet we have no bounds on what the relevant dimensions of variation are. They may, for example, include seemingly irrelevant aspects of the description of the game (such as the labels of the strategies or the ordering of the strategies in a matrix) or they may depend on cultural factors that are not included in the formal description of the game.

The **Pareto efficient Nash equilibrium solution** is a variation of the focal equilibrium idea that restricts attention to aspects that are included in the formal definition of the game, and in particular to features of the players' preferences over strategy combinations. A Nash equilibrium, σ, is Pareto efficient relative to other Nash equilibria if there is no other Nash equilibrium, σ', such that all players prefer σ' to σ. This removes possible ambiguities associated with the focal point concept and may in some cases lead to a tighter set of predictions then the Nash equilibrium concept. Hence in the example below the principle would suggest that the outcome will be (L, L) and not (R, R)—a finding consistent with the empirical literature.

However, the justification for selecting (L, L) as the solution relies on a focal point argument and a behavioral assumption. It is still the case that if, for whatever reason, one player felt that another player was "very likely" to play R, then we would expect that that player would play R.

Aha! In fact the seed of a rival solution concept–Risk Dominance–is contained in the words "very likely"...

9.8 Strong equilibrium

The idea of a "**strong equilibrium**," due to Aumann (1959) attempts to narrow down the set of Nash equilibria by looking for criteria to dismiss Nash equilibria from which some joint deviations are beneficial to *more than one player*. Here is the formal definition:

Definition 117 *A strategy profile σ^* is a strong equilibrium if for every subset M of N, there is no sub-profile $(\sigma_i')_{i\in M}$ such that $u_j((\sigma_i')_{i\in M},(\sigma_i^*)_{i\in N\setminus M}) > u_j(\sigma^*)$ for all j in M.*

The following two properties follow very quickly from the definition:

- Strong equilibrium implies Nash equilibrium
- Strong equilibrium implies Pareto efficiency

These properties indicate how strong the equilibrium concept is. In fact it is *so* strong that it may often fail to exist In particular, it does not exist in games in which all Nash equilibria are Pareto dominated—as for example in the Prisoners' Dilemma. They also help us to see that the set of Pareto efficient Nash equilibria (where efficiency is defined within the class of nash equilibria) is not a subset of the set of strong equilibria.

You could imagine generating a series of weaker spin-off solution concepts based on this idea; for example we could define a "k-strong equilibrium" as above where the cardinality of sets M used in the definition are restricted to be less than or equal to k. In this case, Nash equilibrium is simply a "1-strong equilibrium" and a strong equilibrium is an $|N|$-strong equilibrium.

A second problem with the notion of strong equilibrium–with which we deal more directly in our discussion of cooperative game theory–is that insofar as it uses groups rather than individuals as a unit of analysis, it requires somewhat different microfoundations to the Nash concept.

		I	
		L	R
II	L	100, 100	0, 0
	R	0, 0	1, 1

FIGURE 9.7. Pareto Dominance

9.9 Coalition Proof equilibrium

The notion of a "Coalition Proof Equilibrium," due to, Bernheim, Peleg and Whinston (1987), also weakens the notion of strong equilibrium somewhat, by restricting the set of admissible deviations to those that are not themselves vulnerable to deviations within the ranks of the deviators.

The idea is that a strategy profile σ^* is coalition proof if there are no profitable k-player deviations that themselves admit no profitable deviations by any m players from among the k players for $1 \leq m \leq k$. This rules out a deviation by some pair from consideration, of one of the players from the pair then has an incentive to deviate unilaterally from the deviation; and it rules out deviations by trios that induce pairs or individuals to deviate, and so on. From this we can see that the Nash equilibrium in the Prisoners' Dilemma is also coalition proof (since the coalition that can improve upon it, is itself subject to deviation by its own members).

9.10 Perfect Equilibrium (Trembling Hand Perfection)

Another approach, aimed at reducing the set of equilibria, is to use the idea of trembles in a system–the possibility that players make mistakes, and ask the question: is an equilibrium robust to trembles? If one Nash equilibrium is robust, while another is not, this can serve as a good way to distinguish between them.

A good way to think about the possibility of trembles is to require that an acceptable strategy profile σ should have the property that it includes each player's best response to some (but not every) slightly perturbed version of the game. This gives rise to the following definition:

Definition 118 *Strategy profile σ is a "**trembling hand perfect equilibrium**" if there exists a sequence $(\sigma^n)_{n=0}^{\infty} \gg 0$ that converges to σ such that for each i, σ^i is a best response to each σ^n.*[4]

If this is true then the strategy σ is robust to some set of trembles.[5] **Existence**: A nice property is that trembling hand perfect equilibria exist in finite games. Also, they correspond, in 2 player games, to the set of

[4] The notation "$(\sigma^n)_{n=0}^{\infty} \gg 0$" should be interpreted as: every element in σ^n (the set of probability distributions at each information set) accorded strictly positive weight to all actons in its support.

[5] Note that all that is required is that there exists *some* such sequence, the fact that you can also find some sequence such that σ^i is not a best response to σ^n does not rule out the possibility that σ^i is trembling hand perfect. Hence showing that σ is trembling hand perfect is easier than showing that it is not.

mixed strategy Nash equilibria for which neither players strategy is weakly dominated. In the 2-player case this gives a very easy method for identifying them.

9.11 Proper equilibrium

The notion of a "proper equilibrium" takes the notions of trembles one step further. rather than allowing any type of tremble, it only consider trembles that are in some way inversely proportionate to the payoff loss related to the trembles. A formal definition follows:

Definition 119 *An "ε-proper equilibrium" is a totally mixed strategy profile σ^{ε} such that, if $u_i(s_i, \sigma^{\varepsilon}_{-1}) < u_i(s'_i, \sigma^{\varepsilon}_{-1})$ then $\sigma^{\varepsilon}_i(s_i) \leq \varepsilon\sigma^{\varepsilon}_i(s'_i)$. A "**proper equilibrium**," σ, is any limit of ε-proper equilibria as ε tends to 0.*

We know (Myerson 1978) that proper equilibrium exist in finite strategic form games.

10
Solution Concepts for Evolutionary Games

The basic idea behind equilibrium concepts in evolutionary game theory is that strategies are selected based on how well they have performed in the past or based on how they are expected to perform given the present state of the world, rather than on rational beliefs about the play of others in future periods. Backwards induction for example typically plays little or no role, nor do beliefs about the best reply functions of other players. In a way these models assume that players exercise parametric rationality rather than strategic rationality (the environment is choice theoretic rather than game theoretic), but all agents do so in environments in which they respond to the past (and under some interpretations, the expected) actions of others, treating these actions as part of the givens of their choice problem—the state of the world— rather than as something that depends on their own choice.

There are many ways to model the way that individual actions affect the state of the world. It could be for example that we think of a population of players playing particular combinations of strategy. We can then introduce different types of rules that allow the relative frequency of plays of a given strategy yo rise if that strategy does "well" relative to other strategies. Different stories can be used to justify this, one is very literal—that people playing these strategies are successful and have relatively more offspring, another might reflect the relative importance of a player in the population, such as his market share or the extent of his territory or voting rights, another might work through emulation—that when players do well, other players copy them. To get a flavor of the logic we begin with some relatively simple solution concepts closely related to the Nash equilibra we have already seen; we then consider teh ESS solution concept and end with a brief discussion of stochastic stability.

10.1 Resistance, Risk Dominance and Viscosity

For the game in Figure 9.7 above let μ_I denote player I's belief regarding the probability that II plays R. Note now that L would be an optimal strategy for I for any values of μ_I in the range $[0, \frac{100}{101}]$, choosing R is only optimal for the "smaller" range $[\frac{100}{101}, 1]$. Hence it seems that a wider range of beliefs support (L, L) than support (R, R). It turns out that this consideration can be used to derive alternative solution concepts.

In one approach this intuition corresponds to the notion of the "resistance" of an equilibrium. The idea can be motivated by an evolutionary model of strategy selection. Assume that a large set of agents plays a symmetric two player game Γ in which there are k rival equilibrium strategy profiles, $\{\sigma^1, \sigma^2, ..., \sigma^k\}$. Assume that some proportion of agents, μ^i is programmed to play the strategies associated with each σ^i. Then (again assuming von Neumann-Morgenstern utility) a player of type i has expected utility given by: $\sum_{j=1}^{k} \mu^i u_i(\sigma^i, \sigma^j)$. In evolutionary games of this form it is assumed that given some initial distribution of types, the distribution evolves after each round of play with each group growing in proportion to its expected payoff.

We are now ready to describe the resistance of one strategy against another. Formally, the "**resistance**" of strategy σ^i against σ^j is given by the largest number $\mu \in [0, 1]$ such that for all $i \in N$:

$$u_i(\sigma^i, \mu\sigma^j + (1-\mu)\sigma^i) \geq u_i(\sigma^j, \mu\sigma^j + (1-\mu)\sigma^i)$$

In words: μ is is the maximal fraction of σ^j types that can be introduced into a σ^i population such that the reproductive capacity of the σ^is is at least as good as the reproductive capacity of the σ^js. If it is the case that when the proportion of σ^j types is less than μ we have that the inequality in the last expression is strict, then this may provide grounds for us to expect the σ^j types to "drive out" the σ^j types over time.[1]

Hence, in this context the notion of resistance can be used to rank pairs of strategies against each other. In the case where each player only has two pure strategies available to her in Γ, we say that σ^i "**risk dominates**" σ^j if the resistance of σ^i against σ^j is greater than the resistance of σ^j against σ^i. An equilibrium then is "**risk dominant**" if it is not dominated by any other equilibrium.

[1] Although establishing such a claim formally requires a richer sescription of the game and the set of all strategies employed in the initial generation.

It is easy to check that the Pareto optimal outcome in the game in Figure 9.7 risk dominates the Pareto inferior outcome. As the next example shows however, the notion of risk dominance is independent of Pareto dominance: of two Nash equilibria, the risk dominant outcome may be Pareto dominated by the risk dominated outcome.

		I	
		L	R
II	L	10, 10	-1000,0
	R	0,-1000	9, 9

FIGURE 10.1. Risk Dominance

Problem 120 *For Figure 10.1, calculate the resistance of* (L, L) *relative to* (R, R)*. Calculate the resistance of the mixed strategy equilibrium relative to each of these pure strategy equilibria.*

Even in the context of risk dominance, Pareto optimality has some appeal; in evolutionary settings a simple way to reintroduce a role for Pareto dominance is to loosen the assumption that agents meet randomly. In particular assume that there is an element of *viscosity* in the sense that players are more likely to meet players within their own subgroup (herd, ethnic group, party...) than they are to meet agents outside their group. If this is the case, then we may expect that given some initial random distribution of strategy types across subgroups, most subgroups will end up playing uniformly risk dominant strategies, but those that *do* end up playing Pareto dominant (although possible risk dominated) strategies will grow faster in the long run and eventually dominate overall. Following Myerson, we construct a viscosity parameter δ and say that a symmetric strategy profile σ of a two-person symmetric game is a "**δ-viscuous equilibrium**" if for each pure strategy $a \in A$:

$$\text{if } \sigma_i(a) > 0 \text{ then } a \in \arg\max_{a' \in A} [u(a', (1-\delta)\sigma_i + \delta a')]$$

And so if a is employed by an individual in a group then it should be a best response in a situation in which that individual encounters other individuals in her group with probability δ and other members of the population (that employ σ_i) with probability $1 - \delta$. Clearly as δ goes to 0 this condition becomes a necessary condition for symmetric Nash equilibrium.

10.1.1 Identifying Risk Dominant strategy profiles

The following fact is useful for the class of two person, two strategy coordination games representable by Figure 10.2 and in which (L, L) and (R, R) are the only two pure strategy Nash equilibria.

		I	
		L	R
II	L	a_{11}, b_{12}	a_{12}, b_{12}
	R	a_{21}, b_{21}	a_{22}, b_{22}

FIGURE 10.2. General 2 × 2 coordination game

In these cases we have that (L,L) is **risk dominant** if it maximizes the product of the gains from unilateral deviation, that is, if $(a_{11} - a_{21})(b_{11} - b_{12}) \geq (a_{22} - a_{12})(b_{22} - b_{21})$. (See H. Peyton Young, Chapter 4 and Harsanyi and Selten 1988)

Problem 121 *Prove it.*

Remark 122 *We have a solution concept that identifies particular Nash equilibria based on the range of expectations about the behavior of other players that may support it. We motivated this by considering an evolutionary process. But it turns out that the risk dominant equilibrium solution concept has equivalency properties with another solution concept,* ***stochastically stable states****, that is based on adaptive learning in stochastic environments.*[2]

10.2 Evolutionarily Stable Strategies

We now consider perhaps the most important evolutionary solution concept in the study of normal form games: the "evolutionary stable strategies.". The idea is to consider a large population of agents that all play a given strategy σ and ask how will this population fare if a "mutant" strategy σ' is introduced. Again we assume that players play two player symmetric games after being paired through some random process. The stability

[2] For more on this see H. Peyton Young, 1998.

requirement is this: the stable strategy should do strictly better than the mutant strategy on average and so the mutant should die out over time. Specifically, this means that given the introduction of fraction ε of mutants, for σ to be an **Evolutionarily Stable Strategy (ESS)**, we require that for all σ' and for ε sufficiently small:

$$\varepsilon.u(\sigma,\sigma') + (1-\varepsilon)u(\sigma,\sigma) > \varepsilon.u(\sigma',\sigma') + (1-\varepsilon)u(\sigma',\sigma) \quad (*)$$

Now if the *profile* in which all players play σ is not a Nash equilibrium then there exists some σ' for which $(1-\varepsilon)u(\sigma,\sigma) < (1-\varepsilon)u(\sigma',\sigma)$ and hence, for ε sufficiently small condition $(*)$ can not be satisfied. So a necessary condition for satisfying $(*)$ is that the profile in which all players play σ be a Nash equilibrium. Studying condition $(*)$, we can then write the necessary and sufficient conditions as follows:

- **Case 1**: If σ is not the unique best response to itself and so for some $\sigma' \neq \sigma$, $(1-\varepsilon)u(\sigma,\sigma) = (1-\varepsilon)u(\sigma',\sigma)$ then a necessary and sufficient condition for σ to be ESS is that $\varepsilon.u(\sigma,\sigma') > \varepsilon.u(\sigma',\sigma')$, or equivalently, $u(\sigma,\sigma') > u(\sigma',\sigma')$.

- **Case 2**: If σ is the *unique* best response to itself then we have that $(1-\varepsilon)u(\sigma,\sigma) > (1-\varepsilon)u(\sigma',\sigma)$ and so, choosing sufficiently small ε, we can dispense with the condition that $u(\sigma,\sigma') > u(\sigma',\sigma')$.

The conditions from these two cases can be summarized in a single necessary and sufficient condition: σ is an ESS if and only if $u(\sigma,\sigma') > u(\sigma',\sigma')$ for every σ' that is a best response to σ (but which is not the same as σ).

The ESS notion is a strong solution concept and it may fail to exist. Furthermore, although we already had that every evolutionarily stable strategy profile is nash, we also have that is *proper*.

Identifying ESS's. From the condition just given, to identify an ESS you first need to locate the set of symmetric Nash equilibria. A good next step is to check to see whether for any of these Nash equilibria each element of the equilibrium is a unique best response to the other elements, if so then that those elements are ESS. (Note that there is no point trying this shortcut if the equilibrium in question is not a pure strategy since as we saw above that mixed strategies are never unique best responses (unless they are pure)). If the equilibrium is not a unique best response then you need to search for strategies σ' among the set of best responses to σ for which $u(\sigma,\sigma') \leq u(\sigma',\sigma')$

Exercise 123 *For the game in Figure 10.3, identify all the Nash equilibria and identify which (if any) of these are ESS.*

			I	
		L	M	R
II	L	1, 1	3, -3	-3, 3
	M	-3, 3	1,1	3, -3
	R	3, -3	-3, 3	1, 1

FIGURE 10.3. Find the ESS

10.3 Stochastically Stable Equilibria

The final solution concept that we shall consider for normal form games is the notion of "Stochastically Stable Equilibrium." One critique brought to bear on ESS is that it only considers isolated mutations—that is, the solution concept is useful for situations where there is a small error, but does not consider the possibility that such errors can cumulate. There might not be simply instances of isolated mutations that eventually die out, but rather a continuous process of perturbations; in such cases these accumulated effects may have, albeit with low probability, large effects on a system.

A stochastically stable equilibrium is a *state*, P, with the property that in the long run it is 'almost certain' that the system will lie in open set containing P as the noise tends slowly towards 0 (Foster and Young, 1990). A stochastically stable *set*, SSS, is a set of states, S, such that the system will lie in every open set containing S as the noise tends slowly towards 0.

We are used to thinking of equilibria as being a feature of profiles of strategies. For our purposes here, *states* are not strategy profiles, but they are functionas of a strategy profile that summarizes the relevant information. Assume that there exist k pure strategies. Let a state at time t be given by a k-dimensional vector, $s(t) = \{s_1(t), s_2(t)...s_k(t)\}$ where each element $s_j(t)$ reports the relative number of agents playing strategy s_j.

Now, assume that due to random matching (or other stochastic pressures) there is uncertainty over the value of $s_h(t+1)$, can be constructed using *Markov Chain Theory* with respect to discrete time Markov processes.

To do this we now let $p_{ss'}$ denote the probability of moving from state s to state s'.

Clearly we can now populate a $k \times k$ matrix P with elements $p_{ss'}$. We assume that P is *time invariant* (time homogenous).

Let μ_t be a probability distribution over states in time t. Hence μ_t lies on the unit simplex. For example, if the state is s_1 in period 0, then $\mu_0 = (1, 0, ..., 0)$.

In this case we can work out the *distribution* of states in time 1. It is given simply by: $\mu_1 = \mu_0 P$. In our example with $\mu_0 = (1, 0, ..., 0)$ we have $\mu_1 = \mu_0 P = (p_{00}, p_{01}, ..., p_{0k})$.

Repeating the process we have that the distribution at time 2 is $\mu_2 = \mu_1 P$. In our example we have that:

$$\mu_2 = (p_{00}p_{00} + p_{01}p_{10} + p_{02}p_{20}..., p_{00}p_{01} + p_{01}p_{11}..., ..., p_{00}p_{0k} + p_{01}p_{1k}...)$$

Which is clearly more compactly, and more generally, written as $\mu_2 = \mu_0 PP$ or $\mu_0 P^2$.

More generally, at time t we have that the distribution of states is $\mu_0 P^t$. Using these ideas we now define a useful set of terms:

- The **asymptotic frequency distribution** of a process beginning at μ_0 is $\mu_0 P^\infty$.[3]
- A process is **ergodic** (not path dependent) if $\mu_0 P^\infty$ does not depend on μ_0
- State s' is **accessible** from s' if for some t, $(P^t)_{ss'} > 0$
- States s and s' **communicate** if each is accessible from the other
- A set of states is a **communication class** if each pair communicates
- A set of states is a **recurrent class** if it is a communication class from which no state outside the class is accessible
- A state is a **recurrent state** if it is in a recurrent class; otherwise it is transient
- A state is **absorbing** if it is a single recurrent class
- A process is **irreducible** if there is only one recurrent class, consisting of the entire state space.
- A distribution, μ, is a **stationary distribution** if $\mu P = \mu$
- let N_s denote the state of integers $n > 0$ for which there is a positive probability of moving from s back to s in n periods. If the greatest common divisor of N_s is 1 for each s, then the process is **aperiodic**. As a counter example: a process that returned a state to itself with positive probability only in every second period would be periodic.

The following features are useful:

- Stationary distributions exist, and, furthermore they are unique if and only if P has a unique recurrent class.
- If P has a unique recurrent class, then the stationary distribution μ is also the time average asymptotic behavior of the process, independent of the initial state.

[3] μP^t converges almost surely at $t \to \infty$.

- If however, P has more than one recurrent class, then it is non-ergodic.
- If P is irreducible and aperiodic the stationary distribution μ not only approximates time average behavior but also approximates the actual state for large t.

For notions of stochastic stability we want to allow for the possibility that errors accumulate; in particular we want to allow for the possibility that movements from anywhere to anywhere are possible. Perhaps not in any given period, but certainly over time.

Definition 124 *A process is a "**Regular Perturbed Markov Process**" of the process P^0 if (i) for some ε^*, P^ε is irreducible for every $\varepsilon \in (0, \varepsilon^*]$ (ii) $\lim\limits_{\varepsilon \to 0} P^\varepsilon_{ss'} = P^0_{ss'}$ and (iii) if $P^\varepsilon_{ss'} > 0$ then for some $r(s, s') \geq 0$, we have $0 < \lim\limits_{\varepsilon \to 0} \frac{P^\varepsilon_{ss'}}{\varepsilon^{r(s,s')}} < \infty$.*

The term $r(s, s')$ describes the resistance of the move from s to s'. It can be shown that if such an $r(s, s')$ exists it is uniquely defined. Furthermore $r(s, s') = 0$ if and only if $P^0_{ss'} > 0$. that is there is no resistance between a and b if, even without any noise you can move freely from a to b.

Note that since P^ε is irreducible it has a unique stationary distribution and hence a unique stationary distribution that is a solution to $\mu^\varepsilon P^\varepsilon = \mu^\varepsilon$.

We are now (finally) ready to define stochastic stability:

Definition 125 (Young 1993) *A state, s, is **stochastically stable** if $\lim\limits_{\varepsilon \to 0} \mu^\varepsilon(s) > 0$*

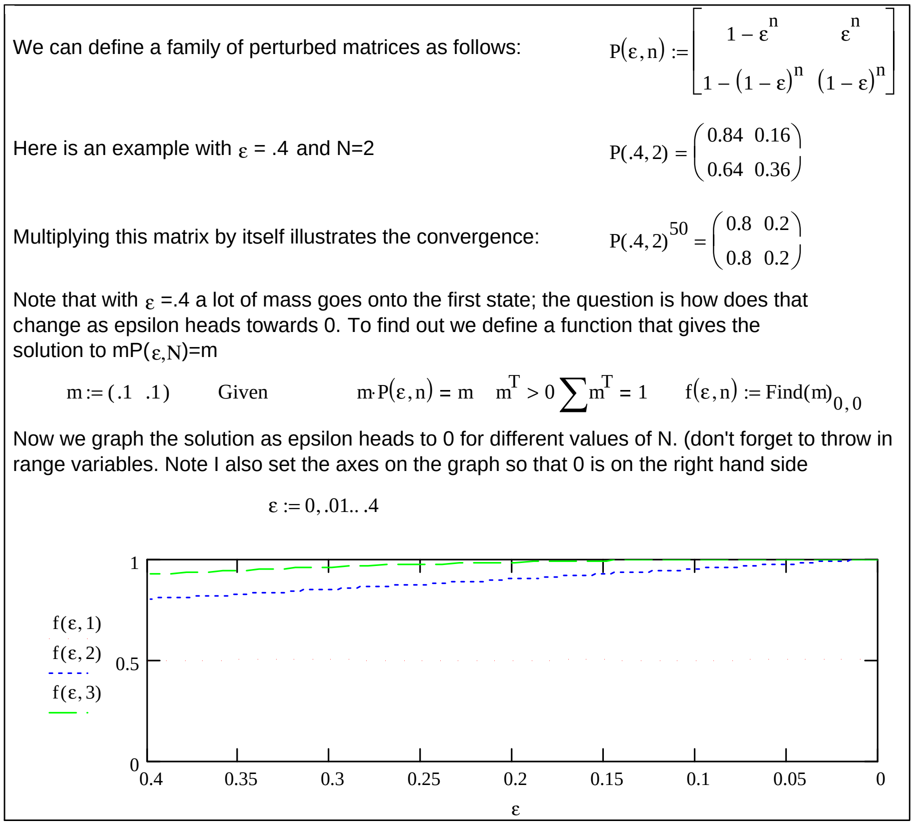
The best way to get a feel for stochastic stability is to work through some examples. Here I discuss one simple example and leave a slightly more involved one as a problem

Example 126 *Consider a process with only two states. The transition matrix, P, is then a 2×2matrix given by: $P = \begin{bmatrix} p_{11} & 1 - p_{11} \\ p_{12} & 1 - p_{12} \end{bmatrix}$. Assume that P is aperiodic and irreducible. Note that aperiodicity rules out matrices like $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, irreducibility rules out matrices like $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. Observe that if $\mu = [\mu_1, 1 - \mu_1]$ is a stationary distribution over states 1 and 2, then $\begin{bmatrix} \mu_1 & 1 - \mu_1 \end{bmatrix} \begin{bmatrix} p_{11} & 1 - p_{11} \\ p_{21} & 1 - p_{21} \end{bmatrix} = \begin{bmatrix} \mu_1 & 1 - \mu_1 \end{bmatrix}$ or equivalently: $\mu_1 p_{11} + (1 - \mu_1) p_{21} = \mu_1$ and $\mu_1(1 - p_{11}) + (1 - \mu_1)(1 - p_{21}) = 1 - \mu_1$. Although this looks like two distinct equations, manipulation of the second one should convince you that these are both the same equation. Hence we* *have one equation in one unknown. Solving for μ_1 gives $\mu_1 = \frac{p_{21}}{(1-p_{11})+p_{21}}$. This then is the stationary distribution over states. What does P^{∞} look like? It is worth multiplying P out a number of times to get a sense of how its shape evolves from $P^1, .P^2 ..., P^t$. However, since P is aperiodic and irreducible we already know that it converges and we also know what it must converge to! Since (from ergodicity) for all μ', $\mu' P^{\infty} = \mu$, it must be the case that each row of P^{∞} is the same. Furthermore it must be that $P^{\infty}_{11} = \mu_1$ and $P^{\infty}_{12} = 1 - \mu_1$. Hence $P^{\infty} = \begin{bmatrix} \frac{p_{21}}{(1-p_{11})+p_{21}} & 1 - \frac{p_{21}}{(1-p_{11})+p_{21}} \\ \frac{p_{21}}{(1-p_{11})+p_{21}} & 1 - \frac{p_{21}}{(1-p_{11})+p_{21}} \end{bmatrix}$. (What then is $P^{\infty}P$?) With P irreducible ($p_{11}, p_{22} \neq 1$) and aperiodic ($p_{11} + p_{22} \neq 0$) it is easy to check from $\mu_1 = \frac{p_{21}}{(1-p_{11})+p_{21}}$, that $\mu_1 > 0$ and $\mu_2 > 0$. Hence from our definition we can see from Definition 125 that because $\lim\limits_{\varepsilon \to 0} \mu^{\varepsilon} = \mu$, both states 1 and states 2 are stochastically stable. This follows because we have assumed irreducibility without introducing any noise at all. The somewhat unsatisfying result then is that in systems like this where without noise there is no stability (there are no recurrent classes except for the whole state), everything is stochastically stable.*

The real bite of SSS however comes for situations in which there are too many stable equilibria or the unperturbed game, not too few. That is, if P is not irreducible and consists instead of a couple of recurrent classes. In such cases we have multiple equilibria. We now want to know: can stochastic stability distinguish between these cases? The good news is that it can. Assume that there are two equilibria. In this case we have $P = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. Clearly, P does not have a unique recurrent class and so there is not a unique stationary distribution μ. Now we introduce the noise. Define the perturbed matrix $P^{\varepsilon} = \begin{bmatrix} 1 - \varepsilon^{|N|} & \varepsilon^{|N|} \\ 1 - (1-\varepsilon)^{|N|} & (1-\varepsilon)^{|N|} \end{bmatrix}$. The idea of this perturbation (many different perturbations are imaginable) is that from state 1 you need a compounding of $|N|$ errors—everybody has to make a mistake—in order to move to state 2. If the probability that one person makes an error is ε, then $|N|$ (independent) errors will occur with probability $\varepsilon^{|N|}$. However a shift from state 2 to state 1 only requires that one person makes an error: this occurs with probability $1 - (1-\varepsilon)^{|N|}$. (Note, you should think of other types of relevant perturbations and see how they would affect the outcome to follow)

*We now need to know: Is P^{ε} a **regular perturbed Markov process** of P^0? The answer is "Yes" since (i) for $\varepsilon \in (0, \varepsilon^*]$, the process is irreducible as long as $\varepsilon^* \in (0,1)$ (ii) clearly $\lim\limits_{\varepsilon \to 0} P^{\varepsilon}_{ss'} = P_{ss'}$ and, finally, although perhaps less obviously, (iii) if $P^{\varepsilon}_{ss'} > 0$ then for some $r(s,s') \geq 0$, we have $0 < \lim\limits_{\varepsilon \to 0} \frac{P^{\varepsilon}_{ss'}}{\varepsilon^{r(s,s')}} < \infty$. In particular the resistances are given by: $r(1,2) = |N|$ and $r(2,1) = 1$. Note that the proposed resistances here reflect directly* *the number of errors it takes to move between states. To check that these reisstances work for the definition, note that* $\frac{P^{\varepsilon}_{12}}{\varepsilon^{r(1,2)}} = \frac{\varepsilon^{|N|}}{\varepsilon^{|N|}} = 1$*, which does not depend on* ε*, hence* $\lim_{\varepsilon \longrightarrow 0} \frac{P^{\varepsilon}_{ss'}}{\varepsilon^{r(s,s')}} = 1$*. Similarly* $\frac{P^{\varepsilon}_{22}}{\varepsilon^{r(1,1)}} = \frac{1-(1-\varepsilon)^{|N|}}{\varepsilon} = \frac{\alpha_1\varepsilon^{|N|}+\alpha_2\varepsilon^{|N|-1}+\alpha_3\varepsilon^{|N|-2}...\alpha_{|N|}\varepsilon}{\varepsilon}$*, which converges to* $\alpha_{|N|} > 0$*. So far so good, we have our perturbed process.*

We can now use the work we have already done to work out what μ^{ε} *looks like. In particular we have* $\mu^{\varepsilon}_1 = \frac{p^{\varepsilon}_{21}}{(1-p^{\varepsilon}_{11})+p^{\varepsilon}_{21}} = \frac{1-(1-\varepsilon)^{|N|}}{(1-(1-\varepsilon^{|N|}))+1-(1-\varepsilon)^{|N|}} = \frac{1}{\frac{1}{\alpha_1+\alpha_2\varepsilon^{-1}+\alpha_3\varepsilon^{-2}...\alpha_{|N|}\varepsilon^{-|N|}}+1}$*, which for* $N > 1$ *converges to* 1 *as* ε *tends to 0. Hence* ***all*** *the mass goes on to the first state—the state with the higher resistance. For N=1* $\mu^{\varepsilon}_1 = \frac{1-(1-\varepsilon)^{|N|}}{(1-(1-\varepsilon^{|N|}))+1-(1-\varepsilon)^{|N|}} = \frac{\varepsilon}{\varepsilon+\varepsilon} = .5$*. We confirm our calculations rapidly with a modelling of* P^{ε} *and* μ^{ε} *in Mathcad (figure 10.4).*



We can define a family of perturbed matrices as follows: $P(\varepsilon,n) := \begin{bmatrix} 1-\varepsilon^{n} & \varepsilon^{n} \\ 1-(1-\varepsilon)^{n} & (1-\varepsilon)^{n} \end{bmatrix}$

Here is an example with ε = .4 and N=2 $P(.4,2) = \begin{pmatrix} 0.84 & 0.16 \\ 0.64 & 0.36 \end{pmatrix}$

Multiplying this matrix by itself illustrates the convergence: $P(.4,2)^{50} = \begin{pmatrix} 0.8 & 0.2 \\ 0.8 & 0.2 \end{pmatrix}$

Note that with ε =.4 a lot of mass goes onto the first state; the question is how does that change as epsilon heads towards 0. To find out we define a function that gives the solution to mP(ε,N)=m

$m := (.1\ \ .1)$ Given $m \cdot P(\varepsilon,n) = m$ $m^{T} > 0$ $\sum m^{T} = 1$ $f(\varepsilon,n) := \text{Find}(m)_{0,0}$

Now we graph the solution as epsilon heads to 0 for different values of N. (don't forget to throw in range variables. Note I also set the axes on the graph so that 0 is on the right hand side

$\varepsilon := 0, .01 .. .4$

FIGURE 10.4.

Problem 127 *In Mathcad or otherwise, study a process with 3 players playing the game in Figure 9.7. Assume that in each period one player is chosen at random and that with probability* $1-\varepsilon$ *this player chooses his best response to the actions of the other two players, with probability* ε *he chooses the other strategy. There are four states in this game, one with nobody playing left, one with one, one with two and one with three. Write out the transition probabilities for the* 4×4 *transition matrix. Find the stochastically stable state. How would the process change with small changes in the payoffs for each action?*

Exercise 128 *Consider any normal form game that interests you in which there exists more than one Nash equilibrium (or any dynamic process in which actors can take actions at any stage to alter the state). For example a public goods provision game (n-player chicken) with heterogenous actors or a Battle of the Sexes game with different weights across players in the value they play on sociabilty. Model it as a stochastic dynamic process in which agents draw random samples of observations and play best responses to these samples; graph the evolution of the process over time. Find a summary statistic for a single process an graph the distribution of this summary statistic.*

Problem 129 *Same as above but allow for the sample from which players learn to be non-random. For example players might "talk" only to people that live near them or think like them. They may even place different weights on the "representativeness" of different people when calculating their beliefs about the strategies of other agents...*

10.4 Readings for the week after next

Next week we will look at tools for studying the general class of extensive form games and for the special case of "repeated" games—in which a normal form game is repeated many times. If you are uncomfortable with the notion of subgame perfect equilibrium, read section 6.2 of Osborne and Rubinstein before beginning the week's readings, it's just 5 pages long but provides the definitions you need and will complement the notes on the one deviation property. Of the assigned readings, read Muthoo first, then Ferejohn and then the Osborne and Rubinstein Chapter 8. In Muthoo, concentrate on pages 42-50. Muthoo describes a classic non-cooperative bargaining game. Read the set up of the model first and write down the game tree for a finite version of the game (say with 2 periods). Try and work out the subgame perfect equilibrium of the two period game: who gets the bigger share of the pie? How much bigger is it? To get a sense of what the solution would be for the infinite version you may now want to add on a third period (to make life easier add the third period to the front of the game tree rather than to the back). How does the outcome change? Now read the solution to the game and check your intuitions. You will see that Muthoo also provides a uniqueness result. Before reading the proof of this result think of how you would go about proving uniqueness. Part 3.2.3 is lighter reading but provides a good example that you may wish to follow in your papers of how to analyze a model that you have constructed. It also provides ammunition for our mythical debate on cooperative versus non-cooperative approaches. The Ferejohn model asks how electorates are able to use elections to control politicians, given that politicians are not required to do what they said they would do before they got elected. Think about what a reasonable answer to that question might look like before reading the model. In terms of picking up tools, the main thing to look out for in reading this piece is the way terms are used (in this case V^O, V^I) to represent the value of future streams of utility, conditional upon optimal behavior by all actors. The key results in Chapter 8 of Osborne and Rubinstein that we will discuss are in 8.5. You should note however that the results in this section are with respect to Nash equilibrium and not subgame perfect Nash equilibrium. If you are concerned by this, do continue to read section 8.8.

11 Solution Concepts for Extensive Form Games

11.1 The Problem With Nash Equilibrium in Extensive Form Games

As before, we treat the set of strategies as the domain of preference relations, utility functions then map from Σ into $\mathbb{R}^1$. In games of perfect information, a given combination of strategies is uniquely associated with a terminal history; hence we can treat the set of strategy profiles as the domain of the individuals' preference relations.

Assuming that preferences can be represented by a utility function, we say that a strategy profile σ is a **Nash equilibrium** if $u(\sigma_i, \sigma_{-i}) \geq u(\sigma_i', \sigma_{-i})$ for every strategy σ_i' and for every player $i \in N$.

Our first observation is that we can apply the concept of Nash equilibrium to the context of extensive form games. that's the good news.

The bad news is our second observation: Nash equilibrium is an unsatisfactory solution concept in these contexts because, by ignoring the sequencing of actions, it excludes considerations of a simple version of what's called the principle of "**sequential rationality**"—that a player's strategy should be optimal at every decision node. The following simple game illustrates the problem and why we should care.

In the game depicted in Figure 11.1, an armed group decides whether or not to kidnap the president. Congress responds by deciding whether or not to negotiate with the group. Now, the following is a Nash equilibrium: the armed group decides not to kidnap the president, the congress refuses

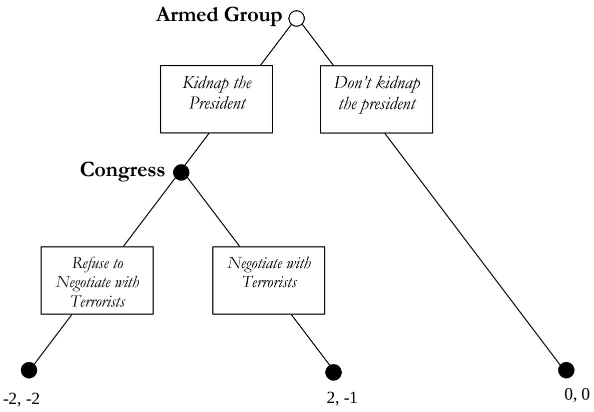

Armed Group
Kidnap the President
Don't kidnap the president
Congress
Refuse to Negotiate with Terrorists
Negotiate with Terrorists
-2, -2
2, -1
0, 0

FIGURE 11.1. One Shot Kidnapping Game

to negotiate in the event that the president is kidnapped. This is a Nash equilibrium because, conditional upon the other player's strategy, neither player has an incentive to deviate: in particular the Congress does not have an incentive to change their strategy because, in equilibrium they never in fact have to choose whether or not to negotiate.

This solution is unsatisfactory however because we would expect that *if* the armed group were in fact to kidnap the president, *then* the Congress would be better off negotiating. This line of thought however requires that we consider deviations by *both* player, but by definition the concept of Nash equilibrium only requires us to check that no single player has an incentive to deviate.

The logic suggests however that a more plausible outcome is that the president is kidnapped and the congress negotiates. This is also a Nash equilibrium, but one in which no player has an incentive to deviate at any stage in the game tree. It seems then that perhaps the problem is not with the whole class of Nash equilibria but with determining *which* Nash equilibria are consistent with sequential rationality. We have then a problem of *refinement* of the Nash equilibrium concept.

The first, and most basic refinement, is subgame perfection...

11.2 Subgames and Subgame Perfection

The logic that we discussed for using backwards induction in games of perfect information to identify sequentially rational strategies provides the basis for a solution concept known as "subgame perfection. " It makes sense to introduce the concept here because the solution concept is defined for games of imperfect information as well as games of perfect information. We begin by defining the notion of a subgame formally:

Definition 130 *A subgame of an extensive form game is a subset of the game with the following properties: (i) it begins with an information set that contains a single decision node, it contains all successors to these nodes and contains no other nodes (ii) if decision node x is in the subgame, then all nodes in the information set of x are also in the subgame.*

Problem 131 *Draw a game tree and identify a subset of the game tree where condition (i) above is satisfied but condition (ii) is not.*

Given this definition we have that every subgame is itself an extensive form game (of perfect or imperfect information). We are now in a position to define subgame perfection formally:

Definition 132 *A profile of strategies, σ, in an extensive form game, Γ, is a "**subgame perfect Nash equilibrium**" if it induces a Nash equilibrium in every subgame of Γ.*

By "σ induces a Nash equilibrium in a subgame" we mean that if we create a strategy profile σ' that specifies the same actions by all players as σ at all nodes shared by the game and the subgame, then σ' is a Nash equilibrium of the subgame.

Our example of the kidnapping game given above demonstrates that not all Nash equilibria are subgame perfect Nash equilibria; however from the definition of subgame perfection we have that all subgame perfect Nash equilibria are Nash equilibria (why?).

For finite extensive form game with perfect information a useful property of the subgame perfect solution concept is that it exists:

Theorem 133 (Kuhn) *Every finite extensive form game with perfect information has a subgame perfect equilibrium*

11.3 When Subgame perfection is not enough

So we have a better solution concept than Nash equilibrium for extensive form games. So far so good. It turns out however that in games of imperfect information, requiring subgame perfection is not sufficient to satisfy sequential rationality.

Consider the problem illustrated in Figure 11.2.

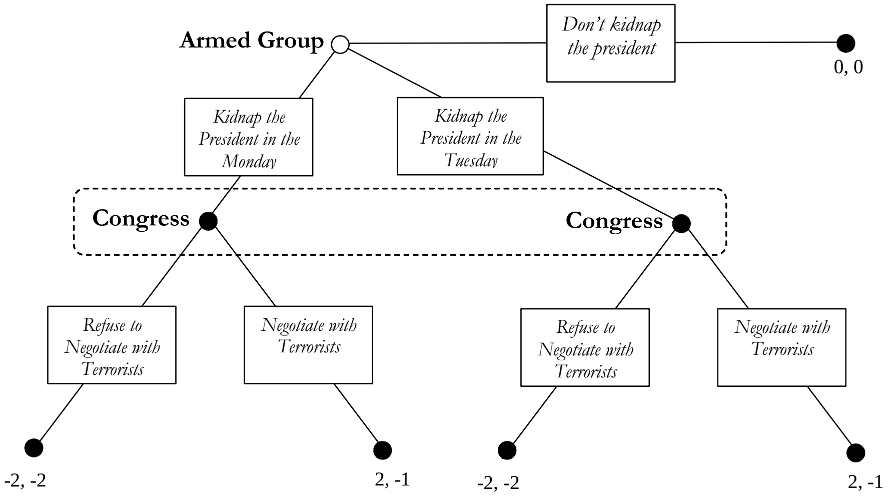

Armed Group
Don't kidnap the president
0, 0
Kidnap the President in the Monday
Kidnap the President in the Tuesday
Congress
Congress
Refuse to Negotiate with Terrorists
Negotiate with Terrorists
Refuse to Negotiate with Terrorists
Negotiate with Terrorists
-2, -2
2, -1
-2, -2
2, -1

FIGURE 11.2.

This problem is identical to the previous problem except for some irrelevant uncertainty over the timing of the capture of the president. The problem though is that, from the point of view of analyzing the game, this uncertainty prevents us from employing backwards induction: the problem is that because of the new information set there are no subgames to this game (other than the game itself). In such contexts we may often have to specify beliefs regarding which node player's believe they are at, and calculate their optimal actions conditional upon their beliefs; we can then require that player's be sequentially rational given their beliefs. in taking steps of this form we are entering the world of subgame perfect equilibrium refinements. Solution concepts that refine the idea of subgame perfection include, "fuzzy subgame perfection," "trembling hand perfection," "perfect Bayesian equilibrium," and "sequential equilibrium." We leave off the discussion of these concepts until after we have discussed tools for modelling how player's learn during the course of a game.

12
Solving Extensive Form Games

12.1 Backwards Induction

12.1.1 Procedures and Properties

The procedure that we use to identify sequentially rational Nash equilibria in finite extensive form games is called "backwards induction." The procedure is as follows:

1. Calculate optimal actions at the final decision nodes of the game tree
2. Then calculating optimal actions in the next to last set of nodes conditional upon the optimal actions from the final set of nodes being implemented.
3. Continue in this manner right up the game tree until there are no choices left to make.
4. At this point we will have identified a set of actions to be taken at each decision node in the game tree

Three properties of this procedure are worth knowing (see: "**Zermelo's theorem**"):

1. In a finite game of perfect information, every strategy that can be identified using backwards induction is a Nash equilibrium.
2. In every finite game of perfect information there is at least one Nash equilibrium that can be identified using backwards induction.

3. If no player has the same payoffs at any two terminal nodes then there exists a *unique* Nash equilibrium that can be identified using backwards induction

.

12.1.2 Backwards Induction with Continuous Action Spaces

Before introducing the closely related (but more general) notion of subgame perfect equilibrium, let's consider how to apply backwards induction in situations where players have continuous strategy sets. In such situations we often do not use trees to model the game explicitly, yet the principle of backwards induction is basically the same: simply consider the final stage of the game, work out the optimal strategies at the final stage conditional upon whatever actions have been taken before (even though you do not yet know what those values are) and then work out optimal strategies at earlier stages, taking into account that optimal choices at later stages can be foreseen. Here's a simple example:

An investor has to choose some share, α, of \$1 to invest in a country. The government then decides what tax rate, $t \in [0,1]$, to apply to the investment. The investor's payoff is given by $\alpha(1-t)+\frac{1-\alpha^2}{2}$—that is, the investor gets the non-taxed income from its investment, given by $\alpha(1-t)$, but loses the opportunity cost of investing elsewhere; its gains from investing in some other project are given by $\frac{1-\alpha^2}{2}$. The government's payoff is given by $\frac{1}{\alpha}\ln(t\alpha)+\ln((1-t)\alpha)$, that is, it gains directly from the tax revenue but also benefits from the direct returns to the country.[1]

Solving the government's problem first (first, because they move last) we choose t to maximize $\frac{1}{\alpha}\ln(t\alpha)+\ln((1-t)\alpha)$. First order conditions yield: $\frac{\alpha}{\alpha(t\alpha)}-\frac{\alpha}{(1-t)\alpha}=0$ and so $\frac{1}{t\alpha}=\frac{1}{1-t}$. Solving for t (hopefully) gives $t=\frac{1}{1+\alpha}$. Note that for $\alpha\in[0,1]$ this lies between $\frac{1}{2}$ and 1 and it is decreasing in α: if only a small amount is invested the tax rates will be high, but they fall as more is invested. The key thing to note here is that we find a solution to t that conditions upon α *even though we do not yet know the value of* α. Turning now to the investor's decision we want to find the α that maximizes $\alpha(1-t)+\frac{1-\alpha^2}{2}$. The key here is that we do not simply maximize this function with respect to α. Rather we take account of the fact that the

[1] Note that this utility function is chosen for simplicity not for its realism. For α and t strictly beween 0 and 1 the first part is increasing in t and (less obviously, α). The second part is increasing in α but decreasing in t. This captures the effects the mixed motives we may expect a government to have with regard to the direct and indirect benefits of investment. The key problems with the utility function is that they are much too specific, and, more of a technical problem, that it is not defined if α or t are 0 (this can be fixed by adding an arbitrarily small term inside the ln function).

investor knows that whatever α she chooses, this choice will be observed by the government and will affect the government's choice of tax rate: the more he invests the lower will be the tax rate on his investment.

So, looking ahead, a strategic investor sees her problem as finding an α to maximize $\alpha(1-\frac{1}{1+\alpha})+\frac{1-\alpha^2}{2}=\frac{\alpha^2}{1+\alpha}+\frac{1-\alpha^2}{2}$. First order conditions are given by: $\frac{2\alpha}{1+\alpha}-\frac{\alpha^2}{(1+\alpha)^2}-\alpha=0$. The solutions to this are given by $\alpha=0$, $\frac{\sqrt{5}-1}{2}$ and $\frac{-\sqrt{5}-1}{2}$.[2] Among these the investor's utility is maximized at $\alpha^*=\frac{\sqrt{5}-1}{2}$.[3]

With the investor's action now fully determined, we can return to determine the tax rate that the government sets in equilibrium. It is given by $t^*=\frac{1}{1+\alpha^*}=\frac{1}{1+\frac{\sqrt{5}-1}{2}}=\frac{\sqrt{5}-1}{2}$.

Remark:

Substituting for stage t decisions before maximizing at $t-1$ is key.[4] This is the most important principle of backwards induction. If the investor simply maximized $\alpha(1-t)+\frac{1-\alpha^2}{2}$ (thereby ignoring the effect of α on t) then she would choose $\alpha=1-t$. Together with the condition $t=\frac{1}{1+\alpha}$ we would then have a solution at $a=0$ and $t=1$—and hence we would underestimate the amount of investment and overestimate the amount of taxation. Doing this would get the maths wrong and the politics wrong.

Problem 134 *Find the solution to this game if the government can set the tax rate first and the investor then chooses how much to invest. Who does better and who does worse? What are the policy implications?* $\alpha=1-t$, $\frac{1}{\alpha}\ln(t\alpha)+\ln((1-t)\alpha)$, $\frac{1}{1-t}\ln(t(1-t))+\ln((1-t)(1-t))$

Exercise 135 *Design a two person game extensive form game of perfect information in which each person can take an action from some continuous (possibly multidimensional) action set and the utility to each is a function of the actions of both. For example in the spatial model or in a trade model you could imagine that each player controls one dimension of policy (such as domestic tariffs) but both care about both. Choose some utility functions and*

[2] That $\alpha=0$ is a solution can be seen immediately. Assumig $\alpha\neq 0$ and fividing across by α reduces the first order condition to the quadratic $1-\alpha-\alpha^2=0$ which is easily solved to find he other roots.

[3] Unfortunately the investors objetive function is not globally concave so we cannot rely on second order conditions. We can however check that her utility is increasing in α for $\alpha\in[0,\frac{\sqrt{5}-1}{2})$ and decreasing fro $\alpha\in(\frac{\sqrt{5}-1}{2},1]$.

[4] **Tool**: Actually substituting requires you to be able to solve for t. We could do that in this example but sometimes we cannot. If we cannot substitute for t directly then instead we can replace t with $t(\alpha)$ in order to make it explicit that t depends upon α. When we differentiate with respect to α in order to find the investor's optimal α we then also differentiate $t(\alpha)$.

explore how the outcome looks if the players choose their optimal policies sequentially.

12.2 Identifying Subgame Perfect Nash Equilibria

12.2.1 Generalized Backwards Induction

Subgame Perfect Nash Equilibria for games of complete or incomplete information can be identified using a generalized version of backwards induction. Do do it as follows:

1. Identify the Nash equilibria of all of the final subgames (these final subgames may contain many nodes).

2. Then identify the Nash equilibria in the next to last set of subgames conditional upon one of the Nash equilibria that you identified in Step 1 being played in each of the final set of subgames.

3. Continue in this manner right up the game tree until there are no subgames left.

4. At this point we will have identified a complete set of strategies for all players and this set of strategies forms a subgame perfect Nash equilibrium.

This procedure lets us identify a subgame perfect Nash equilibrium, note though that this does not imply that there exists a unique subgame perfect Nash equilibrium. For example if there are multiple Nash equilibria in any of the final subgames then there are multiple sub game perfect Nash equilibria in the game itself. We also have however that if an extensive form game, Γ, consists of a finite series of games in strategic form in which the outcome of each strategic form game is observed after it is played and players' utilities are given by the sum of the utilities derived from the strategic form games, then, if each strategic form game has a unique Nash equilibrium, the game Γ has a unique subgame perfect Nash equilibrium.

In practice when implementing this procedure for games with continuous strategy spaces you will need to follow the same principles as those given in the discussion of backwards induction. You assume maximizing actors that, when optimizing, take account of future plays and the impact of actions taken now on future strategies.

12.2.2 The One Stage Deviation Principle

The generalized backwards induction procedure given above is guaranteed to work. The problem with it is that it can be cumbersome to implement, especially when games are large. In such cases subgame perfect Nash equilibria can sometimes be readily identified use a principle variously called the One-Stage-Deviation Principle, the No-Single Improvement Principle, The One Deviation Property... this principle tells us that when we check for subgame perfection, we need not check for all possible deviations from an equilibrium, rather, if we can satisfy ourselves that no deviation in just one stage is profitable, then no deviation is profitable. the idea works by showing that if any deviation is possible, then a single deviation is possible. The intuition behind the proof is rather simple: if playing any rival strategy is profitable then consider the string of points at which deviations add to the profitability of the rival strategy, take the last of this string and consider the subgame starting there: that is a subgame in which deviating only in the first stage is profitable. Here is a more formal statement and proof of the principle:

Proposition 136 (The One Stage Deviation Principle) *A profile of strategies forms a subgame perfect equilibrium if and only if there is no stage of the game from which any player can gain by changing her strategy there, keeping it fixed at all other stages. That is: the strategy profile σ is a subgame perfect equilibrium of a finite horizon extensive form game if and only if whenever any player i moves at history h, $u_i(\sigma_i(h), \sigma_{-i}(h)) > u_i(\sigma_i'(h), \sigma_{-i}(h))$ where $\sigma_i'(h)$ differs from $\sigma_i(h)$ only in the action prescribed after history h.*

Proof. The "if" part follows from the definition of sub-game perfection. For the "only if" part we need to show that in every case in which a player has an incentive to deviate from strategy σ_i there is a subgame such the player profits by deviating only at the initial node of that subgame. So, assume that after history h, i has an incentive to deviate from σ_i by playing some other strategy (that may be altogether different from σ_i after history h). In particular, consider the rival strategy σ_i' that, among all those that are profitable deviations from σ_i, is the one (or one of the ones) that is different from σ_i after as few histories as possible.[5] In particular, σ_i' differs from σ_i after a finite number of histories (since we are considering finite games). Let h^* be the longest history for which σ_i' differs from σ_i (that is choose h^* such that σ_i' and σ_i specify the same actions for all histories after h^*). Now we then have that in the subgame beginning at h^*, there is a profitable strategy that differs from σ_i only at h^* (note that we have if

[5] For example if there are two deviant strategies σ_i'' and σ_i''' that are both profitable deviations from σ_i; with σ_i'' specifying different actions to σ_i after two histories and σ_i''' specifying different actions after three histories, then let $\sigma_i' = \sigma_i''$.

it were not profitable to deviate at h^* then σ_i' would not be the deviation from σ_i that has the smallest number of deviations). Hence, whenever a player has an incentive to deviate from strategy σ_i there is some subgame (namely the one beginning at h^*) in which she has an incentive to deviate only at h^*, conversely, if in every subgame the player does not have an incentive to deviate only at the beginning of that subgame then he never has an incentive to deviate at all. ■

Intuitively, this principle is useful because it tells us that we do not have to search for complex deviations that involve taking a hit in one period in order to regain the loss in a later period. Once we can establish that at no point is a simple deviation profitable, the principle then tells us that no deviations are profitable.[6] The principle extends to many infinite horizon games (see Fudenberg and Tirole 1995, 108-10) and to other solution concepts for games of incomplete information (see below).

Example 137 *The proof of Proposition 3.1 in Muthoo implicitly used a version of the one stage deviation principle for infinite horizon games. The principle can be used to provide easier proofs more general theories. Consider for example the following proposition that makes use of the finding of Exercise 44.*

Proposition 138 *There exists a pair* $(\bar{x},\bar{y}) \in \mathbb{R}^n \times \mathbb{R}^n$ *such that:*
$\bar{x}$ *maximizes* $\succsim_1$ *subject to* $(x,0) \succsim_2 (\bar{y},1)$ *and*
$\bar{y}$ *maximizes* $\succsim_2$ *subject to* $(y,0) \succsim_1 (\bar{x},1)$
Furthermore the following constitutes a sub-game perfect equilibrium set of strategies: Player 1 always proposes $\bar{x}$ *and accepts* $\bar{y}$ *and any proposal* y *if and only if* $(y,0) \succ_1 (\bar{x},1)$; *Player 2 always proposes* $\bar{y}$, *accepts* $\bar{x}$ *and any proposal* x *if and only if* $(x,0) \succ_2 (\bar{y},1)$.

Proof. Existence was established by you in Exercise 44. Now, using the one stage deviation principle, we check that a deviation from the prescribed strategies in any single stage does not improve the payoff of any player. There are two types of deviations to consider: (i) deviations in which in a single stage an offerer offers some outcome other than the prescribed offer; and (ii) deviations in which the receiver uses a different acceptance rule in a given stage.

(i) For Player 1, offering any $\tilde{x}$ that does not maximizes $\succsim_1$ subject to $(x,0) \succsim_2 (\bar{y},1)$ must either be an offer for which $\tilde{x} \prec_1 \bar{x}$ or else it must be

[6] For a useful analogous principle that you can use for other solution concepts such as sequential equilibrium or perfect Bayesian equilibrium see Ebbe Hendon, Hans Jorgen Jacobsen, and Birgitte Sloth. "The One-Shot-Deviation Principle for Sequential Rationality." Games and Economic Behavior 12, 274–282 (1996) http://sv5.vwl.tuwien.ac.at/literatur/GEB/Vol12-2/0018a.pdf

that $\tilde{x}$ is not acceptable to Player 2. The former is clearly sub-optimal. In the latter case, Player 1 receives $(\bar{y}, 1)$ instead of $(\bar{x}, 0)$. We now show that this is sub-optimal. Note that if $(\bar{y}, 0) \succ_1 (\bar{x}, 0)$ then (since $(\bar{y}, 0) \succsim_2 (\bar{y}, 1)$), $\bar{x}$ does not maximizes $\succsim_1$ subject to $(x, 0) \succsim_2 (\bar{y}, 1)$, a contradiction. It follows then that $(\bar{y}, 1) \prec_1 (\bar{y}, 0) \precsim_1 (\bar{x}, 0)$ and hence that choosing $(\bar{y}, 1)$ over $(\bar{x}, 0)$ is suboptimal.

(ii) Deviation in any stage where Player 1 has to choose whether to accept or reject an offer made by Player 2 occurs if Player 1 accepts an offer y with $(y, 0) \precsim_1 (\bar{x}, 1)$, or rejects an offer y for which $(y, 0) \succ_1 (\bar{x}, 1)$. In the former case, Player 1 does not improve upon his return from playing rejecting, since rejecting yields $(\bar{x}, 1)$ and $(y, 0) \precsim_1 (\bar{x}, 1)$. In the latter case accepting y rather than waiting one period and receiving $\bar{x}$ yields a lower return for sure since $(y, 0) \succ_1 (\bar{x}, 1)$.

An analogous argument demonstrates that one stage deviation is also sub-optimal for Player 2. This establishes that no single stage deviation is profitable for either player and hence the strategies form a subgame perfect Nash equilibrium. ■

Remark 139 *Note that all we have done is established that a one-stage deviation is not profitable; we have not checked for more complex deviations. Fortunately, the one stage deviation principle tells us that we do not have to. Note also that the result here is more general than that in Muthoo insofar as it allows for a much more general class of preferences (Muthoo's is essentially for linear preferences and zero-sum games).*

12.3 The Principle of Optimality

An entire field of mathematics, "Dynamic Programming" exists to solve problems that involve maximizing streams of utility over time. Dynamic programming is difficult, even if the problems involve only one player. For multiple players it is more complex. However, a core but conceptually simple principle from dynamic programming can be used for infinitely repeated games: the principle of optimality. The idea is that actions taken in a given period affect your welfare in that period but they also affect the "state" that you will be in for the next period, and hence the choices available to you (or to others) in future periods, and hence your future welfare.

In this context we define a set of possible states, S_t, with typical element s_t. For example, S_t may be given by $S_t = \{s_1 =$"I am President in time t", $s_2 =$ "I am not President in time t"$\}$. These states are themselves a function of the profile of actions taken by all players, a_t, with some value determined by a transition function of the form $s_{t+1} = g(a_t | s_t)$ (continuing our example, g is the function that gets you elected and has as argument action profile a and whether or not you were President last period).

In the one player case then, where a player has an action set A_i, assuming some exogenous s_0, her intertemporal maximization problem is of the form:

$$\begin{aligned}\max_{(a_t)_{t=0\ldots\infty}} \sum_{t=0}^{\infty} \delta^t u(a_t, s_t) &= \max_{(a_t)_{t=0\ldots\infty}} \sum_{t=0}^{\infty} \delta^t u(a_t, g(a_{t-1}|s_{t-1})) \\ &= \max_{(a_t)_{t=0\ldots\infty}} \sum_{t=0}^{\infty} \delta^t u(a_t, g(a_{t-1}|g(a_{t-2}|s_{t-1}))) \\ &= etc.\end{aligned}$$

You can see how this problem gets complicated. The way to deal with this problem is to assume that a solution exists[7] and then see what properties of the solution are implied by this, in order to pin down what the solution is. We define a value function associated with each possible state $v(s)$. Where $v(s_0)$ is the solution to $\max\limits_{(a_t)t=1\ldots\infty} \sum_{t=0}^{\infty} u(a_t, s_t)$ and describe the maximum possible utility that a player can get given that the present state is s_0. The principle of optimality, or the Bellman equation, then states that if a solution to this problem exists, we then have:

$$v(s_0) = \max\{u(a_0, s_0) + \delta v(g(a_0|s_0))\}$$

This equation provides simple conditions on what v looks like and leads to the identification of optimal actions $(a_t)_{t=0\ldots\infty}$. The procedure is to find a solution to the right hand side (which will contain the first derivative of v as an argument) and then to use the Bellman equation to pin down v.

The method is even easier to apply if there is a small set of possible states and we are interested in knowing under what conditions equilibrium actions leads to a stationary outcome—that is when will individuals take actions to preserve the status quo? This is essentially the approach used in the Ferejohn piece on the readings for this week.

Example 140 *Ignoring the complications associated with the stochastic nature of the policy process, the core of the Ferejohn model uses the principle of optimality as follows. A individual i can receive payoff V^I, if she is in office and acts optimally in this and all future periods. She receives payoff V^O if she is out of office, and acts optimally in this and all future periods. Note that these payoffs are assumed to be time independent. An agent in office (known as "the government") has Action set $A_i = \{w =$work hard, $s =$slack off$\}$. If she works hard she gets returned to office, if she slacks she gets booted out of office.*

[7] For this you will need to impose conditions that guarantee that a solution will exist. Such conditions may include assuming that A is a nonempty, time-invariant, compact subset of $\mathbb{R}^n$, that u is continous and has an upper limit $\bar{u}$, and that $\delta \in (0, 1)$.

Hence, using the Bellman equation the agent who is in office can expect an optimal payoff of:

$$V^I = \max\{u(w) + \delta V^I, u(s) + \delta V^O\} \quad (*)$$

Look hard at this and convince yourself that it is true. Now, assume that V^O is exogenous and that the government can unilaterally find a solution to maximizing $u(s)$, now let's ask the question: what does the electorate need to do to fix the utility associated with w to be such that the government will want to work hard and stay in office? The answer is that the government should want to choose w over s. This answer, together with the Bellman equation $(*)$ provides us immediately with two conditions (make sure you can see where these two conditions are coming from!):

(i) $V^I = u(w) + \delta V^I$

(ii) $u(w) + \delta V^I \geq u(s) + \delta V^O$

The first condition gives: $V^I = \frac{u(w)}{1-\delta}$

Substituting this into (ii) gives: $u(w) + \delta\frac{u(w)}{1-\delta} \geq u(s) + \delta V^O$ *and hence:*

$$u(w) \geq (1-\delta)[u(s) + \delta V^O]$$

This then is the minimum amount of utility that the government has to get from performing well in order to want to stay in office every period. It is the core condition that drives the Ferejohn model.

Problem 141 *Design any simple 2 player game (for example with 2 parties, 2 politicians, a lobbyist and a politician, two countries...), infinitely repeated, in which each person has a continuous action set but there are a finite number of states. Defend the assumption that the values achievable in each state are time independent (if they are!) and use the Bellman equation to solve the game. Keep it simple.*

12.4 Repeated Games

12.4.1 Identifying Nash Equilibria in Repeated Games

Repeated games are a special case of extensive form games. Here I focus on infinitely repeated games, although I include one exercise in which you are asked to find an equilibrium also for a finite game.

Let Γ denote a game and let Γ^* denote the game which has Γ repeated an infinite number of times. The game Γ^* is called the "supergame of Γ."

Payoffs from the supergame are typically given by some average of the payoffs of the stage game. One possibility is the Cesaro limit: the limit of $\sum_{t=1}^{h} \frac{u_i(t)}{h}$as $h \to \infty$. A second, more common, possibility is to use: $V((u_t)_{t=0,1,\ldots\infty}) = (1-\delta)\sum_{t=0}^{\infty} \delta^t u_i(t)$.

Problem 142 *Why is $(1-\delta)$ placed before the $\sum_{t=0}^{\infty}\delta^t u_i(t)$?*

Definition 143 *Let a payoff vector x be called "**feasible**" in Γ if it is a convex combination of payoff vectors to pure strategy profiles in Γ. The set of feasible payoffs is then a convex set: it is the set of payoffs that could be achieved if all players mixed jointly (in contrast to the non-convex set that we illustrated in Figure 14.1 that resulted from independent mixing).*

Definition 144 *Let a payoff vector x be called "**individually rational**" in Γ if for each player i, $x_i \geq \min_{\sigma_{-i}} \max_{\sigma_i} u_i(\sigma_i, \sigma_{-i})$. For player i, x_i is termed i's* $\min\max$*; other players can not force him below this level.*[8]

Definition 145 *Let a payoff vector x be called "**Nash dominating**" in Γ if for each player i, $x_i > \min\{u_i(\sigma)|\sigma$ is a Nash equilibrium$\}$.*[9]

12.4.2 Nash Equilibrium (Folk theorem)

Using these definitions we now state without proof the following theorem:

Theorem 146 (A Folk Theorem for Nash Equilibria) *The payoff vectors to Nash equilibria points in the supergame Γ^* is the set of feasible, individually rational payoffs in the game Γ.*

Hence to establish that a payoff vector u can be supported as a Nash equilibrium in your game you simply need to show that the payoff vector is feasible and individually rational. More constructively, you can demonstrate that such a Nash equilibrium is indeed a Nash equilibrium by demonstrating that for both players and some discount rates, player i's maximum payoff from deviating in one period plus the value to her of receiving her minimax in all future periods exceeds (weakly) her (average) valuation of u for this and all future periods.

More formally, letting u_i^* denote the highest payoff that an individual can achieve in a stage game, given the strategies employed by the other players, letting u_i^{mm} denote the player's minimax payoff, you simply need to establish that for each player i, for some δ_i and for some u_i we have:

$$u_i^* + \frac{\delta_i}{1-\delta_i}u_i^{mm} \geq \frac{1}{1-\delta_i}u_i$$

[8] In this definition the function "$\min_{\sigma_{-i}}(x)$" should be interpreted as "choose the minimum value that x can take when all players other than i choose a profile of actions σ_{-i}".

[9] The term "Nash dominating is *not* a standard term in the literature, but will be useful for what follows. Note a Nash dominating payoff vector does not necessarily Pareto dominate any Nash equilibrium payoff vector, since it only requires that each person does better than they would do under *some* Nash equilibrium.

equivalently

$$u_i \leq \delta_i u_i^{mm} + (1 - \delta_i) u_i^*$$

which in English is: that u_i is less than some convex combination of u_i^{mm} and some u_i^*.

Exercise 147 *Design a* 2 *person* 3 *pure strategy normal form game. Graph the set of feasible and individually rational payoffs to the repeat play version of your game. Choose one element in the* interior *of this set and suggest how this could be supported as a Nash equilibrium of the repeated game. Find a game in which no point in the interior of the set can be supported as a Nash equilibrium of the repeated game.*

Although in political science, we often focus on Nash equilibria in the repeated game that are Pareto superior to payoffs from repeating Nash equilibria from the stage games, Theorem 146 says nothing about Pareto optimality. In fact the theorem tells us that it is possible to sustain equilibria that are Pareto *inferior* to Nash equilibria from the stage game. These seem perverse but, because a player's individually rational payoff is never greater than, and possibly lower than, the minimum payoff that she can achieve in a Nash equilibrium (prove this!), these perverse outcomes are specifically admitted by the Theorem. This is illustrated in the following exercise.

Problem 148 *Consider the following example (Figure 12.1). The Figure illustrates a game in which each player has a dominant strategy to play A or a.*

		II		
		A	*B*	*C*
I	*a*	10 10	5 4	0 2
	b	4 5	3 3	0 1
	c	2 0	1 0	0 0

FIGURE 12.1. Folk Theorem and Pareto Dominated equilibria

Identify the unique Nash equilibrium. Show that outcome (b, B)*, can be sustained as a Nash equilibrium of the supergame for sufficiently patient players.*

The previous problem relies on the fact that the minimax payoffs are lower than any Nash dominating payoffs (indeed, lower than the lowest Nash equilibrium payoffs for each player). The gap between these two can be used to sustain inefficient equilibria through threats of pushing players into receiving payoffs that hurt them even more than the Nash equilibrium payoffs hurt. There is reason to wonder whether such threats of punishments are in fact credible, and this leads us to a discussion of subgame perfection in repeat play games.

Before we consider subgame perfection however, we note one advantage of the existence of a "gap" between minimax payoffs and payoffs from repeated stage game Nash equilibria: the gap allows us to support Pareto improvements over Nash equilibria even in *finite* games (see Benoit and Krishna, 1987). Consider the following example.

Exercise 149 *Assume that players I and II are to play the following game for T periods. The game is a kind of Prisoner's Dilemma but with an option to provide some for of self-incriminating evidence that hurts all players. Assume that each has a discount factor of $\delta = 1$ (hence we are assuming full patience, but since this is a finite game utility is finite). Show that a Nash equilibrium exists in which for some T, average payoffs (given by $\sum_{t=1}^{T} \frac{u_i(t)}{T}$ are arbitrarily close to the cooperative payoff from the stage game,* 3.

		II		
		A	*B*	*C*
I	*a*	3 3	4 1	0 1
	b	1 4	2 2	0 0
	c	1 0	0 0	0 0

FIGURE 12.2. Augemented Prisoners' Dilemma

12.4.3 Subgame Perfect Nash Equilibrium (Folk theorem)

Working through Problem 148 and Exercise 149 you will likely be concerned that the proposed solutions involve strategies that are not Nash equilibria off the equilibrium path; in other words, there is a concern of subgame perfection. In fact, since it is a stronger solution concept,subgame perfection is harder to achieve in these repeated games than is simple Nash equilibrium. For sufficiently patient players, we have the following result:

Theorem 150 (A Folk Theorem for subgame perfect Nash equilibria) *The payoff vectors to subgame perfect Nash equilibria in the supergame* Γ^* *is the set of feasible, Nash dominating payoffs in the game* Γ.

Importantly, the set of subgame perfect equilibria that is established in Theorem 150 (due to Friedman 1971) is a subset, and possibly a strict subset, of those identified in Theorem 146. Using the method described above to identify and characterize sub-game perfect Nash of the repeated game will not necessarily work for these games, although a similar method can be employed in which punishment involves employing a Nash equilibrium, and possibly a *particular* Nash equilibrium that especially punishes deviant players.

Other, stronger results than Theorem 150 exist that increase the set of identifiable subgame perfect Nash equilibria; these however typically involve very complex strategies on the parts of all players and they also, typically, require the existence of outcomes that discriminate finely across players, allowing for punishment or reward of given players without that implying a punishment (or reward) of other players.[10]

[10] For more on this, see Section 8.8 of Osborne and Rubinstein. The problem is one that has produced considerable work over the past decade.

13
Solving Extensive Form Games of Incomplete Information

The study of games with different types of incomplete information has been a growing and exciting area of research. In many contexts, political actors may have incentives to hide things: to hide what they do or to hide what they want. When they don't want to hide them, they may still be incapable of communicating their true desires because of fears others have that they may be lying. Similarly, when engaging with others, they may have uncertainty about the strategies available to their opponents, about the constraints the opponent face, over what the opponents value, over what they do, or are likely to do, and even over the rationality of their opponents: over how they reason and how they make use of information. The results and approaches we discuss this week are designed within the context of rational actors, worth thinking about though is how you might try to use some of these ideas to model situations in which you do not think that players chose strategies optimally or learn efficiently.

13.1 Identifying Equilibrium in Bayesian Extensive Games

The considerations raised in our discussion of uncertainty over player types at the end of section 8.1.1, are especially germane for the class of "Bayesian Extensive Games with Observable Actions." In these games, uncertainty is centered on the information individuals have over each others' preferences.

In particular, as discussed above, in these, the "type" of each player, i, is assumed to be drawn from some distribution Θ_i.The particular value θ_i is chosen and is private information to player i. All other information is public. This set-up is obviously appropriate for modelling situations where you do not know a person's preferences; but it can also be used to model many different types of uncertainty, such as uncertainty about the environment in which an agent is operating. For such situations we simply treat individuals in different environments as if they were different types of individuals.

In such games players choose optimally conditional upon their beliefs about the "types" of all other players. They need to take account of what each possible type might in fact do or want at any relevant information set.

A notion of equilibrium then requires both a specification of strategies, σ, and a specification of beliefs, μ. The beliefs can be treated either as a belief about the types of the players, or, in practice, as a belief about what decision node has been reached within any given information set.[1]

In the context of such games we have the following equilibrium concept:

Definition 151 *A pair (σ, μ) where σ is a profile of strategies for each type of player (not simply for each player) and μ is a collection of probability measures, with typical element $\mu_i(\iota)$ specifying a probability measure over the nodes in information set ι, is a "**Weak Perfect Bayesian Equilibrium**" if:*

[Sequential Rationality] For each player of type θ_i, moving at information set ι, player i's expected payoff from playing $\sigma_i(\theta_i)$, $\mathsf{E}[u_{i(\iota)}|\iota, \mu, \sigma_{i(\iota)}, \sigma_{-i(\iota)}]$ is at least as good for type θ_i as the expected payoff from playing any rival strategy σ_i' available to player i, $\mathsf{E}[u_{i(\iota)}|\iota, \mu, \sigma_{i(\iota)}', \sigma_{-i(\iota)}]$.

[Bayesian updating on the Equilibrium Path] On the equilibrium path, beliefs, μ, are determined by Bayes' rule and by the players' equilibrium strategies wherever possible. That is, whenever $\Pr(\iota|\sigma) > 0$, we have:

$$\Pr(x|\iota, \sigma) = \frac{\Pr(x|\sigma)}{\Pr(\iota|\sigma)}$$

It is very important to emphasize that the strategy specifies an action for each type, even though in any real situation only one type really exists. The reason is that if others do not know what the true type is they have to act and form beliefs under the assumption that they may be playing with any one of some set of possible opponents.

This notion of equilibrium can be applied to both normal and extensive form games. In normal form games, we do not have a process of learning over time and so do not need to use updating. But games are still interesting

[1] See figure 13.C.1 in MWG, p451 for a graphic illustrattion of this equivalence.

in their own right but also useful to work with to develop skills that you will use in more complex games. before moving onto the cases with learning, we consider here an example of an application with no learning.

Example 152 (incomplete information but no learning) *Consider the following variation of the problem we looked at in Exercise 109. As before, let* $N = \{1, 2, ...n\}$ *and let* $A_i = \{L, R\}$*. This time let utilities be given by:*

$$u_i(a) = \begin{cases} 0 & \textit{if all players play } L \\ 2 - \theta_i & \textit{if at least one player, including } i\textit{, plays } R \\ 2 & \textit{if at least one player plays } R \textit{ but } i \textit{ plays } L \end{cases}$$

Hence this is the same as in Exercise 109 if $\theta_i = 1$ *for all players. In this case however we assume* θ_i *is drawn from a uniform distribution on* $[0, 3]$*, and, although the distribution is common and is common knowledge, each* θ_i *is private information for each* i*.*

Exercise 153 *In this game the following is a Weak Perfect Bayesian equilibrium of the two player game. Each player of type* $\theta_i < \theta^*$ *plays R, and each player of type* $\theta_i \geq \theta^*$ *plays L, where* $\theta^* = \frac{6}{5}$*. Players' beliefs about the types of the other players are given simply by their priors.*

It's relatively straightforward to check that this is an equilibrium. We do that as follows. We fix the strategy of Player 1 and check that the strategies of Player 2 are indeed equilibrium strategies given her type. If Player 1 plays according to this equilibrium, then player 2's belief that player 1 will play R is given by $\frac{\theta^*}{3}$ *(recall that* θ_i *is drawn from a uniform distribution on* $[0, 3]$*), or for* $\theta^* = \frac{6}{5}$*, by* $\frac{2}{5}$*. If Player 2 of type* θ_i' *plays R she gains* $2 - \theta_i'$*, if she plays L she gains* $\frac{3}{5} \times 0 + \frac{2}{5} \times 2 = \frac{4}{5}$*. Playing R is better than playing L then if* $2 - \theta_i' > \frac{4}{5}$*, or* $\theta_i' < \frac{6}{5} = \theta^*$*; Playing L is at least as good as playing R if* $2 - \theta_i' \leq \frac{4}{5}$*, or* $\theta_i' \geq \frac{6}{5} = \theta^*$*. Hence the equilibrium strategies we have described do indeed characterize optimal play. Note that as in our discussion of Harsanyi's justification of mixed strategies, we now have a pure strategy equilibrium in a game that is a perturbation of the game in Exercise 109 in which we studied mixed strategies.*

Now let's get behind the strategy of proof here. The form of equilibrium is very common in these games: all types less than θ^* *do one thing and all those greater do another thing. This is a property called monotonicity and can be established ex ante by considering properties of the game.*[2] *In the proof above we verify monotonocity ex post. The trick though is to pin down*

[2] Especially useful for establishing monotonoicity is the following:

Theorem 154 (Topkis Theorem (1978)) *Consider the problem* $max_{a \in A} f(a, \theta)$*. If all of the crosspartials of* f *are strictly positive, then the optimal solution has monotone comparative statics.*

For more on this see Ashworth and Bueno de Mesquita, 2004. "Monotone Comparative Statics in Models of Politics."

θ^*. This is done as follows. We know that for any player with type $\theta_i < \theta^*$ we have that a given type will be indifferent if and only if $2-\theta_i' = \frac{\theta^*}{3} \times 2$, or $\theta_i' = 2 - \frac{2\theta^*}{3}$. Assuming monotonicity, we now focus on the indifferent type: this will be the cutoff type in the set of equilibrium strategies: all players below this indifferent type will play one strategy and all those above play the other strategy. But this simply means that the indifferent type is one for whom $\theta_i' = \theta^*$. To identify this type we then use $\theta_i' = 2 - \frac{2\theta^*}{3}$, and let $\theta_i' = \theta^*$ and then solve for θ^*. Doing this we find $\theta^* = \frac{6}{5}$.

Exercise 155 *Solve for the weak Perfect Bayesian equilibrium for the N-person game.*

We now consider a case of an extensive form game. In such games we have to consider behavior both on and off the equilibrium path. However, note that the beliefs that form part of the equilibrium are *not* constrained *off the equilibrium path*: hence players are free to hold *any* beliefs about what node they are at when at information sets that they never reach. In practice this means that you can specify equilibria that are sustained by off-the-equilibrium path beliefs that you have constructed in order to sustain the equilibrium. This is often not very compelling. We use the following example then both to illustrate both how to search for Weak Perfect Bayesian equilibria and to demonstrate why the solution concept may not be entirely satisfactory.

Example 156 *There is a diplomatic incident in which the President of country A may or may not have insulted the President of country B. The problem is that no one is altogether sure. Country A must decide whether to follow up on the possible insult with a genuine attack, or wait it out. If A attacks it beats B handily enough. If it waits it out, it may be that B tries to attack A. Now, if B attacks A, B can beat A handily enough. If it doesn't then there is no fighting, but political scientists eventually discover whether or not offense was meant by the remark. If offense was meant then A gets to gloat about the one in got away with; if not then A, but especially B gets to gloat about how level headed they are. Assume that there is a 50:50 chance that the remark was an insult. Then the game is as represented below:*

There are a number of possible Nash equilibria to this game. Here however we focus on a particular Weak Perfect Bayesian Equilibrium: one in which A always attacks, and B always attacks if A does not attack. These behavioral strategies are marked with bold lines on the game tree. Recall that weak perfect Bayesian equilibria also need to specify a set of beliefs for the likelihood of each node given that it's information set is reached. These are indicated by numbers at each of the nodes in each of the information sets.

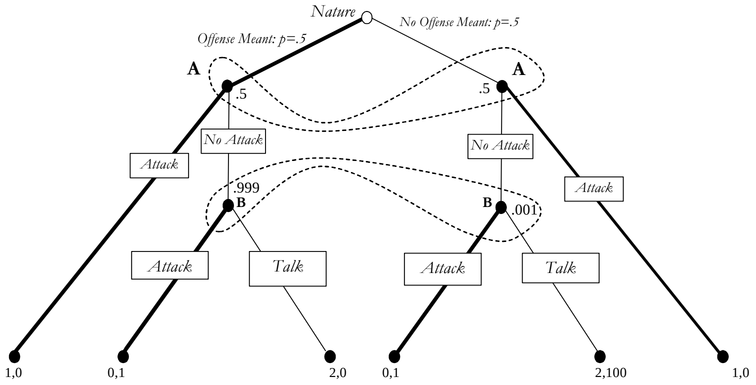


FIGURE 13.1. What did he mean by that exactly? A security dilemma.

In the last example, I proposed a Weak Perfect Bayesian equilibrium. You need to check to see if it really is one. This requires (1) Checking all the beliefs: given the strategies employed by all, do these beliefs satisfy Bayes' rule wherever possible? (2) Checking sequential rationality: given the beliefs players have about what nodes they are at within any information set, work out what the *expected* payoff of each actor is from choosing each action: are their choices optimal?

In fact verifying that a pair is a Weak Perfect Bayesian Equilibrium is not so hard; more difficult is identifying the set of Weak Perfect Bayesian Equilibria. For that I suggest the following approach: Go to the ends of the game tree and consider each strategy choice for each player. Assuming that they are choosing at a non-singleton information set, consider the set of beliefs that they would need to opt to use each of the pure strategy options available to them. For some you will find that there are reasonable beliefs, for others you will find that there are not. If there is more than one outcome that could be optimal given some beliefs, work out what set of beliefs would make the individual indifferent between all of these outcomes: these are the beliefs that are required for the individual to be willing to mix. The beliefs that you identify should be written as a function of the individual's priors and of past actions by other players, much as we did when looking at backwards induction. (For example we might have: Player i will choose x at information set ι if $\mu(x|\iota,\sigma) > .5$. And from our discussion of Bayes' rule we may have $\mu(x|\iota,\sigma) = \frac{\mu(x|\sigma)}{\Pr(\iota|\sigma)}$, where $\mu(x|\sigma)$ is a function of the strategies of other players.) Equipped with this you can now move up the game tree and see which strategies by other players will both be optimal given that the final player plays x and *also* induce the beliefs that $\mu(x|\iota,\sigma) > .5$. This method can be continued up the game tree to collect a set of equations that simultaneously need to be satisfied for all strategies to be optimal and all beliefs consistent with Bayes" rule.

Returning now to our uncertain insults game: Once you satisfy yourself that this set of strategies and beliefs form a Weak Perfect Bayesian equilibrium look to see whether you think it is reasonable. In fact it relies on player B being very confident about a particular interpretation of past events, without having *any* information to support his beliefs. He can do this because in equilibrium he should never have to make use of these beliefs, nonetheless they determine what lies on the equilibrium path. However, if in fact A placed even the tiniest weight on not going to war, then B could employ Bayes' rule and work out that, since A took his action in ignorance of the state of the world (specifically, with a 50:50 prior), then this is the correct set of beliefs for B also.[3]

We will now work through two slightly more complex games; one in the study of bargaining and one from a study of the politics of judicial decision making. In the first case a player has a continuous action space but the problem is made a little easier by discrete types. In the second the parameter over which there is uncertainty is distributed according to a continuous distribution function f over some range and so Bayesian updating will require using integration. The example is however rendered tractable by relatively simple (and non continuous) utility functions.

13.2 Applications

13.2.1 An Application with Continuous Actions and Discrete Types: A Two Period Bargaining Model

The first example is based on the first proposition in Fudenberg and Tirole (1983) ("Sequential bargaining with Incomplete Information").

Consider a case where a government and secessionist rebels are at war. The government is considering offering autonomy level $a \in [0, 1]$. The government and rebels have discount rates δ_g and δ_r which are common knowledge. If the government's offer is accepted, the government gains payoff $1-a$ and the rebels have payoff a with probability $\frac{1}{2}$ (the rebels are moderate) and $a - \theta$ with probability $\frac{1}{2}$ (the rebels are hard core), where $\theta \in (0, \frac{1}{2})$.

What should the government offer in a take it or leave it environment? Obviously if the government offered $a = \theta$ the offer would be accepted by either type (for convenience we assume acceptance in the case of indifference). If she offered something less than θ it would only be accepted by the moderates, in which case she would do better by offering $a = 0$. The

[3] Such a condition is called "**structural consistency**." It requires that there exists *some* subjective probabilities over strategy profiles that could justify the beliefs. For more, see Osborne and Rubinstein 228.1 and the following discussion.

question then is whether to offer θ or 0. Which to do depends on the government's beliefs about the rebels being hard core or not. Let q denote the government's belief that the rebels are moderate. The government is better off offering $a = \theta$ if getting payoff $1 - \theta$ for sure is better than getting $1 - 0$ with probability q and getting 0 with probability $1 - q$. That is, if $1 - \theta > q$. We then have:

Lemma 157 *For the one period game with beliefs q, the unique optimal strategy for the government is to offer θ if $q < 1 - \theta$ and 0 otherwise.*

How about the two period case? The second period of the two period case is like the one period game except that the beliefs may differ. A number of the features are the same: the government should only offer 0 or θ; the rebels should always accept θ; 0 should be accepted only by the moderates.

Moving back to the first period, we consider first the responses of the rebels to any first round offer a_1. Since in the second period nothing more than θ will be offered, accepting autonomy level θ, or more, always makes sense for all rebel types. For the hard core types, acceptance in the first round makes sense if and only if $a_1 \geq \theta$. The problem then is with the moderates who may have an incentive in rejecting some otherwise acceptable offer in the hopes of a better return in the next round.

Let $\tilde{a}_1$ denote the lowest amount of autonomy that the moderates will accept in the first round if they believe that θ will be offered in the second round should they refuse. By definition then we have: $\tilde{a}_1 = \delta_r \theta$.

We now use these elements to establish the following claim:

Claim 158 *There exists an (essentially) unique perfect Bayesian equilibrium. In the first period the government plays either θ or $\tilde{a}_1$ in the first period (depending on whether $\theta \overset{\leq}{\underset{>}{=}} \frac{1-\delta_g}{2-\delta_g-\delta_r}$). The hard-core rebels accept a_1 if and only $a_1 \geq \theta$; the moderates accept offer a_1 if and only if $a_1 \geq \tilde{a}_1$. If a second period is reached, the government believes that the rebels are hard-core with probability greater than $\frac{1}{2}$ and offers autonomy level $a_2 = \theta$ in the second period and this is accepted by all rebels.*

Proof. (i) We begin by working out the government's beliefs about the rebels in all cases in which a second round is reached.

Assume first that a first round offer less than θ is made. In every equilibrium the hard core always rejects such offers.

Let $\sigma_m(p_1)$ denote the probability with which the moderate rebel rejects any offer $a_1 < \theta$. In the event that some some offer $a_1 < \theta$ is rejected. Note that such an offer is always rejected by the hard core. The government's posterior probability that the rebel is moderate is then:

$$\begin{aligned} Pr(M|rejection) &= \frac{Pr(rejection|M) \times Pr(M)}{Pr(rej'n|H) \times Pr(H) + Pr(rej'n|M) \times Pr(M)} \\ &= \frac{\sigma_m(a_1) \times \frac{1}{2}}{\sigma_m(p_1) \times \frac{1}{2} + \frac{1}{2}} \\ &\leq \frac{1}{2} \\ &< 1 - \theta \end{aligned}$$

Hence the posterior that the rebel is a moderate (for any σ_h function) is below $1 - \theta$.

Although in no equilibrium should an offer greater than θ be made, for completeness we examine the case in which such an offer is made and is rejected (this part of the game tree is only reached if errors are made by both players!). In this case we simply assume that the posterior that the rebel is a moderate (for any σ_h function) is still below $1 - \theta$.[4]

(ii) Next we consider period 2 actions given these beliefs. From Lemma 157, the government's unique best response is to offer autonomy level θ in the second period. This offer is always accepted by all rebel types.

(iii) Finally, we consider actions in the first period.

We have already seen that the hard core rebels always reject offers below θ. Since θ is always acceptible to rebels in a second round, nothing greater than θ is ever offered in equilibrium in the second round and hence in any equilibrium a first period offer of $a_i \geq \theta$ is always accepted by rebels in the first round and hence no offer strictly greater than θ should be made.

Hence in period 1 the government must choose between offering $a_1 = \theta$ which will be accepted for sure, or offering some $a_1 < \theta$, which may be rejected, in which case she will offer θ in period 2.

If she offers $a_1 < \tilde{a}_1$ this will be rejected by both types (by the definition of $\tilde{a}_1$). If she offers any $a_1 \in (\tilde{a}_1, \theta)$, this will be accepted only if the buyer is moderate, in which case so too would $\tilde{a}_1$ be accepted, which is a better deal for the government.

The only question then is whether to offer $\tilde{a}_1$ or θ. The government does better by offering $\tilde{a}_1$ iff $\frac{1}{2}\delta_g(1-\theta) + \frac{1}{2}(1-\tilde{a}_1) > 1 - \theta$, or, substituting for $\tilde{a}_1$: iff $\frac{1}{2}\delta_g(1-\theta) + \frac{1}{2}(1-\delta_r\theta) > 1 - \theta$. Solving for θ this condition is satisfied, iff $\theta > \frac{1-\delta_g}{2-\delta_g-\delta_r}$. ■

Note that if $\theta > \frac{1-\delta_g}{2-\delta_g-\delta_r}$, then the bargaining process may involve a one period delay as first offers are rejected by hard core types. A necessary

[4] Note, although not very reasonable, it is possible that in this case we could posit that the government has other unreasonable beliefs that would alter his strategy in the second stage but these would only be relevant if the government makes an error in the first stage.

condition for this to happen is that the government is more patient than the rebels in the sense that $\delta_g > \delta_r$. To see this note that $\delta_g \leq \delta_r$ implies $\frac{1-\delta_g}{2-\delta_g-\delta_r} \geq \frac{1}{2}$, and hence delay requires that $\theta > \frac{1}{2}$, a possibility that we ruled out at the outset. The logic is that if the rebels are very patient than the government, then $\tilde{a}_1$ is close to θ; but if the government is impatient then it is needlessly taking a risk of losing delay to make a small marginal gain, relative to offering θ at the outset. In contrast if the rebels are not patient, then there is a bigger spread between $\tilde{a}_1$ and θ, and so the riskier strategy of offering only $\tilde{a}_1$ may work out to the government's advantage.

Remark 159 *With a little change in notation you can see that this problem also represents a standard bargaining problem in which a seller offers a price in two periods to a prospective buyer with unknown valuations. Simply replace a with* $1-p$. *Hence whereas the government's utility is* $1-a$, *the seller's utility is p. The moderate rebels gain a, which corresponds to buyers with utility* $1-p$, *which can be interpreted as utility when the buyers valuation is high but the buyer pays price p. The hard core types gain* $a-\theta$, *which corresponds to buyers with utility* $(1-\theta)-p$, *which can be interpreted as low valuation types, who value the good at only* $(1-\theta)$.

13.2.2 An Application with Discrete Actions and Continuous Types: A Model of Strategic Auditing

Consider the following game, based on Cameron, Segal and Songer (2000). There are two players, $N = \{L, H\}$, where L denotes the lower court and H the higher court. The game involves the decision whether to accept the facts of a case as being admissible or not, where admissibility depends on how intrusive the search was that produced the facts. If L accepts or rejects the case, H can then decide whether to audit L's decision or not. More formally, the strategy sets are $A_L = \{\text{accept, reject}\}$ and $A_H = \{\text{audit, not audit}\}$.[5] Each player wants a final outcome that is consistent with their view of acceptability, but auditing comes at a cost—both to the auditor and the audited, The true facts of a case are given by some value $x \in \mathbb{R}^1$. Each player $i \in \{L, H\}$ has a notion of maximally intrusive facts, given by $x_i \in \mathbb{R}^1$ and deems a case admissible if $x < x_i$. We assume that $x_H > x_L$, and hence anything that L finds admissible, H also finds admissible. Utilities are give simply by:

$$\begin{aligned} u_H &= I(\text{correct ruling from } H\text{'s perspective}) \\ &\quad -k_H \times I(H \text{ undertook audit}) \end{aligned}$$

[5] In the language of the original paper, L, the Lower court, can accept or reject the evidence and the higher court can grant or deny certiorari.

$$\begin{aligned} u_L &= I(\text{correct ruling from } L\text{'s perspective}) \\ &\quad -k_L \times I(L\text{'s ruling was overturned by audit}) \end{aligned}$$

Where $k_i \in (0,1)$ denotes the cost incurred on the parties by the audit.

There are publicly observable facts about the case, denoted by $\hat{x}$. There are also unobservable facts, denoted by t,where t is distributed over a space T with density f. The true facts about the case, x, are given by $x = \hat{x} + t$. The unobservable facts t, are available to L, but are available to H only if she undertakes a costly audit.

General properties of the equilibrium strategies

General features of L's Strategy.

We know the following two aspects of L's strategy for sure:

- If $x > x_H$ then L plays "reject"
- If $x < x_L$ then L plays "accept"

In the first case, playing "accept" is dominated for L: if L plays "reject" she will receive $u_L = 1$ whether or not she is audited; but if she chooses "accept" she receives either 0 or $-k_L$ (depending on whether or not she is audited).

In the second case playing "reject" is dominated for L: if x were in fact below x_L, playing "accept" would yield $u_L = 1$, no matter what H does, but if she chooses "reject" she receives either 0 or $-k_L$ (depending on whether or not she is audited).

More difficult is the situation where x lies between x_L and x_H. These are the ranges in which L and H are in disagreement over the best outcome. For this range, let us simply denote L's probability of rejecting as $s(x)$, or $s(\hat{x} + t)$, where, at the moment, we are unsure whether $s(x)$ is a pure or mixed strategy in this range

General Features of H's Beliefs.

For H to decide on her best action, she needs to work out what the chances are that x lies above x_H or below x_H, given L's actions.

In this game, it is easy to check that observing "accept" implies that the x is *below* x_H, since, as we saw above, if x were above x_H, playing "accept" would always be dominated for L. Similarly, observing "reject" implies that the x is *above* x_L, since if x were below x_L, playing "reject" would always be dominated for L.

Hence the difficult part for H is working out what are the chances that x lies *between* x_L and x_H, given that she L playing "reject." Her expectation (belief) that x lies in this range can then be denoted by:

$$\mu(x_L \leq x < x_H|reject, \hat{x})$$

Now, we have from the above that:

$$\begin{aligned}\mu(x_L &\leq x < x_H|\text{reject, } \hat{x}) = \mu(x_L \leq \hat{x} + t < x_H|\text{reject, } \hat{x}) \\ &= \mu(x_L - \hat{x} \leq t < x_H - \hat{x}|\text{reject, } \hat{x}) \\ &= \mu(t_a \leq t < t_b|\text{reject, } \hat{x})\end{aligned}$$

So really we just want to find the probability that t lies in a critical range between $t_a = x_L - \hat{x}$ and $t_b = x_H - \hat{x}$, since if $t \in [t_a, t_b]$ then $x \in [x_L, x_H]$.

From Bayes' rule we have:

$$\begin{aligned}\Pr(t_a &\leq t < t_b|\text{reject, } \hat{x}) \\ &= \frac{\Pr(\text{reject}|t_a \leq t < t_b, \hat{x}) \times \Pr(t_a \leq t < t_b, \hat{x})}{\Pr(reject|\hat{x})} \\ &= \frac{\left[\begin{array}{c}\text{probability}\\ \text{of rejection for}\\ t \in [t_a, t_b]\end{array}\right]\Pr(t \in [t_a, t_b])}{\left[\begin{array}{c}\text{prob}\\ \text{of rej'n for}\\ t < t_a\end{array}\right]\Pr(t < t_a) + \left[\begin{array}{c}\text{prob}\\ \text{of rej'n for}\\ t \in [t_a, t_b]\end{array}\right]\Pr(t \in [t_a, t_b]) + \left[\begin{array}{c}\text{prob}\\ \text{of rej'n for}\\ t \geq t_b\end{array}\right]\Pr(t \geq t_b)}\end{aligned}$$

Now, using $s(\hat{x} + t)$ to denote the probability with which L rejects given any $x = \hat{x} + t$ and using the fact that we know the equilibrium probability of rejection and acceptance outside of the critical range, we have:

$$\begin{aligned}\Pr(t_a &\leq t < t_b|\text{reject, } \hat{x}) \\ &= \frac{\frac{\int_{t_a}^{t_b} s(\hat{x}+t)f(t)dt}{\int_{t_a}^{t_b} f(t)dt} \times \int_{t_a}^{t_b} f(t)dt}{F(t_a) \times 0 + \frac{\int_{t_a}^{t_b} s(\hat{x}+t)f(t)dt}{\int_{t_a}^{t_b} f(t)dt} \times \int_{t_a}^{t_b} f(t)dt + (1 - F(t_b)) \times 1} \\ &= \frac{\int_{t_a}^{t_b} s(\hat{x}+t)f(t)dt}{\int_{t_a}^{t_b} s(\hat{x}+t)f(t)dt + 1 - F(t_b)} \\ &= \frac{\int_{x_L-\hat{x}}^{x_H-\hat{x}} s(\hat{x}+t)f(t)dt}{\int_{x_L-\hat{x}}^{x_H-\hat{x}} s(\hat{x}+t)f(t)dt + 1 - F(x_H - \hat{x})}\end{aligned}$$

General features of H's strategy.

From our discussion of L's strategy we know that whenever H observes acceptance she knows that x is no greater than x_H and so she should accept the lower courts decision. We have then that H's best response to "accept" is not to audit.

We need then to work out her best response to "reject." Her expected utility from auditing is simply $1-k_H$; her expected utility from not auditing is a function of her beliefs: she expects to receive 0 with probability $\mu(x_L \leq x < x_H|reject, \hat{x})$ and 1 with probability $1-\mu(x_L \leq x < x_H|reject, \hat{x}))$.

Hence she will be willing to audit if and only if:

$$1 - k_H \geq (1 - \mu(x_L \leq x < x_H|reject, \hat{x}))$$

or

$$\mu(x_L \leq x < x_H|\text{reject, } \hat{x}) \geq k \tag{13.1}$$

She will be willing to play a mixed strategy if and only if:

$$\mu(x_L \leq x < x_H|\text{reject, } \hat{x}) = k$$

or:

$$\Pr(t_a \leq t < t_b|\text{reject, } \hat{x}) = k$$

In the case where she mixes, let r denote the probability with which she audits.

Equilibrium

We now have all the elements we need to characterize equilibrium.

Characterizing all the Pure Strategy Equilibria

Consider first the cases with pure strategies. There are two types two consider:

(i) Assume first that L rejects whenever $x > x_L$. In this case upon observing "reject", H's belief that x is in the zone of disagreement is exactly: $\mu = \frac{\int_{x_L-\hat{x}}^{x_H-\hat{x}} s(\hat{x}+t)f(t)dt}{\int_{x_L-\hat{x}}^{x_H-\hat{x}} s(\hat{x}+t)f(t)dt+1-F(x_H-\hat{x})} = \frac{F(x_H-\hat{x})-F(x_L-\hat{x})}{1-F(x_L-\hat{x})}$. We have then that, H's best response depends on the relative size of k and $\frac{F(x_H-\hat{x})-F(x_L-\hat{x})}{1-F(x_L-\hat{x})}$. In cases where $k > \frac{F(x_H-\hat{x})-F(x_L-\hat{x})}{1-F(x_L-\hat{x})}$, we have $k > \mu$ and so from Equation 13.1, H will not be willing to audit. In such cases L is indeed wise to reject whenever $x > x_L$ and we have an equilibrium. However, if $k < \frac{F(x_H-\hat{x})-F(x_L-\hat{x})}{1-F(x_L-\hat{x})}$ then H's best response to L's pure strategy is to audit (again from Equation 13.1). In this case L receives payoff $0 - k_L$ and she clearly could have done better by accepting off the bat. So this is not an equilibrium. We have then that the strategies

- L rejects if and only if $x > x_L$
- H never audits
- $\mu = \frac{F(x_H - \hat{x}) - F(x_L - \hat{x})}{1 - F(x_L - \hat{x})}$

is an equilibrium if $k > \frac{F(x_H - \hat{x}) - F(x_L - \hat{x})}{1 - F(x_L - \hat{x})}$. (This is Part 1 of Proposition 1)

Furthermore this characterizes essentially *all* equilibria in which L plays the pure strategy: reject if and only if $x > x_L$. In particular, there is no such equilibrium when $k < \frac{F(x_H - \hat{x}) - F(x_L - \hat{x})}{1 - F(x_L - \hat{x})}$.

(ii) Assume next that L accepts whenever $x_L < x < x_H$. Let's see if this can be part of an equilibrium set of strategies. In this case we have: $\mu = \frac{\int_{x_L - \hat{x}}^{x_H - \hat{x}} s(\hat{x}+t) f(t) dt}{\int_{x_L - \hat{x}}^{x_H - \hat{x}} s(\hat{x}+t) f(t) dt + 1 - F(x_H - \hat{x})} = \frac{0}{1 - F(x_H - \hat{x})}$ and so, upon seeing rejection H believes that $x > x_H$ and so does not audit. But if H does not audit in this supposed equilibrium, then L clearly does better by rejecting whenever $x_L < x < x_H$ and tricking H. Given these incentives to deviate we have that there is no pure strategy equilibrium in which L accepts whenever $x_L < x < x_H$.

Having characterized the pure strategy equilibria, we are left with the problem that we do not have an equilibrium for the cases in which $k \leq \frac{F(x_H - \hat{x}) - F(x_L - \hat{x})}{1 - F(x_L - \hat{x})}$. There remains however the possibility of mixed strategy equilibria. We consider these next.

Characterizing the Mixed Strategy Equilibria

As always, to sustain mixing, we need to ensure that players are indifferent over the alternative over which they mix.

Supporting mixing by L. In cases where in fact $x_L \leq x < x_H$, L will be willing to mix if the payoff to excluding given H's mixing strategy $(r_H \times (0 - k_L) + (1 - r_H) \times 1)$ is equal to her payoff from admitting (0).

That is, if:

$$-k_L r_H + 1 - r_H = 0$$

or

$$r_H = \frac{1}{1 + k_L}$$

Note, a feature that simplifies the analysis is that the probability with which H needs to mix in order to keep L indifferent does not depend on any aspect of L's own strategy.[6]

Supporting mixing by H. In order for H to be willing to mix, we need to find a function $s(\hat{x}+t)$ such that:

$$\Pr(t_a \leq t < t_b|\text{reject}, \, \hat{x}) = k$$

We can do this by working backwards . By setting

$$\frac{\int_{x_L-\hat{x}}^{x_H-\hat{x}} s(\hat{x}+t)f(t)dt}{\int_{x_L-\hat{x}}^{x_H-\hat{x}} s(\hat{x}+t)f(t)dt + 1 - F(x_H-\hat{x})} = k \tag{13.2}$$

we can try to solve for some function s that always satisfies Equation 13.2. The tricky part of solving for s using Equation 13.2 is that we have to deal with all these integrals over s. It turns out however that there is an easy solution if we simply assume that the function s does not depend on t. In particular if we do that, then we can take $s(\hat{x}+t)$ outside of the integrals, yielding:

$$\begin{aligned} k &= \frac{s(\hat{x}+t)\int_{x_L-\hat{x}}^{x_H-\hat{x}} f(t)dt}{s(\hat{x}+t)\int_{x_L-\hat{x}}^{x_H-\hat{x}} f(t)dt + 1 - F(x_H-\hat{x})} \qquad (13.3) \\ &= \frac{s(\hat{x}+t)\left[F(x_H-\hat{x}) - (x_L-\hat{x})\right]}{s(\hat{x}+t)\left[F(x_H-\hat{x}) - (x_L-\hat{x})\right] + 1 - F(x_H-\hat{x})} \qquad (13.4) \end{aligned}$$

and so:

$$k(s(\hat{x}+t)\left[F(x_H-\hat{x}) - (x_L-\hat{x})\right] + 1 - F(x_H-\hat{x})) = s(\hat{x}+t)\left[F(x_H-\hat{x}) - (x_L-\hat{x})\right]$$

and

$$s(\hat{x}+t) = \frac{\frac{k}{1-k}(1-F(x_H-\hat{x}))}{F(x_H-\hat{x}) - F(x_L-\hat{x})}$$

Note that, in fact, as we assumed, $s(\hat{x}+t)$ does not depend on t.

Given the steps we have just followed, you should be able to prove the following claim without much difficulty:

[6] Exercise 162 turns out to be a bit trickier than this because in that case the probability with which one player (the government mixes) at the end of the first period itself depends on the strategies that the other player takes (the investor) at the beginning of the first period.

Claim 160 *The function $s(\hat{x}+t)$ given by $s(\hat{x}+t) = \frac{\frac{k}{1-k}(1-F(x_H-\hat{x}))}{F(x_H-\hat{x})-F(x_L-\hat{x})}$ guarantees that* $\Pr(t_a \leq t < t_b|reject,\ \hat{x}) = k$*; furthermore, this function is independent of* t.

This claim you will see, finds some confirmation in the example given on p106 of the original text. It also makes sense intuitively: it wouldn't make sense for L's mixed strategy–which, recall from our discussion on mixed strategies, is a function of H's expected payoffs, (designed to make H indifferent)–to depend on something that H can't even observe: t. the fact that s need not depend on t (in the range in which L mixes) makes it easy to solve for s explicitly.[7]

Now, since mixing probabilities must lie between 0 and 1, an equilibrium strategy of this form exists if and only if $0 \leq s(\hat{x}+t) \leq 1$ That is, if and only if: $0 \leq \frac{\frac{k}{1-k}(1-F(x_H-\hat{x}))}{F(x_H-\hat{x})-F(x_L-\hat{x})} \leq 1$. It is easy to check that $0 \leq \frac{\frac{k}{1-k}(1-F(x_H-\hat{x}))}{F(x_H-\hat{x})-F(x_L-\hat{x})}$; furthermore $\frac{\frac{k}{1-k}(1-F(x_H-\hat{x}))}{F(x_H-\hat{x})-F(x_L-\hat{x})} \leq 1$ if and only if $k \leq \frac{F(x_H-\hat{x})-F(x_L-\hat{x})}{1-F(x_L-\hat{x})}$. Since clearly there is always a range for which $0 < \frac{F(x_H-\hat{x})-F(x_L-\hat{x})}{1-F(x_L-\hat{x})} < 1$ then there is always range of $k \in (\frac{F(x_H-\hat{x})-F(x_L-\hat{x})}{1-F(x_L-\hat{x})}, 1)$ in which k is too large for this mixed strategy equilibrium to be supported.[8] In these cases the pure strategy identified above can be supported. However, in the range $k \in (0, \frac{F(x_H-\hat{x})-F(x_L-\hat{x})}{1-F(x_L-\hat{x})})$, we can support the mixed strategy equilibrium (and the pure strategy identified above can not be supported).

In the latter case, where $k \in (\frac{F(x_H-\hat{x})-F(x_L-\hat{x})}{1-F(x_L-\hat{x})}, 1)$ we have then a pair of strategies that involves mixing by H and a mixture of pure strategies and mixed strategies for L (depending on the true value of t) that are each best responses to each other given Bayesian updating (this is Part 2 of Proposition I).

Furthermore we have established an essentially exhaustive set of equilibria for this game.

13.2.3 *An Exercise with Continuous Action and Type Spaces: Bargaining Again*

Consider now a bargaining situation between government and rebels with the same utility functions as given above except this time the hard-core parameter θ is distributed uniformly over $[0,1]$.

We want to work out propertyless of the WPBE.

[7] The "latitude' that the authors refer to at this point in the text is a little trivial—L may choose to have another functional form for $s(\hat{x}+t)$ that does depend on t, but only in such a way that it won't actually change H's beliefs (that's the idea of "belief equivalence.")

[8] See Equation (3) in CS&S.

We do so by jumping to the end of the game tree, placing names on the actions on the first period, and given these, and working out the beliefs and actions that are rational responses to the first period actions.

So: Say that in round 1 the government offers a_1^* and that this is accepted, for sure, by type θ_i if and only if $\theta_i \leq \theta^*$ (hence we assume monotonicity and check this is OK later on). Note that we do not know what a_1^* and θ^* are. Nonetheless we can work out the government's posterior distribution as a function of θ^*, $f(\theta^*)$.

- [Q1] What is $f(\theta^*)$?

Given this distribution, the government in the second stage should offer a_2^* to maximize expected payoffs in the second stage: $\int_{\theta^*}^{a_2^*} f(\theta^*)(1-a_2^*)d\theta$. If we take the first order conditions of this problem, we work out what will be offered in the second period, conditional upon θ^*. Thus we find $a_2^*(\theta^*)$.

- [Q2] What is $a_2^*(\theta^*)$?

Next we want to find θ^*. We now know that in time 1, a player with type θ^* will be indifferent between accepting a_1^* and being offered (and accepting[9]) $a_2^*(\theta^*)$ the following period if $a_1^* - \theta^* = \delta(a_2^*(\theta^*) - \theta^*)$. This condition lets us solve for θ^*.

- [Q3] What is θ^*?
- [Q4] Be sure to confirm that the monotonicity assumption holds: that players with $\theta_i < \theta^*$ will prefer to accept a_1^*, and players with $\theta_i > \theta^*$ will reject a_1^*.

We now want to find a_1^*. In time 1 then, the government chooses a_1^* knowing that this will be accepted by types for whom $\theta_i < \theta^*$ and rejected by the rest; that in the case of rejection, $a_2^*(a_1^*)$ will be offered and that this will only be acceptable to players with $\theta_i \leq a_2^*(a_1^*)$. Hence a_1^* should be chosen to maximize:

$$\int_0^{\theta^*(a_1^*)} (1-a_1^*)d\theta + \int_{\theta^*(a_1^*)}^{a_2^*(a_1^*)} (1-a_2^*(a_1^*))d\theta + \int_{a_2^*(a_1^*)}^{1} (0)d\theta$$

- [Q5] Maximize this and solve for a_1^*.

Exercise 161 *Identify a WPBE for this game and prove that it is in fact an equilibrium.*

[9] To check that acceptance is the right benchmark, you should ensure at the solution that $a_1^* - \theta^* > 0$.

Problem 162 *Player I can invest some share, α_1, of \$1 in period 1 and some share, α_1, of another \$1 in period 2. After each investment, the government has the opportunity to predate. Whether or not the government is disposed to do so depends on information that is private to the government: in particular while the government knows its type, the investor believes that the government is a non-predatory type with probability p.*

Non-predatory governments never predate. Their utility can be written independent of investments as any decreasing function of the probability that they predate.

Predatory types are willing to predate (but might not) and take some share, β_t, of the player's investment; in this case their utility in that period is given by $\beta_t\alpha_t$.

Player 1 gets a per period return of $\ln\alpha_t + (1-\alpha_t)$ for any investment in time t that is not predated. Otherwise he receives $(1-\alpha_t)$. Assume that all player know p, and that their overall utility is the sum of period 1 utility plus period 2 utility, discounted by common discount factor δ with $\delta > p$.

Search for a weak Perfect Bayesian equilibrium of the form: Player 1 optimally invests $\alpha_1 > 0$ in the first period and α_2 in the second period and holds beliefs about the government's type in both periods consistent with Bayes' rule. The good government never predates (by assumption); the bad government predates in the final period but with probability q, does not predate in the first period. How much is invested in each period? Does the likelihood of predation rise or fall over time? How do the investor's beliefs about the Government's type change over time?

13.3 Refinements

Part of the problem raised in the last discussion is that we did not have reasonable restrictions to exclude strategies that require unreasonable actions to be taken "off the equilibrium path."

A stronger equilibrium requirement is that in each continuation game (the game that begins at an information set (not strictly a subgame)), the restriction of (σ, μ) to that continuation game must induce a Nash equilibrium in the continuation game. This in a sense ensures that *if* a subgame is reached (that shouldn't be reached!) we have that players play optimally. A second is that at all information sets ι, all players (other than player j) have common beliefs, $\mu_j(\iota)$ about the type of player j. In the context of the game above, this means that Player B, should have the same beliefs about whether or not an insult was meant when he comes to make his choice as does Player A. In imposing such extra conditions we can augment the notion of Weak Perfect Bayesian Equilibrium by introducing restrictions on beliefs and on strategy at points *off* the equilibrium path, such refinements produce the notion of "Perfect Bayesian Equilibrium."[10]

If you examine the example given in the last section of judicial decision making you will see that in this case all actions that are off the equilibrium path for some type are on the equilibrium path for another type. This allows the second player to apply Bayes' rule even of the equilibrium path.

However, forming such beliefs is not possible in all games. In response a series of refinements have been developled. We examine two.

13.3.1 Extensive Form Trembling Hand Perfection

The following approach is intuitively appealing and has the advantage of being well defined for Bayesian as well as more general games of incomplete information.

The approach involves checking to see what sort of strategies are robust to the possibility that players make *mistakes* that randomly throw play off the equilibrium path. The idea is that all players know that "hands tremble" (and have common beliefs over the distribution of these trembles) and will take account of this possibility when they make their choices. In particular they may be slow to choose options whose payoffs rely on a partner implementing his strategy perfectly, even if it is in the partner's interest to do so.

In our discussion of the solution concepts for normal form games, we introduced a solution concept that is robust to such trembles: "perfect" or "trembling hand perfect" equilibrium. We gave the following definition:

Definition 163 *Strategy profile σ is a "**trembling hand perfect equilibrium**" if there exists a sequence $(\sigma^n)_{n=0}^{\infty} \gg 0$ that converges to σ such that for each i, σ^i is a best response to each σ^n.*[11]

Now, for extensive form games we might expect that trembling hand perfection would also ensure sub-game perfection since, given trembles, all subgames are reached with positive probability.

However, this is not true. By requiring only that σ^i be a best response to σ^n we endow players with a certain arrogance: they choose strategies without taking into account that they themselves might tremble. Hence when they choose their strategies they may exclude the possibility that some subgames are reached, namely those subgames that they can eliminate due to their own steady hand strategies. As a result, this arrogance can

[10] Note however that the definition of perfect Bayesian equilibrium varies and the term is often defined with respect to particular games being modelled.

[11] The notation "$(\sigma^n)_{n=0}^{\infty} \gg 0$" should be interpreted as: every element in σ^n (the set of probability distributions at each information set) accorded strictly positive weight to all actons in its support.

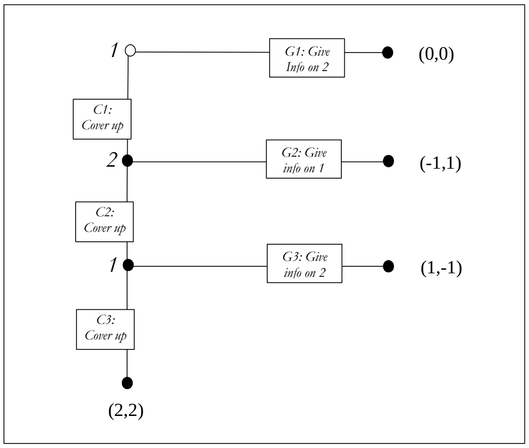


FIGURE 13.2. Tembling Hand Perfection does not guarantee Subgame Perfection

lead to the selection of some Nash Equilibrium strategies that satisfy our trembles restriction but that are not subgame perfect.

Example 164 *Interrogators look for information from Prisoner 1, then from Prisoner 2, then from Prisoner 1 again. Prisoner 1 has two chances to cover up or to talk; Player 2 has one chance. If both keep quiet then they both do well. If one of them talks then the payoffs depend on who talked and when. The game is represented below in Figure 13.2.*

In this game it is easy to see that $\sigma_1 = (C1, C3)$, $\sigma_2 = (C2)$ *is the unique subgame perfect Nash equilibrium; it also satisfies our trembles concern. However the Nash equilibrium* $(\sigma_1 = (G1, G3), \sigma_2 = (G2))$ also *satisfies our restriction on trembles. If Player 1 believes that there is high probability that Player 2 will play* $\sigma_2 = (G2)$ *then the strategy* $\sigma_1 = (G1, G3)$ *is a best response. Furthermore, if Player 2 thinks that there is a high chance that Player 1 will play* $\sigma_1 = (G1, G3)$ *(and thinks that, should, 1 play* $C1$ *by mistake that* $\sigma_1 = (C1, G3)$ *is much more likely than* $\sigma_1 = (C1, C3)$*), then* $\sigma_2 = (G2)$ *is a best response for him. The problem is that conditional upon playing C1, player 1 rules out the possibility that he will have to choose between* $C3$ *and* $G3$*. A little more modesty about his possible errors at his first choice node may lead him to make better decisions.*

We need to find a way then to allow Players to choose optimal strategies even as they take account of the possibility that they might tremble. Conceptually this is a little difficult because we want a given agent to choose decide to implement a strategy conditional upon the possibility that he will not in fact implement that strategy. The solution that has been proposed to the conceptual problem is to create an "**agent normal form**" of an extensive form game.

Definition 165 *The "**agent normal form**" of an extensive form game is a normal form game played by a set of "agents" representing the players: in particular, each player has one agent that can take actions at each of his information sets but each agent has the same preferences as the player he represents.*

The idea of the agent representation is to break down some of the internal dependencies that a single player may impose on his own strategies; in particular with each agent acting independently, but all in the interests of the player, one agent can "fix up" the errors made by another (hence, the optimality of a player's behavior strategy is equivalent to optimality for each of his agents, treating the strategies of the player's other agents as fixed).

Equipped with this definition we can now define a stronger notion of trembling hand perfection:

Definition 166 *Strategy profile σ is an "**extensive form trembling hand perfect equilibrium**" of a finite extensive form game, Γ, if it is a trembling hand perfect equilibrium of the agent normal form of Γ.*

Since the condition requires that each agent optimally chooses an action at every information set, the condition guarantees that any extensive form trembling hand perfect equilibrium is also sub game perfect.[12] We will see next that not all subgame perfect equilibria are extensive form trembling hand perfect: this then implies that trembling hand perfection is a *refinement* of subgame perfection. Furthermore it is not such a strong refinement that it prevent us from locating equilibrium, since we can show **existence**: every finite extensive form game has an extensive form trembling hand perfect equilibrium.[13]

Let's now look at an example of where this might be useful. The following example is based on work by Selten and Picasso.

Example: Selten's Horse

The president wants the leader of the opposition dead. He can choose between personally hiring an assassin to do it, or asking some loyal stooge to hire the assassin. Either way, the assassin will receive the contract anonymously and have to decide whether to carry out the requested assassination

[12] Indeed we can show a stronger result, that if σ is extensive form trembling hand perfect, then it is also part of a "sequential equilibrium". More on this below.

[13] In fact, this existence result follows directly from the existence of a trembling hand perfect equilibrium. Since for any extensive form game an agent normal form representation exists, and as all normal form games have trembling hand perfect equilibria, we have that extensive form tembling hand perfect equilibria exist.

without knowing whether the President gave the order directly or not. And once the job is done it will become clear to all who gave the orders.

The assassin, we assume, has a particular dislike of the stooge but depends on the President for survival.

Now if the assassin carries out a direct order from the President he is in a great position, having done the President's bidding (plus, if the President ever tries to turn on him, he's got some dirt on the President!). If he refuses to do it he's still got some dirt on the president, which is always useful and at least make it difficult for the president to turn on him.

If in fact the message comes from the stooge, then the situation is different: in that case, by accepting, the assassin doesn't get into the President's good books. But if he refuses, he gets some dirt on the stooge, a good thing.

The stooge and the President have the same preferences over outcomes: Both would like to see the opposition assassinated and both would rather that the order come from the stooge rather than from the president. Both will be particularly upset though if the order is given but not followed through, especially in the case where the president and his regime get compromised by having given the order

The game is illustrated in Figure 13.3.

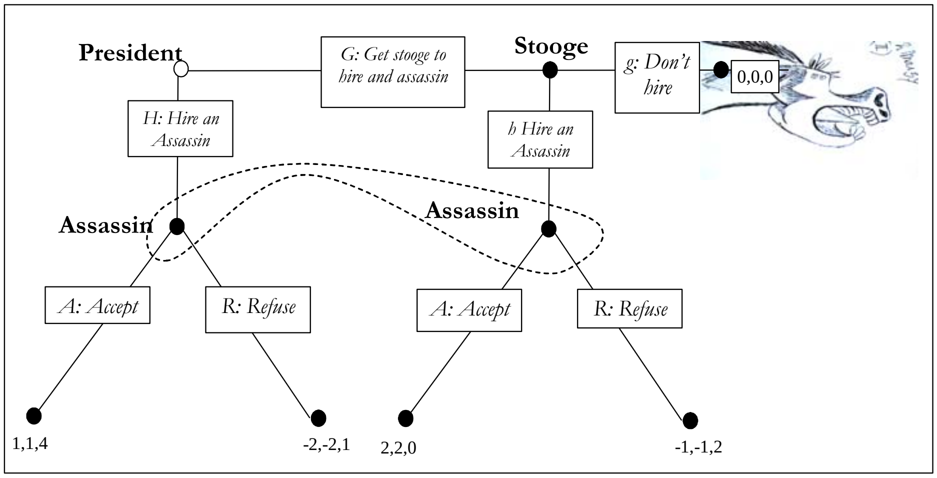


FIGURE 13.3. Selten's Horse

Now for the game in Figure 13.3, each player only has one information set and has two pure strategies: given by (H, G), (h, g), (A, R) for the President, the Stooge and the Assassin respectively. Let us write out the behavioral strategies of the players explicitly in terms of the probability distributions over these pure strategies. Hence for example a strategy of he form $\sigma_{\text{President}} = (.2)$ should be interpreted as: "The President plays H with probability .2 and plays G with probability .8." A strategy profile might then be $\sigma = (.3, .1, .2)$ interpreted as "President plays H with probability .3, the Stooge plays h with probability .1 and the assassin plays A with probability .2. "

We can start the job of identifying the extensive form trembling hand prefect equilibria by looking within the class of subgame perfect equilibria (since trembling hand is a refinement).

Observe first that there are no strict subgames to this game, hence any Nash equilibrium for the game as a whole is also a Subgame Perfect Nash Equilibrium. So really we just need the set of Nash equilibria.

Note next that there are no pure strategy Nash equilibria. This can be verified readily. So we have to consider equilibria in which players employ mixed strategies. Using the approach in which we look for conditions under which a player will be indifferent between any two of his pure strategies we find that there are two families of mixed strategy Nash equilibria:

Type A equilibria: $\sigma = (1, p, 1)$ for $p < \frac{1}{2}$

Type B equilibria: $\sigma = (0, 0, q)$ for $q < \frac{1}{3}$.

(You can work out why these are Nash equilibria, here we will try to work out which of them are trembling hand perfect)

The type B equilibria are trembling hand perfect. To see this, consider the perturbation for small $\varepsilon \in (0, q)$ given by $\sigma = (\varepsilon, \frac{3}{2}\frac{\varepsilon}{(1-\varepsilon)}, q - \varepsilon)$ for $q < \frac{1}{3}$. Under any such a perturbation the Assassin is even *less* likely to carry out the order and so the President and stooge have less incentive to deviate. We now simply need to check that the Assassin will be willing to mix. In fact, in order to show better how such sequence are determined, I will show how we construct σ in order to *ensure* that the assassin will mix.

Keeping $\sigma_{\text{president}}$ and σ_{assassin} fixes, let us define the strategy profile $\sigma = (\varepsilon, z, q-\varepsilon)$. We now determine what value z has to take in order to ensure that the Assassin is willing to mix.

We first work out that Assassin's beliefs given that he finds himself with a choice to make and all players are using $\sigma = (\varepsilon, z, q - \varepsilon)$. He needs to work out the probability that the president sent the order. He know that the probability that the President sends an order is ε, and the probability that the Stooge sends an order is $(1-\varepsilon)(z)$. Given that he has observed an order he can then use Bayes' rule to work out that the probability that it was the President who sent the order is: $\frac{\varepsilon}{\varepsilon+(1-\varepsilon)(z)}$. Under these conditions the stooge will be willing to mix if his payoff from Accepting is equal to his payoff from rejecting. That is:

$$\frac{\varepsilon}{\varepsilon+(1-\varepsilon)(z)}4 + (1 - \frac{\varepsilon}{\varepsilon+(1-\varepsilon)(z)})0 = \frac{\varepsilon}{\varepsilon+(1-\varepsilon)(z)}1 + (1 - \frac{\varepsilon}{\varepsilon+(1-\varepsilon)(z)})2$$

or

$$z = \frac{3}{2}\frac{\varepsilon}{(1-\varepsilon)}$$

Hence, we have determined z as that value for which the assassin will be willing to mix (and so we know that if $z = \frac{3}{2}\frac{\varepsilon}{(1-\varepsilon)}$, the assassin *will* be willing to mix). Hence the type B equilibria are trembling hand perfect.

What of the type A equilibria?

They are *not* trembling hand perfect because if the President accidentally asks the Stooge to send the order (no matter how unlikely this is), and if the assassin accepts with probability anywhere close to 1, then the Stooge has a unique best response to send the order. The problem is that the conditions under which the Stooge is willing to randomize are fragile: they require that his actions don't make any difference in equilibrium. If he had to make the choice in a context where the assassin is likely to do the job, then he would no longer be prepared to randomize.

In this case the solution concept helps us distinguish between two very different outcomes. In the one unreasonable outcome the order is given and carried out by the assassin because he has perfect confidence that the President will not accidentally give the stooge the order to send and that the Stooge would, given the option, not elect to send such an order. In the equilibrium that satisfies trembling hand perfection however, neither player passes on the order because they can't be sure the assassin will do the job.

13.3.2 Sequential Equilibrium

The final solution concept that we consider is the notion of "sequential equilibrium."

The formal description of sequential equilibrium is similar to that of Bayesian equilibrium but the logic of the solution is in fact closer to trembling hand perfection.

As in our discussion of Bayesian games, our solution is composed both of a set of beliefs and a set of behavioral strategies.

Let $\mu(\iota)$ denote a probability measure over the set of nodes in information set ι. This describes the beliefs of the player taking a decision in information set ι about what history has been implemented. We use μ to denote a complete set of such probability measures, one for each information set, that assigns a probability to each event in that information set. As before, we let σ denote a profile of behavioral strategies.

Armed with these concepts we definite the family from which Sequential equilibria are drawn: "

Definition 167 *An "**assessment**" in an extensive form game is a pair, (σ, μ), that contains a profile of behavioral strategies, σ, and a belief system, μ.*

Now comes the key concept to deal with the 0 probability events problem:

Definition 168 *An assessment (σ, μ) is "**consistent**" if there is a sequence of assessments $((\sigma^n, \mu^n))_{n=1}^{\infty}$that converges to (σ, μ) (in Euclidean space[14]) but in which $\sigma^n \gg 0$ and each μ^n is derived using Bayes' rule.*

The most important feature here is that for every element in the sequence, $\sigma^n \gg 0$, this means that every branch in the tree can be followed with *some* strictly positive probability. As a result, every information set can be reached with some probability and so Bayes' rule can be employed. We now have the following solution concept:

Definition 169 *An assessment (σ, μ) is a "**sequential equilibrium**" if it is consistent and sequentially rational in the sense that for each player i, moving at information set ι, player i's expected payoff from playing $\sigma(\theta_i)$, $\mathsf{E}[u_{i(\iota)}|\iota, \mu, \sigma_{i(\iota)}, \sigma_{-i(\iota)}]$ is at least as good for type θ_i as the expected payoff from playing any rival strategy σ' available to player i, $\mathsf{E}[u_{i(\iota)}|\iota, \mu, \sigma'_{i(\iota)}, \sigma_{-i(\iota)}]$.*

The bite of the requirement then is this: if (σ, μ) is a sequential equilibrium then there must be beliefs that are *really really* close to μ that are derived using Bayes' rule strategies, σ^n, that are completely mixed and that are *really really* close to σ. (Note that we don't require that these σ^n strategies themselves be optimal in any sense).

Problem 170 *What are the sequential equilibria for the Selten's horse example discussed above?*

The solution concept however not perfect insofar as it still allows for a very wide class of beliefs that sometimes may lead to the failure to select reasonable equilibria. In response a very large number of "refinements" have been developed. Nonetheless, it has many nice properties and along with trembling hand perfection is a good check on any equilibria that you identify (indeed the last point indicates that it may be sufficient to check *either* sequential equilibrium *or* trembling hand perfection)

The solution concept has nice properties:

- it exists (See Osborne and Rubinstein Proposition 249.1)
- it is a subset of Nash equilibrium
- in an extensive form game of perfect information (σ, μ) is a sequential equilibrium if and only if σ is a Nash equilibrium

[14] Recall that each element in σ and μ is a vector of probability distributions for each information set and hence just contains numbers between 0 and 1.

- there is a one stage deviation principle also for sequential equilibria (for more, see Osborne and Rubinstein Proposition 227.1)

- if there is no sequential equilibrium involving behavioral strategy σ, then σ is not trembling hand perfect. (The converse is *almost* true: if σ is not trembling hand perfect, then there exists no sequential equilibrium involving behavioral strategy σ (see Kreps and Wilson (1982b))).

14
Solution Concepts for Cooperative Games

14.1 Randomization and the Negotiation Set

Consider again the mixed strategy equilibrium identified for the battle of the sexes game discussed in Section 9.5. The expected payoff at the mixed strategy equilibrium is given by $u_i = \frac{2}{3}\frac{1}{3}2 + \frac{2}{3}\frac{1}{3}1 + \frac{2}{3}\frac{2}{3}0 + \frac{1}{3}\frac{1}{3}0 = \frac{2}{3}$. In this instance, the payoff is lower than what could be achieved for either player at either of the pure strategy outcomes. In other words, it is Pareto inefficient.

Exercise 171 *Is this always the case for the class of 2-player games where each player has 2 pure strategies and there are 3 equilibria? Prove it either way.*

Assuming that players play (independently) mixed strategies (p_I, p_{II}), the set of possible payoffs from this game is given by the set:

$$\{2p_Ip_{II} + 1.(1 - p_I)(1 - p_{II}), p_Ip_{II} + 2(1 - p_I)(1 - p_{II})|p_I, p_{II} \in [0, 1]\}$$

If we draw this set it looks like this:

Eyeballing this set you should be able to identify the payoffs associated with each of the Nash equilibria. What's striking is that you can also see that the *only* Pareto efficient points in this graph are given by the pure strategy Nash equilibria: the independent randomization introduces a *lot*

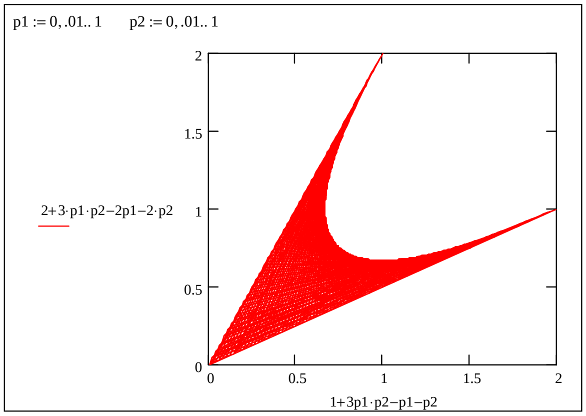


FIGURE 14.1. Expected utilities from independent randomizations.

of inefficiency because players select outcomes that yield (0,0) with positive probability. In contrast, (and assuming von Neumann Morgenstern utility functions) randomization over the Nash equilibria allows for payoffs along the line joining points (1,2) and (2,1) that Pareto dominate all points that are achievable through independent randomization. But to achieve these points (in this example) requires that players agree on a *joint randomized strategy*.

In other words, they have to cooperate somehow or other. In our discussion of correlated equilibria we saw cases of implicit cooperation, executed in a non-cooperative environment. But even there we saw that fuly efficient outcomes were missed. Full cooperation would require mechanisms that incite people to play in a particular way even when this can not be achieved through joint randomization. Many believe that humans are smart enough to get over problems like this, even when we don't know exactly how they do it.

This concern motivated von Neumann and Morgenstern to suggest that a plausible solution concept is quite simply the set of Pareto optimal points that could be achieved from joint randomization. They called this set the "**negotiation set**," although now it is more commonly called the "**Pareto set**."

Besides the fact that it is unclear how players succeed in reaching the Pareto set, the most evident problem with the Pareto set as a solution concept is that it is imprecise: it is typically "large." Von Neumann and Morgenstern argue that we are not in a good position to pick and choose from within this set, rather, which element of a Pareto set is selected depends on psychological features of the players that lie outside of the scope of formal models.

John Nash disagreed and offered a bargaining solution as a rival cooperative solution that is a refinement of the Pareto set. As such the Nash bargaining solution can be thought of as a general solution concept to games in which players can make binding commitments to each other.

14.2 The Nash Bargaining Solution

Nash's study of "the bargaining problem" is a response to a concern similar to that motivating Arrow.[1] Nash solves the problem directly in a "utility space" (rather than solving it in "outcome space" and then remapping back to utility space) which he "anchors" in a status quo utility; in effect he is looking to select some optimal improvement over the status quo (Arrow assumes no such anchoring). Nash's justification for using utility space is that if any two outcomes deliver the same utility to all players, then for the purposes of studying social choices, they should be thought of as equivalent outcomes. In order to work in utility space, Nash needs to use richer utility information than Arrow; he does so by assuming that players have preferences that can be represented by **von Neumann-Morgenstern** (**Bernoulli)** utility functions (henceforth **VNM**). The key things to note about this assumption for the bargaining problem are the following:

1. Since the representation of an individual's preferences by a VNM utility function is "unique up to a monotonically increasing affine transformation" (that is, if the utility function u represents an individual's preferences, so too does the utility function $\alpha + \beta \times u$ for $\beta > 0$) we are free to change the *origin* (α) and the *units* (β) of an individual's utility. Given that we have this freedom, the least that we can expect of any theory is that its predictions do not depend on the particular choice of utility origins or units.

2. This assumption also allows Nash to assume that the set X (whose typical element is an $n-$tuple of utility values $(u_1, u_2 ... u_n)$) is *convex*.[2]

[1] This reading was not on the syllabus, if you can though do try and read Nash's short 1950 article on bargaining: "The Bargaining Problem" *Econometrica*, 18:2, 155-162, on-line at: http://links.jstor.org/sici?sici=0012-9682%28195004%2918%3A2%3C155%3ATBP%3E2.0.CO%3B2-H

[2] A convention that I ignore here is to use U to denote utility space to distiguish it from X, the outcome space. The argument for assuming that X is convex is this: even if the utilites that result from the set of "primitive" outcomes do not produce a convex set in utility space, players could consider randomizing across these outcomes and evaluate all resulting lotteries; from the assumption of VNM utilities we would then have that the utility set over these lotteries would be convex.

3. Finally, this assumption allows Nash to normalize the units of utility space such that the *status quo* utility of each person is 0 (*if you ever want to use the Nash bargaining solution you will also have to make this normalization!*).

4. **What the assumption does not allow**: while cardinality is introduced, the assumption of VNM utility functions does not allow us to make inter-personal comparability of utility

Like Arrow, Nash selects reasonable sounding conditions that he wants to impose on the choice rule. Two conditions that we have not yet encountered are:

- **IIA** (Note that this IIA condition is somewhat different to the IIA condition we saw last week.) The choice function f satisfies IIA if $X' \subset X$ and $f(X) \in X'$ implies $f(X) = f(X')$. This means that if you chose chicken when beef was available, you're not going to change your choice when you find out that they're out of beef.

- **Symmetry** The choice function f satisfies symmetry if whenever X is a symmetric set then f chooses the egalitarian solution.[3] That is, if the choice set does not distinguish between individuals then neither does the choice rule.

Nash's question then is: Can we find a social choice function that can pick an element of X in such a way that the choice:

1. Is independent of origins and units
2. Is Paretian
3. Satisfies the symmetry condition and
4. Satisfies IIA ?

His answer is "Yes! We can find one *and only one!*"

That's a nice answer as it means that the theory gives determinate predictions. Nash's result is even more useful however because his existence proof is constructive: he tells us not just that such a choice function exists, he also tells us what it is, namely: $f(X) = \arg \max\limits_{x \in X_+} (\prod_{i=1}^{n} x_i)$. That is,

[3] To be a symmetric set we require that if x is in X so too is x' where x' differs from x only by a permutation of its entries; for example, if $(1, 2, 0)$ is in X so are $(1, 0, 2)$, $(0, 1, 2)$, $(0, 2, 1)$, $(2, 1, 0)$ and $(2, 0, 1)$.

select the outcome that maximizes the product of the utilities of all the individuals. Equivalently we may write: $f(X) = \arg\max_{x \in X_+} (\sum_{i=1}^{n} \ln(x_i))$. Hence we have:

Proposition 172 *A choice function f is independent of utility origins and units, Paretian, symmetric and independent of irrelevant alternatives iff it maximizes $n(x) := \sum_{i=1}^{n} \ln(x_i)$.*

Proof. I will only prove the "only if" part; try do the "if" part yourself. To prove the "only if" part we need to show that no solution other than Nash's satisfies the four conditions; or equivalently, if any solution satisfies the four conditions, that solution is equivalent to Nash's solution. Define a rival function g that satisfies the four conditions. Consider the two possible choice sets X' and X'' defined by:

$X' = \{x \in \mathbb{R}^N : \sum\limits_{i \in N} \frac{x_i}{(f(X))_i} \leq N\}$

$X'' = \{x \in \mathbb{R}^N : \sum\limits_{i \in N} x_i \leq N\}$

The set X'' is symmetric and so from Pareto optimality and symmetry we have $g(X'') = (1, 1, ..., 1)$. From the independence of utility units we may transform the set X'' into the set X' by multiplying each person's utility by $(f(X))_i$, without this changing the (appropriately scaled) outcome. Hence we have $g(X') = (1 \times (f(X))_1, 1 \times (f(X))_2, ..., 1 \times (f(X))_n) = f(X)$. Note next that $X \subset X'$. To see this note that for some $x^* \in X$ if $x^* \notin X'$ then $\sum\limits_{i \in N} \frac{x_i^*}{(f(X))_i} > N$ and hence, since $\sum\limits_{i \in N} \frac{x_i^*}{(f(X))_i} = x^*.\nabla n(f(X))$ and since $N = \sum\limits_{i \in N} (f(X))_i \frac{1}{(f(X))_i} = f(X).\nabla n(f(X))$, we have that $x^*.\nabla n(f(X)) > f(X).\nabla n(f(X))$. But so $(x^* - f(X)).\nabla n(f(X)) > 0]$ Finally, since $X \subset X'$ and $g(X') \in X$, we have from IIA that $g(X) = g(X') = f(X)$ and so we have what we were looking for: $g(X) = f(X)$. ■

Math Interlude: What is going on with the triangles?

This last part of the proof is worth unpacking. The mathematics used here will be useful later when we look at more bargaining problems and when we turn to voting equilibria. Geometrically the idea is as follows: consider what we know about the point $f(X)$: First, it lies on the boundary of a convex set X; Second, it lies on the boundary of a second convex set of points, namely, the set $\{x : n(x) \geq n(f(X))\}$; this set is termed the upper contour set of x, let us call it Y.[4] Third, from maximization we can show that $f(x)$ lies on a hyperplane H that separates X and Y; that is: all the points in X are on one side of H, all the points in Y are on the other

[4] **Problem**: How do we know that Y is convex?

and X and Y only intersect at $f(X)$.[5] The Hyperplane H can be written $H := \{x : x.v = b\}$, where v is a vector, orthogonal to the hyperplane, called the "normal" vector of the hyperplane and b is a scalar and $x.v$ is the dot product of x and v. Any point y, lying "above" this hyperplane has the property that $y.v > b$ while any point z lying below has $z.v < b$. Now, from maximization we can show that the gradient of $n(.)$, written $\nabla n(x)$, is orthogonal to the hyperplane and hence can serve as the normal vector for H. Also, since we know that $f(X)$ is *on* the hyperplane we know that $b = \nabla n(x).f(X)$. Hence any point y, above H, including all the points in Y (other than $f(X)$) has $y.\nabla n(x) > \nabla n(x).f(X)$; while any point z, below H, including all points in X, have $z.\nabla n(x) < \nabla n(x).f(X)$

Problem 173 *Draw two convex sets in any position that touch at a point and draw the line between them; mark the normal to the hyperplane as a vector pointing in some direction away from the point of intersection of the sets and orthogonal to the line. Using the definitions of "above" and "below"given above, work out which area is "above" the line and which is "below" it.*

Reality Check

But will people apply the Nash bargaining solution? The Nash bargaining solution, like Arrow's result, assumes that we have information about the preferences of individuals. What would we expect to see if we relied upon individuals to volunteer this information themselves? To consider this, imagine the situation where two players are bargaining over the division of a pie of "size 1,"[6] that is, the utility possibility set is given by $\{u_1, u_2 : u_1 + u_2 \leq 1\}$. Now, consider the case where although each player has status quo utility of $\underline{u}_1 = \underline{u}_2 = 0$, in practice $u_1 = 0$ is not public knowledge.

In this situation the Nash solution is found by maximizing: $(1-u_1)(u_1 - \underline{u}_1)$. First order conditions then yield: $u_1 = \frac{1+\underline{u}_1}{2}$, and $u_2 = \frac{1-\underline{u}_1}{2}$.

With $\underline{u}_1 = 0$ reported truthfully, this gives the symmetric egalitarian division that you would expect. But consider a mechanism whereby Player 1 must declare her valuation of the status quo and then the Nash bargaining solution is implemented (whenever implementing it produces an improvement for both players). What should Player 1 declare? Assuming that she is a utility maximizer with no scruples about lying, she should try to maximize $u_1 = \frac{1+\underline{u}_1}{2}$ conditional upon $u_2 = \frac{1-\underline{u}_1}{2}$ being positive. The solution

[5] **Problem**: Show that if the intersection of X and Y contained any other points then $f(X)$ does not maximize $n(x)$ on X, contrary to our assumption.

[6] Why is "size 1" in quotes?

is to declare $\underline{u}_1 = 1 - \varepsilon$ (for some tiny number ε) and gain the full share of the pie. That is, that she should declare herself quite happy with the status quo and very unsatisfied with almost everything but the status quo; in doing so she can argue that the only possible gains in trade occur when she receives the full pie. The implication of the argument is that if we are relying on players to report information about their preferences, then we can not expect them to do it truthfully.

This observation introduces a large field of game theoretic work, that of mechanism design or implementation theory; we discuss this next week.

14.2.1 The Nash Program

There has been a fairly active research program aiming to link cooperative and non-cooperative game theory. Much of this work is aimed at deriving the class of non-cooperative games that support outcomes from cooperative games. The study of repeated games can be seen as one part of this program. One of the interesting results of this program is the demonstration that there is an equivalence between the Nash bargaining solution and the outcome of the Rubenstein model, as discount factors tend to unity.

14.3 The Core

The Nash bargaining solution uses very little information about the structure of the underlying game being played, making use simply of the utility possibility set that arises if all players strike deals collectively. We now consider other solution concepts that share the assumption that players can strike agreements with each other but that use richer information about the ability of different *sub*groups of players to achieve beneficial outcomes on their own.

Perhaps the most prominent solution concept for cooperative games is "**the Core**."

Definition 174 *"**The Core**" is the set of outcomes $Y \subset X$ with the property that no coalition, C, can, by acting alone, achieve an outcome y that every member of C prefers to some member of Y.*

The Core concept is similar in spirit to the idea of Nash equilibrium except that to be in the Core we require not just that no individual deviations are beneficial to any individual but that no group deviations are beneficial for any group. In this regard we may expect it to be more difficult to find points in the Core than it is to find Nash equilibria. A second difference is that the core is defined over outcomes (or utilities) rather than over strategies.

Problem 175 *Consider a game in which some set of players $i \in N = \{1, 2, ..., n\}$, (with n an odd number) have single peaked preferences over outcomes in $X = [0, 1]$ with ideal points given by $(p^i)_{i \in N}$. Assume that any coalition of $\frac{n+1}{2}$ players can implement any policy in X but that no transfers are possible. Which outcomes in X are in the Core?*

Exercise 176 *Consider a game in which some set of players $i \in N = \{1, 2, 3, 4\}$ have Euclidean preferences over outcomes in $X = [0, 1] \times [0, 1]$ with ideal points given by $(p^i)_{i \in N}$ where each p^i is in X and no three ideal points are collinear. Assume that any coalition of* 3 *or more players can implement any policy in X but that no transfers are possible. Which outcomes in X are in the Core?*

14.4 The Core in Voting Games: Plott's Theorem

The following is a discussion of Theorems 1 and 2 from Plott (1967). One aspect that makes this model both quite different from most simpler models, but also lends the model its generality, is that the key information that Plott uses about voter utilities is the "gradient vectors of their utility functions" as defined at some status quo point, rather than their utility functions directly or their "ideal points." So, the first thing is to be *really* clear about what those are.

The gradient vectors of utility function $u : \mathbb{R}^n \to \mathbb{R}^1$, are the vectors that are found by taking the derivative of $u(x)$ (where $x \in \mathbb{R}^n$) with respect to each component of x, and evaluating that derivative at x. The gradient vector is written alternatively as $\nabla u(x)$ or Du_x or, less compactly, as $(\frac{\partial u(x)}{\partial x_1}, \frac{\partial u(x)}{\partial x_2}, ..., \frac{\partial u(x)}{\partial x_n})$. It is alternatively referred to as the "derivative of u at x" or as the "Jacobian derivative of u at x." Each element of the vector gives the sign and magnitude of the increase in utility from a change in that element of x. Note that that gradient vectors can themselves be represented as points in $\mathbb{R}^n$.

Example 177 *So let's take an example: let $u(x) = x_1 x_2 + x_1$. Then $\frac{\partial u(x)}{\partial x_1} = x_2 + 1$ and $\frac{\partial u(x)}{\partial x_2} = x_1$. We then have $\nabla u(x) = (x_2 + 1, x_1)$.*

Example 178 *Here's another example: let $u(x) = -\frac{1}{2}(p_1 - x_1)^2 - \frac{1}{2}(p_2 - x_2)^2$. This is a quadratic utility function for someone with a ideal point at (p_1, p_2). Then $\frac{\partial u(x)}{\partial x_1} = (p_1 - x_1)$ and $\frac{\partial u(x)}{\partial x_2} = (p_2 - x_2)$. Evaluated at $x = 0$, we then have $\nabla u(0) = (p_1, p_2)$. This means that with quadratic utilities, the gradient vector evaluated at 0* is *the player's ideal point.*

Gradient vectors are then used to evaluate attitudes to small changes to x. In particular if we think of a tiny change to x in the direction b, the change in player's i's utility is given by $\nabla u(0).b$ (where the "." is the dot product, although in Plott's notation this is typically left out). This has a simple interpretation: write out $\nabla u(0).b$ fully as $\frac{\partial u(x)}{\partial x_1}b_1 + \frac{\partial u(x)}{\partial x_2}b_2 + ... + \frac{\partial u(x)}{\partial x_n}b_n$ and you can interpret this as the sum of the per unit gains in each direction multiplied by the amount of change in that direction: in some directions there may be a gain, in some a loss, but $\nabla u(0).b$ adds them all up together and tells you whether there is a net gain or a net loss. We can say that a player is "**satiated**" at x if $\nabla u(x) = 0$ (note that this just means that the first order conditions for x to maximize u are satisfied). Plott assumes that an individual supports a "motion" b away from x if $\nabla u(x).b > 0$, that is, if it increases the individual's utility.

So now, getting back to Plott's notation: the status quo is taken as fixed, let's just assume that it is 0. The gradient vector for a given individual i, evaluated at 0, is written simply as a_i instead of $\nabla u(0)$. This vector is just a point in $\mathbb{R}^n$. in fact, from the last example, with quadratic utility, this point *is* the player's ideal point.

We then have that the condition for one person to want a change in direction b from 0 is that $a_i.b > 0$. For a person to be indifferent we have $a_i.b = 0$, we need and for a person to be against $a_i.b < 0$. Indifferent people vote against. Geometrically this means that people on one side of a hyperplane with directional vector b and intercept a support motion b, people on the other side oppose, and people on the hyperplane are indifferent. See Figure 14.2 for examples.

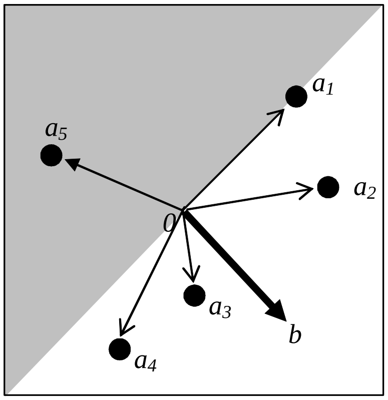


FIGURE 14.2. In this example for $a_i.b > 0$ for each of players 2, 3 and 4. But $a_i.b = 0$ for player 1 and $a_i.b < 0$ for player 5.

So now, how about the conditions for more than one person to support a motion...

Plott and Decision-making Under Unanimity

The condition for *all* m individuals to support such a change is that:

$$\begin{array}{c} a_1.b > 0 \\ a_2.b > 0 \\ : \\ a_m.b > 0 \end{array} \tag{14.1}$$

In fact though if you define the matrix A by:

$$A = \begin{bmatrix} a_1 \\ a_2 \\ : \\ a_m \end{bmatrix}$$

You can write the m conditions in line (14.1) more simply as:

$$Ab > 0 \tag{14.2}$$

The first result then is that a point (in this case 0) will not be changed by a group operating under unanimity unless there is a $b \in \mathbb{R}^n$ such that $Ab > 0$. Hence a point is an equilibrium point if there is no b such that $Ab > 0$.

To characterize the set of points for which this is true, Plott uses the following result from linear programming: there is be *no* b such that $Ab > 0$ if and only if there *is* a y (with $y \in \mathbb{R}_+$ and $y \neq 0$) such that $y.A = 0$.[7] The set of equilibrium points under unanimity then is the set of points for which there is a (semipositive) solution to $y.A = 0$.

Example 179 *To be sure we got this, let's take an example: What's the set of equilibrium points under unanimity to a game with three players,* 1, 2 *and* 3 *with quadratic utility and ideal points given by* $(0,0)$, $(0,1)$, $(1,0)$ *respectively?*

Well, for any status quo, q, the matrix A is given by:

$$A = \begin{bmatrix} -x_1 & -x_2 \\ -x_1 & 1-x_2 \\ 1-x_1 & -x_2 \end{bmatrix}$$

[7] If you want more on this last part, look up linear programming and duality theory. Or write out some examples to convince yourself that it's true. To think about this problem geometrically its useful to note that $yA = 0 \Leftrightarrow A^T y^T = 0^T$ and hence the condtion is that there is no hyperplane through the origin with directional vector $y \in \mathbb{R}_+$ (and $y \neq 0$) such that all the vectors formed by the rows of A^T (which list the gradient of each individual along a particular dimension) lie on the hyperplane.

If a semipositive y satisfies $yA = 0$ this means there exist $y_1, y_2, y_3 \geq 0$ with at least one of them strictly positive for which:

$$\begin{aligned} -x_1y_1 - x_1y_2 + (1-x_1)y_3 &= 0 \\ -x_2y_1 + (1-x_2)y_2 - x_2y_3 &= 0 \end{aligned}$$

A little rearranging gives:

$$x_1 = \frac{y_3}{y_1+y_2+y_3} \qquad x_2 = \frac{y_2}{y_1+y_2+y_3}.$$

Note that with y semipositive the right hand side of these equations are non-negative numbers lying between 0 and 1. So we know for sure that we must have $x_1, x_2 \in [0,1]$. Also we must have $x_1 + x_2 = \frac{y_2+y_3}{y_1+y_2+y_3} \leq 1$. Necessary conditions for $yA = 0$ to have a semipositive solution then are: $x_1, x_2 \in [0,1]$, $x_1 + x_2 \leq 1$. This, you can see describes exactly the convex hull of the player's ideal points! furthermore, these necessary conditions are also sufficient (prove it!). This is a proof of the claim that the Pareto set of these three players is simply the convex hull of their ideal points.

Problem 180 *Use this method to locate the set of equilibria under unanimity rule for 3 players who do* not *have Euclidean preferences.*

Plott and Decision-making Under Majority Rule

The logic for the majority rule case is similar. Now we replace the condition that there be no b such that $Ab > 0$ with the condition that there be no b such that there exists a submatrix M of A, where M has $\frac{m+1}{2}$ rows (and the same number of columns) such that $Mb > 0$. All this is saying is that we should not be able to find any direction for which we can find some majority that supports a change in that direction. The remainder of Plott's proof is about locating the conditions that must hold for A for there not to be any submatrix M such that there exists no b with $Mb > 0$.

His strategy of proof is the following:

(Lemma 1) Just confirms that an equilibrium requires that no majority will support a change in some direction. This is almost definitional.

(Lemma 2) This confirms that at an equilibrium point, no matter what direction you choose, one person will always be "indifferent" to a change in that direction. Why? Because if some majority all *strictly opposed* a change in direction b, then that same majority would all strictly *support* a change in direction $-b$. And that can't be (at an equilibrium point).

(Theorem 1) An equilibrium point has to be at the ideal point of one player. The proof for this is a little awkward but the idea is quite easy. First of all note that if an equilibrium is at the ideal point of a player, then that person has a gradient vector of $a_i = (0, 0, ..., 0)$ and hence is indifferent to *any* small change away from his ideal (since $a_ib = 0$ for all

b). Now that sounds odd, but the idea is just that people's utility functions are flat at the top, so a tiny move doesn't make any difference at all. Now Lemma 2 tells us that for *every* direction *someone* has to be indifferent. Clearly this is true if one person has a zero gradient vector. We want to check that there is no other way that it could be true. Well, there is one "rival explanation": perhaps for each direction, b, there is someone with a non-zero gradient vector a_j that is orthogonal to b (that is for all b, there exists an $a_j \neq 0$ such that $a_j b = 0$) (see Figure 14.3). Now for each person, j, with non zero a_j, there is only a finite number of directions for which $a_j b = 0$. So, with a finite number of players, this orthogonality argument can only give the right answer for a finite number of directions. We are free however to consider a change in *any* direction (see Figure 14.4). If we just choose a change that is not orthogonal to the gradient vectors of any of the players we can rule out this rival explanation. (Note that when reading Plott's proof, there is at this point an error on the third last line of p. 797: $a_{in} b_n \rightarrow -\alpha_i$ should read: $a_{in} b_n \neq -\alpha_i$).

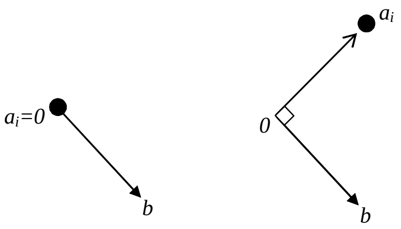


FIGURE 14.3. Two ways in which we can have $a_i b = 0$. In the left figure $a_i b = 0$ because $a_i = 0$; in the right $a_i \neq 0$ but a_i is orthogonal to b and so $a_i b = 0$ again.

(Lemma 3 and its corollary) These simply provide conditions for a system of equations to have a solution.

(Lemma 4) States that at an equilibrium point, for any person i that is *not* satiated, if that person is indifferent to a move, b, from the *status quo*, then so is someone else (and someone other than the one person that we already know to be satiated). The idea is that if this is not true, then you can find some other movement that will be acceptable to some majority that includes i plus some group of people that supported b. See Figure 14.5.

(Lemma 5) Lemma 5 pushes the logic of Lemma 4 one step further to claim that at an equilibrium if you choose any person i (other than the one person who we know to be satiated) then there is at least one other person whose gradient vector lies on a *line* through a_i and 0 (although not necessarily on different sides of 0).

For the two dimensional case we already have this from Lemma 4. And in any number of dimensions it will be true that if two non zero vectors lie on

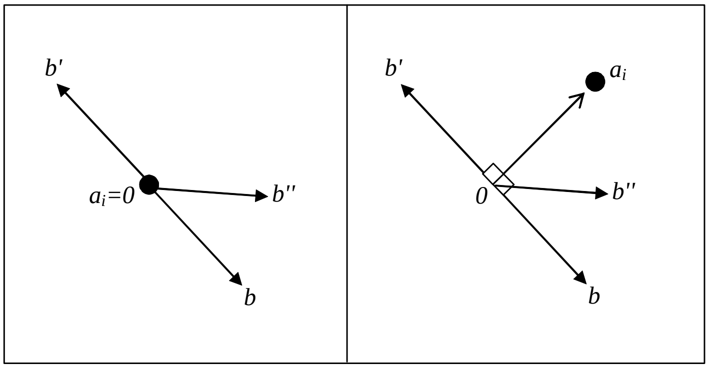


FIGURE 14.4. In the left panel where $a_i = 0$ we see that $a_i b = 0$ for all of b, b' and b''. In the right panel we have for $a_i \neq 0$ that $a_i b = 0$ only for the two vectors, b and b' that are orthogonal to a_i but not for b'' which is not. There are clearly lots of vctors like b'' that we can choose.

the same line then if one is orthogonal to b, so will be the other (formally, if $a_i = \lambda a_j$ and if $a_j b = 0$, then, clearly, $a_i b = \lambda a_j b = \lambda 0 = 0$). In more than two dimensions it is possible that two non zero vectors are both orthogonal to b without that implying that they lie on the same line. This then is a rival hypotheses much like the one we saw in Theorem 1. However, as in Theorem 1, since we are free to choose any b from a whole infinite set of them, the conditions for this rival hypothesis are just too hard to satisfy.

(Theorem 2) Theorem 2 just pushes the logic in Lemma 5 a final step further to say that pairs are not just sitting on the same line, but that they are on opposite sides of the same line: if you take the whole collection of people (except for the one person who we know to be satiated) and if none of these are satiated themselves, then you can divide them all up into pairs for whom each person in each pair lies on the opposite side of a line through 0 as the other person in his pair. The logic of the proof is illustrated in Figure 14.6.[8]

The result of Theorem 2 then is that we require exceptionally strong conditions upon the distribution of these gradient vectors in order to have an equilibrium point.

Finally, its pretty easy to see that if these (exceptionally tough) conditions are satisfied, that this sufficient for the *status quo* to be an equilibrium. This follows because for every person in favor of any movement you have another person against. That guarantees $\frac{m}{2}$ people against each movement.

[8] Note that in these figures, Plott's set (q) is given by a_1 and a_3 in the left panel and by a_1 in the right panel. Plott's set (l) is empty in the left panel and contains a_3 in the right panel. Note that there is the same number of elements in q and l in the right panel but not in the left panel.

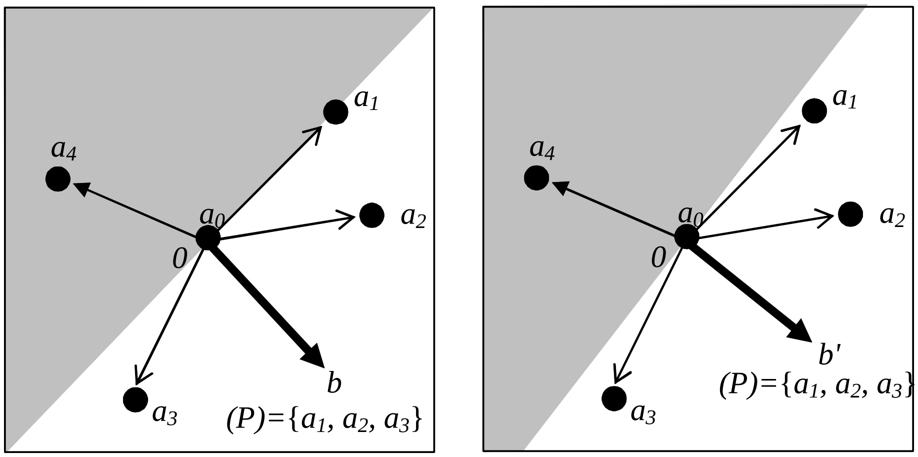


FIGURE 14.5. In the left panel $a_1b = 0$ but there is no one else (except 0) for whom $a_ib = 0$. Hence we can find a movement that everyone in (P) supports. Such a direction is illustrated in the right panel.

And of course the guy in the middle is against everything. so that gives a majority against every movement.

14.5 The Shapley Value

The final solution concept that we will consider for cooperative games, albeit briefly, tries to relate the size of a share of the pie that accrues to an individual in a cooperative setting to that individual's contribution to the size of the pie. This sounds reasonable. The tricky part though is that the contribution of an individual to the size of a coalition's pie will likely depend on what other people are already in the coalition. Perhaps a player's contribution is large if some other player is also contributing but small otherwise, if so then perhaps the measurement of that other player's contribution should take account of his impact on the first player's contribution. And so on. The idea of the Shapley value (for cooperative games with transferable utility) is that a value can be associated with each individual if we average over the whole set of marginal contributions that she could make to every coalition that does not contain her already. Letting $\mathcal{R}$ denote the set of all possible ordered sets of the players in N, with typical element R, and letting $S_i(R)$ denote the set of players that precede player i in the ordered set R, we can write the Shapley value, ψ, formally as:

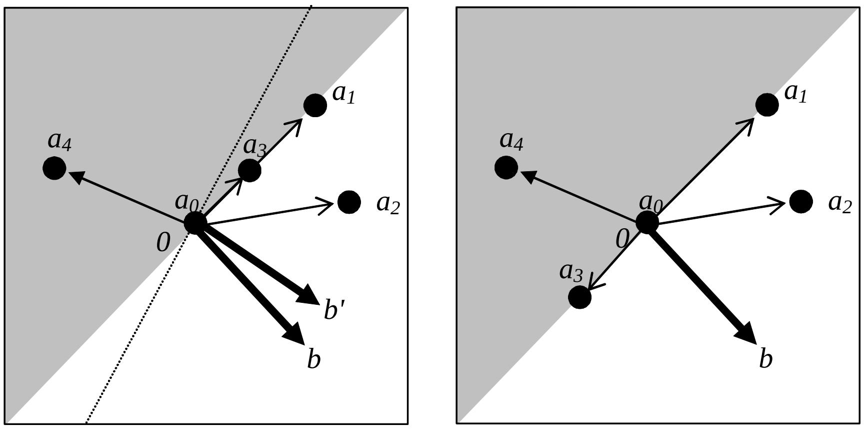


FIGURE 14.6. In the left panel a_1 and a_3 are on the same side of 0. This lets us find a b' that 1, 2 and 3 all support. Not so in the right panel: here, for any b', if 1 supports it then 2 is against it. (We do not claim that a_0 is therefore an equilibrium; why?)

$$\psi_i(N,v) = \frac{\sum\limits_{R\in\mathcal{R}} [v(S_i(R)\cup\{i\}) - v(S_i(R))]}{|N|!}$$

Remark 181 *Note that*

$$\begin{aligned}\sum_{i\in N}\psi_i(N,v) &= \frac{\sum\limits_{R\in\mathcal{R}}\left[\sum\limits_{i\in N} v(S_i(R)\cup\{i\}) - v\sum\limits_{i\in N}(S_i(R))\right]}{|N|!} \\ &= \frac{\sum\limits_{R\in\mathcal{R}}[v(N)]}{|N|!} \\ &= v(N)\end{aligned}$$

Assuming that player's do not make negative *contributions to groups, this observation allows us to define a "power index" in which we accord to each individual i a power score, α_i, where $\alpha_i \in [0,1]$ and $\sum\alpha_i = 1$:*

$$\alpha_i = \frac{\psi_i(N,v)}{v(N)}$$

This index (just one of many power indices) is called the Shapley-Shubik power index.

Exercise 182 *Consider a three person voting game in which players have transferable utility and a majority of votes is sufficient to determine any allocation of a pie of size 1. Write down the Shapley values for the games in which the voting weights of players 1, 2 and 3 are given respectively by* $\{51, 48, 1\}$*,* $\{50, 49, 1\}$*, and* $\{49, 49, 2\}$*. (note that the Shapley value is a vector in each case!*

Remark 183 *The Shapley value can be uniquely derived axiomatically although the axioms are arguably less compelling than Nash's. For details see Osborne and Rubinstein 14.4.2.*

There are a number of known relations between the Shapely value and the core. On of these is the following:

Definition 184 *A game* $\langle N, v\rangle$ *is "convex" if for every* $i \in N$*, for all* $B \subset A$*, with* $i \notin B$*:*
$v(B \cup \{i\}) - v(B) \leq v(A \cup \{i\}) - v(A)$

Note that the term convex here is used somewhat in analogy to the notion of a convex function: The marginal affect of adding any player to a coalition is larger when the coalition is already large.[9] As you may expect from convexity, the coalition in which each individual's marginal effect is largest is the grand coalition. This feature leads to the following result:

Proposition 185 (MWG 18.AA.1) *If a game* $\langle N, v\rangle$ *is convex, then the core of* $\langle N, v\rangle$ *is non-empty, and in particular,* $\psi(N, v)$ *lies in the core.*

Proof. It is enough to show that there is no strict subgroup, M of N, such that $v(M) > \sum_{i \in M} \psi_i(N, v)$, that is that the total resources obtainable by M are higher when they receive Shapley allocations from the full game than they can achieve on their own. In fact, relying on the fact that $v(M) = \sum_{i \in M} \psi_i(M, v)$ we are done if we can show that for each i, $\psi_i(M, v) \leq \psi_i(N, v)$.

Consider the case where M differs from N only by the exclusion of player j: $N = M \cup \{j\}$. Let $\mathcal{R}$ denote the set of all possible orderings of the players in N and let $\mathcal{R}'$ denote the set of all possible orderings of the players in M. Let $\mathcal{R}' \subset \mathcal{R}$ denote the set of $N!/2$ orderings in which player j appears after player i. Note that the set , and let there correspond N orderings in $\mathcal{R}$

Note that from convexity we have that for any $R \in \mathcal{R} \backslash \mathcal{R}'$:

[9] To see the intuition for the analogy, note that were there a continuum of homogenous individuals, and a twice differentiable characteristic function that depends only on the number of individuals in a coalition, n, then we would have $\partial^2 v(n)/\partial n^2 \geq 0$.

$$[v(S_i(R) \cup \{i\}) - v(S_i(R))] \geq [v(S_i(R \backslash \{j\}) \cup \{i\}) - v(S_i(R \backslash \{j\}))] \tag{14.3}$$

Note also that for each element R' in $\mathcal{R}$ there exist N elements in $\mathcal{R}$ (including R) that differ from R' only in the position of player j. [For example, for three players, only (1,2,3), (1,3,2) and (2,1,3) differ from (1,2,3) only in the position of player 1]. Call the set of such elements $C(R')$. Using Equation 14.3 N times, we then have that for any, R':

$$\frac{\sum_{R \in C(R')} \left[v(S_i(R') \cup \{i\}) - v(S_i(R'))\right]}{|N|} \geq [v(S_i(R \backslash \{j\}) \cup \{i\}) - v(S_i(R \backslash \{j\}))]$$

Summing up over $R' \in \mathcal{R}$ and dividing by $|M|!$ gives:

$$\frac{\sum_{R' \in \mathcal{R}} \sum_{R \in C(R')} \left[v(S_i(R') \cup \{i\}) - v(S_i(R'))\right]}{|N|!} \geq \frac{\sum_{R' \in \mathcal{R}} \left[v(S_i(R' \backslash \{j\}) \cup \{i\}) - v(S_i(R' \backslash \{j\}))\right]}{|M|!}$$

Now, letting $\mathcal{R}'$ denote the set of all permutations of $\{N \backslash \{j\}\}$, we can remove double counting on both sides, and divide across by N to observe:

$$\frac{\sum_{R' \in \mathcal{R}} [v(S_i(R') \cup \{i\}) - v(S_i(R'))]}{|N|!} \geq \frac{\sum_{R' \in \mathcal{R}'} [v(S_i(R') \cup \{i\}) - v(S_i(R'))]}{|M|!}$$

and hence

$$\psi_i(N, v) \geq \psi_i(M, v)$$

These steps can be repeated for any subset M' that differs from N by more than one individual. ■

15
Turning Game Theory on Its Head: Mechanism Design

So far we have concentrated on finding solutions to games. A lot of politics however is about finding games that produce particular solutions. For example, in designing constitutions, law makers attempt to devise a set of rules that produce outcomes that they deem good in some way (where the procedures themselves may formally be considered part of the outcome). In organizing campaigns, contracts or auctions, designers choose among models on the basis of the incentives that they will have on the actions of other players. The study of mechanism design is about how this is done, and when mechanisms can be created that will produce outcomes deemed desirable. We have already touched on the topic in our study of Arrow and of the Gibbard-Sattertwaithe theorem in week 1, we now explore these themes in greater depth, equipped with the solution concepts developed in later weeks.

Formally, the **design problem** that we care about is a 4-tuple: $\langle N, X, \mathcal{E},\ f\rangle$, where:

- N denotes the set of players

- X denotes the outcome space over which players' preferences are defined, with typical element x;

- $\mathcal{E}$ is a set of environments, that describe the various ways the world could be (there may be weapons of mass destruction in Iraq, there may not; same sex marriages may be more important to voters than foreign policy, or vice versa) If $\mathcal{U}$ denotes the set of all von Neumann

Morgenstern utility functions, then we can have $\mathcal{E} = \mathcal{U}^{|N|}$, with typical element U, a preference profile. Henceforth we will restrict attention to environments that describe variation in preference profiles only.

- f a choice rule that maps from $\mathcal{E}$ into X. Hence for example we have $x = f(U)$.

The design problem is the *given*, and what we want to identify is a game form or mechanism. The game form, $G = \langle (A_i)_{i\in N}, g\rangle$ is a pair: where $(A_i)_{i\in N}$ is a collection of strategy spaces and g is an outcome function $\times_{i\in N}A_i \to X$. The game form then clearly includes information on N and X. From our formal definitions of a game, we can see that to fully describe a game, we need to combine the game form with information of player preferences over outcomes. We can then have a particular game $\Gamma = \Gamma\langle G, U\rangle$.

Given these definitions, we can choose any solution concept, S, we like, and describe the set of solutions to the game Γ, under this concept. Call the set of solutions to the game Γ under solution concept S, $S(\Gamma\langle G, U\rangle)$, we then have that the set of outcomes when Γ is played is given by $g(S(\Gamma\langle G, U\rangle))$. Our task now is to design a game form G that produces the desired outcome. That is to choose G such that $g(S(\Gamma\langle G, U\rangle)) = f(U)$.

In what follows I begin with an important result for political scientists due to Gibbard and Satterthwaite (independently). I then introduce two useful tools for the study of mechanism design: the revelation principle and monotonicity. These are used to establish conditions and properties of solutions to the design problem for voting games and auctions and to establish general conditions for mechanisms to be "incentive compatible." We end the section with a final, pessimistic result on the possibility of designing voluntary mechanisms for efficient bargaining in the presence of uncertainty.

15.1 Manipulation and Voting: The Gibbard-Satterthwaite Theorem

As an introduction to the idea of "mechanism design" problems we consider a setting very similar to one studied by Arrow. Now instead of asking "given information on players' preferences, what social welfare function satisfies desirable properties" we ask "what kind of social welfare function would get people to reveal their preferences in the first place?" So now, rather than assuming that all players' preferences were public knowledge, as in Arrow's framework, we assume that they are private information and that the aggregation mechanism has to rely on players' reporting of their preferences. Under what situations will players have an incentive to behave strategically? Or equivalently, when will the social choice function be such that players have a dominant strategy to reveal their true preferences? If players do not have a dominant strategy to reveal their preferences, then we call the social choice function "**manipulable.**" The Gibbard-Satterthwaite result tells us that players almost always have an incentive to misrepresent their preferences in these collective choice settings.

Definition 186 *A "**social choice function**" is a function that assigns a single collective choice to each possible profile of agents' types.*

Proposition 187 *(The Gibbard-Satterthwaite Theorem) Suppose that X is finite and contains at least three elements and that all players have strict preference relations over elements in X. Then a weakly Pareto efficient social choice function is non-manipulable if and only if it is dictatorial.*

Proof. (**Informal**[1]) Consider just two players $\{A, B\}$ with strict preference orderings over three options, $\{a, b, c\}$. And consider any determinate non-manipulable decision rule that selects one Pareto optimal outcome given any pair of preference orderings. (Such a rule should take any pair of declared preferences and select some Pareto optimal outcome, and given the selection method of the rule no player should want to misrepresent her preferences.) We now show that the only such rule is *Dictatorship.*

Consider first the situation where Players A and B declare themselves to have the orderings given in Figure 15.1.

A	B
a	b
b	a
c	c

FIGURE 15.1. GS-0

What outcome will be chosen? Option c is Pareto dominated, and so either a or b must be selected by the rule. We don't know which, but, assume without loss of generality that option a is selected (a symmetrical argument to what follows can be made if b is selected).

Now, fixing A's preferences we note that Player B has a total of 6 "types" that he can declare himself to be. These six types refer to each of the six possible strict orderings of the three outcomes, marked B-I – B-VI.

[1] A more formal statement and proof of the theorem excludes the notion of Pareto efficiency from the statement of the theorem but derives it from non-manipulability.

A	B-I	B-II	B-III	B-IV	B-V	B-VI
a	*a*	*a*	*b*	*b*	*c*	*c*
b	*b*	*c*	*a*	*c*	*a*	*b*
c	*c*	*b*	*c*	*a*	*b*	*a*

FIGURE 15.2. GS-1

Step 1:We want to see what the rule will select given each type for B. For most of the types that B might declare, we can use the Pareto Principle to eliminate outcomes.

A	B-I	B-II	B-III	B-IV	B-V	B-VI
a	*a*	*a*	*b*	*b*	*c*	*c*
b	~~*b*~~	~~*c*~~	*a*	~~*c*~~	*a*	*b*
c	~~*c*~~	~~*b*~~	~~*c*~~	*a*	~~*b*~~	*a*

FIGURE 15.3. GS-2

Step 2: We already know that if B declared Type III then option *a* would be chosen. It follows from this however that *a* would also be chosen if B declared Type IV as otherwise (i.e. were *b* chosen if Type IV were declared) a Type III B would declare himself to be Type IV, contrary to the assumption of non-manipulability.

But if that is true then it must also be that Declarations B-V and B-VI also lead to outcome *a*, since if they led to anything else, a Type B-IV would clearly be better off declaring himself a B-V or a B-VI to get *anything* but *a*.

A	B-I	B-II	B-III	B-IV	B-V	B-VI
a	$\underline{\boldsymbol{a}}$	$\underline{\boldsymbol{a}}$	~~*b*~~	~~*b*~~	~~*c*~~	~~*c*~~
b	~~*b*~~	~~*c*~~	$\underline{\boldsymbol{a}}$	~~*c*~~	$\underline{\boldsymbol{a}}$	~~*b*~~
c	~~*c*~~	~~*b*~~	~~*c*~~	$\underline{\boldsymbol{a}}$	~~*b*~~	$\underline{\boldsymbol{a}}$

FIGURE 15.4. GS-3

Hence no matter what rule is used, Player A's most preferred option, *a*, is selected irrespective of B's preferences. Hence A is a dictator. ■

15.2 The Revelation Principles

In our discussion of the Gibbard-Satterthwaite we saw a negative result for a game form in which individuals declare their "types"—we saw that unless the choice rule is dictatorial, then individuals will sometimes have an incentive to misrepresent their own preferences.

This might not worry you very much if you think that in most games player's don't report "types" but rather they have the option to play much more complex and interesting strategies. And perhaps it won't worry you if what you care about is outcomes, not about truthful strategies.

The family of Revelation Principles tells you that in fact you *should* be worried. These results provide one way of dramatically increasing the field of application of results of this form.

The principles state that if a social choice function f can be S-implemented by some mechanism in the sense that every equilibrium (using solution concept "S") is one of the social choices given preferences U, and every social choice $x \in f(U)$ can be attained as a solution to the induced game (using solution concept "S"), then there exists a mechanism that "truthfully S-implements" f. The notion of "truthfully S-implementation" is simply that the strategy sets provided by the mechanisms is simply the set of preference profiles and that there is an S-equilibrium in which all players (a) tell the truth about U and (b) when they do this they select an element from f.[2]

Let us state three of them formally first before considering why they are useful:

Proposition 188 [Revelation Principle for Dominant Strategy Equilibrium] *Suppose there exists a mechanism* $\left\langle (A_i)_{i\in N}, g(.)\right\rangle$ *that implements* f *in dominant strategies, then* f *is truthfully implementable in dominant strategies.*

Proposition 189 [Revelation Principle for Nash Equilibrium O&R 185.2] *Suppose there exists a mechanism* $\left\langle (A_i)_{i\in N}, g(.)\right\rangle$ *that Nash-implements* f*, then* f *is truthfully Nash-implementable.*

Proposition 190 [Revelation Principle for Bayesian Nash Equilibrium MWG 23.D.1] *Suppose there exists a mechanism* $\left\langle (A_i)_{i\in N}, g(.)\right\rangle$ *that implements* f *in Bayesian Nash Equilibrium, then* f *is truthfully implementable in Bayesian Nash equilibrium.*

These principles claim that if any game form (be it a one stage or multistage game) has an equilibrium outcome, then that same outcome *could* be achieved in equilibrium if players played some game corresponding to the original game but in which their strategy options are declarations of their types (that is, if they took part in a "direct" mechanism). Furthermore, if

[2] The language is a little confusing, but the fact that a mechanism truthfully S-implements f does not mean that is actually implements f. The reason for this is that "implementation" requires that all elements of f are S-equilibria and only elements of f are S-equilibria. Truthful implementation simply requires that some S-equilibria are elements of f. Others may not be. we see two cases below which emphasize this point, one in the case of auctions and the second in teh case of majority rule.

an equilibrium outcome is achievable in the original game then, the same outcome can be achieved in equilibrium in a corresponding game in which all players reveal their types truthfully.

But the real punchline comes with the contrapositive...

Contrapositive: [Revelation Principle for Bayesian Nash Equilibrium] *Suppose that $f(.)$ is not truthfully implementable in Bayesian Nash equilibrium, then there exists* no *mechanism $\left\langle (A_i)_{i\in N}, g(.)\right\rangle$ that implements the social choice function $f(.)$ in Bayesian Nash Equilibrium.*

From the contrapositive, the revelation principle then implies that if we can identify a property of all outcomes in games (in a given class) induced by direct truth revealing mechanisms (such as dictatorship, inefficiency, or non-existence), then this property extends to the outcomes of all possible games (in that class). In the next two sections we get some mileage out of this principle.

Problem 191 ***A game form for the median voter result.*** *We know that if any game form implements the median voter result in dominant strategies then there exists a game form in which each player has a dominant strategy truthfully to reveal her "type". For the median voter result one game form is simply: let each player declare her ideal point. Then the median declaration is implemented (Moulin 1980). Check that players do not have an incentive to deviate.*

15.3 Monotonicity

Some general results on implementability have been found using only information on the properties of the choice rules. Some of the most important results use information on the monotonicity of the choice rule.

Definition 192 *Consider strategy profiles $U = \{u_i\}$ and $U' = \{u'_i\}$. A choice rule $f:\mathcal{U} \to \mathcal{X}$ is monotonic if for some x and y, if $u_i(x) \geq u_i(y)$ implies $u'_i(x) \geq u'_i(y)$ for every $i \in N$, then $x \in f(U)$ implies $x \in f(U')$.*

This is equivalent to the following somewhat more intuitive definition:

Definition 193 *Consider strategy profiles $U = \{u_i\}$ and $U' = \{u'_i\}$. A choice rule $f:\mathcal{U} \to \mathcal{X}$ is "**monotonic**" if whenever $x \in f(U)$ and $x \notin f(U')$ then there is some player $i \in N$ and some outcome y, such that $u_i(x) \geq u_i(y)$ but $u'_i(x) < u'_i(y)$.*

Hence, if x is selected by the rule given preference profile U but not profile U', then x must have 'moved down' in the rankings of at least one player. For example the following rule is monotonic:

Example 194 *"$x \in f$ if x is the preferred outcome of at least k players."*

This example includes simple majority rules and the Pareto rule. The rule in this example uses information on preferences in a monotonic fashion. But for many implementation problems, the rule is not based on the preferences of individuals. For example imagine a situation where the rule is aimed at eliciting whether a criminal is guilty of a crime with the following choice rule:

Example 195 *"$f =$"guilty" if i is guilty; $f =$"innocent" if i is innocent."*

In this case the rule is not monotonic if the preferences of guilty and innocent players over the court's two possible rulings does not depend on whether or not they are guilty.

We can use the notion of monotonicity to establish the following powerful result (due to Maskin, 1977):

Theorem 196 *If f is Nash implementable then it is monotonic.*

Proof. (Direct Proof) Assume that the game form $\mathcal{G} = \langle N, (A_i), g\rangle$ Nash-implements f. For some U choose $x \in f(U)$ and assume there exists another profile U' with $x \notin f(U')$. The fact that x is a Nash equilibrium of the game $\langle \mathcal{G}, U\rangle$ implies that there exists a Nash equilibrium strategy profile, a, with $g(a) = x$ and $u_i(x) \geq u_i(g(a_i', a_{-i}))$ for every player i and every alternative action a_i', in i's action set. However the fact that $x \notin f(U')$ implies that x is not a Nash equilibrium of the game $\langle \mathcal{G}, U'\rangle$ and hence there must exist at least one player for whom there exists some action, a' such that $u_i'(x) < u_i(g(a_i', a_{-i}))$. For such an a_i' define $y = g(a_i', a_{-i})$. We then have that for this player, $u_i(x) \geq u_i(y)$ but $u_i'(x) < u_i(y)$. ■

The contrapositive to this result is strong: if a rule is *not* monotonic, then it is *not* Nash-implementable. In this case, as in Example 195 above, a rule that depends on the characteristics of players other than on their preferences may not be Nash-implementable. The following example provides a celebrated case of a non Nash-implementable choice function.

Example 197 (Solomon's Problem) . *Wise King Solomon is asked to arbitrate a dispute between two women, A and B, both claiming a child to be theirs. There are three outcomes, outcome x, in which the child is awarded to A; outcome y, in which the child is awarded to B, and outcome z, in which the child is awarded to neither (and executed). Solomon believes that there are two possible preference profiles:*

U in which $u_A(x) \geq u_A(y) \geq u_A(z)$ and $u_B(y) \geq u_B(z) \geq u_B(x)$
and
U' in which $u'_A(x) \geq u'_A(z) \geq u'_A(y)$ and $u'_B(y) \geq u'_B(x) \geq u'_B(z)$

The idea behind these profiles is that under U, Woman A is the real mother and she would rather have the child go to Woman B than to have it killed; Woman B would rather have it killed than to go to Woman A. Under U', Woman B is the real mother and the preferences are the other way round. The social choice rule simply gives the child to the real mother; formally, $f : \mathcal{U} \to \mathcal{X}$ is given by is the following:

$$f = \begin{cases} f(U) = x \\ f(U') = y \end{cases}$$

It is evident to check that this choice rule is not monotonic. The implication from Theorem 196 is that King Solomon could not develop a mechanism that would produce his desired result if the women play Nash.[3]

What of rules like that in Example 194? This rule is monotonic, but is it implementable? Theorem 196 cannot answer the question. However, it turns out that a close converse of Theorem 196 also holds. consider the following notion of veto power:

Definition 198 *A choice rule, f, has "**no veto power**" if $x \in f(U)$ whenever $|\{i \in N | u_i(x) \geq u_i(y)\}| \geq N - 1$ for all y in X.*

This definition simply states that if $N - 1$ players all feel that x is one of their most preferred outcomes, then x is one of the chosen outcomes.

With three or more players, monotonicity is enough to ensure Nash-implementability for rules with no veto power. Formally we have the following:

Theorem 199 *[Maskin 1977]Assume that $|N| \geq 3$, and that f is monotonic and has no veto power, then f is Nash implementable.*

The proof of this theorem is constructive. I do not give the proof here but the following example provides the main intuition behind it. Consider a case of Example 194 in which there are just three player, deciding by majority rule. This rule satisfies all the conditions for Theorem 199 and so we have that it is Nash implementable.

Furthermore, we then have from the revelation principle that the social choice is implementable directly, by players simply revealing their types (or profles of types).

Sure enough, consider a case where the three players have strict preferences over A and B, assume without loss of generality that at least two

[3] It does turn out to be possible to construct extensive form games in which there is a solution to the problem that is implementable in subgame perfect Nash equilibrium.

players prefer A to B. Now consider the following direct mechanism: all players state a preference profile and if one outcome receives a majority of supporters, then that outcome is selected. At least one Nash equilibrium of this game involves two players that prefer A to B stating that they prefer A to B. Hence this direct mechanisms selects the outcome that satisfies the choice rule. *However*, this mechanism does not Nash-implement the choice rule! Why? Because there exist other Nash equilibria of this game that produce outcomes different to those that would be selected by the choice rule. For example, the strategy profile in which all players state that they prefer B (even though at least 2 prefer A to B) is a Nash equilibrium, since any individual deviation from this strategy does not change the result.

What then would a mechanism look like that did in fact implement the rule in Example 194? In Maskin's proof he uses a rule of the following form (written here in terms of the example at hand). Each players strategy set is a triple: a statement about the profile of preferences, $U \in \mathcal{U}$; a statement about their preferred outcome, $x \in X$; and a real number $k \in \mathbb{R}$. In the example we are considering if we use a to label the preference of a player that prefers A to B, and b the preferences of a player that prefers B to A, then a declared profile may take the form $\langle a, b, a\rangle$ or $\langle b, b, a\rangle$; an example of a strategy for a given player might be $\langle \langle a, b, a\rangle, A, 5\rangle$. Note that all players have the same strategy sets.

Now consider the following rule: If $N-1$ players all play strategy $\langle U^*, x^*, k^*\rangle$, where $x^* = f(U^*)$, and the N^{th} player plays some $\langle U', x', k'\rangle$, then implement x^* if, according to U^*, the N^{th} player does better out of x' then she does out of x^*, but implement x' if, according to U^*, the N^{th} player actually does no better out of x' than she does out of x^*. In all other cases, identify the player (or some player) for whom $k_i \geq k_j$ for all $j \in N$ and x_i (in other words, play the outcome identified by whoever chose the highest number).

In this case all Nash equilibria implement the rule. In particular, the strategy profile in which all players play $\langle \langle b, b, b\rangle, B, k\rangle$ can not be a Nash in this game since for any player that prefers A, some deviation of the form deviation to $\langle \langle a, b, b\rangle, A, k\rangle$ will result in A being selected, and hence is preferable for this player. assuming that the true preference profile for this game is $\langle a, a, b\rangle$All equilibria for this game are of the form $\langle \langle a, a, b\rangle, A, k\rangle$ (note that deviations by the third player in a bid to propose B will always be ignored since the majority has signalled that this player is a minority player that prefers B).

15.4 Incentive Compatibility

The previous results were centered on conditions for which players will in fact play according to a strategy that is associated with their type in equilibrium. This is the question of (Bayesian) incentive compatibility.

The next proposition helps pin down some quite general conditions for Bayesian incentive compatible (a.k.a truthfully implementable) mechanism design questions for the class of cases where players have quasi-linear utility and their possible "types" can be represented on a single dimension of variation. The results have some surprising implications for Auction theory, summarized below in a version of the "**revenue equivalence theorem**."

The proposition that follows states first (roughly) that when an incentive compatible mechanism is used to regulate competition over some good, people that value the good more are more likely to get it in equilibrium, *ceteris paribus*; the second is that an individual's expected payoff from participating in the mechanism depends on the payoff of the lowest type in his distribution of types and on the function that assigns the probability that different types will win; it does not depend on other features of the mechanism. More formally:

Proposition 200 (Bayesian Incentive Compatibility) *Given:*

- *Let θ denote the vector of true types in the population and $\tilde{\theta}$ the vector of declared types.*
- *Let $p_i(\tilde{\theta})$ denote the probability that player i receives the good as a function of the declared types of all of the players.*
- *And let $t_i(\tilde{\theta})$ denote player i's expected transfer as a function of the declared types of all of the players.*
- *Assume that the expected utility of player i of type θ_i when the vector of declared types is $\tilde{\theta}$ is given by $u_i(\theta_i|\tilde{\theta}) := \theta_i p_i(\tilde{\theta}) - t_i(\tilde{\theta})$.*

Then the mechanism given by $\langle (p_i(.))_{i\in N}, (t_i(.))_{i\in N}\rangle$ is Bayesian incentive compatible if and only if:

1. *$p_i(\tilde{\theta}_i|\theta_{-i})$ is non decreasing in $\tilde{\theta}_i$*
2. *the equilibrium expected utility for each i when all players report their true types is given by:*

$$u_i(\theta_i|\theta) = u_i(\underline{\theta}_i) + \int_{\underline{\theta}_i}^{\theta_i} p_i(x)dx$$

Proof. In what follows we concentrate on one player and assume that all other players play truthfully. This allows us to drop player subscripts for the proof of this proposition. [Hence we now use θ to denote the type of a particular player rather than the vectors of all types.]

For the "only if" part: Incentive compatibility means that a player with type θ does better when declaring θ than she would from declaring any other type, say $\tilde{\theta}$. That is:

$$\begin{aligned} u(\theta|\theta) &\geq u(\theta|\theta+\Delta) \\ &= \theta p(\theta+\Delta) - t(\theta+\Delta) \\ &= (\theta+\Delta)p(\theta+\Delta) - (\theta+\Delta)p(\theta+\Delta) + \theta p(\theta+\Delta) - t(\theta+\Delta) \end{aligned}$$

or:

$$u(\theta|\theta) \geq u(\theta+\Delta|\theta+\Delta) + p(\theta+\Delta)(-\Delta) \tag{15.1}$$

By the same logic a type $\theta+\Delta$ does not want to imitate a type θ.

Hence:

$$u(\theta+\Delta|\theta+\Delta) \geq u(\theta|\theta) + p(\theta)\Delta$$

Combining these two conditions gives:

$$p(\theta+\Delta) \geq \frac{u(\theta+\Delta|\theta+\Delta) - u(\theta|\theta)}{\Delta} \geq p(\theta)$$

This gives the monotonicity we need to establish Part (1). Now taking the limit of this expression as Δ goes to 0 gives:

$$p(\theta) = \frac{d}{d\theta}u(\theta|\theta) = p(\theta)$$

Integrating over $[\underline{\theta}, \theta]$ gives:

$$\int_{\underline{\theta}}^{\theta} p(x)dx = \int_{\underline{\theta}}^{\theta} \frac{d}{dx}u(x|x)dx = u(\theta|\theta) - u(\underline{\theta}|\underline{\theta})$$

And so (as we are assuming that all players play truthfully we can removing the conditioners):

$$u(\theta) = u(\underline{\theta}) + \int_{\underline{\theta}}^{\theta} p(x)dx$$

This establishes Part 2.

For the "if" part we essentially work backwards. Note that with $\Delta > 0$, Part 2 of the proposition tells us that: $u(\theta) - u(\theta+\Delta) = u(\underline{\theta}) + \int_{\underline{\theta}}^{\theta} p(x)dx - u(\underline{\theta}) - \int_{\underline{\theta}}^{\theta+\Delta} p(x)dx = \int_{\theta+\Delta}^{\theta} p(x)dx$. From monotonicity (Part 1 of the proposition) we then have: $\int_{\theta+\Delta}^{\theta} p(x)dx \geq \int_{\theta+\Delta}^{\theta} p(\theta+\Delta)dx = p(\theta+\Delta)(-\Delta)$ and so $u(\theta) \geq u(\theta+\Delta) + p(\theta+\Delta)(\theta+\Delta)$ but this is precisely condition (15.1) above that guarantees that the player does not want to declare $\theta+\Delta$ instead of θ. The same logic holds for $\Delta < 0$. ■

15.5 Application: The Revenue Equivalence Theorem

Proposition 200 provides most of what we need to prove a simple version of the **revenue equivalence theorem**. Revenue equivalence theorems state that when players are risk neutral, the expected payment by each individual and the expected revenue earned by an auctioneer do not depend on the details of the auction design. For example the auctioneer can expect the same revenue whether he chooses to use an English auction a Dutch auction, a first or a second price sealed bid auction, an all-pay auction and so on. More outlandish auction designs could also be considered.[4]

Proposition 201 (Revenue Equivalence Theorem) *Consider the class of auctions in which:*

- *A single object is on offer to be sold*
- *A set of n risk neutral buyers have types drawn independently from the same distribution F with density function $f(.)$.*
- *In equilibrium, the object always goes to the player with the highest type*
- *If any player has the lowest possible valuation of the object then she has a zero probability of winning it in equilibrium .*

Then the expected payments by all individuals and the expected revenue of the seller is independent of the details of the auction design.

Proof. From the revelation principle we can restrict our attention to incentive compatible direct mechanisms. Under the conditions of the proposition we have from Proposition 200 that:
(1) $u(\theta) = u(\underline{\theta}) + \int_{\underline{\theta}}^{\theta} p(x)dx$
(2) $u(\theta) = \theta p(x) - t(\theta)$
Together these imply that $u(\underline{\theta}) + \int_{\underline{\theta}}^{\theta} p(x)dx = \theta p(x) - t(\theta)$ and so:

$$\begin{aligned} t(\theta) &= \theta p(x) - u(\underline{\theta}) - \int_{\underline{\theta}}^{\theta} p(x)dx \\ &= \theta p(x) - 0 - \int_{\underline{\theta}}^{\theta} p(x)dx \\ &= \theta(F(x))^{n-1} + \int_{\underline{\theta}}^{\theta} (F(x))^{n-1}dx \end{aligned}$$

[4] Although see the next exercise for a caveat on the notion that the highest type winner "always" wins in equilibrium. More carefully, in these games there is guaranteed to be an equilibrium in which the hiughest type wins...

And hence the expected payment by the player depends only on features of the distribution of types and not on any features of the auction. But this is true for all player types. The expected revenue is also constant since it is the sum of these individual payments. ■

Exercise 202 *(based on Osborne and Rubinstein 18.3) The second price sealed bid auction truthfully implements (in dominant strategies) equilibrium outcomes from other auction mechanisms, such as the Japanese auction, with independent private values. Show first that it is a weakly dominant strategy equilibrium for all players to declare their types truthfully. Does this imply that this mechanism* necessarily *results in truthful type revelation? Search for an example of an equilibrium from a two player (two buyers) second price sealed bid auction in which the highest type does not win the auction and players do not truthfully reveal their types.*

15.6 Application II: Bargaining (Myerson-Satterthwaite)

We now consider a general result on the universe of incentive compatible bargaining mechanisms that could be used to structure a trade between two parties when the values of the two parties are not known to each other. Myerson and Satterthwaite (1983) ask: in such situations, is there any mechanism, such as alternating offers type bargaining protocols, take-it-or-leave it offers protocols, or any mechanisms that we have not yet thought of, that can guarantee that trade will be ex-post efficient in situations where there is *ex ante* uncertainty that gains can be made from trade?

Their answer is “No! At least not unless you introduce some outside actor that either forces players to take part in the mechanism even if they think it could be harmful for them, or some outside factor that intervenes to pump in extra money or extract surplus monies.”

Proposition 203 (Myerson-Satterthwaite 1983) *Consider the situation in which a buyer and a seller wish to engage in trade negotiations over an object. The value of the object to the buyer is given by β where β is distributed over the range $[\underline{\beta}, \overline{\beta}]$ with strictly positive density b (and corresponding cumulative density B) The value of the object to the seller is given by σ where σ is distributed over the range $[\underline{\sigma}, \overline{\sigma}]$ with strictly positive density s (and corresponding cumulative density S). Assume furthermore that there is a positive probability that trade would be efficient (i.e. $\underline{\sigma} < \overline{\beta}$). But assume that there is also a positive probability that trade would be inefficient (i.e. $\underline{\beta} < \overline{\sigma}$).Then there is no* ***efficient*** *trading mechanism that* *satisfies* ***individual rationality****,* ***incentive compatibility*** *and* ***budget balance****.*

Proof. A buyer of type β has expected utility of:

$$U_\beta(\beta) = \mathsf{E}_\sigma p(\beta,\sigma)\beta - \mathsf{E}_\sigma t(\beta,\sigma)$$

A seller of type σ has expected utility of:

$$U_\sigma(\sigma) = \mathsf{E}_\beta t(\beta,\sigma) - \mathsf{E}_\beta p(\beta,\sigma)\sigma$$

Adding the seller's and buyer's expected utilities together gives:

$$U_\beta(\beta) - U_\sigma(\sigma) = \mathsf{E}_\sigma p(\beta,\sigma)\beta - \mathsf{E}_\sigma t(\beta,\sigma) + \mathsf{E}_\beta t(\beta,\sigma) - \mathsf{E}_\beta p(\beta,\sigma)\sigma$$

However, from Preposition 200 above we also have that the **incentive compatibility** condition allows us to write buyer's expected utility can be written as a function of the utility of the buyer type with the lowest value; the seller's expected utility can be written in terms of the payoff to the seller with the highest type.[5] That is:

$$U_\beta(\beta) = U_\beta(\underline{\beta}) + \int_{\underline{\beta}}^{\beta} \mathsf{E}_\sigma p(x,\sigma)dx$$

$$U_\sigma(\sigma) = U_\sigma(\overline{\sigma}) + \int_{\sigma}^{\overline{\sigma}} \mathsf{E}_\beta p(\beta,y)dy$$

And hence:

$$U_\beta(\beta) - U_\sigma(\sigma) = U_\beta(\underline{\beta}) + \int_{\underline{\beta}}^{\beta} \mathsf{E}_\sigma p(x,\sigma)dx + U_\sigma(\overline{\sigma}) + \int_{\sigma}^{\overline{\sigma}} \mathsf{E}_\beta p(\beta,y)dy$$

Using these two equations we can write net transfers as:

$$\begin{aligned} \mathsf{E}_\beta t(\beta,\sigma) - \mathsf{E}_\sigma t(\beta,\sigma) \quad = \quad & \left(\int_{\underline{\beta}}^{\beta} \mathsf{E}_\sigma p(x,\sigma)dx - \mathsf{E}_\sigma p(\beta,\sigma)\beta\right) \\ & + \left(\int_{\sigma}^{\overline{\sigma}} \mathsf{E}_\beta p(\beta,y)dy + \mathsf{E}_\beta p(\beta,\sigma)\sigma\right) + U_\beta(\underline{\beta}) + U_\sigma(\overline{\sigma}) \end{aligned}$$

[5] Note that the application of the proposition is slightly different depending on whether we are dealing with the buyer or the seller. This simply requires an interpretation of what we mean by the "highest" and "lowest" types. For more on this see Fudenburg and Tirole section 7.3 or MWG section 23.D.

Net *expected* transfers then can be written:

$$\begin{aligned}\mathsf{E}_\sigma\mathsf{E}_\beta t(\beta,\sigma)-\mathsf{E}_\sigma\mathsf{E}_\beta t(\beta,\sigma) &= \mathsf{E}_\beta\left(\int_{\underline{\beta}}^{\beta}\mathsf{E}_\sigma p(x,\sigma)dx-\mathsf{E}_\sigma p(\beta,\sigma)\beta\right)\\ &\quad+\mathsf{E}_\sigma\left(\int_{\sigma}^{\overline{\sigma}}\mathsf{E}_\beta p(\beta,y)dy+\mathsf{E}_\beta p(\beta,\sigma)\sigma\right)+U_\beta(\underline{\beta})+U_\sigma(\overline{\sigma})\end{aligned}$$

But from **budget balance** we have that net expected transfers is equal to 0, or:

$$\begin{aligned}0 &= \mathsf{E}_\beta\left(\int_{\underline{\beta}}^{\beta}\mathsf{E}_\sigma p(x,\sigma)dx-\mathsf{E}_\sigma p(\beta,\sigma)\beta\right)+\mathsf{E}_\sigma\left(\int_{\sigma}^{\overline{\sigma}}\mathsf{E}_\beta p(\beta,y)dy+\mathsf{E}_\beta p(\beta,\sigma)\sigma\right)\\ &\quad+U_\beta(\underline{\beta})+U_\sigma(\overline{\sigma})\end{aligned}$$

and hence we can write the sum of the expected utility of the lowest type bargainer plus the highest type seller as:

$$\begin{aligned}U_\beta(\underline{\beta})+U_\sigma(\overline{\sigma}) &= -\mathsf{E}_\beta\left(\int_{\underline{\beta}}^{\beta}\mathsf{E}_\sigma p(x,\sigma)dx-\mathsf{E}_\sigma p(\beta,\sigma)\beta\right)\\ &\quad-\mathsf{E}_\sigma\left(\int_{\sigma}^{\overline{\sigma}}\mathsf{E}_\beta p(\beta,y)dy+\mathsf{E}_\beta p(\beta,\sigma)\sigma\right)\end{aligned}$$

Now from the **individual rationality** constraint we have $U_\beta(\underline{\beta})\geq 0$ and $U_\sigma(\overline{\sigma})\geq 0$ and hence, $U_\beta(\underline{\beta})+U_\sigma(\overline{\sigma})\geq 0$. We then have :

$$-\mathsf{E}_\beta(\int_{\underline{\beta}}^{\beta}\mathsf{E}_\sigma p(x,\sigma)dx-\mathsf{E}_\sigma p(\beta,\sigma)\beta)-\mathsf{E}_\sigma(\int_{\sigma}^{\overline{\sigma}}\mathsf{E}_\beta p(\beta,y)dy+\mathsf{E}_\beta p(\beta,\sigma)\sigma)\geq 0$$

Integrating by parts yields:

$$\mathsf{E}_\beta\left(\beta-\frac{1-B(\beta)}{b(\beta)}\right)\mathsf{E}_\sigma p(\beta,\sigma)-\mathsf{E}_\sigma\left(\sigma+\frac{S(\sigma)}{s(\sigma)}\right)\mathsf{E}_\beta p(\beta,\sigma)\geq 0$$

[**Note**: That was the hardest step technically so far. For the first part you need to use the chain rule: $[\int_a^b f(x)\frac{dg(x)}{dx}dg=[fg]_a^b-\int_a^b\left[g\frac{df}{dx}\right]dx]$ to show $\mathsf{E}_\beta\int_{\underline{\beta}}^{\beta}[\mathsf{E}_\sigma p(x,\sigma)]\,dx=\mathsf{E}_\beta\left(\frac{1-B(\beta)}{b(\beta)}\right)[\mathsf{E}_\sigma p(\beta,\sigma)]$; and similarly for the second part. Try it. If you run into problems I can send you the full steps.]

And collecting gives:

$$\mathsf{E}_\beta \mathsf{E}_\sigma p(\beta,\sigma)\left[\left(\beta - \frac{1-B(\beta)}{b(\beta)}\right) - \left(\sigma + \frac{S(\sigma)}{s(\sigma)}\right)\right] \geq 0$$

Writing the expectations out explicitly we have:

$$\int_{\underline{\beta}}^{\overline{\beta}} \int_{\underline{\sigma}}^{\overline{\sigma}} p(\beta,\sigma)\left[\left(\beta - \frac{1-B(\beta)}{b(\beta)}\right) - \left(\sigma + \frac{S(\sigma)}{s(\sigma)}\right)\right] s(\sigma)b(\beta)d\sigma d\beta \geq 0$$

Next we use our final assumption: that **bargaining is efficient.** This implies that $p(\beta,\sigma) = 1$ whenever $\sigma < \beta$ and $p(\beta,\sigma) = 0$ whenever $\beta < \sigma$. This lets us write our condition as:

$$\phi := \int_{\underline{\beta}}^{\overline{\beta}} \int_{\underline{\sigma}}^{\min(\beta,\overline{\sigma})} \left[\left(\beta - \frac{1-B(\beta)}{b(\beta)}\right) - \left(\sigma + \frac{S(\sigma)}{s(\sigma)}\right)\right] s(\sigma)b(\beta)d\sigma d\beta \geq 0$$

Note: We have now used up all our assumptions, this means that to complete the proof we have to find a **contradiction**. The contradiction is this: contrary to our claim that $\phi \geq 0$, we will show that $\phi < 0$. Here we go...

Integrating over the seller's type gives:

$$\phi = \int_{\underline{\beta}}^{\overline{\beta}} \left[\left(\beta - \frac{1-B(\beta)}{b(\beta)} - \sigma\right) S(\sigma)\right]_{\underline{\sigma}}^{\min(\beta,\overline{\sigma})} b(\beta)d\beta$$

...which can then be expanded out to give:

$$\begin{aligned}\phi \quad = \quad & \int_{\underline{\beta}}^{\overline{\beta}} \left[\left(\beta - \frac{1-B(\beta)}{b(\beta)} - \min(\beta,\overline{\sigma})\right) S(\min(\beta,\overline{\sigma}))\right] b(\beta)d\beta \\ & - \int_{\underline{\beta}}^{\overline{\beta}} \left[\left(\beta - \frac{1-B(\beta)}{b(\beta)} - \underline{\sigma}\right) S(\underline{\sigma})\right] b(\beta)d\beta\end{aligned}$$

But using $S(\underline{\sigma}) = 0$ the second part vanishes, yielding:

$$\phi = \int_{\underline{\beta}}^{\overline{\beta}} \left[\left(\beta - \frac{1-B(\beta)}{b(\beta)} - \min(\beta,\overline{\sigma})\right) S(\min(\beta,\overline{\sigma}))\right] b(\beta)d\beta$$

To deal with the awkward $\min(\beta,\overline{\sigma}))$ term we split the range of integration over β into two parts, that above and that below $\overline{\sigma}$. Within each of these two parts we know the value of $\min(\beta,\overline{\sigma})$.

$$\phi = \int_{\underline{\beta}}^{\overline{\sigma}} \left(\beta - \frac{1-B(\beta)}{b(\beta)} - \beta \right) S(\beta)b(\beta)d\beta + \int_{\overline{\sigma}}^{\overline{\beta}} \left(\beta - \frac{1-B(\beta)}{b(\beta)} - \overline{\sigma} \right) S(\overline{\sigma})b(\beta)d\beta$$

Tidying up gives:

$$\begin{aligned} \phi &= -\int_{\underline{\beta}}^{\overline{\sigma}} \left[(1-B(\beta))\, S(\beta) \right] d\beta + \int_{\overline{\sigma}}^{\overline{\beta}} \left[(\beta - \overline{\sigma})\, b(\beta) - 1 + B(\beta) \right] d\beta \\ &= -\int_{\underline{\beta}}^{\overline{\sigma}} (1-B(\beta))\, S(\beta) d\beta + \left[(\beta - \overline{\sigma})\, (B(\beta) - 1) \right]_{\overline{\sigma}}^{\overline{\beta}} d\beta \end{aligned}$$

Note that as $B(\overline{\beta}) = 1$ the second term drops out, leaving:

$$\phi = -\int_{\underline{\beta}}^{\overline{\sigma}} (1-B(\beta))\, S(\beta) d\beta$$

With $\underline{\beta} < \overline{\sigma}$ we have that $\int_{\underline{\beta}}^{\overline{\sigma}} (1-B(\beta))\, S(\beta)d\beta$ is positive and hence the whole expression is negative (with $\underline{\beta} < \overline{\sigma}$ the integral spans a positive range, the quantities $(1-B(\beta))$ and $S(\underline{\beta})$take on a positive values for $\underline{\beta} < \beta < \overline{\sigma} < \overline{\beta}$).[6] And this provides our contradiction. ■

*How "**general**" is this result?*

The result is general insofar as it applies to a very large class of mechanisms. One limitation on its generality is that it assumes that the densities of player types is everywhere positive over their range. We can show that if this condition fails the result no longer holds (In fact, look at the next exercise!). Another seeming limitation is that the proposition assumes that players have quasi-linear utility. This could however be seen as a strength of the model—we have seen that positive results can be achieved in quasi-linear environments even though they cannot be achieved in more general environments. The Myerson-Satterthwaite theorem tells us that positive results cannot be attained *even* in quasi-linear environments, hence they cannot be guaranteed in more general environments in which quasi-linear environments are a subset (although possibly their are some types of preferences such that some efficient mechanisms do exist in the presence of uncertainty). Arguably too the model is limited to the extent that it only applies to bilateral bargains rather than to the more general multilateral case. This type

[6] We can interpret the part within the integral as the probability that the buyer's valuation is greater than β multiplied by the probability that the seller's is less than β, which is the probability that type β lies in the range where trade is inefficient.

of limitation is not severe to the extent that the domain of the limitation is known—in applications we typically know whether we are dealing with two person rather than multilateral bargaining; we might not know if the densities of players' types are everywhere strictly positive. For generalizations to settings with more players see [P. Cramton, R. Gibbons and P. Klemperer, "Dissolving a partnership efficiently." *Econometrica* 55 (1987), 615-632 http://www.cramton.umd.edu/papers1984-1989/87econ-dissolving-a-partnership-efficiently.pdf].

Exercise 204 *Find a mechanism that is efficient, budget-balancing, incentive compatible and that satisfies participation constraints when buyers and sellers are drawn from "atomic" distributions (e.g.* $Prob(s = \underline{s}) = p$*,* $Prob(s = \overline{s}) = 1 - p$*;* $Prob(b = \underline{b}) = q$ *,* $Prob(b = \overline{b}) = 1 - q$*. (And prove that your mechanism has these properties!)*