Voting, Arbitration, and Fair Division

The mathematics of social choice

Marcus Pivato
Trent University

March 10, 2007

Copyright

This is a work in progress. Updates and improvements are available at the author's website:

http://xaravve.trentu.ca/pivato

Colophon

All text was prepared using Leslie Lamport's LaTeX2e typesetting language. Pictures were generated using William Chia-Wei Cheng's excellent `TGIF` object-oriented drawing program. This book was prepared entirely on computers using the `RedHat Linux` and `Ubuntu Linux` operating systems.

Contents

Part I

Voting and Social Choice

Chapter 1

Binary Voting Procedures

Democracy is the worst form of government except all those other forms that have been tried from time to time. —Winston Churchill

The word 'democracy' has been used to describe many different political systems, which often yield wildly different outcomes. The simple rule of 'decision by majority' can be made complicated in several ways:

- Granting veto power to some participants (e.g. the permanent members of the UN Security Council or the President of the United States), possibly subject to 'override' by a sufficiently large majority of another body (e.g. the United States Senate). .
- Requiring a majority by two different measures (e.g. in a federal system, a constitutional ammendment might require the support of a majority of the population and a majority of states/provinces).
- Giving different 'weight' to different voters (e.g. different shareholders in a publically traded corporation, or different states in the European Union).
- Forcing voters to vote in 'blocs' (e.g. political parties)

In this chapter we will consider the simplest kind of democratic decision-making: that between two alternatives. Nevertheless, we will see that aforementioned complications engender many surprising phenomena.

1A Simple Majority Voting: May's Theorem

Prerequisites: §2C.1 **Recommended:** §2C.3

The most obvious social choice function for two alternatives is the simple majority vote. The more elaborate voting procedures (Borda count, pairwise votes, approval voting, etc.) all reduce to the majority vote when $|\mathcal{A}| = 2$. Indeed, the conventional wisdom says that majority

vote is the 'only' sensible democratic procedure for choosing between two alternatives. The good news is that, for once, the conventional wisdom is right.

Suppose that $\mathcal{A} = \{A, B\}$. In §2C.3 we introduced three desiderata which any 'reasonable' voting procedure should satisfy:

(M) (*Monotonicity*) Let ρ be a profile such that $A \overset{\rho}{\sqsubseteq} B$. Let $v \in \mathcal{V}$ be some voter such that $B \underset{v}{\overset{\rho}{\preceq}} A$, and let δ be the profile obtained from ρ by giving v a new preference ordering $\underset{v}{\overset{\delta}{\succeq}}$, such that $A \underset{v}{\overset{\delta}{\succeq}} B$ (all *other* voters keep the same preferences). Then $A \overset{\delta}{\sqsubseteq} B$.

(A) (*Anonymity*) Let $\sigma : \mathcal{V} \longrightarrow \mathcal{V}$ be a permutation of the voters. Let ρ be a profile, and let δ be the profile obtained from ρ by permuting the voters with σ. In other words, for any $v \in \mathcal{V}$, $\quad \delta(v) \;=\; \rho\Big(\sigma(v)\Big)$. Then $\Big(A \overset{\rho}{\sqsupseteq} B\Big) \iff \Big(A \overset{\delta}{\sqsupseteq} B\Big)$.

(N) (*Neutrality*) Let ρ be a profile, and let δ be the profile obtained from ρ by reversing the positions of A and B for each voter. In other words, for any $v \in \mathcal{V}$,

$$\left(A \underset{v}{\overset{\rho}{\succeq}} B\right) \iff \left(B \underset{v}{\overset{\delta}{\succeq}} A\right).$$

Then the outcome of δ is the reverse of the outcome of ρ. That, is, for any $B, C \in \mathcal{A}$,

$$\Big(A \overset{\rho}{\sqsupseteq} B\Big) \iff \Big(B \overset{\delta}{\sqsupseteq} A\Big).$$

A *semistrict* voting procedure is a function $\Pi : \mathfrak{R}^*(\mathcal{V}, \mathcal{A}) \longrightarrow \mathcal{P}(\mathcal{A})$. Thus, the voters must provide *strict* preferences as input, but the output might have ties. Let $V := \#(\mathcal{V})$. We say Π is a **quota system** if there is some $Q \in [0..V]$ so that, for any $\rho \in \mathfrak{R}(\mathcal{V}, \mathcal{A})$,

(Qa) If $\#\left\{v \in \mathcal{V} \;;\; A \underset{v}{\overset{\rho}{\succ}} B\right\} \;>\; Q$, then $A \overset{\rho}{\sqsupset} B$.

(Qb) If $\#\left\{v \in \mathcal{V} \;;\; B \underset{v}{\overset{\rho}{\succ}} A\right\} \;>\; Q$, then $B \overset{\rho}{\sqsupset} A$.

(Qc) If neither of these is true, then $A \overset{\rho}{\approx} B$.

For example:

- The **simple majority vote** is a quota system where $Q = V/2$.
- The **two thirds majority vote** is a quota system where $Q = \frac{2}{3}V$. If an alternative does not obtain at least two thirds support from the populace, then it is not chosen. If *neither* alternative gets two thirds support, then *neither* is chosen; the result is a 'tie'.

- The **unanimous vote** is a quota system where $Q = V - 1$. Thus, an alternative must be receive unanimous support to be chosen. This is the system used in courtroom juries.
- The **totally indecisive** system is one where $Q = V$; hence condition **(Qc)** is always true, and we always have a tie.

Note that the quota Q must be no less than $V/2$. If $Q < V/2$, then it is theoretically possible to satisfy conditions **(Qa)** and **(Qb)** simultaneously, which would be a contradiction (since **(Qa)** implies a *strict* preference of A over B, and **(Qb)** implies the opposite).

If V is odd, and we set $Q = V/2$ (the simple majority system) then condition **(c)** is never satisfied. In other words, ties never occur, so we get a *strict* voting procedure.

Theorem 1A.1 *Any semistrict binary voting system satisfying* **(A)**, **(N)** *and* **(M)** *is a quota system.*

Proof: For any $\rho \in \mathfrak{R}^*(\mathcal{V}, \mathcal{A})$, let

$$\alpha(\rho) \quad = \quad \#\Big\{ v \in \mathcal{V} \,;\, A \underset{v}{\overset{\rho}{\succ}} B \Big\} \qquad \text{and} \qquad \beta(\rho) \quad = \quad \#\Big\{ v \in \mathcal{V} \,;\, B \underset{v}{\overset{\rho}{\succ}} A \Big\}.$$

Now, suppose $\Pi : \mathfrak{R}^*(\mathcal{V}, \mathcal{A}) \longrightarrow \mathcal{P}(\mathcal{V}, \mathcal{A})$ satisfies **(A)**, **(N)** and **(M)**. Since Π is anonymous, the outcome is determined entirely by the number of voters who prefer A to B, and the number who prefer B to A. In other words, $\Pi(\rho)$ is determined entirely by $\alpha(\rho)$ and $\beta(\rho)$.

However, the voters must provide *strict* preferences, so we also know that $\beta(\rho) = V - \alpha(\rho)$. Thus, $\Pi(\rho)$ is is really determined by $\alpha(\rho)$. Hence, there is some function $\widetilde{\Pi} : \mathbb{N} \longrightarrow \mathcal{P}(\mathcal{A})$ such that $\Pi(\rho) \;=\; \widetilde{\Pi}\Big(\alpha(\rho)\Big)$, for any $\rho \in \mathfrak{R}^*(\mathcal{V}, \mathcal{A})$.

Claim 1: *Suppose $\rho, \delta \in \mathfrak{R}^*(\mathcal{V}, \mathcal{A})$ are two profiles, such that $\alpha(\delta) \geq \alpha(\rho)$. Then*

$$\Big(A \;\overset{\rho}{\sqsupset}\; B \Big) \Longrightarrow \Big(A \;\overset{\delta}{\sqsupset}\; B \Big).$$

Proof: **Exercise 1.1** Hint: use the Monotonicity axiom **(M)**. ◇ Claim 1

Let $Q \quad = \quad \min\Big\{ q \in \mathbb{N} \,;\, \text{there exists some } \rho \in \mathfrak{R}^*(\mathcal{V}, \mathcal{A}) \text{ such that } \alpha(\rho) = q \text{ and } A \;\overset{\rho}{\sqsupset}\; B \Big\}$.

Claim 2: *If $\delta \in \mathfrak{R}^*(\mathcal{V}, \mathcal{A})$ is any profile, then* $\Big(A \;\overset{\delta}{\sqsupset}\; B \Big) \iff \Big(\alpha(\delta) \geq Q \Big)$.

Proof: '$\Longrightarrow$' is true by definition of Q.

'$\Longleftarrow$': By definition, there is some profile $\rho \in \mathfrak{R}^*(\mathcal{V}, \mathcal{A})$ such that $\alpha(\rho) = Q$ and $A \;\overset{\rho}{\sqsupset}\; B$.

Thus, if $\alpha(\delta) \geq Q = \alpha(\rho)$, then Claim 1 implies that $A \;\overset{\delta}{\sqsupset}\; B$. ◇ Claim 2

Claim 3: *If $\delta \in \mathfrak{R}(\mathcal{V}, \mathcal{A})$ is any profile, then* $\Big(B \;\overset{\delta}{\sqsupset}\; A \Big) \iff \Big(\beta(\delta) \geq Q \Big)$.

Proof: **Exercise 1.2** Hint: use Claim 2 and the Neutrality axiom **(N)**. ◇ Claim 3

Claim 2 says Q satisfies property **(Qa)** of a quota system. Claim 3 says Q satisfies property **(Qb)** of a quota system. Claims 2 and 3 together imply that Q satisfies property **(Qc)** of a quota system. □

Corollary 1A.2 May's Theorem (1952) [May52]
Suppose $V = \#(\mathcal{V})$ is odd. Then the only strict binary voting system satisfying **(A)**, **(N)** *and* **(M)** *is the simple majority vote.*

Proof: The previous theorem says that any binary voting system satisfying **(A)**, **(N)** and **(M)** must be a quota system. To ensure that this is a *strict* voting system, we must guarantee that ties are impossible; ie. that condition **(Qc)** never occurs. This can only happen if V is odd and we set $Q = V/2$ —ie. the we have the simple majority vote. □

Further reading: May's theorem first appeared in [May52]. A good discussion is in Taylor [Tay95, §10.3].

1B Weighted Majority Voting Systems

Whenever you find yourself in the majority, it is time to stop and reflect. —Mark Twain

Prerequisites: §2C.1 **Recommended:** §2C.3

Suppose $\mathcal{A} = \{Y\!\!\!/\!\!\!e\!\!\!s, N\!\!\!/o\}$, where we imagine $Y\!\!\!/\!\!\!e\!\!\!s$ to be some proposal (eg. new legislation) and $N\!\!\!/o$ to be the negation of this proposal (eg. the status quo). We can assume that any a binary voting procedure to decide between $Y\!\!\!/\!\!\!e\!\!\!s$ and $N\!\!\!/o$ must be *strict*, because in a tie between $Y\!\!\!/\!\!\!e\!\!\!s$ and $N\!\!\!/o$, the 'default' choice will be $N\!\!\!/o$. Observe that such a procedure will generally *not* satisfy neutrality axiom **(N)**, because the status quo ($N\!\!\!/o$) is favoured over novelty ($Y\!\!\!/\!\!\!e\!\!\!s$).

A **weighted** voting system is a strict binary voting procedure where the votes of different voters have different 'weights'. To be precise, there is a **weight function** $\omega : \mathcal{V} \longrightarrow \mathbb{N}$ so that, for any $\rho \in \mathfrak{R}(\mathcal{V}, \mathcal{A})$, the *total support* for alternative $Y\!\!\!/\!\!\!e\!\!\!s$ is defined:

$$\Upsilon(\rho) \quad = \quad \sum_{y \in \mathcal{Y}(\rho)} \omega(y) \qquad \text{where} \quad \mathcal{Y}(\rho) \; = \; \left\{ v \in \mathcal{V} \; ; \; Y\!\!\!/\!\!\!e\!\!\!s \underset{v}{\overset{\rho}{\succ}} N\!\!\!/o \right\}.$$

The **total weight** of the system is $W = \sum_{v \in \mathcal{V}} \omega(v)$. Finally, we set a **quota** $Q \in [0...W]$ so that

$$\left(Y\!\!\!/\!\!\!e\!\!\!s \; \overset{\rho}{\sqsupset} \; N\!\!\!/o \right) \iff \left(\Upsilon(\rho) \geq Q \right).$$

Most systems favour the 'status quo' alternative ($N\!\!\!/o$), which means that $Q \geq \frac{1}{2}W$.

Example 1B.1: The European Economic Community

The Treaty of Rome (1958) defined a voting procedure with six voters:

$$\mathcal{V} \;=\; \{\text{France}, \text{Germany}, \text{Italy}, \text{Belgium}, \text{Netherlands}, \text{Luxembourg}\}$$

and the following weights:

$$\begin{aligned} \omega(\text{France}) \;=\; \omega(\text{Germany}) \;=\; \omega(\text{Italy}) \;&=\; 4 \\ \omega(\text{Belgium}) \;=\; \omega(\text{Netherlands}) \;&=\; 2 \\ \omega(\text{Luxembourg}) \;&=\; 1 \end{aligned}$$

Thus, $W \;=\; 4+4+4+2+2+1 \;=\; 17$. The quota was set at $Q = 12$. Thus, for example, $Y\!es$ would be chosen if it had the support of France, Germany, and Italy because

$$\Big(\mathcal{Y}(\rho) \;=\; \{\text{France}, \text{Germany}, \text{Italy}\}\Big) \Longrightarrow \Big(\Upsilon(\rho) \;=\; 4+4+4 \;=\; 12 \;\geq\; Q\Big) \Longrightarrow \Big(Y\!es \;\overset{\rho}{\sqsupset}\; N\!o\Big).$$

However, $Y\!es$ would not be chosen if it only had the support of France, Germany, Belgium, and Luxembourg, because

$$\Big(\mathcal{Y}(\rho) \;=\; \{\text{France}, \text{Germany}, \text{Belgium}, \text{Luxembourg}\}\Big) \Longrightarrow$$
$$\Big(\Upsilon(\rho) \;=\; 4+4+2+1 \;=\; 11 \;<\; Q\Big) \Longrightarrow \Big(Y\!es \;\overset{\rho}{\sqsubset}\; N\!o\Big). \qquad \diamondsuit$$

Example 1B.2: United Nations Security Council

According to the Charter of the United Nations, the UN Security Council consists of five *permanent members*

$$\mathcal{V}^* \quad = \quad \{\text{U.S.A.}, \text{U.K.}, \text{France}, \text{Russia}, \text{China}\}$$

along with ten *nonpermanent members* (positions which rotate amongst all member nations). Thus, the set $\mathcal{V}$ has fifteen members in total. Approval of a resolution requires two conditions:

(a) The support of at least 9 out of 15 Security Council members.

(b) The support of *all five* permanent members (any permanent member has a veto).

The 'veto' clause in condition **(b)** suggests that the Security Council is *not* a weighted voting system, but it actually is. First, note that conditions **(a)** and **(b)** can be combined as follows:

(c) A resolution is approved if and only if it has the support of *all five* permanent members, and *at least four* nonpermanent members.

We assign *seven* votes to every permanent member, and *one* vote to every nonpermanent member, and set the quota at 39. That is:

$$\begin{aligned} \omega(v) &= 7, \qquad \text{for all } v \in \mathcal{V}^* \\ \omega(v) &= 1, \qquad \text{for all } v \in \mathcal{V} \setminus \mathcal{V}^*. \\ Q &= 39. \end{aligned}$$

Now let $\mathcal{Y}(\rho)$ be the set of voters supporting proposal Yes. Suppose $\mathcal{Y}(\rho) \not\supset \mathcal{V}^*$ (ie. at least one permanent member is 'vetoing'). Even if *every* other nation supports the resolution, we still have

$$\Upsilon(\rho) \leq 4 \times 7 + 10 \times 1 = 28 + 10 = 38 < 39 = Q.$$

Hence the resolution is not approved.

Now suppose $\mathcal{Y}(\rho) \subset \mathcal{V}^*$. Then

$$\Upsilon(\rho) = 5 \times 7 + N \times 1 = 35 + N,$$

where N is the number of nonpermanent members supporting the resolution. Thus,

$$\Big(\Upsilon(\rho) \geq Q\Big) \iff \Big(35 + N \geq 39\Big) \iff \Big(N \geq 4\Big)$$

Thus, once all five permanent members support the resolution, it will be approved if and only if it also has the support of at least four nonpermanent members, exactly as required by **(c)**. ♢

Example 1B.3: Factionalism: Block voting and party discipline

Sometimes even supposedly 'nonweighted' voting systems can behave like weighted systems, because the voters organize themselves into factions, which synchronize their votes on particular issues. In this case, it is no longer correct to model the electorate as a large population of individual voters with equal weight; instead, we must model the electorate as a small number of competing factions, whose votes are 'weighted' in proportion to the size of their membership. Two examples of this phenomenon are *block voting* and *party discipline.*

Block Voting: Ideological groups (eg. labour unions, religious organizations, etc.) with a highly dedicated membership often dictate to their members how to vote in particular issues. Assuming that the group members are mostly obedient to the voting instructions of their leadership, the entire group can be treated as a single voting block.

Party Discipline: Modern elections involve hugely expensive advertising campaigns, and it is difficult to be elected without access to a powerful campaign finance machine. Thus, an individual politician must affiliate herself to some political *party*, which she depends upon to bankroll her campaigns. Political parties thus effectively 'own' their member politicians, and can dictate how their members vote on particular issues. A politician who defies her party might be denied access to crucial campaign funding. In a parliamentary system, the currently

governing party can also reward 'loyalty' through prestigious cabinet positions and patronage appointments. These mechanisms guarantee that the party can generally be treated as a unified voting block. ♢

Theorem 1B.4 *Let* Π *be a weighted voting system.*

(a) Π *always satisfies the* **monotonicity** *axiom* **(M)** *and* **Pareto** *axiom* **(P)**.

(b) Π *satisfies* **anonymity** *axiom* **(A)** *if and only if* ω *is a constant (ie. all voters have the same weight).*

(c) Π *satisfies* **neutrality** *axiom* **(N)** *if and only if* $Q = \frac{1}{2}W$.

(d) Π *is a* **dictatorship** *if and only if there is some* $v \in \mathcal{V}$ *whose vote 'outweighs' everyone else combined; ie.* $\rho(v) \ > \ \sum_{w \neq v} \omega(w)$.

Proof: **Exercise 1.3** □

In any binary voting system (weighted or not), a **winning coalition** is a collection of voters $\mathcal{W} \subset \mathcal{V}$ so that

$$\Big(\mathcal{Y}(\rho) \ = \ \mathcal{W}\Big) \Longrightarrow \Big(\text{Yes} \ \stackrel{\rho}{\sqsupset} \ \text{No}\Big).$$

Thus, for example, in a weighted voting system, $\mathcal{W}$ is a winning coalition iff $\sum_{w \in \mathcal{W}} \omega(w) \ \geq \ Q$.

A voting system is called **trade robust** if the following is true: If $\mathcal{W}_1$ and $\mathcal{W}_2$ are two winning coalitions, and they 'swap' some of their members, then at least one of the new coalitions will still be winning. In other words, given any subsets $\mathcal{U}_1 \subset \mathcal{W}_1$ and $\mathcal{U}_2 \subset \mathcal{W}_2$ (where $\mathcal{U}_1$ is disjoint from $\mathcal{W}_2$, while $\mathcal{U}_2$ is disjoint from $\mathcal{W}_1$) if we define

$$\mathcal{W}_1' \quad = \quad \mathcal{U}_2 \sqcup \mathcal{W}_1 \setminus \mathcal{U}_2 \quad \text{and} \quad \mathcal{W}_2' \quad = \quad \mathcal{U}_2 \sqcup \mathcal{W}_2 \setminus \mathcal{U}_1$$

...then at least one of $\mathcal{W}_1'$ or $\mathcal{W}_2'$ is *also* a winning coalition.

Theorem 1B.5 (Taylor & Zwicker, 1992) [TZ92]
Let Π *be a binary voting system. Then* $\Big(\Pi \text{ is a weighted system}\Big) \iff \Big(\Pi \text{ is trade robust}\Big)$.

Proof: '$\Longrightarrow$' If $\mathcal{W}_1$ and $\mathcal{W}_2$ are winning coalitions, then $\sum_{w \in \mathcal{W}_1} \omega(w) \ \geq \ Q$ and $\sum_{w \in \mathcal{W}_2} \omega(w) \ \geq \ Q$.

Let $\mathcal{U}_1 \subset \mathcal{W}_1$ and $\mathcal{U}_2 \subset \mathcal{W}_2$ be arbitrary subsets. Observe that

$$\sum_{w \in \mathcal{W}_1'} \omega(w) \ = \ \sum_{u \in \mathcal{U}_2} \omega(u) \ + \ \sum_{w \in \mathcal{W}_1} \omega(w) \ - \ \sum_{u \in \mathcal{U}_1} \omega(u),$$

$$\text{and} \quad \sum_{w \in \mathcal{W}_2'} \omega(w) \ = \ \sum_{u \in \mathcal{U}_1} \omega(u) \ + \ \sum_{w \in \mathcal{W}_2} \omega(w) \ - \ \sum_{u \in \mathcal{U}_2} \omega(u)$$

Suppose that $\sum_{u\in\mathcal{U}_1}\omega(u) \geq \sum_{u\in\mathcal{U}_2}\omega(u)$. Then $\sum_{u\in\mathcal{U}_1}\omega(u) - \sum_{u\in\mathcal{U}_2}\omega(u) \geq 0$. Thus,

$$\sum_{w\in\mathcal{W}_2'}\omega(w) \quad = \quad \left(\sum_{u\in\mathcal{U}_1}\omega(u) - \sum_{u\in\mathcal{U}_2}\omega(u)\right) + \sum_{w\in\mathcal{W}_2}\omega(w) \quad \geq \quad \sum_{w\in\mathcal{W}_2}\omega(w) \quad \geq \quad Q.$$

so $\mathcal{W}_2'$ is still a winning coalition. On the other hand, if $\sum_{u\in\mathcal{U}_2}\omega(u) \geq \sum_{u\in\mathcal{U}_1}\omega(u)$, then symmetric reasoning shows that $\mathcal{W}_1'$ is still a winning coalition.

'$\Longleftarrow$' See Taylor and Zwicker [TZ92]. □

Not all binary voting procedures are weighted systems, because not all systems are trade robust.

Example 1B.6: Amendment Formula for Canadian Constitution

To approve an amendment to the Canadian Constitution, the amendment must have the support of at least seven out of ten provinces, which together must represent at least 50% of the Canadian population. For the sake of argument, say the populations are as follows:

Ontario: 30%

Quebec: 30%

B.C.: 10%

Alberta, Manitoba & Saskatchewan: 15%

New Brunswick & Nova Scotia: 10%

P.E.I. & Newfoundland: 5%

Now consider the following coalitions:

$\mathcal{W}_1$ = {Ontario, B.C., Alberta, Manitoba, Saskatchewan, P.E.I., Newfoundland }
7 members, **Total weight:** $30 + 10 + 15 + 5 = 60\%$

$\mathcal{W}_2$ = {Quebec, B.C., Alberta, Manitoba, Saskatchewan, New Brunswick, Nova Scotia}
7 members, **Total weight:** $30 + 10 + 15 + 10 = 65\%$.

Now, let $\mathcal{U}_1$ = {P.E.I., Newfoundland} and $\mathcal{U}_2$ = {Quebec}, so that

$\mathcal{W}_1'$ = {Ontario, Quebec, B.C., Alberta, Manitoba, Saskatchewan}
6 members, **Total weight:** $30 + 30 + 10 + 15 = 85\%$.

$\mathcal{W}_2'$ = { B.C., Alberta, Manitoba, Saskatchewan, New Brunswick, Nova Scotia, P.E.I., Newfoundland}
8 members, **Total weight:** $10 + 15 + 10 + 5 = 40\%$

Now, $\mathcal{W}_1'$ is losing because it only has six members, while $\mathcal{W}_2'$ is losing because it only comprises 40% of the population.

Thus, the Canadian constitutional amendment formula is *not* trade robust, so Theorem 1B.5 says that it is *not* a weighted voting system. ♢

1C Vector-Weighted Voting Systems

Prerequisites: §1B

A **vector weighted** voting system is a strict binary voting procedure where the votes of different voters have vector-valued 'weights'. To be precise, there is a **vector weight function** $\boldsymbol{\omega} : \mathcal{V} \longrightarrow \mathbb{N}^D$ (for some $D \geq 2$) so that, for any $\rho \in \mathfrak{R}(\mathcal{V}, \mathcal{A})$, the *total support* for alternative $Y_{\mathscr{A}}$ is the vector $\boldsymbol{\Upsilon}(\rho) \in \mathbb{N}^D$ defined:

$$\boldsymbol{\Upsilon}(\rho) \quad = \quad \sum_{y \in \mathcal{Y}(\rho)} \boldsymbol{\omega}(y) \qquad \text{where} \quad \mathcal{Y}(\rho) \; = \; \left\{ v \in \mathcal{V} \; ; \; Y_{\mathscr{A}} \overset{\rho}{\underset{v}{\succ}} N_{\mathscr{O}} \right\}$$

The **total weight** of the system is the vector $\mathbf{W} = \sum_{v \in \mathcal{V}} \boldsymbol{\omega}(v)$. Finally, we set a vector-valued **quota** $\mathbf{Q} \in \mathbb{N}^D$ so that

$$\left(Y_{\mathscr{A}} \; \overset{\rho}{\sqsupseteq} \; N_{\mathscr{O}} \right) \iff \left(\boldsymbol{\Upsilon}(\rho) \geq \mathbf{Q} \right), \tag{1.1}$$

where "$\boldsymbol{\Upsilon} \geq \mathbf{Q}$" means $\Upsilon_1 \geq Q_1$, $\Upsilon_2 \geq Q_2$, ... , $\Upsilon_D \geq Q_D$.

Most systems favour the 'status quo' alternative ($N_{\mathscr{O}}$), which means that $\mathbf{Q} \geq \frac{1}{2}\mathbf{W}$.

The **dimension** of a vector-weighted voting system Π is the smallest D so that we can represent Π using D-dimensional weight vectors so that eqn.(1.1) is satisfied. Clearly, Π has dimension 1 if and only if Π is a 'scalar' weighted voting system, as described in §1B.

Example 1C.1: Canadian Constitutional Amendment Formula

The Canadian Constitutional Amendment Formula (Example 1B.6) is a vector-weighted voting system of dimension 2. To see this, let $\mathcal{V}$ be the ten provinces of Canada. For any $v \in \mathcal{V}$, let P_v be the population of province v. Define $\boldsymbol{\omega} : \mathcal{V} \longrightarrow \mathbb{N}^2$ so that, for any $v \in \mathcal{V}$, $\boldsymbol{\omega}(v) = (1, P_v)$. Now let $\mathbf{Q} = (7, H)$, where H is half the population of Canada. Thus, for any coalition $\mathcal{U} \subset \mathcal{V}$,

$$\sum_{u \in \mathcal{U}} \boldsymbol{\omega}(u) \quad = \quad \left(\#(\mathcal{U}), \; \sum_{u \in \mathcal{U}} P_u \right)$$

Hence, $\left(\sum_{u \in \mathcal{U}} \boldsymbol{\omega}(u) \; \geq \; \mathbf{Q} \right) \iff \left(\#(\mathcal{U}) \geq 7, \text{ and } \sum_{u \in \mathcal{U}} P_u \; \geq \; H \right)$.

In other words, $\mathcal{U}$ is a winning coalition if and only if $\mathcal{U}$ consists of at least seven provinces, together comprising at least half of Canada's population.

Thus, we can represent the Constitutional Amendment Formula using 2-dimensional vectors. We already know that we can't use 1-dimensional vectors (because Example 1B.6 shows that the Constitutional Amendment Formula is *not* a 'scalar' weighted voting system). Thus, the Amendment Formula has dimension 2. ♢

Theorem 1C.2 *Every monotone binary voting procedure is a vector-weighted procedure.*

Proof: A **losing coalition** is a collection of voters $\mathcal{L} \subset \mathcal{V}$ so that

$$\Big(\mathcal{Y}(\rho) \;=\; \mathcal{L}\Big) \Longrightarrow \Big(\mathit{Yes} \;\overset{\rho}{\sqsubset}\; \mathit{No}\Big).$$

Thus, for example, in a weighted voting system, $\mathcal{L}$ is a losing coalition iff $\sum_{\ell \in \mathcal{L}} \omega(\ell) \;<\; Q$.

Let $\mathfrak{L} = \{\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_D\}$ be the set of all losing coalitions. (We know $\mathfrak{L}$ is finite because $\mathcal{V}$ is finite, and a finite set only has finitely many subsets). For each $\mathcal{L}_d \in \mathfrak{L}$, we define a weight function $\omega_d : \mathcal{V} \longrightarrow \mathbb{N}$ by

$$\omega_d(v) \quad = \quad \begin{cases} 1 & \text{if} \quad v \notin \mathcal{L}_d \\ 0 & \text{if} \quad v \in \mathcal{L}_d \end{cases}$$

Claim 1: *If $\mathcal{U} \subset \mathcal{V}$ is some coalition, then*

$$\Big(\mathcal{U} \textit{ is a losing coalition}\Big) \quad \Longleftrightarrow \quad \left(\sum_{u \in \mathcal{U}} \omega_d(u) \;=\; 0 \textit{ for some } \mathcal{L}_d \in \mathfrak{L}\right)$$

Proof: '$\Longrightarrow$' Clearly, if $\mathcal{U}$ is losing, then $\mathcal{U} \in \mathfrak{L}$. Suppose $\mathcal{U} = \mathcal{L}_d$; then $\sum_{u \in \mathcal{U}} \omega_d(u) \;=\; 0$.

'$\Longleftarrow$' Suppose $\sum_{u \in \mathcal{U}} \omega_d(u) \;=\; 0$, for some $\mathcal{L}_d \in \mathfrak{L}$. Then $\mathcal{U} \subset \mathcal{L}_d$. Thus, $\mathcal{U}$ is also a losing coalition, because Π is monotone, by hypothesis. ♢ Claim 1

Now define $\boldsymbol{\omega} : \mathcal{V} \longrightarrow \mathbb{N}^D$ by

$$\boldsymbol{\omega}(v) \;=\; (\omega_1(v), \omega_2(v), \ldots, \omega_D(v)).$$

Now let $\mathbf{Q} = (1, 1, \ldots, 1)$. Then for any $\mathcal{U} \subset \mathcal{V}$,

$$\begin{aligned} \Big(\mathcal{U} \text{ is a losing coalition}\Big) \quad &\Longleftrightarrow \quad \left(\sum_{u \in \mathcal{U}} \omega_{\mathcal{L}_d}(u) \;=\; 0 \text{ for some } \mathcal{L}_d \in \mathfrak{L}\right) \\ &\Longleftrightarrow \quad \left(\sum_{u \in \mathcal{U}} \omega_d(u) \;=\; 0 \text{ for some } d \in [1..D]\right) \\ &\Longleftrightarrow \quad \left(\sum_{u \in \mathcal{U}} \boldsymbol{\omega}(u) \;\not\geq\; \mathbf{Q}\right). \end{aligned}$$

Thus, $\Big(\mathcal{U} \text{ is a winning coalition}\Big) \;\Longleftrightarrow\; \left(\sum_{u \in \mathcal{U}} \boldsymbol{\omega}(u) \;\geq\; \mathbf{Q}\right)$, as desired. □

Further reading: Much of the material in this section was drawn from Chapters 3 and 8 of Taylor [Tay95], which, in turn, are based mainly on papers by Taylor and Zwicker [TZ92, TZ93]. The Canadian constitutional amendment formula was first studied by Kilgour [Kil83]; see also [Str93].

1D Voting Power Indices

Democracy is a process by which the people are free to choose the man who will get the blame.
—Laurence J. Peter

Prerequisites: §1B

The weighted and vector-weighted voting schemes of §1B show how different voters can wield different amounts of 'power' in a voting system. Naively, we would expect the 'power' of a particular voter to be proportional to the 'weight' of her vote, but this is not true. For example, consider a weighted voting system with four voters, Ulrich, Veronique, Waldemar, and Xavier, with the following weights:

$$\omega(u) \;=\; \omega(v) \;=\; \omega(w) \;=\; 3; \qquad \omega(x) \;=\; 2.$$

Thus, $W \;=\; 3+3+3+2 = 11$. Let $Q = 6$; hence the alternative *Yes* will be chosen if and only if *Yes* recieves 6 or more votes. Naively, we would say that Ulrich, Veronique, and Waldemar each have 3/11 of the total power, while Xavier has 2/11. But in fact, Xavier has *no* power, because his vote is irrelevant. In a binary vote of *Yes* vs. *No*, the alternative *Yes* will be chosen if and only if *Yes* recieves the support of at least two out of three members of the set {Ulrich, Veronique, Waldemar}, *regardless* of Xavier's vote. To see this, consider the following table of possible profiles and outcomes:

Individual Votes				Total Score		Outcome
u	v	w	x	*Yes*	*No*	
No	*No*	*No*	*No*	0	11	*No* wins
No	*No*	*No*	*Yes*	2	9	*No* wins
Yes	*No*	*No*	*No*	3	8	*No* wins
Yes	*No*	*No*	*Yes*	5	6	*No* wins
Yes	*Yes*	*No*	*No*	6	5	*Yes* wins
Yes	*Yes*	*No*	*Yes*	8	3	*Yes* wins
Yes	*Yes*	*Yes*	*No*	9	2	*Yes* wins
Yes	*Yes*	*Yes*	*Yes*	11	0	*Yes* wins

(Ulrich, Veronique, and Waldemar have identical weights, so the same outcomes will occur if we permute the 'u', 'v', and 'w' columns of this table.)

This example tells us that 'voting weight' is not the correct measure of 'voting power'. Instead, 'voting power' should answer the question

(P) *How often does Xavier's vote actually make a difference to the outcome?*

In other words, assuming all other voters have already decided their votes, how often will it be the case that *Yes* will be chosen over *No* if and only if Xavier votes for *Yes*? This is the motivation behind the definition of various *Voting Power Indices*. In the previous example, the answer to question **(P)** is 'never'; hence, by any measure, the "voting power" of Xavier is *zero*.

The Shapley-Shubik Index: On any particular issue, we can arrange all the voters in some linear order, from those most in favour of a particular policy to those most opposed to that policy. We can imagine arranging the voters in an *ideological spectrum* from left to right[1], eg:

$$a \quad b \quad c \quad \ldots\ldots \quad m \,|\, n \quad o \quad \ldots\ldots \quad x \quad y \quad z$$

Those at the right end of the spectrum (eg x, y, z) strongly support the policy; those at the left end (eg. a, b, c) are strongly opposed, and those in the middle (eg. m, n, o) are more or less neutral, perhaps with some slight bias one way or the other.

We can then draw a line so thate the voters *right* of this line vote 'yes' and the voters *left* of this line vote 'no'. In this case, the line is between m and n. If the voters to the *right* of the line form a winning coalition, then the policy will be chosen. Otherwise it will be rejected.

We say that voter n is a **critical** if the set $\{n, o, p, \ldots, x, y, z\}$ *is* a winning coalition, but the set $\{o, p, \ldots, x, y, z\}$ is *not* a winning coalition. In other words, Yes will be chosen if and only if n votes for Yes.

For example, in a *weighted* voting system, the policy will be chosen if and only if $\omega(n) + \omega(o) + \cdots + \omega(x) + \omega(y) + \omega(z) \;\geq\; Q$, where Q is the quota and $\omega(m)$ is the weight of voter m, etc. Thus, the voter n is critical if and only if

$$\omega(o)+\omega(p)+\cdots+\omega(x)+\omega(y)+\omega(z) \quad < \quad Q \quad \leq \quad \omega(n)+\omega(o)+\omega(p)+\cdots+\omega(x)+\omega(y)+\omega(z)$$

which is equivalent to saying

$$0 \quad < \quad Q \;-\; \Big(\omega(o) + \omega(p) + \cdots + \omega(x) + \omega(y) + \omega(z)\Big) \quad \leq \quad \omega(n).$$

Clearly, for a particular ordering of voters into an ideological spectrum, there will be exactly *one* critical voter. The Shapley-Shubik index of voter n is a crude estimate of the probability that voter n will be this unique critical voter. Formally, define

$$\mathcal{O} \;:=\; \{\text{All orderings of the voters}\} \quad \text{and} \quad \mathcal{O}_n \;:=\; \{\text{All orderings where } n \text{ is critical}\}.$$

$$\text{Then we define} \quad \mathtt{SSI}(n) \quad := \quad \frac{\#\mathcal{O}_n}{\#\mathcal{O}} \quad = \quad \frac{\#\mathcal{O}_n}{N!}, \quad \text{where } N \text{ is the number of voters.}$$

Note that this definition assumes that *all orderings of the voters are equally likely.* In the present abstract context, this *a priori* assumption is clearly no worse than any other. However, in a specific concrete setting, it may be clear that some orderings are much more likely than others. (For example, on economic issues, we can expect that the voters will always arrange themselves in roughly the same order, forming a spectrum from socialism to laissez-faire capitalism).

Observe that the Shapley-Shubik indices of all players add up to 1:

$$\mathtt{SSI}(a) + \mathtt{SSI}(b) + \cdots + \mathtt{SSI}(z) \quad = \quad 1. \tag{1.2}$$

This follows from the fact that $\mathcal{O}_a \sqcup \mathcal{O}_b \sqcup \cdots \sqcup \mathcal{O}_z \;=\; \mathcal{O}$.

[1] 'Left' and 'right' here do not necessarily correspond to the vague political categories of 'left-wing' vs. 'right-wing', although of course they might.

Example 1D.1: United Nations Security Council

Recall that the U.N. Security Council has five 'permanent' members and ten 'nonpermanent' members. A resolution passes if and only if it is supported by *all five* permanent members, and, in addition, by *at least four* nonpermanent members.

Let $\mathcal{P} = \{p_1, p_2, p_3, p_4, p_5\}$ be the set of permanent members, and let $\mathcal{N} = \{n_1, \ldots, n_{10}\}$ be the set of nonpermanent members, so that $\mathcal{C} = \mathcal{P} \sqcup \mathcal{N}$ is the entire Council.

Let's compute the Shapley-Shubik Index of a 'permanent' member, say, p_1. First, note that, if $\mathcal{W} \subset \mathcal{C}$ is *any* winning coalition, then p_1 is critical for $\mathcal{W}$, because any permanent member has a veto. Thus, to count the orderings where p_1 is critical, we must simply count all orderings of all possible winning coalitions, such that p_1 is the 'borderline' member in the ordering.

Now, $\mathcal{W}$ is a winning coalition only if $\mathcal{P} \subset \mathcal{W}$ and $\mathcal{W}$ also contains at least four nonpermanent members. Hence if $\widetilde{\mathcal{W}} = \mathcal{W} \setminus \mathcal{P}$, then $\widetilde{\mathcal{W}}$ must have $m \geq 4$ elements.

Suppose $\widetilde{\mathcal{W}}$ is chosen, and let $\widehat{\mathcal{W}} = \mathcal{W} \setminus \{p_1\}$; then $\#(\widehat{\mathcal{W}}) = 4 + m$, so there are $(4+m)!$ ways that $\widehat{\mathcal{W}}$ can be ordered. If $\mathcal{V} = \mathcal{C} \setminus \mathcal{W}$ is the set of remaining Council members (those 'opposed'), then $\#(\mathcal{V}) = 10 - m$, so there are $(10-m)!$ ways to order $\mathcal{V}$. Thus, having fixed $\widetilde{\mathcal{W}}$, there are a total of $(10-m)! \cdot (4+m)!$ orderings of $\mathcal{C}$ where p_1 is the 'boundary' element between $\widehat{\mathcal{W}}$ and $\mathcal{V}$.

There are $\left(\begin{array}{c} 10 \\ m \end{array} \right)$ subsets of $\mathcal{N}$ which contain m elements. Hence, there are $\left(\begin{array}{c} 10 \\ m \end{array} \right) = \dfrac{10!}{m!(10-m)!}$ choices for $\widetilde{\mathcal{W}}$. This gives a total of

$$\frac{10!}{m!(10-m)!} \cdot (10-m)! \cdot (4+m)! \quad = \quad \frac{10!(4+m)!}{m!}$$

possible orderings involving m nonpermanent members where p_1 is critical. Thus, the total number of orderings where p_1 is critical is given:

$$\#(\mathcal{O}_{p_1}) \quad = \quad \sum_{m=4}^{10} \frac{10!(4+m)!}{m!}$$

Now, $\#(\mathcal{C}) = 15$, so there are 15! orderings in total. Thus, the Shapley-Shubik Index for p_1 is given:

$$\texttt{SSI}(p_1) \quad = \quad \frac{10!}{15!} \sum_{m=4}^{10} \frac{(4+m)!}{m!} \quad = \quad \left(\frac{70728}{360360} \right) \quad = \quad \left(\frac{421}{2145} \right) \quad \approx \quad 0.19627\ldots$$

which is roughly 19.6%.

To compute the Shapley-Shubik index of a *nonpermanent* member, we could go through a similar combinatorial argument. However, there is a simpler way, using eqn.(1.2). First, note

that, by symmetry, all permanent members have the same SSI. Let P denotes the SSI of a permanent member; then we've just shown that $P = \frac{421}{2145}$. Likewise, by symmetry, all nonpermanent members have the same SSI, say N. Then eqn.(1.2) implies that

$$5 \cdot P + 10 \cdot N = 1$$

Substituting $P = \frac{421}{2145}$ and simplifying, we have:

$$\begin{aligned} N &= \frac{1}{10}(1 - 5P) = \frac{1}{10}\left(1 - 5 \cdot \frac{421}{2145}\right) = \frac{1}{10}\left(1 - \frac{421}{429}\right) = \frac{1}{10}\left(\frac{8}{429}\right) \\ &= \frac{8}{4290} = \frac{4}{2145}. \approx 0.001865\ldots \end{aligned}$$

Hence, each of the five permanent members has about 19.6% of the total voting power, while each of the ten nonpermanent members has about 0.187% of the total voting power. ♢

The Shapley-Shubik index is only one of many 'voting power indices' which have been developed; others include the Banzhaf index, the Johnston index, and the Deegan-Packel index. All the indices measure the power of a voter v by counting the scenarios where v is 'critical'; the indices differ in how they define and enumerate these scenarios. For example, the Shapley-Shubik index defines 'scenarios' as linear orderings of the voters; hence counting critical scenarios means counting orderings. The Banzhaf index, on the other hand, simply counts the total number of *unordered* winning coalitions where v is a critical member; hence, one 'critical' coalition in the Banzhaf model corresponds to many 'critical' coalitions in the Shapley-Shubik model (because it could have many orderings). This yields a different formula, with a different numerical value.

In some cases, the various voting power indices roughly agree, while in others, they wildly disagree, suggesting that at least one of them must be wrong. Like Shapley-Shubik, the other indices are based on some *a priori* estimate of the probability of various voting scenarios (eg. Shapley-Shubik assumes that all $N!$ orderings are equally probable). To the extent that these *a priori* estimates are unrealistic (because some orderings are much more ideologically plausible than others), *none* of the indices will be perfectly accurate. Nevertheless, they are valuable as a rough estimate of how (non)egalitarian a voting system is. For example, our analysis of the U.N. Security Council confirms the conventional wisdom that the ten nonpermanent members have virtually no power.

Further reading: The Shapley-Shubik index first appeared in [SS54]. Felsenthal and Machover have recently published a highly respected book on voting power indices [FM98]. A recent paper by Laruelle and Valenciano introduces probabilistic definitions of a voter's 'success' and 'decisiveness', and shows how how many power indices are special cases of this definition [LV05]. For an elementary introduction to the various power indices, see Chapters 4 and 9 of Taylor [Tay95].

Chapter 2

Multi-option Voting Systems

Democracy is a device that insures we shall be governed no better than we deserve.

—George Bernard Shaw

Strictly speaking, democracy only insures that the *majority* will be governed no better than they deserve; the rest of us will also be governed no better than they deserve. Even this is only true when an absolute majority has chosen the government or policies in question; we shall see that this is rarely the case when there are three or more alternatives to choose from.

2A Plurality voting & Borda's Paradox

Most people would agree that in a democracy, the laws are decided by the Will of the People. The 'Will of the People' is usually determined through an election or referendum. However, all of us have seen how the electoral process can fail to adequately measure the Will of the People. A classic example is 'vote splitting'. For instance, suppose that in the hypothetical country of Danaca, the political breakdown of the population is roughly as follows:

Left	15%
Centre	34%
Right	51%

One would expect that Danacians would normally elect a right-wing government. However, suppose that more than three political parties compete in the election. For simplicity, we will assume that each political party falls into neatly one of the three categories 'Left', 'Centre', and 'Right'

Left	*Ultracommunists*	1%	
	New Demagogues	14%	
	Total:		15%
Centre	***Literal Party***		**34%**
Right	*Regressive Coercitives*	26%	
	Behaviour & Compliance	25%	
	Total:		51%

Although they do not have an absolute majority, the the centrist *Literal* party has a ***plurality***; that is they receive the largest proportion of popular support amonh all the parties. Thus, in many electoral systems, the *Literals* will win the election, despite the fact that an absolute majority (51%) of Danacians would prefer a right-wing government. The *Literal* party does not have an absolute majority; it only has a *plurality*, ie. the largest (minority) vote of any party. The standard multiparty electoral system is called the *Plurality System*, and the Plurality System has clearly somehow failed in this scenario.

The problem here is that the two right-wing parties have 'split' the right-wing vote between them. One possible solution is to 'unite the right': the *Regressive Coercitives* and the *Behaviour & Compliance* party could unite to form a single *Coercitive* party, which would then win with a 51% majority:

Left	*Ultracommunists*	1%	
	New Demagogues	14%	
	Total:		15%
Centre	*Literal Party*		34%
Right	***Coercitive***		**51%**

If it is not possible to 'unite the right', another option is to split the centrist support of the *Literals*. For example, the *Regressive Coercitives* could covertly support the emergence of 'fringe' centrist parties, fracturing the *Literal* support:

Left	*Ultracommunists*	1%	
	New Demagogues	14%	
	Total:		15%
Centre	*Qubekistan Liberation Front*	2%	
	Earth First! Party	2%	
	Popular Front for the Liberation of Qubekistan	2%	
	Ganja Legalization Party	3%	
	Literal Party	24%	
	People's Democratic Front for Qubekistan Liberation	1%	
	Total		34%
Right	***Regressive Coercitives***	**26%**	
	Behaviour & Compliance	25%	
	Total:		51%

Now the *Regressive Coercitives* will win (barely) with a plurality of 26%. In both cases, the election outcome changed, not because of the 'Will of the People', but because of clever manipulation of the electoral process. This is a simple example of *election manipulation*.

Next, consider a hypothetical gubernatorial election in the state of Kolifönia. The candidates are Ahnold, Bustamante, Carey, and Davis. We will write, for example, $A \succ B$ to mean that a voter prefers Ahnold to Bustamante. The voter's preferences are as follows:

15%	A	$\succ$	D	$\succ$	B	$\succ$	C
15%	A	$\succ$	C	$\succ$	D	$\succ$	B
15%	B	$\succ$	C	$\succ$	D	$\succ$	A
10%	B	$\succ$	D	$\succ$	C	$\succ$	A
25%	C	$\succ$	D	$\succ$	B	$\succ$	A
20%	D	$\succ$	B	$\succ$	C	$\succ$	A

Assuming people vote for their first-place choice, we get the following results:

Ahnold	**30%**
Bustamante	25%
Carey	25%
Davis	20%

Hence, Ahnold wins the plurality vote, and becomes the new governor of Kolifönia, despite the fact that fully 70% of Kolifönians despise him and ranked him *last* of the four possible candidates. The incumbent, Davis, is defeated by an overwhelming vote of nonconfidence, with the smallest support of any candidate (despite the fact that 70% of Kolifönians prefer him to Ahnold).

2B Other voting schemes

Because of pathologies like Borda's Paradox, many people have proposed many replacements for the standard "plurality" voting scheme.

2B.1 Pairwise elections

In the two examples from §2A, the pathology seems to arise from the fact that, with more than two candidates, there is often no candidate who obtains an *absolute majority* (ie. greater than 50% of the vote) so we must instead choose the candidate who achieves a *plurality* (the largest *minority* share, eg. 30%). The obvious solution is to only allow two-candidate elections. With more than two candidates, however, we need more than one such election. For example. we might have the following agenda of elections:

1. First, Ahnold competes with Bustamante.
2. Next, the winner of this election takes on Carey.
3. Finally, the winner of *this* election takes on Davis.

For convenience, we reprint the profile of Kolifönia voter preferences from §**??**, with extra columns showing how each voter group votes in each pairwise election.

%	Preference	A v B	B v C	B v D	D v C	C v A
15%	$A \succ D \succ B \succ C$	A	B	D	D	A
15%	$A \succ C \succ D \succ B$	A	C	D	C	A
15%	$B \succ C \succ D \succ A$	B	B	B	C	C
10%	$B \succ D \succ C \succ A$	B	B	B	D	C
25%	$C \succ D \succ B \succ A$	B	C	D	C	C
20%	$D \succ B \succ C \succ A$	B	B	D	D	C
	Final tallies:	30/70	60/40	25/75	45/55	70/30

By inspecting the columns ' A v B', 'B v C' and 'B v D' of this table, it is clear the election outcomes will be as follows:

$$\boxed{\begin{array}{ll} \text{A} & 30\% \\ \textbf{B} & \textbf{70\%} \end{array}} \xrightarrow{B} \boxed{\begin{array}{ll} \textbf{B} & \textbf{60\%} \\ \text{C} & 40\% \end{array}} \xrightarrow{B} \boxed{\begin{array}{ll} \text{B} & 25\% \\ \textbf{D} & \textbf{75\%} \end{array}} \xrightarrow{D} \textbf{D}$$

At stage 1, Bustamante defeats Ahnold by a landslide. At stage 2, Bustamante easily defeats Carey. But at stage 3, Bustamante *loses* against Davis. Thus, Davis 'wins' the election, despite the fact that Davis had the *smallest support of any candidate* in the original plurality vote. The People have spoken.

Or have they? The problem here is the election *agenda* —ie. the order in which candidates are compared. If we use a different agenda, we get a different outcome. For example, suppose we use the following agenda:

1. First, Bustamante competes with Davis
2. Next, the winner of this election takes on Carey.
3. Finally, the winner of *this* election takes on Ahnold.

$$\boxed{\begin{array}{ll} \text{B} & 25\% \\ \mathbf{D} & \mathbf{75\%} \end{array}} \xrightarrow{D} \boxed{\begin{array}{ll} \text{D} & 45\% \\ \mathbf{C} & \mathbf{55\%} \end{array}} \xrightarrow{C} \boxed{\begin{array}{ll} \mathbf{C} & \mathbf{70\%} \\ \text{A} & 30\% \end{array}} \xrightarrow{C} \mathbf{C}$$

Davis wins the first round against Bustamante, but is then defeated by Carey in the second round. Carey goes on to soundly defeat the much-reviled Ahnold, so now it is *Carey* who wins the election.

Clearly, an electoral scheme isn't very good if the decision varies depending upon the agenda of pairwise elections. The outcome is then not the result of the 'will of the People' but instead an artifact, a consequence of a technicality. This scheme is also vulnerable to manipulation. For example, the incumbent, Davis, can decide the agenda, and he will choose the *first* agenda, which will ensure his re-election.

2B.2 The Condorcet Scheme

The problem with pairwise elections is that even if candidate X beats candidate Y, she may lose to candidate Z. Depending on the order of the agenda, a different person may end up winning the last round. The Marquis de Condorcet's response was that someone can only claim legitimate victory if they can beat *every* other candidate. The *Condorcet scheme* works as follows:

- For each possible pair of candidates, determine who wins in an election between that pair.
- The Condorcet winner is the candidate who beats *every other candidate* in a pairwise match.

For example, suppose the profile of voter preferences was as follows:

%	Preference	A v B	A v C	A v D	B v C	B v D	C v D
30%	$A \succ B \succ C \succ D$	A	A	A	B	B	C
15%	$B \succ C \succ D \succ A$	B	C	D	B	B	C
10%	$B \succ D \succ C \succ A$	B	C	D	B	B	D
25%	$C \succ B \succ D \succ A$	B	C	D	C	B	C
20%	$D \succ B \succ C \succ A$	B	C	D	B	D	D
	Final tallies:	30/70	30/70	30/70	75/25	80/20	70/30

The Condorcet scheme would yield the following outcomes:

	vs. A		*vs. B*		*vs. C*		*vs. D*	
A			$30\backslash^{70}$	**B**	$30\backslash^{70}$	*C*	$30\backslash^{70}$	*D*
B	$70\backslash^{30}$	**B**			$75\backslash^{25}$	**B**	$80\backslash^{20}$	**B**
C	$70\backslash^{30}$	*C*	$25\backslash^{75}$	**B**			$70\backslash^{30}$	*C*
D	$70\backslash^{30}$	*D*	$20\backslash^{80}$	**B**	$30\backslash^{70}$	*C*		

Thus, B beats every other individual candidate in a pairwise race, so B is the Condorcet winner. It is easy to prove:

Theorem 2B.1 *Suppose X is the Condorcet winner. Then X will be the ultimate victor of a sequence of pairwise elections, no matter what the order of the agenda.*

Proof: **Exercise 2.1** □

The problem with this method is that there may not *be* a Condorcet winner, in general. Indeed, Theorem 1 immediately implies that there is no Condorcet winner in the Kolifönia election (because otherwise different agendas wouldn't have yielded different outcomes). For an even more extreme example, consider the *Condorcet Paradox:*

%	Preference					A v B	A v C	B v C
33%	*A*	$\succ$	*B*	$\succ$	*C*	*A*	*A*	*B*
33%	*B*	$\succ$	*C*	$\succ$	*A*	*B*	*C*	*B*
34%	*C*	$\succ$	*A*	$\succ$	*B*	*A*	*C*	*C*
					Final tallies:	67/33	33/67	66/34

This yields the following pairwise results:

	vs. A		*vs. B*		*vs. C*	
A			$67\backslash^{33}$	*A*	$33\backslash^{67}$	*C*
B	$33\backslash^{67}$	*A*			$66\backslash^{34}$	*B*
C	$67\backslash^{33}$	*C*	$34\backslash^{66}$	*B*		

Thus, although A beats B, he loses to C. Likewise, B beats C, but loses to A, and C beats A, but loses to B. Like a game of 'Scissors, Rock, Paper', there is no clear winner. This has the following 'paradoxical' consequence:

> No matter which alternative is chosen as leader, this leader will be opposed by a *majority* of voters. Furthermore, this opposing majority can always identify a *specific* alternative they prefer to the current leader.

This clearly has highly undesirable consequences for political stability. A 'Condorcet Paradox' society is a society where a majority of voters are always dissatisfied with the status quo, and constantly seek to replace the existing regime with a new one.

Exercise 2.2 (a) Show that, in the Condorcet paradox, a sequence of pairwise elections will always elect the *last* candidate named in the agenda. For example, if the agenda is: 'First A vs. B; next the winner takes on C', then the ultimate victor will be C.

(b) Generalize the Condorcet paradox to four or more candidates. Is the analog of part (a) true? Why or why not?

2B.3 Borda Count

Perhaps the problem with the traditional plurality vote, pairwise elections, and the Condorcet method is that they all attempt to reduce a complicated piece of information (the complete ordering of the voters' preferences) to a sequence of simplistic binary choices (eg. A vs. B). A more sophisticated method would try to take into account the complete *order structure* of a voter's preferences. One such method is the *Borda count*. Suppose we are choosing amongst N alternatives.

- Assign $(N-1)$ points to each voter's *first* choice, $(N-2)$ points to her *second* choice, and so on, assigning $(N-k)$ points to her kth choice, and 0 points to her *last* choice.
- Add up all the points for each alternative. The winner is the alternative with the highest score.

For example, in the *Condorcet paradox* example (page 22), we get the following scores:

%	Preferences					Points		
						A	B	C
33%	A	$\succ$	B	$\succ$	C	2	1	0
33%	B	$\succ$	C	$\succ$	A	0	2	1
34%	C	$\succ$	A	$\succ$	B	1	0	2
					Total score:	100	99	101

(2.1)

Thus, C is the winner (barely) with a total Borda count of 101 points.

The Borda count has three shortcomings:

Strategic voting, where voters 'lie' about their preferences to manipulate the outcome.

Failing the Condorcet criterion An alternative can lose in the Borda count, even though it is the Condorcet winner.

Sensitivity to irrelevant alternatives, where a Borda loser can become a Borda winner when extra (losing) alternatives are introduced to the election.

Strategic Voting

To see how vote manipulation can occur, consider the following profile of voter preferences in a competition between Arianne, Bryn, and Chloe:

%	Preferences					Points		
						A	B	C
45%	A	$\succ$	B	$\succ$	C	2	1	0
45%	B	$\succ$	A	$\succ$	C	1	2	0
10%	C	$\succ$	A	$\succ$	B	1	0	2
					Total score:	145	135	20

In the Borda count election, Arianne narrowly defeats Bryn, and both candidates obliterate the pitifully unpopular Chloe.

However, suppose that a pre-election survey compiles data on voter's preferences, and predicts this outcome. Armed with this knowledge, the 45% who support Bryn decide to manipulate the results. They reason as follows: "We hate Chloe, but there's clearly no danger of her winning. We prefer that Bryn win rather than Arianne, and we are inadvertently contributing to Arianne's victory by ranking her second (rather than third) in our preferences. So, to ensure that Bryn wins, let's *pretend* that we like Chloe more than Arianne, and rank Arianne third."

The new (dishonest) profile of voter preferences is as follows:

%	Preferences					Points		
						A	B	C
45%	A	$\succ$	B	$\succ$	C	2	1	0
45%	B	$\succ$	C	$\succ$	A	0	2	1
10%	C	$\succ$	A	$\succ$	B	1	0	2
					Total score:	100	135	65

If the people vote according to these preferences on election day, then Bryn emerges as the clear winner.

However, the 45% who support Arianne have spies in the Bryn campaign, and they discover this dastardly plot. They plan a counterattack: "We hate Chloe, but there's clearly no danger of her winning. We prefer that Arianne win rather than Bryn, and we are inadvertently contributing to Bryn's illegitimate victory by ranking her second (rather than third) in our preferences. So, to ensure that Arianne wins, let's *pretend* that we like Chloe more than Bryn, and rank Bryn third."

The new (doubly dishonest) profile of voter preferences is as follows:

%	Preferences					Points		
						A	B	C
45%	A	$\succ$	C	$\succ$	B	2	0	1
45%	B	$\succ$	C	$\succ$	A	0	2	1
10%	C	$\succ$	A	$\succ$	B	1	0	2
					Total score:	100	90	110

The machinations of Arianne and Bryn have cancelled out, and annihilated their political advantage; Chloe scrapes by and wins the race!

It should be pointed out that Borda is not the only method vulnerable to strategic voting. Indeed, the traditional Plurality vote is notoriously vulnerable. This is why, during elections, you often hear people talk about voting 'against' one candidate, instead of voting 'for' another candidate, and why you often get the advice, 'Don't waste your vote; vote for X'; which is essentially asking you to vote strategically.

Failing the Condorcet Criterion

Consider the following profile of voter preferences in a competition between Arianne, Bryn, and Chloe:

Arianne vs. Bryn vs. Chloe								
%	Preferences					Points		
						A	**B**	C
60%	A	$\succ$	B	$\succ$	C	2	**1**	0
40%	B	$\succ$	C	$\succ$	A	0	**2**	1
					Total score:	120	**140**	40

(2.2)

Clearly Bryn wins the Borda count, with a score of 140. However, observe that Arianne is the Condorcet winner: she beats both Bryn and Chloe in pairwise races. Thus, the Borda Count does *not* satisfy Condorcet's criterion.

This isn't necessarily a fatal flaw, but it will certainly cause political instability if the winner of the Borda count is a Condorcet loser: this guarantees that a *strict majority* of voters will react to the outcome by saying, 'Why did Bryn win? I preferred Arianne.'

Sensitivity to irrelevant alternatives

The Borda count also yields scenarios where the electoral outcome can change when a (losing) candidate is added or removed. Imagine that Arianne, Bryn, and Chloe are mathematicians shortlisted for the Fields Medal. A committee has been convened to compare the candidates and decide the winner. At the press conference, the committee chair stands up and begins, "We have decided to award the Fields Medal to Bryn..." At that moment, an aide bursts into

the room and announces that Chloe has withdrawn from the competition because a subtle but fatal error was found in her proof of the Beiberbach Conjecture. "Ah," says the committee chair. "In that case, the winner is Arianne." You can imagine that the assembled dignitaries would find this somewhat peculiar.

And yet, this is exactly the outcome of the following scenario. Suppose that at the beginning, the profile of voter preferences is as in table (2.2) from the previous section. Thus, even at the beginning, Chloe is not a serious contender; the race is basically between Arianne and Bryn, and Bryn is the overall winner, with a Borda count of 140.

Now, with the news of Chloe's withdrawal, she drops to bottom place in everyone's preferences, effectively out of the running. This yields the following profile:

Arianne vs. Bryn vs. Chloe								
%	Preferences					Points		
						A	*B*	*C*
60%	*A*	$\succ$	*B*	$\succ$	*C*	**2**	1	0
40%	*B*	$\succ$	*A*	$\succ$	*C*	**1**	2	0
					Total score:	**160**	140	0

(2.3)

Now Arianne wins the Borda count! What's going on?

Perhaps the introduction of Chloe has revealed information about the 'intensity' of support for Arianne and Bryn. Arianne's 60% like her only 'slightly' more than Bryn (which is why they rank Bryn over Chloe). However, Bryn's 40% like her a *lot* more than Arianne (so they even rank Chloe ahead of Arianne), as illustrated in the following figure:

$$\begin{array}{ll} & \text{Good} \longleftarrow - - - (\text{Cardinal utility}) - - \longrightarrow \text{Bad} \\ 60\% & A - B - C \\ 40\% & B - - - - - - C - - - - - - - - - - - - - - - - A \end{array} \tag{2.4}$$

Thus, Bryn's supporters prefer her to Arianne much more 'intensely' than Arianne's supporters prefer her to Bryn, and this tips the balance in Bryn's favour. However, it takes the presence of a third candidate (even a losing candidate) to reveal this 'intensity'. If we apply this reasoning to the 'Fields Medal' parable, then the last minute withdrawal of Chloe should *not* change the outcome, because her presence in the competition has already revealed this 'intensity' information, and that information is still valid even after she withdraws. Hence, the award should *still* go to Bryn.

However, this 'intensity' defence of the Borda count is debatable. According the 'intensity' defence, the positional ranking of the alternatives acts as a crude proxy for *cardinal utility*; hence, the Borda count approximates Bentham's *Utilitarian* system (see §3). But it's easy to construct scenarios where an alternative's positional ranking is a *very* poor proxy for its cardinal utility; in these situations, the presence of a losing 'Chloe alternative' really *is* irrelevant, or worse, actually misleading.

For example, perhaps Arianne's 60% prefer Arianne to Bryn to exactly the same degree that Bryn's 40% prefer Bryn to Arianne. It's just that Arianne's 60% really despise Chloe, whereas

Bryn's 40% are indifferent to Chloe, as illustrated in the following figure:

$$\begin{array}{rl} & \text{Good} \longleftarrow - - - (\text{Cardinal utility}) - - \longrightarrow \text{Bad} \\ 60\% & A - - - - - - - - - - - B - - - - - - - - - - C \\ 40\% & B - - - C - - - - - - A - - - - - - - - - - - \end{array} \tag{2.5}$$

In this case, it seems wrong that the presence/absence of Chloe in the election can affect the choice between Arianne and Bryn.

We will return to the issue of 'irrelevant alternatives' when we discuss Arrow's Impossibility Theorem (§2E).

2B.4 Approval voting

Democracy is being allowed to vote for the candidate you dislike least. —Robert Byrne

The problems with the Borda Count seem to arise from the assignment of variable numbers of points to candidates depending on their rank. The introduction of a new candidate (even an unpopular one) changes the ranks of all other candidates, thereby differentially inflating their scores, and possibly changing the outcome. Perhaps a solution is to only assign *one* point to each 'preferred' candidate. This is the rationale behind *Approval voting*. Approval voting is similar to the Borda count, except that each voter can only assign 1 or 0 points to each candidate. However, the voter can give a point to *several different* candidates. The candidate with the highest score wins.

There are two versions of approval voting:

Fixed allotment: Each voter is given a fixed number of points, which she must spend. For example, there might be 4 candidates, and each voter must vote for exactly 2 of them.

In the limit case, when each voter is given exactly *one* point, this is just the traditional Plurality vote.

Variable allotment: A voter can vote for any number of candidates (including none of them or all of them).

The analysis of *fixed allotment* approval voting is simpler, so that is what we'll consider here. Suppose that there are four candidates, Arianne, Bryn, Chloe, and Dominique, and thirteen voters, with the following preferences:

#	Preferences
4	$A \succ D \succ C \succ B$
1	$B \succ A \succ D \succ C$
2	$B \succ A \succ C \succ D$
3	$C \succ B \succ D \succ A$
2	$D \succ B \succ C \succ A$
1	$D \succ C \succ B \succ A$

This example (from Riker [Rik82, §4E]) shows how different voting procedures produce different winners. First if each voter casts a *single* approval vote, so that we basically have a traditional plurality competition, then Arianne wins, with 4 votes:

#	Preferences	Approval Points			
		A	B	C	D
4	$A \succ D \succ C \succ B$	1	0	0	0
1	$B \succ A \succ D \succ C$	0	1	0	0
2	$B \succ A \succ C \succ D$	0	1	0	0
3	$C \succ B \succ D \succ A$	0	0	1	0
2	$D \succ B \succ C \succ A$	0	0	0	1
1	$D \succ C \succ B \succ A$	0	0	0	1
	Total score:	**4**	3	3	3

Next, if each voter casts *two* approval vote, then Bryn wins, with 8 points:

#	Preferences	Approval Points			
		A	B	C	D
4	$A \succ D \succ C \succ B$	1	0	0	1
1	$B \succ A \succ D \succ C$	1	1	0	0
2	$B \succ A \succ C \succ D$	1	1	0	0
3	$C \succ B \succ D \succ A$	0	1	1	0
2	$D \succ B \succ C \succ A$	0	1	0	1
1	$D \succ C \succ B \succ A$	0	0	1	1
	Total score:	7	**8**	4	7

However, if each voter casts *three* approval vote, then Chloe wins, with 12 points:

#	Preferences	Approval Points			
		A	B	C	D
4	$A \succ D \succ C \succ B$	1	0	1	1
1	$B \succ A \succ D \succ C$	1	1	0	1
2	$B \succ A \succ C \succ D$	1	1	1	0
3	$C \succ B \succ D \succ A$	0	1	1	1
2	$D \succ B \succ C \succ A$	0	1	1	1
1	$D \succ C \succ B \succ A$	0	1	1	1
	Total score:	7	9	**12**	11

Finally, if we use the Borda count method, then Dominique wins with 21 points:

%	Preferences	Borda Points			
		A	B	C	D
4	$A \succ D \succ C \succ B$	3	0	1	2
1	$B \succ A \succ D \succ C$	2	3	0	1
2	$B \succ A \succ C \succ D$	2	3	1	0
3	$C \succ B \succ D \succ A$	0	2	3	1
2	$D \succ B \succ C \succ A$	0	2	1	3
1	$D \succ C \succ B \succ A$	0	1	2	3
	Total Score:	18	20	19	**21**

So the question is, who is *really* the democratically legitimate choice of the People?

Further reading: This section contains only a few of the multitude of voting systems which have been proposed. Others include: the Copeland and Black rules (Examples (3.4⟨c⟩) and (3.4⟨d⟩) on page 38); the Gocha, Dodgson, Peleg, and Lexmin rules [KR80, §4.1]; the Nash rule [Rik82, §2B], the Schwartz and Kemeny rules [Rik82, §4C], and a plethora of quasi-economic methods. An inventory of these and other voting procedures can be found in any book on voting theory, such as [Bla58, Fis73, Str80a, Tay95].

The disturbing and counterintuitive examples presented in this section are examples of *voting paradoxes*. Other voting paradoxes include *Ostrogowski's paradox* [Ost03, RD76], *Anscombe's paradox* [Ans76, Wag84], the *Referendum paradox* [Nur98, §2.4], the *Divided government paradox* [BKZ93], and *Simpson's paradox* [Saa88, Saa90]. See Hannu Nurmi's excellent survey article [Nur98] and monograph [Nur99] for more information.

The Condorcet method was first proposed [ClMdC85] by the French mathematician and revolutionary political theorist Jean-Antoine-Nicolas Caritat, the Marquis de Condorcet (1743-1794). The Borda count was originally proposed [Bor81] by the French mathematical physicist, political theorist, and naval captain Jean-Charles Borda (1733-1799). An extensive discussion of the merits of the Borda count can be found in Saari [Saa95], which also contains a nice account of Borda's life and achievements [Saa95, §1.3.3]. The pros and cons of approval voting are examined in [BF78] and [KR80, §4.6].

2C Abstract Voting Procedures

Democracy: The election by the incompetent many instead of the appointment by the corrupt few. —George Bernard Shaw

The pathologies in §2B raise the question: is there *any* voting procedure which will not produce the 'wrong' answer in certain circumstances? To answer this question, we must mathematically define what we mean by a *voting procedure* and precisely specify the sorts of 'wrong' answers we want to avoid.

2C.1 Preferences and Profiles

Recommended: §??

Preference orderings: In the examples of Chapter **??**, we began with a hypothetical population of *voters*, $\mathcal{V}$, and a collection of *alternatives*, $\mathcal{A}$, and we assumed that each voter is capable of ranking the various alternatives in $\mathcal{A}$ in some 'linear' way, eg.

$$A \succ B \succ C \succ D.$$

In other words, each voter's preferences determine a **preference ordering**: a relation ' $\succeq$ ' on $\mathcal{A}$ which satisfies three axioms:

Completeness: For any pair of alternatives X and Y, either $X \succeq Y$ or $Y \succeq X$.

Reflexiveness: For any alternative $X \in \mathcal{A}$, $\quad X \succeq X$.

Transitivity: For any alternatives X, Y and Z,

$$\Big(X \succeq Y \text{ and } Y \succeq Z\Big) \Longrightarrow \Big(X \succeq Z\Big).$$

If $X \succeq Y$, then we say the voter **prefers** X to Y. Note that it is possible to have $X \succeq Y$ and $X \preceq Y$. In this case, we say the voter is **indifferent** between X and Y, and write $X \approx Y$. We then say that $\{X, Y\}$ is an **indifferent pair** of alternatives. If $X \succeq Y$ but $X \not\approx Y$, then the voter **strictly prefers** X to Y, and we write $X \succ Y$.

In some situations, however, it may be necessary for the voter to make a choice; she cannot be indifferent between two alternatives. You can either have your cake later or eat it now; you can't do both. A **strict preference ordering** is a relation ' $\succ$ ' on $\mathcal{A}$ which satisfies three axioms:

Completeness: For any pair of alternatives X and Y, either $X \succ Y$ or $Y \succ X$.

Antisymmetry: For any pair of alternatives X and Y, we *cannot* have both $X \succ Y$ and $Y \succ X$.

Transitivity: For any alternatives X, Y and Z,

$$\Big(X \succ Y \text{ and } Y \succ Z\Big) \Longrightarrow \Big(X \succ Z\Big).$$

Thus, strict preference is like preference, except that we replace the *Reflexiveness* axiom with an *Antisymmetry* axiom. Clearly, any strict preference ordering ' $\succ$ ' can be expanded to a (nonstrict) preference ordering ' $\succeq$ ' by defining

$$\Big(X \succeq Y\Big) \iff \Big(X \succ Y \text{ or } X = Y\Big)$$

The converse is half true: A (nonstrict) preference ordering ' $\succeq$ ' can be reduced to a strict preference ordering ' $\succ$ *if and only if* there exist no indifferent pairs of alternatives. In this case, we can define a strict preference ' $\succ$ ' by:

$$\Big(X \succ Y\Big) \iff \Big(X \succeq Y \text{ and } X \neq Y\Big).$$

2C.2 Voting procedures

Intuitively, a *voting procedure* is some method which takes a collection of voters (each with some preference ordering), and chooses a single alternative in $\mathcal{A}$ as the 'collective choice' of the group. Actually, a voting procedure must be more complicated than this. Suppose the group elects alternative A as president, but A then dies in a mysterious accident. There must be someone else who is 'next in line' for the the presidency. To put it another way: if A withdrew from the election at the last moment, after all the voters had finalized their preference orderings, then who would they elect instead? Suppose it was B, and suppose that B also withdrew. Who would the *third* choice be?

Reasoning in this manner, it is clear that a voting procedure doesn't just pick a single 'first' choice, it actually implicitly defines a *preference order* on the set of alternatives; a preference order which supposedly reflects the 'collective will of the People'.

Thus, we could define a voting procedure as a function which takes a *collection* of preference orders as input, and produces a *single* preference order as output. To be more precise, let $\mathcal{P}(\mathcal{A})$ be the set of all possible preference orderings on the set of alternatives $\mathcal{A}$. For example, if $\mathcal{A} = \{A, B, C\}$ is a set of three alternatives, then $\mathcal{P}(\mathcal{A})$ has thirteen elements:

$$\begin{aligned}
\mathcal{P}(\mathcal{A}) \;=\; \{&A \succ B \succ C, \quad B \succ C \succ A, \quad C \succ A \succ B, \\
&A \succ C \succ B, \quad C \succ B \succ A, \quad B \succ A \succ C \\
&A \succ B \approx C, \quad B \succ C \approx A, \quad C \succ A \approx B, \\
&A \approx C \succ B, \quad C \approx B \succ A, \quad B \approx A \succ C \\
&A \approx B \approx C\}.
\end{aligned}$$

Let $\mathcal{V}$ be a collection of voters. A **profile** is a function $\rho : \mathcal{V} \longrightarrow \mathcal{P}(\mathcal{A})$ assigning a preference ordering to each voter. Let $\mathfrak{R}(\mathcal{V}, \mathcal{A})$ be the set of all profiles for the voters in $\mathcal{V}$ and alternatives in $\mathcal{A}$. A **voting procedure**[1] is a function

$$\Pi : \mathfrak{R}(\mathcal{V}, \mathcal{A}) \longrightarrow \mathcal{P}(\mathcal{A}).$$

In other words, Π is a function which takes any profile as input, and produces a single **collective** (or **social**) preference ordering $\Pi(\rho)$ as output. We will indicate the preference ordering $\Pi(\rho)$

[1] Sometimes called a **social choice function** or a **social welfare function.**

with the relation ' $\stackrel{\rho}{\sqsupseteq}$ ' (or simply ' $\sqsupseteq$ ', when ρ is clear from context). Thus, $B \stackrel{\rho}{\sqsupseteq} C$ means that, given profile ρ, the voting procedure has ranked alternative B over alternative C.

Example 2C.1:

⟨a⟩ (a) **Plurality Vote:** Let $\rho \in \mathfrak{R}(\mathcal{V}, \mathcal{A})$ be a profile. For each $A \in \mathcal{A}$, let

$$N(A) \quad = \quad \#\Big\{ v \in \mathcal{V} \,;\, A \stackrel{\rho}{\underset{v}{\succeq}} B, \text{ for all } B \in \mathcal{A} \Big\}$$

be the number of voters who rank A 'first' in their preferences. Define ranking $\stackrel{\rho}{\sqsupseteq}$ by:

$$\Big(A \stackrel{\rho}{\sqsupseteq} B \Big) \iff \Big(N(A) \geq N(B) \Big).$$

Thus, the winner is the alternative which is ranked 'first' by the most voters.

⟨b⟩ (b) **Borda Count:** Let $\rho \in \mathfrak{R}(\mathcal{V}, \mathcal{A})$ be a profile. For each $v \in \mathcal{A}$, define $U_v^\rho : \mathcal{A} \longrightarrow \mathbb{N}$ as follows: for any $A \in \mathcal{A}$, $\quad U_v^\rho(A) \quad = \quad \#\Big\{ B \in \mathcal{A} \,;\, A \stackrel{\rho}{\underset{v}{\succeq}} B \Big\} - 1$ is the number of alternatives which voter v deems 'no better' than A (not including A itself).

Then define $U^\rho : \mathcal{A} \longrightarrow \mathbb{N}$ by $\quad U^\rho(A) \quad = \quad \sum_{v \in \mathcal{V}} U_v^\rho(A)$ $\quad$ (the 'Borda Count' of A).

Then define ranking $\stackrel{\rho}{\sqsupseteq}$ by: $\Big(A \stackrel{\rho}{\sqsupseteq} B \Big) \iff \Big(U^\rho(A) \geq U^\rho(B) \Big)$. Thus, the winner is the alternative with the highest Borda Count.

⟨c⟩ (c) **Approval Voting:** Suppose $\#(\mathcal{A}) = N$ and $M < N$, and suppose each voter must vote for exactly M out of N alternatives. Let $\rho \in \mathfrak{R}(\mathcal{V}, \mathcal{A})$ be a profile, and for each $v \in \mathcal{A}$, define $T_v^\rho : \mathcal{A} \longrightarrow \mathbb{N}$ as follows: for any $A \in \mathcal{A}$, let

$$T_v^\rho(A) \quad = \quad \#\Big\{ B \in \mathcal{A} \,;\, B \stackrel{\rho}{\underset{v}{\succeq}} A \Big\}.$$

be the number of alternatives which voter v prefers to A.

Define $f : \mathbb{N} \longrightarrow \{0, 1\}$ by $\quad f(t) \quad = \quad \begin{cases} 1 & \text{if} \quad t \leq M \\ 0 & \text{if} \quad t > M \end{cases}$.

Thus, $f\Big(T_v^\rho(A) \Big) = 1$ if and only if alternative A is in the 'top M' alternatives for voter v. Now define $T^\rho : \mathcal{A} \longrightarrow \mathbb{N}$ as follows: for any $A \in \mathcal{A}$, let

$$T^\rho(A) \quad = \quad \sum_{v \in \mathcal{V}} f\Big(T_v^\rho(A) \Big),$$

be the total number of 'approval votes' for A.

Define ranking $\stackrel{\rho}{\sqsupseteq}$ by: $\Big(A \stackrel{\rho}{\sqsupseteq} B \Big) \iff \Big(T^\rho(A) \geq T^\rho(B) \Big)$. Thus, the winner is the alternative with the most approval votes.

⟨d⟩ (d) **Condorcet:** Suppose $\mathcal{A} = \{A, B, C, D\}$. If there *is* a Condorcet winner (say A), then we have preference order $A \overset{\rho}{\sqsupset} B \overset{\rho}{\approx} C \overset{\rho}{\approx} D$. If there is *no* Condorcet winner, then we have preference order $A \overset{\rho}{\approx} B \overset{\rho}{\approx} C \overset{\rho}{\approx} D$. ♢

Remarks: **(a)** The four voting procedures described here are not the *only* ways to implement Plurality Vote, Borda Count, Approval Voting, and the Condorcet method (but they are arguably the most 'natural'). For example, there are many voting procedures which will identify a Condorcet winner (if one exists); see if you can invent another one.

(b) Observe that a voting procedure can produce an outcome which is *indifferent* between two or more alternatives. For example, suppose the Viking Longboat Society is using the Condorcet vote to decide whether to serve *Ale*, *Beer*, or *Cider* at their annual fundraiser, but the outcome is the *Condorcet Paradox* (page 22). The only reasonable response is to serve all three beverages!

Strict voting procedures: However, sometimes we cannot accept 'indifferent' outcomes; sometimes our procedure *must* give a strict ordering with a maximal element. For example, in a presidential election, there must be a *unique* choice; we can't have a scenario where three people tie for first place and share the job. Hence, the output must not only be a preference ordering, but a *strict* preference ordering. However, we cannot expect a strict preference as output if we do not provide strict preferences as input. Let $\mathcal{P}^*(\mathcal{A})$ be the set of strict preference orderings on $\mathcal{A}$. For example, if $\mathcal{A} = \{A, B, C\}$, then $\mathcal{P}(\mathcal{A})$ has six elements:

$$\begin{aligned}\mathcal{P}(\mathcal{A}) \;=\; \{&A \succ B \succ C, \quad B \succ C \succ A, \quad C \succ A \succ B,\\ &A \succ C \succ B, \quad C \succ B \succ A, \quad B \succ A \succ C\}.\end{aligned}$$

A **strict profile** is a function $\rho : \mathcal{V} \longrightarrow \mathcal{P}^*(\mathcal{A})$ assigning a strict preference ordering to each voter. Let $\mathfrak{R}^*(\mathcal{V}, \mathcal{A})$ be the set of all strict profiles for the voters in $\mathcal{V}$ and alternatives in $\mathcal{A}$. A **strict voting procedure** is a function $\Pi : \mathfrak{R}^*(\mathcal{V}, \mathcal{A}) \longrightarrow \mathcal{P}^*(\mathcal{A})$. We will indicate the preference ordering $\Pi(\rho)$ with the relation '$\overset{\rho}{\sqsupset}$' (or simply '$\sqsupset$', when ρ is clear from context). Thus, $B \overset{\rho}{\sqsupset} C$ means that, given strict profile ρ, the strict voting procedure strictly prefers alternative B to alternative C.

Example 2C.2: Agenda of Pairwise Elections

Suppose that $\#(\mathcal{V})$ is *odd*, so that it is impossible for a pairwise election to result in a tie (assuming all voters have *strict* preferences). Given a particular agenda of pairwise elections, we define a strict preference ordering on $\mathcal{A}$ as follows:

1. Find the ultimate winner of the agenda of pairwise elections; rank this candidate *first.*
2. Eliminate this candidate from $\mathcal{A}$. From the *remaining* alternatives, find the ultimate winner of the agenda of pairwise elections; rank this candidate *second.*

3. Eliminate the *second* candidate from $\mathcal{A}$. From the *remaining* alternatives, find the ultimate winner of the agenda of pairwise elections; rank this candidate *third*.
4. Proceed in this fashion until you run out of alternatives.

We illustrate with an example. Suppose that $\mathcal{A} = \{A, B, C, D\}$. Let $\rho \in \mathfrak{R}^*(\mathcal{V}, \mathcal{A})$, and suppose that A is the ultimate winner of the pairwise election agenda with profile ρ. Let $\mathcal{A}_1 = \mathcal{A} \setminus \{A\} = \{B, C, D\}$, and let ρ_1 be the profile we get by restricting each voter's ρ-preferences to $\mathcal{A}_1$. For example if voter v had ρ-preferences $B \overset{\rho}{\underset{v}{\succ}} A \overset{\rho}{\underset{v}{\succ}} C \overset{\rho}{\underset{v}{\succ}} D$, then v would have ρ_1-preferences $B \overset{\rho_1}{\underset{v}{\succ}} C \overset{\rho_1}{\underset{v}{\succ}} D$.

Suppose that B is the ultimate winner of the pairwise voting agenda with profile ρ_1 (we skip the election where A would have been introduced). Let $\mathcal{A}_2 = \mathcal{A}_1 \setminus \{B\} = \{C, D\}$, and let ρ_2 be the profile we get by restricting each voter's ρ_1-preferences to $\mathcal{A}_2$.

Suppose that C is the ultimate winner of the pairwise voting agenda with profile ρ_2 (we skip the elections where A and B would have been introduced). Then $\mathcal{A}_3 = \mathcal{A}_2 \setminus \{C\} = \{D\}$. We define order $\overset{\rho}{\sqsupset}$ by: $A \overset{\rho}{\sqsupset} B \overset{\rho}{\sqsupset} C \overset{\rho}{\sqsupset} D$. ♢

Strict vs. nonstrict procedures: A **tiebreaker rule** is a function $\tau : \mathcal{P}(\mathcal{A}) \longrightarrow \mathcal{P}^*(\mathcal{A})$ which converts any nonstrict preference ordering into a strict ordering in an order-preserving way. That is, if $A, B \in \mathcal{A}$ are two alternatives and $A \succ B$, then τ preserves this. However, if $A \approx B$, then τ forces either $A \succ B$ or $B \succ A$.

In general, the tiebreaker rule can be totally arbitrary (eg. flipping a coin, trial by combat, putting alternatives in alphabetical order, etc.), because if a voter is *indifferent* about two alternative then *by definition* it doesn't matter which one we put first.

Given τ, any *nonstrict* voting procedure Π can be turned into a *strict* voting procedure Π^* as follows:

1. Apply Π to a (strict) voter profile.
2. Use τ to convert the resulting (nonstrict) preference ordering to a strict ordering.

Conversely, any *strict* voting procedure Π^* can be extended to a *nonstrict* voting procedure Π as follows:

1. Apply τ to convert each voter's (nonstrict) preference ordering into a strict preference ordering, thereby converting society's nonstrict voter profile into a strict voter profile.
2. Now apply Π^* to the 'strictified' profile.

We formalize this:

Proposition 2C.3 *Let $\tau : \mathcal{P}(\mathcal{A}) \longrightarrow \mathcal{P}^*(\mathcal{A})$ be a tie-breaker rule.*

(a) *Suppose* $\Pi : \mathfrak{R}(\mathcal{V}, \mathcal{A}) \longrightarrow \mathcal{P}(\mathcal{A})$ *is a (nonstrict) voting procedure. Let* $\Pi^* := \tau \circ \Pi$. *Then* $\Pi^* : \mathfrak{R}^*(\mathcal{V}, \mathcal{A}) \longrightarrow \mathcal{P}^*(\mathcal{A})$ *is a strict voting procedure.*

(b) *Suppose* $\Pi^* : \mathfrak{R}^*(\mathcal{V}, \mathcal{A}) \longrightarrow \mathcal{P}^*(\mathcal{A})$ *is a strict voting procedure. Define* $\tau : \mathfrak{R}(\mathcal{V}, \mathcal{A}) \longrightarrow \mathfrak{R}^*(\mathcal{V}, \mathcal{A})$ *by applying* τ *separately to each voter's (nonstrict) preference ordering. Now define* $\Pi = \Pi^* \circ \tau$. *Then* Π *is a (nonstrict) voting procedure.*

Proof: **Exercise 2.3** □

However, it is not always desirable to cavalierly convert nonstrict procedures into strict ones. For one thing, by using an arbitrary tie-breaker rule, we will likely violate desirable properties such as the *Neutrality* axiom **(N)** (see §2C.3). Also, if the nonstrict procedure *too often* produces indifferent outcomes, then we will *too often* end up making decisions based on some arbitrary rule. For example, the (nonstrict) Condorcet procedure *rarely* yields a strict order. If τ is the 'alphabetical order' rule, and Π is the Condorcet rule, then $\Pi^* = \tau \circ \Pi$ will, in practice, end up meaning 'arrange candidates in alphabetical order' 90% of the time. This is hardly a good way to pick the president.

2C.3 Desiderata

Prerequisites: §2C.1

We began by asking: are there any voting procedures which produce sensible results? We will formalize what we mean by a 'sensible result' by requiring the voting procedure to satisfy certain axioms.

Pareto (Unanimity): The *Pareto* (or *Unanimity*) axiom is the following:

(P) If $B, C \in \mathcal{A}$, and ρ is a profile where *all* voters prefer B to C (ie. for all $v \in \mathcal{V}$, $B \underset{v}{\overset{\rho}{\succeq}} C$), then $B \overset{\rho}{\sqsupseteq} C$.

This seems imminently sensible. Any voting scheme which chose a unanimously unpopular alternative over a unanimously popular alternative would be highly undemocratic!

Exercise 2.4 Check that the following voting procedures satisfy axiom **(P)**:

1. Plurality Vote (Example 2C.1(a)).
2. Borda Count (Example 2C.1(b)).
3. Approval Voting Example 2C.1(c).
4. Agenda of Pairwise votes (Example 2C.2).

Monotonicity: If the voting procedure selects a certain alternative C as the 'collective choice' of the society, and some voter changes his preferences to become *more* favourable toward C, then surely C should *remain* the collective choice of the society. This is the content of the *Monotonicity* axiom:

(M) Let $B, C \in \mathcal{A}$, and let ρ be a profile such that $B \overset{\rho}{\sqsubseteq} C$. Let $v \in \mathcal{V}$ be some voter such that $C \underset{v}{\overset{\rho}{\preceq}} B$, and let δ be the profile obtained from ρ by giving v a new preference ordering $\underset{v}{\overset{\delta}{\succeq}}$, such that $C \underset{v}{\overset{\delta}{\succeq}} B$ (all *other* voters keep the same preferences). Then $B \overset{\delta}{\sqsubseteq} C$.

Exercise 2.5 Check that the following voting procedures satisfy axiom **(M)**:

1. Plurality Vote (Example 2C.1(a)).
2. Borda Count (Example 2C.1(b)).
3. Approval Voting Example 2C.1(c).
4. Agenda of Pairwise votes (Example 2C.2).

Anonymity: A basic democratic principle is *political equality:* all voters have the same degree of influence over the outcome of a vote. To put it another way, the voting procedure is incapable of distinguishing one voter from another, and therefor treats all their opinions equally. In other words, the voters are *anonymous*. To mathematically encode this, we imagine that all the voters exchange identities (ie. are *permuted*). A truly 'anonymous' voting procedure should be unable to tell the difference...

(A) Let $\sigma : \mathcal{V} \longrightarrow \mathcal{V}$ be a permutation of the voters. Let ρ be a profile, and let δ be the profile obtained from ρ by permuting the voters with σ. In other words, for any $v \in \mathcal{V}$, $\delta(v) = \rho\Big(\sigma(v)\Big)$. Then ρ and δ yield identical collective preference orderings. In other words, for any alternatives $B, C \in \mathcal{A}$,

$$\Big(B \overset{\rho}{\sqsupseteq} C\Big) \iff \Big(B \overset{\delta}{\sqsupseteq} C\Big).$$

Exercise 2.6 Check that the following voting procedures satisfy axiom **(A)**:

1. Plurality Vote (Example 2C.1(a)).
2. Borda Count (Example 2C.1(b)).
3. Approval Voting Example 2C.1(c).
4. Agenda of Pairwise votes (Example 2C.2).

If we wish to impose axiom **(A)** as a blanket assumption, then we can restrict our attention to *anonymous voting procedures*. An **anonymous profile** is a function $\alpha : \mathcal{P}(\mathcal{A}) \longrightarrow \mathbb{N}$. For example, any profile $\rho \in \mathfrak{R}(\mathcal{V}, \mathcal{A})$ yields an anonymous profile $\alpha_\rho : \mathcal{P}(\mathcal{A}) \longrightarrow \mathbb{N}$, where for any $P \in \mathcal{P}(\mathcal{A})$,

$$\alpha(P) \quad = \quad \#\{v \in \mathcal{V} \,;\, \rho(v) = P\}$$

is the number of voters with preference ordering P. Let $\widetilde{\mathfrak{R}}(\mathcal{A})$ be the set of anonymous profiles. An **anonymous voting procedure** is a function $\widetilde{\Pi} : \widetilde{\mathfrak{R}}(\mathcal{A}) \longrightarrow \mathcal{P}(\mathcal{A})$, which takes an anonymous profile as input, and yields a preference ordering on $\mathcal{A}$ as output.

Exercise 2.7 Let $\Pi : \mathfrak{R}(\mathcal{V}, \mathcal{A}) \longrightarrow \mathcal{P}(\mathcal{A})$ be a voting procedure. Show that $\Big(\Pi$ satisfies axiom **(A)**$\Big)$ $\iff$ $\begin{pmatrix}\text{There is some anonymous procedure } \widetilde{\Pi} \\ \text{so that } \Pi(\rho) \;=\; \widetilde{\Pi}(\alpha_\rho) \text{ for any } \rho \in \mathfrak{R}(\mathcal{V}, \mathcal{A})\end{pmatrix}$.

Neutrality: Just as a voting procedure should be impartial amongst voters, it should be impartial amongst alternatives. An 'unbiased' procedure does not favour the 'first' alternative over the 'second' alternative, and so on. In other words, if we permute the alternatives, we should get the same outcome.

(N) Let $\sigma : \mathcal{A} \longrightarrow \mathcal{A}$ be a permutation of the alternatives. Let ρ be a profile, and let δ be the profile obtained from ρ by permuting the alternatives with σ. In other words, for any $B, C \in \mathcal{A}$, and any $v \in \mathcal{V}$,

$$\Big(B \underset{v}{\overset{\rho}{\succeq}} C\Big) \iff \Big(\sigma(B) \underset{v}{\overset{\delta}{\succeq}} \sigma(C)\Big).$$

Then the preference ordering produced by δ is obtained by similarly permuting the preference ordering produced from ρ. That, is, for any $B, C \in \mathcal{A}$,

$$\Big(B \underset{}{\overset{\rho}{\sqsupseteq}} C\Big) \iff \Big(\sigma(B) \underset{}{\overset{\delta}{\sqsupseteq}} \sigma(C)\Big).$$

Exercise 2.8 Show that an *agenda* of pairwise elections (Example 2C.2) between three or more alternatives does *not* satisfy the axiom **(N)**.

Exercise 2.9 Check that the following voting procedures *do* satisfy axiom **(N)**:

1. Plurality Vote (Example 2C.1(a)).
2. Borda Count (Example 2C.1(b)).
3. Approval Voting Example 2C.1(c).

Exercise 2.10 Show that, if a voting procedure satisfies the *Monotonicity* axiom **(M)** and the *Neutrality* axiom **(N)**, then it must satisfy the *Pareto* axiom **(P)**.

Exercise 2.11 Suppose a *nonstrict* voting procedure Π is transformed into a *strict* voting procedure Π^* using Proposition 2C.3(b) on page 34. Show that Π^* might not satisfy axiom **(N)**, even if Π did.

Condorcet: As we saw in §2B.2, the Condorcet scheme, while laudable, is somewhat unsatisfactory as a voting procedure because it usually doesn't produce a clear winner. Nevertheless, it's based on a good idea, the *Condorcet criterion*, which we might desire in other voting methods.

If $A, B \in \mathcal{A}$ are two alternatives, let $\#[A \succ B]$ be the number of voters who strictly prefer A to B:

$$\#[A \succ B] \quad = \quad \#\left\{ v \in \mathcal{V} \,;\, A \underset{v}{\succ} B \right\}.$$

Let's write "$A \gg B$" if A defeats B in a pairwise vote:

$$\Big(A \gg B\Big) \iff \Big(\#[A \succ B] \;>\; \#[B \succ A]\Big).$$

The Condorcet Criterion states:

(C) If $A \in \mathcal{A}$ is an alternative who beats every other alternative in a pairwise vote, then A is the top-ranked element of the collective preference ordering. That is

$$\Big(\forall\, B \in \mathcal{A}, \quad A \gg B\Big) \quad \Longrightarrow \quad \Big(\forall\, B \in \mathcal{A}, \quad A \;\sqsupset\; B\Big).$$

Procedures satisfying axiom **(C)** are sometimes called *Condorcet extensions*, because they often take the form, "If there is a clear Condorcet winner, then choose her. If not, then choose a winner using the following method instead...'. Thus, Condorcet extensions reduce (but usually do not eliminate) the possibility of a tie or ambiguous outcome.

Example 2C.4:

⟨a⟩ The Borda Count does *not* satisfy axiom **(C)**. See §2B.3.

⟨b⟩ **Sequence of pairwise votes:** Theorem 2B.1 says that any sequence of pairwise votes will choose a Condorcet winner, if one exists. Thus, all such sequences are Condorcet extensions. The problem is, as we saw in §2B.1, different sequences can produce outcomes; hence a pairwise vote sequence violates the *Neutrality* axiom **(N)**.

⟨c⟩ **Copeland Rule:** The **Copeland index** of A is the number of alternatives A defeats in pairwise votes, minus the number which defeat A:

$$i(A) \quad = \quad \#\{B \in \mathcal{A} \,;\, A \gg B\} \;-\; \#\{B \in \mathcal{A} \,;\, B \gg A\}$$

Thus, if A is the Condorcet winner, then $i(A) = \#(\mathcal{A}) - 1$.

The **Copeland rule** tells us to rank the alternatives in decreasing order of their Copeland indices; thus, the Copeland 'winner' is the alternative with the highest Copeland index. The Copeland rule satisfies the Condorcet criterion **(C)** because, if a Condorcet winner exists, he is automatically the Copeland winner.

⟨d⟩ **Black Rule:** The Black Rule is very simple: if a Condorcet winner A exists, then choose A. If there is no Condorcet winner, then use the Borda Count method to order the alternatives. ♢

The Copeland and Black rules are still not *strict* voting procedures, because ties are still possible; they are simply less likely than in the original Condorcet rule.

2D Sen and (Minimal) Liberalism

Prerequisites: §2C.3

Surely in a democratic society, there are certain decisions, concerning your person, over which only you should have control. For example, society can impose some constraints on your actions (eg. *Thou shalt not steal*), but only you should be able to decide what you wear, what you say, and who you choose to associate with.

The idea that individuals have certain inalienable rights is called *liberalism*, and the most minimal form of liberalism is one where a particular voter has control over *one* decision in society. If $v \in \mathcal{V}$ is a voter, and $B, C \in \mathcal{A}$ are alternatives, then we say that v is **decisive** over the pair $\{B, C\}$ if, for any profile ρ,

$$\left(B \stackrel{\rho}{\sqsupseteq} C\right) \iff \left(B \underset{v}{\overset{\rho}{\succeq}} C\right)$$

The axiom of *Minimal Liberalism* requires:

(ML) There are two voters $v_1, v_2 \in \mathcal{V}$, and four alternatives $B_1, C_1,\ B_2, C_2 \in \mathcal{A}$, such that v_1 is decisive over $\{B_1, C_1\}$ and v_2 is decisive over $\{B_2, C_2\}$.

All this says is that there are at least two individuals in the society who have some minimal degree of control over some single aspect of their lives —a very minimal form of liberalism indeed! Nevertheless, we are confronted with...

Sen's Impossibility Theorem: *Suppose there are at least three alternatives in $\mathcal{A}$, and at least two voters in $\mathcal{V}$. Then there is no strict voting procedure which satisfies axioms* **(P)** *and* **(ML)**.

Proof: Suppose that, as in axiom **(ML)**, voter v_1 is decisive over the pair $\{B_1, C_1\}$ and voter v_2 is decisive over the pair $\{B_2, C_2\}$. Given *any* strict profile as input, the voting procedure should produce a strict preference ordering as output. But suppose that voters have the following preferences:

v_1	C_2	$\overset{\diamond}{\succ}$	$\mathbf{B_1}$	$\overset{*}{\succ}$	$\mathbf{C_1}$	$\overset{\star}{\succ}$	B_2
v_2	C_1	$\overset{\star}{\succ}$	$\mathbf{B_2}$	$\overset{\dagger}{\succ}$	$\mathbf{C_2}$	$\overset{\diamond}{\succ}$	B_1
Others	C_1	$\overset{\star}{\succ}$	B_2	and	C_2	$\overset{\diamond}{\succ}$	B_1

(we have accented in bold the choices over which each voter is decisive). Now,

($*$) v_1 is decisive over $\{B_1, C_1\}$, so we must have $B_1 \sqsupset C_1$.

($\star$) Society is unanimous that $C_1 \succ B_2$, so we must have $C_1 \sqsupset B_2$, by axiom **(P)**.

($\dagger$) v_2 is decisive over $\{B_2, C_2\}$, so we must have $B_2 \sqsupset C_2$.

($\diamond$) Society is unanimous that $C_2 \succ B_1$, so we must have $C_1 \sqsupset B_1$, by axiom **(P)**.

We conclude that

$$B_1 \sqsupset C_1 \sqsupset B_2 \sqsupset C_2 \sqsupset B_1.$$

By *transitivity*, it follows that $C_1 \sqsupset B_1$. But since we also have $B_1 \sqsupset C_1$, this contradicts the requirement of *antisymmetry*. We conclude that $\sqsupset$ cannot be a strict preference ordering. □

Remark: The astute reader will notice that the previous proof seems to assume the existence of *four* alternatives (B_1, C_1, B_2 and C_2), despite the fact that Sen's Theorem only hypothesizes *three*. This apparent inconsistency is reconciled by recognizing that the pairs $\{B_1, C_1\}$ and $\{B_2, C_2\}$ may not be *distinct*. For example, we could set $B_1 = C_2$; and then rework the proof of Sen's Theorem without needing to include the 'unanimous' decision that $B_1 \sqsupset C_2$. The details are **Exercise 2.12**.

Exercise 2.13 (a) Show that no voting method can satisfy both the *Anonymity* axiom **(A)** and the *Minimal Liberalism* axiom **(ML)**.

(b) Likewise, show that no voting method can satisfy both the *Neutrality* axiom **(N)** and the *Minimal Liberalism* axiom **(ML)**.

(c) Suggest how you might replace both **(A)** and **(N)** with a modified axiom which allows for **(ML)**, while still encoding the idea that society gives equal political rights to all voters, and decides between conflicting alternatives in an 'unbiased' manner.

Further reading: The original references to Sen's theorem are Sen [Sen70b, Sen70a]. An elementary discussion and proof are given in Saari [Saa97]; other references are Saari [Saa95, §3.4.1] or Kim and Roush [KR80, Thm 4.4.1, p.81]

2E Arrow's Impossibility Theorem

Politics is the art of the possible. —Otto Von Bismarck

Prerequisites: §2C.3

Recall the Danacian election example of §**??**, where the introduction of extra 'fringe' parties into an election 'split' the vote of the ruling *Literal* party, allowing the *Regressive Coercitives* to seize office. Although both the *Literal* and *Coercitive* parties are much more popular than these fringe groups, the existence of fringe parties changes the outcome. The *Plurality* voting procedure is sensitive to the 'irrelevant alternative' of the fringe parties.

Likewise, in §2B.3 we were able to change the outcome of a Borda count election between Arianne and Bryn by introducing a third alternative, Chloe. Despite the fact that *everyone* prefers one of the other alternatives to Chloe, her presence in the race still tips the balance. The choice between Arianne and Bryn is sensitive to the 'irrelevant alternative' of Chloe.

We saw how this sensitivity to irrelevant alternatives makes a process vulnerable to manipulation. The *Coercitives* can manipulate the outcome by covertly supporting the fringe parties. Likewise, the friends of Bryn may encourage Chloe to participate in the election, even though she has no chance of winning, simply to manipulate the results in their favour. We want a procedure which is immune to these machinations. We say that a voting procedure is *Independent of Irrelevant Alternatives* if the following is true.

(IIA) Let $A, B \in \mathcal{A}$ be two alternatives. Suppose ρ, δ are two profiles, such that each voter's ρ-preference concerning the pair $\{A, B\}$ is identical to his δ-preferences concerning $\{A, B\}$. That is, for every $v \in \mathcal{V}$,

$$\left(A \underset{v}{\overset{\rho}{\succeq}} B\right) \iff \left(A \underset{v}{\overset{\delta}{\succeq}} B\right)$$

Then the collective ρ-preference concerning $\{A, B\}$ will be identical to the collective δ-preference concerning $\{A, B\}$. That is: $\left(A \overset{\rho}{\sqsupseteq} B\right) \iff \left(A \overset{\delta}{\sqsupseteq} B\right)$.

To translate this into English, suppose that $\mathcal{A} = \{\text{Arianne}, \text{Bryn}, \text{Chloe}\}$. Let ρ be the profile of table (2.2) on page 25 of §2B.3, and let δ be the profile of table (2.3) on page 26. Then we see that the Borda count does *not* satisfy **(IIA)**, because it says $B \overset{\rho}{\sqsupset} A$ but $A \overset{\delta}{\sqsupset} B$, despite the fact that all voters order A and B the *same* in both profiles. In §2B.3 we showed how this led to an undesirable scenario, where the 'winner' of the Fields Medal changed because a losing candidate (Chloe) dropped out of the race. This is the reason why **(IIA)** is considered a desirable property.

Dictatorship: A **dictatorship** is a voting procedure where one voter makes all the decisions. In other words, there is some voter $v \in \mathcal{V}$ (the **dictator**) so that, for any $B, C \in \mathcal{A}$,

$$\left(B \sqsupseteq C\right) \iff \left(B \underset{v}{\succeq} C\right).$$

We now come to the most famous result in mathematical political science:

Arrow's Impossibility Theorem: *Suppose that $\mathcal{A}$ has at least three alternatives, and $\mathcal{V}$ has at least two voters. Then the only voting procedure which satisfies axioms* **(P)** *and* **(IIA)** *is a dictatorship.*

Proof: First we will show that any procedure satisfying axioms **(P)** and **(IIA)** must also satisfy a version of the *Neutrality* axiom **(N)**.

Claim 1: *Let $A_1, B_1, A_2, B_2 \in \mathcal{A}$ be four alternatives*[2]*, and suppose ρ is a profile such that every voter's preference ordering of the pair $\{A_1, B_1\}$ is identical to her ordering of $\{A_2, B_2\}$. In other words, for every voter $v \in \mathcal{V}$,*

$$\left(A_1 \underset{v}{\succeq} B_1\right) \iff \left(A_2 \underset{v}{\succeq} B_2\right).$$

Then the voting procedure will yield a preference order which also assigns the same order to the pair $\{A_1, B_1\}$ as to the pair $\{A_2, B_2\}$. That is: $\left(A_1 \overset{\rho}{\sqsupseteq} B_1\right) \iff \left(A_2 \overset{\rho}{\sqsupseteq} B_2\right)$.

Proof: Assume WOLOG that $A_1 \overset{\rho}{\sqsupseteq} B_1$. We want to show that $A_2 \overset{\rho}{\sqsupseteq} B_2$. To do this, we will create a new profile δ such that:

(a) Every voter's ordering of $\{A_2, B_2\}$ is identical in δ and ρ. Hence, by **(IIA)**, we have

$$\left(A_2 \overset{\rho}{\sqsupseteq} B_2\right) \iff \left(A_2 \overset{\delta}{\sqsupseteq} B_2\right).$$

(b) δ is structured so that it is clear that $A_2 \overset{\delta}{\sqsupseteq} B_2$.

Combining facts **(a)** and **(b)** yields $A_2 \overset{\rho}{\sqsupseteq} B_2$, as desired.

To obtain δ, take each voter and change her ρ-ranking of A_2 so that it is just above A_1. Likewise, change her ranking of B_2 so that it is just below B_1. We can always do this in a manner which preserves her ordering of the pairs $\{A_1, B_1\}$ and $\{A_2, B_2\}$, as shown by the diagram below:

Before (in ρ)	After (in δ)
$A_1 \overset{\rho}{\succeq} B_1$ and $A_2 \overset{\rho}{\succeq} B_2$	$A_2 \overset{\delta}{\succeq} A_1 \overset{\delta}{\succeq} \ldots \overset{\delta}{\succeq} B_1 \overset{\delta}{\succeq} B_2$
$B_1 \overset{\rho}{\succeq} A_1$ and $B_2 \overset{\rho}{\succeq} A_2$	$B_1 \overset{\delta}{\succeq} B_2 \overset{\delta}{\succeq} \ldots \overset{\delta}{\succeq} A_2 \overset{\delta}{\succeq} A_1$

Now, in δ,

- Every voter prefers A_2 to A_1, so we have $A_2 \overset{\delta}{\sqsupseteq} A_1$ by the Pareto axiom **(P)**.
- Every voter prefers B_1 to B_2, so we have $B_1 \overset{\delta}{\sqsupseteq} B_2$ by the Pareto axiom **(P)**.
- Every voter's ordering of $\{A_1, B_1\}$ is identical in δ and ρ. Hence, $A_1 \overset{\delta}{\sqsupseteq} B_1$, by **(IIA)**.

[2]Here, we assume $A_1 \neq B_1$ and $A_2 \neq B_2$; however, the sets $\{A_1, B_1\}$ and $\{A_2, B_2\}$ might not be disjoint (eg. if $\mathcal{A}$ only has three alternatives in total).

We now have $A_2 \overset{\delta}{\sqsupseteq} A_1 \overset{\delta}{\sqsupseteq} B_1 \overset{\delta}{\sqsupseteq} B_2$. Because $\overset{\delta}{\sqsupseteq}$ is a transitive ordering, it follows that $A_2 \overset{\delta}{\sqsupseteq} B_2$.

However, every voter's ordering of $\{A_2, B_2\}$ is identical in δ and ρ. Hence, by **(IIA)**, we conclude that $A_2 \overset{\rho}{\sqsupseteq} B_2$, as desired.

We have now shown that

$$\left(A_1 \overset{\rho}{\sqsupseteq} B_1\right) \Longrightarrow \left(A_2 \overset{\rho}{\sqsupseteq} B_2\right).$$

By switching A_1 with A_2 and switching B_1 with B_2 throughout the proof, we can likewise show that

$$\left(A_2 \overset{\rho}{\sqsupseteq} B_2\right) \Longrightarrow \left(A_1 \overset{\rho}{\sqsupseteq} B_1\right).$$

This completes the proof. ◇ Claim 1

Now, suppose we number the voters $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$ in some arbitrary way. If $A, B \in \mathcal{A}$ are any two alternatives, we'll write '$A \underset{n}{\overset{\rho}{\succeq}} B$' to mean 'voter v_n prefers A to B in profile ρ'

Claim 2: *There is a 'swing voter' v_m with the following property: Suppose $X, Y \in \mathcal{A}$ are any two alternatives, and ρ is any profile such that*

$$X \underset{n}{\overset{\rho}{\succeq}} Y, \text{ for all } n < m \quad \text{and} \quad X \underset{n}{\overset{\rho}{\preceq}} Y, \text{for all } n > m. \tag{2.6}$$

Then

$$\left(X \overset{\rho}{\sqsupseteq} Y\right) \iff \left(X \underset{m}{\overset{\rho}{\succeq}} Y\right). \tag{2.7}$$

Thus, v_m can 'tip the balance' between X and Y in any profile satisfying eqn.(2.6).

Proof: Fix $A, B \in \mathcal{A}$, and consider the following sets of profiles:

$$\begin{aligned}
\mathfrak{R}_0 &= \{\rho \in \mathfrak{R}(\mathcal{V}, \mathcal{A}) \text{ ; every voter prefers } B \text{ to } A\};\\
\mathfrak{R}_1 &= \left\{\rho \in \mathfrak{R}(\mathcal{V}, \mathcal{A}) \text{ ; } \mathcal{A} \underset{1}{\overset{\rho}{\succeq}} \mathcal{B}, \text{ but for all } n > 1, \ \mathcal{A} \underset{n}{\overset{\rho}{\preceq}} \mathcal{B}\right\};\\
&\vdots \quad \vdots \quad \vdots\\
\mathfrak{R}_m &= \left\{\rho \in \mathfrak{R}(\mathcal{V}, \mathcal{A}) \text{ ; } \begin{array}{c} A \underset{n}{\overset{\rho}{\succeq}} B, \text{ for all } n \leq m \\ \text{but } A \underset{n}{\overset{\rho}{\preceq}} B, \text{ for all } n > m \end{array}\right\};\\
&\vdots \quad \vdots \quad \vdots\\
\mathfrak{R}_N &= \{\rho \in \mathfrak{R}(\mathcal{V}, \mathcal{A}) \text{ ; every voter prefers } A \text{ to } B\}.
\end{aligned}$$

Claim 2.1: *The voting procedure assigns a preference of A vs. B which is constant on each of the sets $\mathfrak{R}_0, \mathfrak{R}_1, \ldots, \mathfrak{R}_N$. In other words, for any $n \in [0..N]$, if there is one $\rho \in \mathfrak{R}_n$ so that $A \overset{\rho}{\sqsupseteq} B$, then it must be the case for all $\delta \in \mathfrak{R}_n$ that $A \overset{\delta}{\sqsupseteq} B$.*

Proof: Axiom **(IIA)** says that the voter's preferences over *other* pairs of alternatives have no effect on the collective preference of A vs. B. Thus, the collective preference of A vs. B is determined only by the pattern of voters' $\{A, B\}$-preferences, and this pattern is constant across each set $\mathfrak{R}_n$, by definition. ◇ Claim 2

We will write "$A \overset{\mathfrak{R}_n}{\sqsupseteq} B$" to mean that $A \overset{\rho}{\sqsupseteq} B$ for all $\rho \in \mathfrak{R}_n$. By the Pareto axiom **(P)**, we know that

$$A \overset{\mathfrak{R}_0}{\sqsubseteq} B \quad \text{and} \quad A \overset{\mathfrak{R}_N}{\sqsupseteq} B.$$

Thus, there must be some $m \in [1..N]$ such that $A \overset{\mathfrak{R}_{m-1}}{\sqsubseteq} B$, but $A \overset{\mathfrak{R}_m}{\sqsupseteq} B$. Hence eqn.(2.7) is true for $X = A$ and $Y = B$.

But the claim says that eqn.(2.7) will be true for *any* alternatives X and Y. To see this, we apply Claim 1, which says, in effect, that if we take any other alternatives A_1 and B_1, and substitute A_1 for A and B_1 for B_0 everywhere in the above construction, we will reach the same result, namely that eqn.(2.7) is true for $X = A_1$ and $Y = B_1$. ◇ Claim 2

Claim 3: *v_m is a dictator.*

Proof: Let $A, B \in \mathcal{A}$. We want to show that $\left(A \overset{\rho}{\sqsupseteq} B\right) \iff \left(A \underset{m}{\overset{\rho}{\succeq}} B\right)$ (regardless of what the other voters think).

Suppose that ρ is some profile. We will first show that

$$\left(A \underset{m}{\overset{\rho}{\succeq}} B\right) \Longrightarrow \left(A \overset{\rho}{\sqsupseteq} B\right) \tag{2.8}$$

To do this, we will construct a new profile δ, so that each voter's ρ-preferences concerning $\{A, B\}$ are identical to her δ-preferences. Thus, by **(IIA)**,

$$\left(A \overset{\rho}{\sqsupseteq} B\right) \iff \left(A \overset{\delta}{\sqsupseteq} B\right).$$

We will build δ so that it is clear that $A \overset{\delta}{\sqsupseteq} B$. To do this, we introduce a third alternative, $C \in \mathcal{A}$. By axiom **(IIA)**, the position of C in the δ-preferences of voters $v_1, \ldots, v_n$ has no effect on whether $A \overset{\delta}{\sqsupseteq} B$ or $B \overset{\delta}{\sqsupseteq} A$. Hence, we can build δ so that:

- For all $n < m$, $\quad C \underset{n}{\overset{\delta}{\succeq}} A$ and $C \underset{n}{\overset{\delta}{\succeq}} B$.
- For all $n > m$, $\quad A \underset{n}{\overset{\delta}{\succeq}} C$ and $B \underset{n}{\overset{\delta}{\succeq}} C$.

- $A \underset{m}{\overset{\delta}{\succeq}} C \underset{m}{\overset{\delta}{\succeq}} B$.
- For every $n \in [1..N]$, $\left(A \underset{n}{\overset{\delta}{\succeq}} B\right) \iff \left(A \underset{n}{\overset{\rho}{\succeq}} B\right)$.

We portray this schematically:

v_1	v_2	v_3	...	v_{m-1}	v_m	v_{m+1}	v_{m+2}	v_{m+3}	...	v_n
C	C	C	...	C	A	A	B	A	...	B
A	B	A	...	B	C	B	A	B	...	A
B	A	B	...	A	B	C	C	C	...	C

By setting $X = C$ and $Y = A$ in Claim 2, we get $A \overset{\delta}{\sqsupseteq} C$.

By setting $X = C$ and $Y = B$ in Claim 2, we get $C \overset{\delta}{\sqsupseteq} B$.

Hence, by transitivity, we conclude that $A \overset{\delta}{\sqsupseteq} B$. Then by **(IIA)** we also have $A \overset{\rho}{\sqsupseteq} B$.

Now, we can do this in any profile ρ where $A \underset{m}{\overset{\rho}{\succeq}} B$; hence we have shown (2.8).

By reversing the roles of A and B throughout the whole argument, we can likewise show that:

$$\left(B \underset{m}{\overset{\rho}{\succeq}} A\right) \Longrightarrow \left(B \overset{\rho}{\sqsupseteq} A\right).$$

Thus, v_m is a dictator. □[Claim 3 & Theorem]

Exercise 2.14 The conclusion of Claim 1 seems superficially different than the *Neutrality* axiom **(N)**, but in fact they are the same.

(a) Show that axiom **(N)** implies Claim 1.

(b) Show that Claim 1, together with axiom **(IIA)**, implies axiom **(N)**.

Discussion: Arrow's Impossibility Theorem says that no 'democratic' procedure can be constructed which is immune to distortion through the introduction of additional alternatives. Arrow's Theorem does *not* say 'democracy is impossible'; it merely says that any democracy will be inherently flawed by 'sensitivity to irrelevant alternatives'.

Indeed, it's not even clear that this sensitivity is a 'flaw'. Despite the terminology of 'impossibility' and 'dictators', Saari [Saa95, §3.4.5,§3.4.9] interprets Arrow's Theorem in a very benign way, as simply saying that **(IIA)** is an unreasonable requirement for a voting procedure; furthermore it isn't even clear that **(IIA)** is desirable. Recall that, in the 'Arianne, Bryn, and Chloe' example on page 25 of §2B.3, the 'irrelevant alternative' Chloe is perhaps not *really* irrelevant, because perhaps her presence in the competition provides information about the 'intensity' with which the voters support Arianne or Bryn (see §2B.3).

Exercise 2.15 Which of the following voting procedures satisfy **(IIA)**? Which don't? Provide a proof/counterexample in each case.

1. Plurality Vote (Example 2C.1(a)).

2. Borda Count (Example 2C.1(b)).

3. Approval Voting Example 2C.1(c).

4. Agenda of Pairwise votes (Example 2C.2).

Further reading: The original exposition of Arrow's Theorem is [Arr63], but it is discussed in pretty much any book on voting theory or social choice, eg. Sen [Sen70a, Chap.3] or Fishburn [Fis73]. For very readable and elementary proofs, see Taylor [Tay95, 10.5] or Hodge and Klima [HK05, Chap.5]. Kim and Roush [KR80, 4.3] contains a proof using boolean matrix theory. Saari [Saa95, §3.4] contains a proof by convex geometry methods, while Saari [Saa97] contains an elementary introduction to the theorem. Luce and Raiffa [LR80, §14.4] gives a classical presentation in the context of game theory. The proof I've given here is adapted from the third of three short, elegant proofs by John Geanokoplos [Gea01]; Geanokoplos also mentions that Luis Ubeda-Rives has a similar (unpublished) proof.

In *Liberalism against Populism*, the political scientist William Riker contrasts two visions of democracy [Rik82]. The *Populist* view (which Riker attributes to Rousseau) asserts that government should reflect some emergent 'General Will' of the people, and that democratic institutions are instruments to express this General Will. In this vision, democratic government has the 'positive' role of formulating policies which embody the collective desires of society. In contrast, the *Liberal* view (which Riker attributes to Madison) regards any form of government as inherently flawed and dangerous, being susceptible to corruption, incompetence, and the abuse of power. Democracy is then an essentially 'negative' mechanism: it is simply a way that the people can remove any government whose corruption and incompetence becomes intolerable. As such, it perhaps provides some incentive for elected officials to maintain at least the appearance of good behaviour. Democracy cannot ensure 'good' government; it can only defend against intolerably 'bad' government.

Riker interprets Arrow's Theorem (along with the voting paradoxes of §2A and §2B) to mean that Rousseau's 'General Will' is not well-defined. Thus, the Populist vision of democracy is incoherent; democracy cannot be an instrument to express the collective will of society, because there is no such thing. Riker therefore concludes that Liberalism is a more realistic democratic project [Rik82, §5A].

2F Strategic Voting: Gibbard & Satterthwaite

Politics is not the art of the possible. It consists of choosing between the disastrous and the unpalatable. —John Kenneth Galbraith

Prerequisites: §2E

Recall from §2B.3 that *strategic voting* means voting *contrary* to your true desires, because you believe that doing so will actually yield a *more* desirable outcome. To be precise, we must introduce some additional terminology.

Social choice functions: Suppose we have a strict voting procedure Π. For any strict profile $\rho \in \mathfrak{R}^*(\mathcal{V}, \mathcal{A})$, the **leader** picked by ρ is the unique maximal element of of the ordering $\overset{\rho}{\sqsupset}$. We indicate the leader by $\chi_\Pi(\rho)$. Thus, if ρ describes the electorate's preferences in a presidential election, then $\chi_\Pi(\rho)$ is the person actually elected president.

The function χ_Π is an example of a social choice function. A **social choice function** is any function $\chi : \mathfrak{R}^*(\mathcal{V}, \mathcal{A}) \longrightarrow \mathcal{A}$. In other words, a social choice function takes a profile of voter preferences as input, and yields a single alternative as output.

Any strict voting procedure yields a social choice function. If $\Pi : \mathfrak{R}^*(\mathcal{V}, \mathcal{A}) \longrightarrow \mathcal{P}^*(\mathcal{A})$ is a strict voting procedure, then the **leadership function** $\chi_\Pi : \mathfrak{R}^*(\mathcal{V}, \mathcal{A}) \longrightarrow \mathcal{A}$ (described in the first paragraph) is a social choice function. Conversely, given a social choice function χ, we can construct many strict voting procedures Π so that $\chi = \chi_\Pi$ (**Exercise 2.16**).

Strategic voting: Suppose $\rho \in \mathfrak{R}^*(\mathcal{V}, \mathcal{A})$ is some strict profile, Let $P \in \mathcal{P}^*(\mathcal{A})$ be a strict preference ordering, and let $v \in \mathcal{V}$ be a voter such that $\rho(v) = P$. Let $P' \in \mathcal{P}^*(\mathcal{A})$ be another strict preference ordering, and let δ be the profile we get if voter v 'pretends' to have preference order P'. In other words, for any $w \in \mathcal{V}$,

$$\delta(w) \;=\; \begin{cases} \rho(w) & \text{if} \quad v \neq w; \\ P' & \text{if} \quad v = w. \end{cases}$$

We say that P' is **strategic vote** for v if $\chi(\delta) \underset{P}{\succ} \chi(\rho)$. In other words, the alternative $\chi(\delta)$ (obtained if v *pretends* to have preference P') is *preferable* for v to the alternative $\chi(\rho)$ (obtained if v is honest).

We say that a social choice function is **nonmanipulable** (or **strategy proof**) if, for any profile $\rho \in \mathfrak{R}^*(\mathcal{V}, \mathcal{A})$, no voter $v \in \mathcal{V}$ has a strategic vote. For example, the Borda Count is *not* strategy proof, as we saw in §2B.3. This fact has often been used to discredit the Borda Count. However, the Gibbard-Satterthwaite Theorem (below) basically says that *any* democratic system is susceptible to strategic voting. Hence, the Borda Count is no worse than any other democratic procedure.

Dictatorship If χ is a social choice function, then a **dictator** for χ is a voter $v \in \mathcal{V}$ so that χ always picks v's favourite choice. In other words, for any profile $\rho \in \mathfrak{R}(\mathcal{V}, \mathcal{A})$, and any $A \in \mathcal{A}$,

$$\Big(\chi(\rho) = A\Big) \iff \Big(A \underset{v}{\overset{\rho}{\succ}} B, \text{ for all } B \in \mathcal{A}\Big).$$

We then say that χ is a **dictatorship**. It is easy to show:

Lemma 2F.1 *Let Π be a strict voting procedure, with leadership function χ_Π. If v is the dictator of Π, then v is also the dictator of χ_Π.* □

Surjectivity The social choice function χ is **surjective** if, for any $A \in \mathcal{A}$, there is some profile ρ so that $A = \chi(\rho)$.

In a sense, surjectivity is a 'nontriviality' requirement: clearly, if the alternative A can never win, under any conditions, then why is A even in the competition? We can thus always assume that χ is surjective, because if it is not, we should simply remove any candidates who are never chosen.

Gibbard-Satterthwaite Impossibility Theorem: *The only surjective, nonmanipulable social choice function is a dictatorship.*

Proof: Suppose $\chi : \mathfrak{R}(\mathcal{V}, \mathcal{A}) \longrightarrow \mathcal{A}$ is nonmanipulable and surjective. We will define a strict voting procedure Π so that $\chi = \chi_\Pi$. We will show that Π satisfies the axioms **(P)** (from §2C.3) and **(IIA)** (from §2E). We will then invoke Arrow's Impossibility Theorem (page 41).

To do this, we first need two technical lemmas. If $P \in \mathcal{P}^*(\mathcal{A})$ is a strict ordering on $\mathcal{A}$, and $A, B \in \mathcal{A}$, with $A \underset{P}{\succ} B$, then we say that A and B are P**-adjacent** if there is no alternative C with $A \underset{P}{\succ} C \underset{P}{\succ} B$. The $\{A, B\}$**-transposition** of P is the new ordering P' obtained from P by simply exchanging A and B in the ordering of P, while keeping all other alternatives the same:

$$\begin{array}{rl} P: & C_1 \underset{P}{\succ} C_2 \underset{P}{\succ} \cdots \underset{P}{\succ} C_j \underset{P}{\succ} A \underset{P}{\succ} B \underset{P}{\succ} D_1 \underset{P}{\succ} \cdots \underset{P}{\succ} D_k \\ P': & C_1 \underset{P'}{\succ} C_2 \underset{P'}{\succ} \cdots \underset{P'}{\succ} C_j \underset{P'}{\succ} B \underset{P'}{\succ} A \underset{P'}{\succ} D_1 \underset{P'}{\succ} \cdots \underset{P'}{\succ} D_k \end{array}$$

Let $\rho \in \mathfrak{R}^*(\mathcal{V}, \mathcal{A})$ be some profile, and let $v \in \mathcal{V}$ be a voter with $\rho(v) = P$. Define ρ' by:

$$\rho'(w) \;=\; \begin{cases} \rho(w) & \text{if} \quad v \neq w; \\ P' & \text{if} \quad v = w. \end{cases}$$

We say that ρ' is a **transposition** of ρ, which **promotes** B, **demotes** A, and **fixes** all other alternatives.

Claim 1: *Let $\rho \in \mathfrak{R}^*(\mathcal{V}, \mathcal{A})$. Let $A, B \in \mathcal{A}$, and let ρ' be a transposition of ρ which promotes B and demotes A.*

(a) *If $\chi(\rho) = A$, then $\chi(\rho') = A$ or B.*

(b) *If $\chi(\rho) \neq A$ (in particular, if $\chi(\rho) = B$) then $\chi(\rho') = \chi(\rho)$.*

Proof: **(a)** By contradiction, suppose $\chi(\rho') = C$, where $C \notin \{A, B\}$. Thus, either $C \underset{P}{\succ} A$ or $C \underset{P}{\prec} A$.

If $C \underset{P}{\succ} A$, then P' is a strategic vote for the voter v (in profile ρ), contradicting nonmanipulability of χ.

If $C \underset{P}{\prec} A$, then also $C \underset{P}{\prec} B$ (because A and B are adjacent) and then $C \underset{P'}{\prec} A$ (by definition of P'). Thus, P is a strategic vote for the voter v (in profile ρ'), again contradicting nonmanipulability of χ.

(b) Suppose $\chi(\rho) = C \neq A$ and $\chi(\rho') = C'$. We want to show that $C = C'$. Suppose by contradiction that $C \neq C'$. Then either $C \underset{P}{\prec} C'$ or $C \underset{P}{\succ} C'$.

If $C \underset{P}{\prec} C'$, then P' is a strategic vote for voter v in profile ρ.

If $A \neq C \underset{P}{\succ} C'$, then also $C \underset{P'}{\succ} C'$ (**Exercise 2.17** Hint: A and B are adjacent); hence P is a strategic vote for voter v in profile ρ'. ◇ Claim 1

Now, if $\rho, \delta \in \mathfrak{R}^*(\mathcal{V}, \mathcal{A})$ are two strict profiles, and $B \in \mathcal{A}$, then a **B-promoting walk** from δ to ρ is a sequence of strict profiles

$$\delta \;=\; \delta_0,\; \delta_1,\; \delta_2, \ldots, \delta_K \;=\; \rho$$

where, for each $k \in [1...K]$, δ_k is a transposition of δ_{k-1} which either promotes B or fixes B.

Claim 2: *Let $\rho, \delta \in \mathfrak{R}^*(\mathcal{V}, \mathcal{A})$ be two strict profiles.*

(a) *Suppose $\chi(\delta) = B$. If there is a B-promoting walk from δ to ρ, then $\chi(\rho) = B$, also.*

(b) *If B is the maximal element in every voter's ρ-preference, then there is a B-promoting walk from δ to ρ.*

Proof: **(a)** By repeated application of Claim 1(b), we have

$$B \;=\; \chi(\delta_0) \;=\; \chi(\delta_1) \;=\; \cdots \;=\; \chi(\delta_K) \;=\; \chi(\rho).$$

(b) follows from the definition of a B-promoting walk. ◇ Claim 2

We use Claim 2 to prove that χ satisfies a kind of 'Pareto' property.

Claim 3: *Let $\rho \in \mathfrak{R}^*(\mathcal{V}, \mathcal{A})$ be a strict profile.*

(a) *If B is the maximal element in every voter's ρ-preference, then $\chi(\rho) = B$.*

(b) *If A and B are the 'top two' alternatives in every voter's ρ-preference, then either $\chi(\rho) = A$ or $\chi(\rho) = B$.*

Proof: **(a)** By hypothesis, χ is *surjective*, so there is *some* profile δ such that $\chi(\delta) = B$. Now, Claim 2(b) yields a B-promoting walk from δ to ρ; then Claim 2(a) says $\chi(\rho) = B$.

(b) Let ρ be a profile so that every voter ranks A and B as her 'top two' alternatives. Let δ be the modified profile where every voter ranks B *first* and ranks A *second*, and ranks all other alternatives the same as in ρ. Thus, B is maximal in every voter's δ-preference, so part **(a)** implies that $\chi(\delta) = B$. But there is an A-promoting walk to get from δ to ρ, so Claim 1(b) implies that either $\chi(\rho) = A$ or $\chi(\rho) = B$. ◇ Claim 3

Now, suppose $\mathcal{A} = \{A_1, \ldots, A_N\}$, where the alternatives are numbered in some entirely arbitrary order. Given any strict profile $\rho \in \mathfrak{R}^*(\mathcal{V}, \mathcal{A})$, and any pair of alternatives $\{B, C\}$, we define the profile $\rho_{B,C}$ as follows:

- Each voter $v \in \mathcal{V}$ places B and C as her top two alternatives, ranked in the same order as she ranked them in ρ.
- Each voter then ranks all remaining candidates in decreasing numerical order.

To illustrate this, suppose $\mathcal{A} = \{A_1, A_2, \ldots, A_7\}$, and that $B = A_3$ and $C = A_6$. Then we have the following:

Before (in ρ)	After (in $\rho_{B,C}$)
$\cdots \succ B \succ \cdots \succ C \succ \cdots$	$B \succ C \succ A_1 \succ A_2 \succ A_4 \succ A_5 \succ A_7$
$\cdots \succ C \succ \cdots \succ B \succ \cdots$	$C \succ B \succ A_1 \succ A_2 \succ A_4 \succ A_5 \succ A_7$

Now, we define a strict voting procedure $\Pi : \mathfrak{R}^*(\mathcal{V}, \mathcal{A}) \longrightarrow \mathcal{P}^*(\mathcal{A})$ as follows. For any strict profile ρ, we define the relation $\overset{\rho}{\sqsupset}$ by

$$\Big(A \overset{\rho}{\sqsupset} B\Big) \iff \Big(\chi(\rho_{A,B}) = A\Big)$$

We don't yet know that $\overset{\rho}{\sqsupset}$ is even a preference ordering. However, we can already show:

Claim 4: *The procedure* Π *satisfies axiom* **(IIA)**. *In other words, if* $A, B \in \mathcal{A}$, *and* ρ *and* δ *are two profiles such that*

$$\text{For all } v \in \mathcal{V}, \quad \Big(A \underset{v}{\overset{\rho}{\succ}} B\Big) \iff \Big(A \underset{v}{\overset{\delta}{\succ}} B\Big), \tag{2.9}$$

then we have: $\Big(A \overset{\rho}{\sqsupset} B\Big) \iff \Big(A \overset{\delta}{\sqsupset} B\Big)$.

Proof: Suppose ρ and δ satisfy eqn.(2.9). Then, by definition, $\rho_{A,B} = \delta_{A,B}$. Hence, $\chi(\rho_{A,B}) = \chi(\delta_{A,B})$. Hence $\Big(A \overset{\rho}{\sqsupset} B\Big) \iff \Big(A \overset{\delta}{\sqsupset} B\Big)$. ◇ Claim 4

To show that Π is a strict voting procedure, we must show:

Claim 5: *For any* $\rho \in \mathfrak{R}^*(\mathcal{V}, \mathcal{A})$, *the relation* $\overset{\rho}{\sqsupset}$ *is a strict preference ordering.*

Proof: Complete: By Claim 3(b), either $\chi(\rho_{A,B}) = A$ or $\chi(\rho_{A,B}) = B$. Hence, either $A \overset{\rho}{\sqsupset} B$ or $B \overset{\rho}{\sqsupset} A$.

Antisymmetric: Clearly, it is impossible to have both $\chi(\rho_{A,B}) = A$ and $\chi(\rho_{A,B}) = B$. Thus, it is impossible to have both $A \overset{\rho}{\sqsupset} B$ and $B \overset{\rho}{\sqsupset} A$.

Transitive: Suppose $A \overset{\rho}{\sqsupset} B \overset{\rho}{\sqsupset} C$. We must show that $A \overset{\rho}{\sqsupset} C$. Suppose by contradiction that $C \overset{\rho}{\sqsupset} A$, so that we have a cycle

$$A \overset{\rho}{\sqsupset} B \overset{\rho}{\sqsupset} C \overset{\rho}{\sqsupset} A. \tag{2.10}$$

Define the strict profile $\delta \in \mathfrak{R}^*(\mathcal{V}, \mathcal{A})$ such that:

- Each voter $v \in \mathcal{V}$ places A, B and C as her top three alternatives, ranked in the same order as she ranked them in ρ.
- Each voter then ranks all remaining candidates in decreasing numerical order.

Then Claim 4 and eqn.(2.10) imply that $A \overset{\delta}{\sqsupset} B \overset{\delta}{\sqsupset} C \overset{\delta}{\sqsupset} A$.

Claim 5.1: $\chi(\delta) \in \{A, B, C\}$.

Proof: Suppose, by way of contradiction, that $\chi(\delta) = D$, where $D \notin \{A, B, C\}$. We define a D-promoting walk from δ to the profile $\rho_{A,B}$, by simply 'walking' C down each voter's list of preferences, one transposition at a time. Each transposition fixes D, except for at most one transposition which switches C and D (thereby promoting D). We conclude, from Claim 2(a), that $\chi(\rho_{A,B}) = \chi(\delta) = D$. But $A \overset{\delta}{\sqsupset} B$, which means, by definition, that $\chi(\rho_{A,B}) = A$. Contradiction. ▽ Claim 5.1

Thus, $\chi(\delta) \in \{A, B, C\}$. Assume $\chi(\delta) = A$. (Since the cycle $A \overset{\delta}{\sqsupset} B \overset{\delta}{\sqsupset} C \overset{\delta}{\sqsupset} A$ is symmetrical, the following argument will also work if $\chi(\delta) = B$ or C).

Claim 5.2: $A \overset{\rho}{\sqsupset} C$.

Proof: We define an A-promoting walk from δ to the profile $\rho_{A,C}$, by simply 'walking' B down each voter's list of preferences, one transposition at a time. Each transposition fixes A, except for at most one transposition, which switches A and B (thereby promoting A). We conclude, from Claim 2(a), that $\chi(\rho_{A,C}) = \chi(\delta) = A$. But, by definition of Π, this means that $A \overset{\rho}{\sqsupset} C$. ◇ Claim 5

Claim 5.2 contradicts the assumption that $C \overset{\rho}{\sqsupset} A$. By contradiction, the cyclical ordering in eqn.(2.10) is impossible. Thus, $\overset{\rho}{\sqsupset}$ must be transitive. ◇ Claim 5

Thus, $\overset{\rho}{\sqsupset}$ is always a strict preference ordering, for any $\rho \in \mathfrak{R}^*(\mathcal{V}, \mathcal{A})$. Hence, Π is a strict voting procedure.

Claim 6: *The procedure* Π *satisfies the* Pareto *axiom* **(P)**.

Proof: Let $\rho \in \mathfrak{R}^*(\mathcal{V}, \mathcal{A})$ and suppose $A, B \in \mathcal{A}$, are such that $A \underset{v}{\overset{\rho}{\succ}} B$ for all $v \in \mathcal{V}$. We must show that $A \overset{\rho}{\sqsupset} B$ —ie. we must show that $\chi(\rho_{A,B}) = A$

But $A \underset{v}{\overset{\rho}{\succ}} B$ for all $v \in \mathcal{V}$, so it follows that, also, $A \underset{v}{\overset{\rho_{A,B}}{\succ}} B$ for all $v \in \mathcal{V}$. Hence, A is the *maximal* element in every voter's preferences under $\rho_{A,B}$. Thus, Claim 3**(a)** says that $\chi(\rho_{A,B}) = A$. Hence $A \overset{\rho}{\sqsupset} B$. ◇ Claim 6

By combining Claims 4 and 6 with Arrow's Impossibility Theorem, we conclude that the procedure Π is a dictatorship.

Claim 7: χ *is the leadership function for* Π.

Proof: Let $\rho \in \mathfrak{R}^*(\mathcal{V}, \mathcal{A})$ and suppose $\chi(\rho) = A$. We must show that A is the maximal element of the ordering $\overset{\rho}{\sqsupset}$ generated by Π. In other words, for any $B \in \mathcal{A}$, we must show that $A \overset{\rho}{\sqsupset} B$, or, equivalently, that $\chi(\rho_{A,B}) = A$. To see this, we define an A-promoting walk from ρ to $\rho_{A,B}$ as follows:

1. For any $v \in \mathcal{V}$ such that $B \overset{\rho}{\underset{v}{\succ}} A$, we 'walk' B to the top of v's preference ordering with a sequence of transpositions that fix A.
2. For any $v \in \mathcal{V}$ such that $A \overset{\rho}{\underset{v}{\succ}} B$, we 'walk' A to the top of v's preference ordering with a sequence of transpositions that promote A.
3. Finally, for every $v \in \mathcal{V}$, we rearrange all remaining alternatives in decreasing numerical order, with a sequence of transpositions that fix both A and B.

Thus, Claim 2(B) implies that $\chi(\rho_{A,B}) = A$. Hence $A \overset{\rho}{\sqsupset} B$. This is true for any $B \in \mathcal{A}$, so A is maximal. ◇ Claim 7

Let $v \in \mathcal{V}$ be the dictator of Π. Then Lemma 2F.1 says v is also the dictator of χ. □

One can actually conduct a more sophisticated analysis, and measure how 'susceptible' various voting procedures are to manipulation. Intuitively, the 'susceptibility' of a voting procedure is measured by the probability of a scenario where a small number of voters can change the outcome through strategic voting. Let $\mathfrak{R}(\mathcal{V}, \mathcal{A})$ be the 'space' of all possible profiles; then any voting procedure partitions $\mathfrak{R}(\mathcal{V}, \mathcal{A})$ into regions corresponding to different outcomes. The strategic voting opportunities occur along the *boundaries* between these regions. By defining the 'susceptibility' of a voting procedure in terms of the size of these boundaries, one can actually prove:

Theorem 2F.2 [Saa95, §5.3] *The voting procedure least susceptible to manipulation is the Borda count. The voting procedure most susceptible to manipulation is the plurality vote.* □

Further reading: The Gibbard-Satterthwaite theorem was proved independently by Gibbard [Gib73] and Satterthwaite [Sat75]. Other discussions are Gärdenfors [Gär77], Saari [Saa95, §5.1], and Kim and Roush [KR80, §4.4].

The proof of Gibbard-Satterthwaite given here was adapted from the proof by Sonnenschein and Schmeidler (1974) as transmitted by Kim and Roush [KR80, Thm 4.4.3]. The original proof doesn't require χ to be surjective, but (at the expense of additional technicalities) uses the weaker assumption that χ takes on at least three distinct values.

Part II

Social Welfare Functions

Chapter 3

Utility and Utilitarianism

Prerequisites: None **Recommended:** §2C.3

3A Utility functions

Let $\mathcal{A}$ be a set of 'alternatives', and let Zara be an individual. Recall that an *ordinal utility function* for Zara is a function $u_0 : \mathcal{A} \longrightarrow \mathbb{R}$ such that, for any $a, b \in \mathcal{A}$, if $u_0(a) \geq u_0(b)$, then this means that Zara 'prefers' alternative a to alternative b. We assume that Zara always seeks to maximize her utility.

Loosely speaking, $u_0(a)$ measures Zara's level of 'happiness' or 'satisfaction' with alternative a. However, this 'psychological' interpretation raises a host of philosophical problems (e.g. can 'happiness' really be quantified using a single numerical parameter? Even if this is true, is maximizing happiness really our sole objective? And even if it is, *should* it be?). Thus, many economists and philosophers prefer a strictly 'behavioural' interpretation of utility: We say that Zara 'prefers' alternative a to alternative b if, empirically, she will always pick a instead of b when offered a choice between the two. Note that this merely describes her observable *behaviour* —we make no assumptions about her emotions. The utility function is then simply a mathematical device for concisely encoding these 'preferences', as revealed through Zara's observed choices. (However, by stripping the utility function of its psychological content in this way, it is possible that we also strip social choice theory of any normative relevance.)

Notwithstanding these philosophical issues, we will assume that people's preferences can be accurately described using utility functions, and that it is morally desirable for society to choose alternatives which grant the largest possible utility to the largest number of members. However, it will be necessary to somewhat enrich our notion of utility to encode 'cardinal' information as well as 'ordinal' information.

A *cardinal utility function* for Zara is an ordinal utility function $U : \mathcal{A} \longrightarrow \mathbb{R}$ with an additional property. Suppose a_1, b_1 and a_2, b_2 are two pairs of alternatives such that $U(a_1) > U(b_1)$ and $U(a_2) > U(b_2)$. Let $\Delta U_1 = U(a_1) - U(b_1)$ and let $\Delta U_2 = U(a_2) - U(b_2)$. If $\Delta U_1 = r \cdot \Delta U_2$ for some $r > 0$, then we interpret this to mean that Zara's preference for a_1 over b_1 is 'r times

as great' as her preference for a_2 over b_2. Our problem is to make sense of this notion.

We do this using a hypothetical gambling game or *lottery*. First we introduce the concept of *expected utility*. Suppose Zara is playing a lottery where she will randomly win one of the alternatives in $\mathcal{A} = \{a, b, c\}$. Suppose the probabilities of these outcomes are $\mathbf{p}(a)$, $\mathbf{p}(b)$ and $\mathbf{p}(c)$ respectively, and suppose Zara assigns them (cardinal) utilities $U(a)$, $U(b)$ and $U(c)$. Then Zara's *expected utility* in this lottery is quantity:

$$\mathbb{E}(U, \mathbf{p}) \quad = \quad \mathbf{p}(a) \cdot U(a) \; + \; \mathbf{p}(b) \cdot U(b) \; + \; \mathbf{p}(c) \cdot U(c)$$

One can interpret this as the *average utility she can expect to gain by playing the lottery many times over.*

More generally, in a lottery with some set of alternatives $\mathcal{A}$, let $\mathbf{p} : \mathcal{A} \longrightarrow [0, 1]$ be a probability distribution (a function assigning a probability to each alternative), and let $U : \mathcal{A} \longrightarrow \mathbb{R}$ be a (cardinal) utility function. Zara's *expected utility* is defined:

$$\mathbb{E}(U, \mathbf{p}) \quad = \quad \sum_{a \in \mathcal{A}} \mathbf{p}(a) \cdot U(a). \tag{3.1}$$

Now we come to the key idea. Given a choice between several lotteries (i.e. several probability distributions over $\mathcal{A}$), *Zara will always pick the lottery which maximizes her expected utility.*

This intuitively plausible observation can then be flipped around, to provide the *definition* of the cardinal utility function. To be precise, a *cardinal utility function* for Zara is a function $U : \mathcal{A} \longrightarrow \mathbb{R}$ so that, if Zara is allowed to choose between various lotteries (with different probability distributions), she will always choose the lottery which yields the highest expected utility, as defined by eqn.(3.1).

To guarantee that such a utility function exists, we must assume that Zara is 'rational' in her choices amongst various lotteries. Let $\mathbb{P}(\mathcal{A})$ be the space of all possible probability distributions on the set of alternatives $\mathcal{A}$. For example, if $\mathcal{A} = \{a_1, \ldots, a_N\}$, then we can identify $\mathbb{P}(\mathcal{A})$ with the N-simplex:

$$\mathbb{P}(\mathcal{A}) \quad = \quad \left\{ \mathbf{p} = (p_1, \ldots, p_N) \in [0, 1]^N \; ; \; \sum_{n=1}^{N} p_n \; = \; 1 \right\}.$$

Thus, $\mathbb{P}(\mathcal{A})$ represents the space of all possible lotteries on these alternatives. We define the relation $\succeq$ on $\mathbb{P}(\mathcal{A})$ as follows: for any $\mathbf{p}, \mathbf{q} \in \mathbb{P}(\mathcal{A})$,

$$\Big(\mathbf{p} \succeq \mathbf{q} \Big) \iff \Big(\text{Zara picks lottery } \mathbf{p} \text{ over } \mathbf{q} \Big).$$

When choosing between lotteries, Zara must satisfy three 'rationality' axioms: *Transitivity*, *Linearity*, and *Continuity*.

(T) (*Transitivity*) For any $\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3 \in \mathbb{P}(\mathcal{A})$, if $\mathbf{p}_1 \succeq \mathbf{p}_2$ and $\mathbf{p}_2 \succeq \mathbf{p}_3$, then $\mathbf{p}_1 \succeq \mathbf{p}_3$.

(L) (*Linearity*) Suppose $\mathbf{p}_0, \mathbf{p}_1 \in \mathbb{P}(\mathcal{A})$ are two lotteries. For any $r \in [0,1]$, let $\mathbf{p}_r$ be the lottery obtained by convex-combining $\mathbf{p}_0$ and $\mathbf{p}_1$:

$$\mathbf{p}_r(a) \quad = \quad r \cdot \mathbf{p}_1(a) \; + \; (1-r) \cdot \mathbf{p}_0(a), \quad \text{for all } a \in \mathcal{A}.$$

Likewise, let $\mathbf{p}'_0, \mathbf{p}'_1 \in \mathbb{P}(\mathcal{A})$ be two lotteries, and for any $r \in [0,1]$, define $\mathbf{p}'_r$ by

$$\mathbf{p}'_r(a) \quad = \quad r \cdot \mathbf{p}'_1(a) \; + \; (1-r) \cdot \mathbf{p}'_0(a), \quad \text{for all } a \in \mathcal{A}.$$

Then

$$\Big(\mathbf{p}_0 \succeq \mathbf{p}'_0 \text{ and } \mathbf{p}_1 \succeq \mathbf{p}'_1\Big) \Longrightarrow \Big(\forall\, r \in [0,1], \quad \mathbf{p}_r \succeq \mathbf{p}'_r\Big).$$

To understand axiom **(L)**, think of $\mathbf{p}_r$ as representing a 'two-stage lottery', where the prize at the *first* stage is a ticket to a *second* lottery held at the *second* stage. At the first stage, there is probability r of winning a ticket to lottery $\mathbf{p}_1$, and probability $(1-r)$ of winning a prize to lottery $\mathbf{p}_0$. The net effect is as if Zara was competing in a *single* lottery with probability distribution $\mathbf{p}_r$. The axiom **(L)** says: if Zara prefers $\mathbf{p}_0$ to $\mathbf{p}'_0$, and prefers $\mathbf{p}_1$ to $\mathbf{p}'_1$, then she will surely prefer a lottery between $\mathbf{p}_0$ and $\mathbf{p}_1$ to a similar lottery between $\mathbf{p}'_0$ and $\mathbf{p}'_1$.

For any alternative $a \in \mathcal{A}$, let $\mathbb{1}_a$ be the lottery which gives probability 1 (i.e. certainty) to outcome a. It is reasonable to assume:

$$\Big(a \succeq b\Big) \iff \Big(\mathbb{1}_a \succeq \mathbb{1}_b\Big).$$

In other words, if choosing between lotteries with 'guaranteed' outcomes, Zara will pick the outcome she prefers. The third axiom of 'rationality' is as follows:

(C) (*Continuity*[1]) Suppose $a_0, b, a_1 \in \mathcal{A}$, and $a_0 \preceq b \preceq a_1$. For every $r \in [0,1]$, let $\mathbf{q}_r \in \mathbb{P}(\mathcal{A})$ be the lottery giving probability r to a_1 and probability $(1-r)$ to a_0. Then there exists a value $r_0 \in [0,1]$ such that

$$\begin{aligned} \Big(r \leq r_0\Big) \quad &\Longrightarrow \quad \Big(\mathbf{q}_r \preceq \mathbb{1}_b\Big); \\ \Big(r = r_0\Big) \quad &\Longrightarrow \quad \Big(\mathbf{q}_r \approx \mathbb{1}_b\Big); \\ \text{and} \quad \Big(r \geq r_0\Big) \quad &\Longrightarrow \quad \Big(\mathbf{q}_r \succeq \mathbb{1}_b\Big). \end{aligned}$$

Furthermore, if $a_0 \prec a_1$, then there is a *unique* $r_0 \in [0,1]$ with these properties.

For example, suppose alternative a_0 represents having \$99.00, alternative b represents having \$100.00, and alternative a_1 represents having \$1,000,099.00. Suppose Zara has \$100.00, and is considering buying a \$1.00 ticket to a lottery with a \$1,000,000 jackpot. The three alternatives then mean the following:

[1] This is sometimes called the *Archimedean* axiom, or the axiom of *Substitutability*.

- a_0: Buy a ticket and lose the lottery; Zara is left with \$99.00.
- b: Don't buy a ticket; Zara is left with \$100.00.
- a_1: Buy a ticket and win; Zara is left with \$1,000,099.00.

If the odds of winning are too low (i.e. $r < r_0$), then Zara considers it a waste of money to buy a ticket. If the odds of winning are high enough (i.e. $r > r_0$), then she considers it a good bet, so she will by a ticket. In between these extremes, there is a critical probability (i.e. $r = r_0$), where Zara can't decide whether or not to buy a ticket. The exact value of r_0 depends upon how much utility Zara assigns to having different amounts of money (i.e. how much she fears bankruptcy, how greedy she is to be rich, etc.)

Clearly, axioms **(T)**, **(L)** and **(C)** are reasonable to expect from any rational person. The next theorem says, loosely speaking, *Any rational gambler has a utility function.*

Theorem 3A.1 (von Neumann and Morgenstern) *Suppose $\succeq$ is a relation on $\mathbb{P}(\mathcal{A})$ satisfying axioms* **(T)**, **(L)**, *and* **(C)**. *Then:*

(a) *There exists a cardinal utility function $U : \mathcal{A} \longrightarrow \mathbb{R}$ so that, if $\mathbf{p}, \mathbf{p}' \in \mathbb{P}(\mathcal{A})$ are two lotteries, then*

$$\Big(\mathbf{p} \succeq \mathbf{p}'\Big) \iff \Big(\mathbb{E}(U, \mathbf{p}) \;\geq\; \mathbb{E}(U, \mathbf{p}')\Big), \tag{3.2}$$

where $\mathbb{E}(U, \mathbf{p})$ is the expected utility defined by eqn.(3.1).

(b) *Furthermore, the function U is unique up to affine transformation. That is, if $\widetilde{U} : \mathcal{A} \longrightarrow \mathbb{R}$ is another function satisfying eqn.*(3.2), *then there exist constants $m > 0$ and $b \in \mathbb{R}$ so that $\widetilde{U} \;=\; m \cdot U + b$.* □

Proof: **(a)** Without loss of generality, suppose $\mathcal{A} = \{0, 1, \ldots, N\}$. Thus, for any $a \in \mathcal{A}$, $\mathbb{1}_a = (0, 0, \ldots, 1, 0, \ldots, 0)$ where the 1 appears in the ath coordinate (i.e. the lottery which awards a with probability 1). By reordering $\mathcal{A}$ if necessary, we can assume that a_0 and a_1 be 'worst' and 'best' elements in $\mathcal{A}$, respectively. That is, for every $a \in \mathcal{A}$, we have $a_0 \preceq a \preceq a_1$.

If $a_0 \approx a_1$, then we must have $a_0 \approx a \approx a_1$ for all $a \in \mathcal{A}$ [by axiom **(T)**]. Thus, if $U : \mathcal{A} \longrightarrow \mathbb{R}$ is *any* constant function, then equation (3.2) is satisfied, and we're done.

Thus, we assume $a_0 \prec a_1$. Define $U(a_1) := 1$ and $U(a_0) := 0$. Then define $U(a) := r$, where $r \in [0, 1]$ is the unique value (which exists, by axiom **(C)**) such that, if $\mathbf{q}_r$ is the lottery giving probability r to a_1 and probability $(1 - r)$ to a_0, then $\mathbf{q}_r \approx \mathbb{1}_a$.

Claim 1: *Let $\mathbf{p}_0$, $\mathbf{p}_1$, $\mathbf{p}'_0$ and $\mathbf{p}'_1$ be any four lotteries. For any $r \in [0, 1]$ define $\mathbf{p}_r$ and $\mathbf{p}'_r$ as in axiom* **(L)**. *Then* $\Big(\mathbf{p}_0 \approx \mathbf{p}'_0 \text{ and } \mathbf{p}_1 \approx \mathbf{p}'_1\Big) \Longrightarrow \Big(\forall\, r \in [0, 1], \quad \mathbf{p}_r \approx \mathbf{p}'_r\Big)$.

Proof: **Exercise 3.1** Hint: use axiom **(L)**. ◇ Claim 1

Claim 2: *For any $u_1, u_2 \in [0,1]$ and any $p_1, p_2 \in [0,1]$ with $p_1 + p_2 = 1$, $p_1\mathbf{q}_{u_1} + p_2\mathbf{q}_{u_2} = \mathbf{q}_{p_1u_1+p_2u_2}$.*

Proof: **Exercise 3.2** ◇ Claim 2

Now, suppose $\mathbf{p} := (p_0, p_1, \ldots, p_N)$. Let $P := p_0 + p_1$ and let $r = p_1/P$.

Claim 3: $\mathbb{E}(U, \mathbf{p}) = Pr + p_2U(2) + p_3U(3) + \cdots + p_NU(N)$.

Proof: **Exercise 3.3** ◇ Claim 3

Claim 4: *Let $E := \mathbb{E}(U, \mathbf{p})$; then $\mathbf{p} \approx \mathbf{q}_E$.*

Proof: **Exercise 3.4** (Hint: Apply Claims 1 and 2 each N times over, and then use Claim 3.) ◇ Claim 4

By identical reasoning, if $E' := \mathbb{E}(U, \mathbf{p}')$, then $\mathbf{p}' \approx \mathbf{q}_{E'}$.

Claim 5: $\Big(\mathbf{p} \preceq \mathbf{p}'\Big) \iff \Big(\mathbf{q}_E \preceq \mathbf{q}_{E'}\Big)$

Proof: **Exercise 3.5** (Hint: Use Claim 4.) ◇ Claim 5

Claim 6: $\mathbb{1}_{a_1} \succeq \mathbf{q}_{E'}$.

Proof: **Exercise 3.6** (Hint: Use Axiom **(L)**.) ◇ Claim 6

Claim 7: $\Big(\mathbf{q}_E \preceq \mathbf{q}_{E'}\Big) \underset{(\dagger)}{\Longleftrightarrow} \Big(E \leq E'\Big)$.

Proof: **Exercise 3.7** (Hint: Use Claim 6 and Axiom **(L)**.) ◇ Claim 7

Thus, $\Big(\mathbf{p} \preceq \mathbf{p}'\Big) \underset{(*)}{\Longleftrightarrow} \Big(\mathbf{q}_E \preceq \mathbf{q}_{E'}\Big) \underset{(\dagger)}{\Longleftrightarrow} \Big(E \leq E'\Big) \iff \Big(\mathbb{E}(U, \mathbf{p}) \leq \mathbb{E}(U, \mathbf{p}')\Big)$, where $(*)$ is by Claim 5 and $(\dagger)$ is by Claim 7. This completes the proof.

(b) Suppose $\widetilde{U}$ is some other function satisfying eqn.(3.2). First note that $\widetilde{U}(a_1)$ must be maximal and $\widetilde{U}(a_0)$ must be minimal over all alternatives in $\mathcal{A}$. That is, for all $a \in \mathcal{A}$, we must have $\widetilde{U}(a_1) \geq \widetilde{U}(a) \geq \widetilde{U}(a_0)$. Define $m := \widetilde{U}(a_1) - \widetilde{U}(a_0)$ and define $b := \widetilde{U}(a_0)$. It is **Exercise 3.8** to verify that $\widetilde{U} = mU + b$. □

Exercise 3.9 One advantage of the 'gambling' interpretation of utility is that it provides an empirical method to *measure* the utility functions of a real person, by performing an experiment where she chooses between various lotteries. Each choice she makes is a data point, and given sufficient data, we can solve a system of linear equations to figure out what her utility function is.

1. Design an experimental protocol to measure the cardinal utilities of three alternatives $\overline{a}, b, \underline{a}$. For simplicity, assume $\overline{a} \succeq b \succeq \underline{a}$ and set $U(\overline{a}) = 1$ and $U(\underline{a}) = 0$. The problem is to find the value of $U(b) \in [0,1]$.

2. Use this protocol in a real experiment with a friend. See if you can measure her cardinal utilities for three alternatives (eg. three restaurants, three movies, etc.)

Exercise 3.10 We say Zara is *risk averse* (or has *concave utility*) if the function U_0 is *concave*, meaning that,

$$\text{For any } x, y \in \mathbb{R}_{\neq}, \qquad \frac{U_0(x) + U_0(y)}{2} \quad \leq \quad U_0\left(\frac{x+y}{2}\right). \tag{3.3}$$

We say that Zara is *strictly risk averse* if U is *strictly concave*, meaning that the "$\leq$" in eqn.(3.3) is always a "$<$". It is pretty much universally assumed in economics that humans are risk-averse with respect to resources like money. Intuitively, this means that a poor person values an single additional dollar more than a rich person does.

(a) Show that the function $U(x) = x^\alpha$ is concave if and only if $0 \leq \alpha \leq 1$, and that U is *strictly* concave iff $0 < \alpha < 1$.

(b) Show that $U(x) = \log(x)$ is strictly concave[2] on $\mathbb{R}_+$.

(c) Suppose the functions U_1 and U_2 are (strictly) concave.

(i) Show that $U_1 + U_2$ is (strictly) concave.

(ii) Show that $U_1 \circ U_2$ is (strictly) concave.

(iii) If $r > 0$, show that the function rU_1 is (strictly) concave.

(d) Suppose that U is twice-differentiable. Show that U is concave if and only if $U''(x) \leq 0$ for all $x \in \mathbb{R}_+$, and that U is *strictly* concave if $U''(x) < 0$ for all $x \in \mathbb{R}_+$.

3B The problem of interpersonal comparisons

Prerequisites: §3A

Now that we have a clear notion of utility, it seems clear that true 'democracy' means that collective social decisions should be made so as to maximize the utilities of all voters. Let $\mathcal{A}$ be a set of alternatives and let $\mathcal{I}$ be a set of individuals. Suppose each individual $i \in \mathcal{I}$ has some cardinal utility function $u_i : \mathcal{A} \longrightarrow \mathbb{R}$. We define the *collective utility function* $\overline{U} : \mathcal{A} \longrightarrow \mathbb{R}$ by:

$$\overline{U}(a) \quad = \quad \sum_{i \in \mathcal{I}} u_i(a), \quad \text{for all } a \in \mathcal{A}. \tag{3.4}$$

The *Utilitarian Procedure* then states

(U) *Society should choose the alternative which maximizes the collective utility function* $\overline{U}$.

[2] Indeed, Daniel Bernoulli first proposed that people had logarithmic utility for money in 1738, as a resolution of the *St. Petersburg Paradox*.

Utilitarianism has several philosophically appealing mathematical properties, such as those given by Harsanyi's *Impartial Observer Theorem* [Har53] and *Social Aggregation Theorem* [Har55b]. It has also been characterized as the only social welfare function satisfying several combinations of axioms encoding 'fairness' and 'rationality'; see [dG77, Mas78, Mye81, Ng75, Ng85, Ng00].

However, Utilitarianism suffers from a critical flaw: the *Problem of interpersonal comparisons of utility*. To illustrate the problem, suppose there are two voters, Zara and Owen, with utility functions u_0 and u_1. Then the collective utility function is defined:

$$\overline{U}(a) \quad = \quad u_0(a) \; + \; u_1(a), \quad \text{for all } a \in \mathcal{A}.$$

Suppose $\mathcal{A} = \{a, b, c\}$, and we have the following utilities:

	u_0	u_1	$\overline{U}$
a	0	1	**1**
b	1	-1	0
c	-1	0	-1

Clearly, society should choose alternative a, which maximizes the collective utility with $\overline{U}(a) = 1$. Now suppose we have the following utilities:

	u_0'	u_1	$\overline{U}$
a	0	1	1
b	10	-1	**9**
c	-10	0	-10

In this case, society should chose alternative b. However, the von Neumann-Morgenstern Theorem (Thm.3A.1) says that the utility function u_0 is only well-defined up to affine transformation, and it is clear that $u_0' = 10 \cdot u_0$. Thus, using the von Neumann-Morgenstern 'gambling' definition of utility, *there is no way to determine whether Zara has utility function u_0 or u_0'.* The two utility functions will produce identical 'gambling' behaviour in Zara, but they clearly yield different outcomes in the collective social choice.

The problem is that there is no natural 'unit' we can use to compare Zara's utility measurements to Owen's. Put another way: there is no way we can empirically determine that Zara prefers alternative b to a 'ten times as much' as Owen prefers a to b. Indeed, in terms of the von Neumann-Morgenstern definition of utility, it doesn't even *make sense* to say such a thing.

This ambiguity makes the 'collective utility' concept meaningless, and worse, subject to manipulation through exaggeration. For example, Owen can regain control of the social choice by exaggerating his preferences as follows:

	u_0	u_1	$\overline{U}$
a	0	1000	**1000**
b	10	-1000	-900
c	-10	0	-10

Zara can then retaliate, and so on. Clearly this rapidly gets ridiculous.

3C Relative Utilitarianism

Prerequisites: §3B

One solution to the problem of interpersonal comparisons is to insist that everyone rescale their personal utility function so that its range lies in a certain compact interval. Typically, all utilities are rescaled to range over a unit interval (e.g. from zero to one). In other words, for all $i \in \mathcal{I}$, we define $r_i := \max_{a\in\mathcal{A}} u_i(a) - \min_{a\in\mathcal{A}} u_i(a)$. We then substitute $\widetilde{U}_i := U_i/r_i$ in eqn.(3.4). This version of utilitarianism has been called *Relative Utilitarianism* (RU), and admits several appealing axiomatic characterizations [Cao82, Dhi98, Kar98, DM99, Seg00].

However, RU is still susceptible to strategic misrepresentation of preferences. The scope for exaggeration of utilities is limited, but if the electorate is large, then each voter might try to maximize the influence of her vote by declaring a value of 'one' for all the alternatives she finds acceptable, and value 'zero' to all the alternatives she finds unacceptable (especially on a hard-fought issue). In this case RU devolves into the 'approval voting' [see Example 2C.1(c)]. Approval voting has many nice properties, but it does not satisfy the same axiomatic characterizations as RU. Furthermore, approval voting is an 'ordinal' voting system, so the Gibbard-Satterthwaite Impossibility Theorem [see §2F] makes it susceptible to further forms of strategic voting.

If the space $\mathcal{A}$ of alternatives is a convex set of feasible allocations of economic resources, then Sobel [Sob01] has shown that the set of Nash equilibria of the resulting 'utility misrepresentation game' for RU contains the set of Walrasian equilibria of a pure exchange economy over these resources with equal initial endowments. However, the misrepresentation game also admits non-Walrasian Nash equilibria which are not even Pareto efficient.

3D The Groves-Clarke Pivotal Mechanism

Prerequisites: §3B

The *Groves-Clarke Pivotal Mechanism* (GCPM) is a hybrid between a referendum and an auction:

1. Each voter i assigns a monetary *valuation* $v_i(a)$ to each alternative $a \in \mathcal{A}$. We regard $v_i(a)$ as a proxy for the value of $u_i(a)$ in eqn.(3.4).

2. Society chooses the alternative $a \in \mathcal{A}$ which maximizes the aggregate valuation:

$$V(a) \quad := \quad \sum_{i\in\mathcal{I}} v_i(a). \tag{3.5}$$

3. Suppose that voter i is *pivotal*, meaning that alternative a wins only because of i's vote. In other words, $V(a) - V(b) < v_i(a) - v_i(b)$, so if i had voted differently (i.e. given a

higher valuation to b and/or a lower one to a), then the alternative b would have won instead. Then voter i must pay a *Clarke tax* t_i defined:

$$t_i \quad := \quad \sum_{j \neq i} [v_j(b) - v_j(a)]. \tag{3.6}$$

Intuitively, $[v_j(b) - v_j(a)]$ is the 'net loss' in utility for voter j because society chose a instead of b; hence the Clarke tax t_i is the 'aggregate net loss' for everyone else besides i. Note that

$$\begin{aligned} t_i &= \sum_{j \neq i} v_j(b) - \sum_{j \neq i} v_j(a) \quad = \quad [V(b) - v_i(b)] - [V(a) - v_i(a)] \\ &= [v_i(a) - v_i(b)] - [V(a) - V(b)] \quad \leq \quad v_i(a) - v_i(b), \end{aligned}$$

(because $V(a) \geq V(b)$ by hypothesis). Thus, the Clarke tax never exceeds i's personal gain in obtaining a rather than b (assuming she expressed her preferences honestly); hence i should always be willing to pay the tax t_i in order to secure alternative a.

In most cases, the winning alternative will win by a margin of victory which far exceeds the valuation assigned by any single voter, so that step #3 will only rarely be implemented. However, in a very close electoral outcome, many voters may find themselves in the position of the 'swing' voter described in step #3 (i.e. each one could have single-handedly changed the outcome), and in these cases, all these voters must pay a Clarke tax.

Because of this possibility, each voter has a strong incentive to express her preferences honestly. If she understates her preference for a particular alternative, then she runs the risk that a less-preferred alternative may be chosen, even though she *could have* changed the outcome to her more preferred alternative had she voted honestly (and would have happily paid the resulting Clarke tax). Conversely, if she *overstates* her value for a particular alternative, then she risks paying more than it is worth for her to 'purchase' her preferred outcome. Thus, the GCPM acts as a kind of 'auction', where each valuation $v_i(a)$ functions not only as a 'vote', but also as a 'bid' for the option to change the referendum outcome. In most cases (e.g. landslide victories), this option will not be exercised, but in a close race, the option *will* be exercised, and the voter must pay her bid value. Just as in an ordinary auction, each voter neither wishes to 'underbid' (and risk unnecessary defeat) nor to 'overbid' (and risk paying too much). Her dominant strategy is always to bid honestly.

Formally, we can model the GCPM as a *Bayesian game*, in which each player $i \in \mathcal{I}$ has a (secret) utility function $u_i : \mathcal{O} \longrightarrow \mathbb{R}$ (where $\mathcal{O}$ is some set of *outcomes*), along with a *strategy set* $\mathcal{S}_i$, and the outcome of the game is determined by a function $o : \prod_{i \in \mathcal{I}} \mathcal{S}_i \longrightarrow \mathcal{O}$. Let $\mathcal{S}_{-i} := \prod_{j \in \mathcal{I} \setminus \{i\}} \mathcal{S}_j$, and regard o as a function $o : \mathcal{S}_i \times \mathcal{S}_{-i} \longrightarrow \mathcal{O}$. We say $s_i \in \mathcal{S}_i$ is a *dominant strategy* for player i if, for any $\mathbf{s}_{-i} \in \mathcal{S}_{-i}$,

$$u_i\left[o(s_i, \mathbf{s}_{-i})\right] \quad \geq \quad u_i\left[o(s_i', \mathbf{s}_{-i})\right], \quad \forall\ s_i' \in \mathcal{S}_i.$$

In other words, s_i is an optimal strategy for player i, given any possible choice of strategies for the other players.

Let $\mathcal{V} := \mathbb{R}^{\mathcal{A}} = \{v : \mathcal{A} \longrightarrow \mathbb{R}\}$ be the set of all monetary *valuations* of the alternatives in $\mathcal{A}$. Consider the Bayesian game where $\mathcal{S}_i = \mathcal{V}$ for all i (each player's strategy is to declare some valuation in $\mathcal{V}$), and where the outcome of the game is a choice of policy in $\mathcal{A}$, and some Clarke tax for each player, as determined by the GCPM. In other words, $\mathcal{O} := \mathcal{A} \times \mathbb{R}^{\mathcal{I}}$, and for any vector of valuations $\mathbf{v} = (v_1, \ldots, v_I) \in \prod_{i \in \mathcal{I}} \mathcal{S}_i$, $o(\mathbf{v}) := (a; \mathbf{t})$, where $a \in \mathcal{A}$ is the alternative with the highest total valuation, and $\mathbf{t} := (t_1, \ldots, t_I) \in \mathbb{R}^{\mathcal{I}}$ is the vector of Clarke taxes computed using eqn.(3.6). Suppose that (after perhaps multiplying by a constant), each voter's utility function has the *quasilinear* form

$$u_i(a, -t_i) \quad = \quad w_i(a) - t_i, \qquad \forall a \in \mathcal{A} \text{ and } \; t_i \in \mathbb{R}, \tag{3.7}$$

where $w_i : \mathcal{A} \longrightarrow \mathbb{R}$ is her utility function over the policy alternatives and t_i is the Clarke tax she must pay. Then it makes sense to say that $w_i(a)$ is the monetary *worth* which voter i assigns to alternative $a \in \mathcal{A}$. Given assumption (3.7), the GCPM is a *dominant strategy implementation* of utilitarianism in the following sense:

Theorem 3D.1 *Suppose all voters have quasilinear utility functions like eqn.*(3.7)*. Then for each $i \in \mathcal{I}$, a dominant strategy is to set $v_i := w_i$. In the resulting dominant strategy equilibrium, the GCPM chooses the same alternative as utilitarianism* (because then maximizing $V = \sum_{i \in \mathcal{I}} v_i$ is equivalent to maximizing $U = \sum_{i \in \mathcal{I}} w_i$)*.*

Proof: See Proposition 23.C.4 on p.877 of [MCWG95] or Lemma 8.1 on p.204 of [Mou88]. □

The GCPM also satisfies other appealing axiomatic characterizations [Mou86, Sjo91]. However, because it links voting to money, the GCPM has several major caveats:

Caveat #1. Theorem 3D.1 only holds if voters have quasilinear utility functions like eqn.(3.7). This is false. Real people are *risk-averse*, which means their utility is highly *concave* as a function of money. At the very least, we should assume utility functions have the 'quasiconcave' form

$$u_i(a, t_i) \quad = \quad w_i(a) + c(E_i + t_i), \qquad \forall a \in \mathcal{A} \text{ and } \; t_i \in \mathbb{R}, \tag{3.8}$$

where c is some concave function (e.g. $c = \log$) and E_i is the initial *endowment* of player i (i.e. her current assets, plus the expected present value of all future earnings). But this leads to further problems:

(a) If c is strictly concave, then the GCPM clearly assigns much more 'voting power' to rich people than poor people. A rich person i might easily be willing to bid \$100,000 to change the outcome of the election from a to b, whereas a poor person j would only bid \$100 to change it from b to a, even though $w_i(a) = w_j(b)$ and $w_i(b) = w_j(a)$.

(b) If c is nonlinear, then Theorem 3D.1 is false; indeed, a voter may not have *any* dominant strategy. For example, suppose $\mathcal{A} = \{a, b, c\}$, and

$$w_i(a) \quad = \quad 0 \quad < \quad w_i(b) \quad = \quad 2 \quad < \quad w_i(c) \quad = \quad 4.$$

Suppose c is a concave function such that $c(E_i) = 0$, $c(E_i - \$2) = -2$ and $c(E_i - \$3) = -4$. Thus, voter i would be willing to pay a \$2 Clarke tax to change outcome a to outcome b, and also \$2 to change outcome b to outcome c, but would only be willing to pay \$3 to change a to c. Suppose voter i declares valuations $v_i(a) = 0$ and $v_i(c) = 3$ (which is a truthful expression of her quasiconcave utility function with respect to a and c). What valuation should she declare for b? If she declares $v_i(b) < 2$, then she has 'undervalued' b versus a; if a ultimately wins by a margin of less than \$2 over b, then she will regret her choice. However, if she declares $v_i(b) > 1$, then she has 'overvalued' b versus c; if b ultimately wins by a margin of less than \$2 over c, then she will still regret her choice.

Suppose, then, that i declares $v_i(a) = 0$, $v_i(b) = 2$, and $v_i(c) = 4$; then she will be satisfied with any referendum outcome of a vs. b or b vs. c. But suppose c beats a by a margin between \$3 and \$4; then i will have to pay a Clarke tax greater than \$3, so once again she will regret her choice. In summary, there is *no* valuation of the alternatives $\{a, b, c\}$ which i will not regret under some circumstances. Her best strategy depends upon her expectations about how other people will vote. In other words, she has no dominant strategy.

In this situation, one or more Nash equilibria may still exist (some of which may even be truth-revealing). But the predictive relevance of a Nash equilibrium depends upon each voter making accurate predictions about the behaviour of every other voter, and in a 'voting game' involving millions of voters (e.g. a modern democracy) this is not very plausible.

(c) Like the quasilinear function (3.7), the quasiconcave function (3.8) 'solves' the problem of interpersonal utility comparison by implicitly assuming that all people have *identical* utility function c for money. This is false. Even if two people have the same initial endowment, their utility for money may differ. For example, a person with modest material needs (e.g. an ascetic monk) will assign less utility to each dollar than a hedonistic playboy. Hence we should assume each person's utility function has the form $u_i(a, t_i) = w_i(a) + c_i(E_i + t_i)$, where c_i is some concave function which may differ from person to person. This further confounds any interpretation of the aggregate monetary valuation of an alternative as its 'aggregate utility'.

Good [Goo77] has proposed a modified pivotal scheme which equalizes voting power between rich and poor or between ascetics and hedonists. Loosely speaking, we redefine $V(a) := \sum_{i \in \mathcal{I}} f_i[v_i(a)]$ in eqn.(3.5), where $f_i[t] := c_i[E_i] - c_i[E_i - t]$ measures the disutility of t lost dollars for voter i (for example, if the function c_i is linear with slope λ_i, then this simplifies to $V(a) := \sum_{i \in \mathcal{I}} \lambda_i v_i(a)$, where presumably the marginal utilities λ_i are smaller for rich people and larger for poor people). The problem, of course, is to

estimate the functions f_i; clearly each person has considerable incentive to misrepresent her marginal utility. Good proposes we use some standard function like $f_i(t) = t/E_i$, but this seems somewhat procrustean. Also, the proof that Good's mechanism is a dominant-strategy truthful implementation of utilitarianism still implicitly assumes that voter's utility functions are quasilinear, so it is vulnerable to Caveat #1(b).

Tideman [Tid97] has proposed that Clarke taxes be paid in *time* (spent, say, in community service) rather than money. This gives the poor the same *a priori* political power as the rich, but it is still far from egalitarian. Different people value their time very differently. The retired and the unemployed have a lot of spare time (and hence, presumably, assign a low marginal utility to this time), whereas working parents and jet-setting professionals have almost no time to spare.

(d) Even the individualized quasiconcave utility functions in #1(c) assume that each person's preferences over the alternatives in $\mathcal{A}$ are totally *separable* from her wealth level E_i. This is false. For example, rich people and poor people have very different preferences concerning redistributive taxation schemes and publicly funded goods.

Caveat #2. Any revenue collected by the Clarke tax must be removed from the economy (e.g. destroyed or donated to a faraway country), because otherwise voters who expect *not* to pay a Clarke tax have an incentive to distort their valuations so as to inflate the amount of revenue which is collected; see [Rik82, p.54] for example. Thus, the GCPM is never Pareto-efficient.

Caveat #3. As Riker [Rik82, p.56] notes, pivotal voting *cannot* be anonymous, because to implement the Clarke tax, we need a public record of each person's valuations of the alternatives. However, anonymity of voting is a crucial feature of modern democracy. Anonymity protects voters from discrimination and political extortion, and also prevents voters from selling their votes for material gain. The GCPM is clearly vulnerable to a scam where I pay a thousand people \$5 each to declare a valuation of \$100 for a particular outcome. If this outcome then wins by a 'landslide' margin of \$100,000 (or indeed, by any margin greater than \$100), then none of my accomplices needs to pay the Clarke tax (so they each profit \$5), and the total cost for me is only \$5,000 (which is much cheaper than personally paying a \$100,000 Clarke tax to swing the outcome in my favour).

3E Further Reading

Origins of Utilitarianism. Utilitarianism was first articulated by British political philosopher Jeremy Bentham (1748-1832), and was later elaborated by other utilitarian philosophers, most notably John Stuart Mill (1806-1873). These early approaches suffered from a fundamental problem: they took for granted that happiness (or 'utility') could be treated as a mathematical quantity, but they gave no concrete way of *quantifying* it. This problem was resolved by game

theorists John von Neumann and Oskar Morgenstern [vM47], who showed how to quantify utility using Theorem 3A.1. Good introductions to the von Neumann-Morgenstern approach are Luce and Raiffa [LR80, §2.4] or Riker [Rik82, §4F].

Interpersonal Comparisons. Some 'solutions' to the problem of interpersonal comparison of utility are Raiffa [Rai53], Braithwaite [Bra55], Goodman-Markowitz [GM52], and Hildreth [Hil53]; see sections 6.11 and 14.6 of Luce & Raiffa [LR80] for a summary. See Roemer [Roe98, §1.1] for a nice formalization of various degrees of interpersonal utility comparison using group theory.

The Groves-Clarke Pivotal Mechanism. The GCPM is a special case of the *demand-revealing mechanism* proposed by Groves [Gro73] and Clarke [Cla71], and later promoted by Tideman and Tullock [TT76]. The GCPM is extensively analyzed in the collection [Tid77] and the monograph [GL79]. See also §8.2 of [Mou88], §23.C of [MCWG95], §5 of [Tid97], and §8.1 of [Mue03]. Another special case of Groves' and Clarke's demand-revealing mechanism is the *Vickrey auction* [Vic61]; for this reason the demand-revealing mechanism is sometimes called the 'Vickrey-Groves-Clarke mechanism'.

Other point voting systems. Systems where citizens vote by allocating a budget of 'voting money' are at least a century old; the earliest known description is Charles Dodgson's (1873) 'Method of Marks' [Dod73, Bla58, Abe02]. Musgrave [Mus59] briefly sketched a system of 'point voting' [p.130-131], while Coleman [Col70] suggested that a currency of 'fungible votes' could supersede vote-trading just as money superseded barter [§III(b), p.1084]. 'Point voting' was also suggested by Mueller [Mue71, Mue73], and is implicit in 'probabilistic' voting schemes [Int73, Nit75], as well as in the 'Walrasian equilibrium' model of vote-trading [Mue67, Mue73, Phi71, Phi72, MPV72].

However, without some mechanism to encourage honesty, each voter will misrepresent her preferences [Dod73, Mue73, Mue77, Lai77, NPL80, Nit85]. For example, in allocating her voting money over the alternatives of a single ballot, each voter might simply pile all her money onto her most-preferred alternative amongst the subset of alternatives she considers most likely to win (in particular, she may not allocate *any* money towards her favourite alternative, if she considers it doomed to lose). Thus, her allocation will not accurately represent her utility function.

Allen [All77, All82] proposed a 'modified method of marks' (MMM), where, instead of allocating a fixed 'budget' of voting money, voters can give each alternative any numerical score within a certain range. Allen claimed his MMM was less susceptible to strategic misrepresentation than Dodgson's Method of Marks, but this was refuted by Hardin [Har82]. (Indeed, we argued in §3C that the MMM would in fact devolve into 'approval voting').

Hylland and Zeckhauser [HZ79] have proposed another 'point-based' voting system which truthfully reveals each voter's preferences for public goods. In the Hylland-Zeckhauser system, each voter has a budget of 'points' (or 'voting money') which she can allocate towards voting for various public expenditures. The amount of government money spent on each public expenditure is then proportional to the sum of the *square roots* of the point scores it receives from all voters; see [Mue03, §8.3, p.170] or [Tid97, §4] for more information.

Like the Groves-Clarke mechanism of §3D, the Hylland-Zeckhauser mechanism makes it

optimal for voters to truthfully reveal their preferences, and implements a utilitarian outcome. Furthermore, the Hylland-Zeckhauser mechanism relies on 'voting money' rather than real money, so it does not favour wealthy voters. However, the Hylland-Zeckhauser mechanism is only designed for allocating a finite budget of resources amongst various 'preapproved' public expenditures; it is not appropriate for making discrete, all-or-nothing choices between policies or between government candidates. Also, the Hylland-Zeckhauser mechanism relies on an iterative process (where voters repeatedly receive feedback and modify their votes), and it is not guaranteed that this iterative process will converge.

Part III

Bargaining and Arbitration

Chapter 4

Bargaining Theory

Prerequisites: §3B; Basic linear algebra

> *All government —indeed, every human benefit and enjoyment, every virtue and every prudent act —is founded on compromise and barter.* —Edmund Burke (1729-1797)

Bargaining and *arbitration* are forms of group decision-making which differ from voting in two ways:

- Each participant begins with an initial 'endowment' or 'bargaining position' which we call the *status quo*. If a mutually satisfactory agreement cannot be reached, than any participant may terminate negotiations and maintain her status quo. (However, if an agreement *is* reached, then all participants must honour it. In other words, contracts can be enforced.)
- Rather than a finite ballot of alternatives, there exists a *continuum* of possible outcomes (normally represented by a subset of $\mathbb{R}^N$). For example, when trading 'infinitely divisible' commodities such as money and oil, there exist a continuum of possible trading positions, each representing a particular quantity of oil purchased at a particular price.

In a *Bargaining* scenario, we regard the participants as playing an unrefereed 'game', and we then apply game theory to predict what sorts of outcomes are (im)possible or (un)likely, assuming all players are rational strategists. It is not important whether the outcome is 'fair'; only whether it is strategically realistic. In an *Arbitration* scenario, however, we assume the existence of an 'arbiter', who tries to propose a 'fair' outcome. The question then becomes: what is fair?

4A The von Neumann-Morgenstern Model

Let $\mathcal{I}$ be a set of two or more individuals who are bargaining. Almost all of the concepts and results we will present extend easily to groups of three or more bargainers (except for the theory

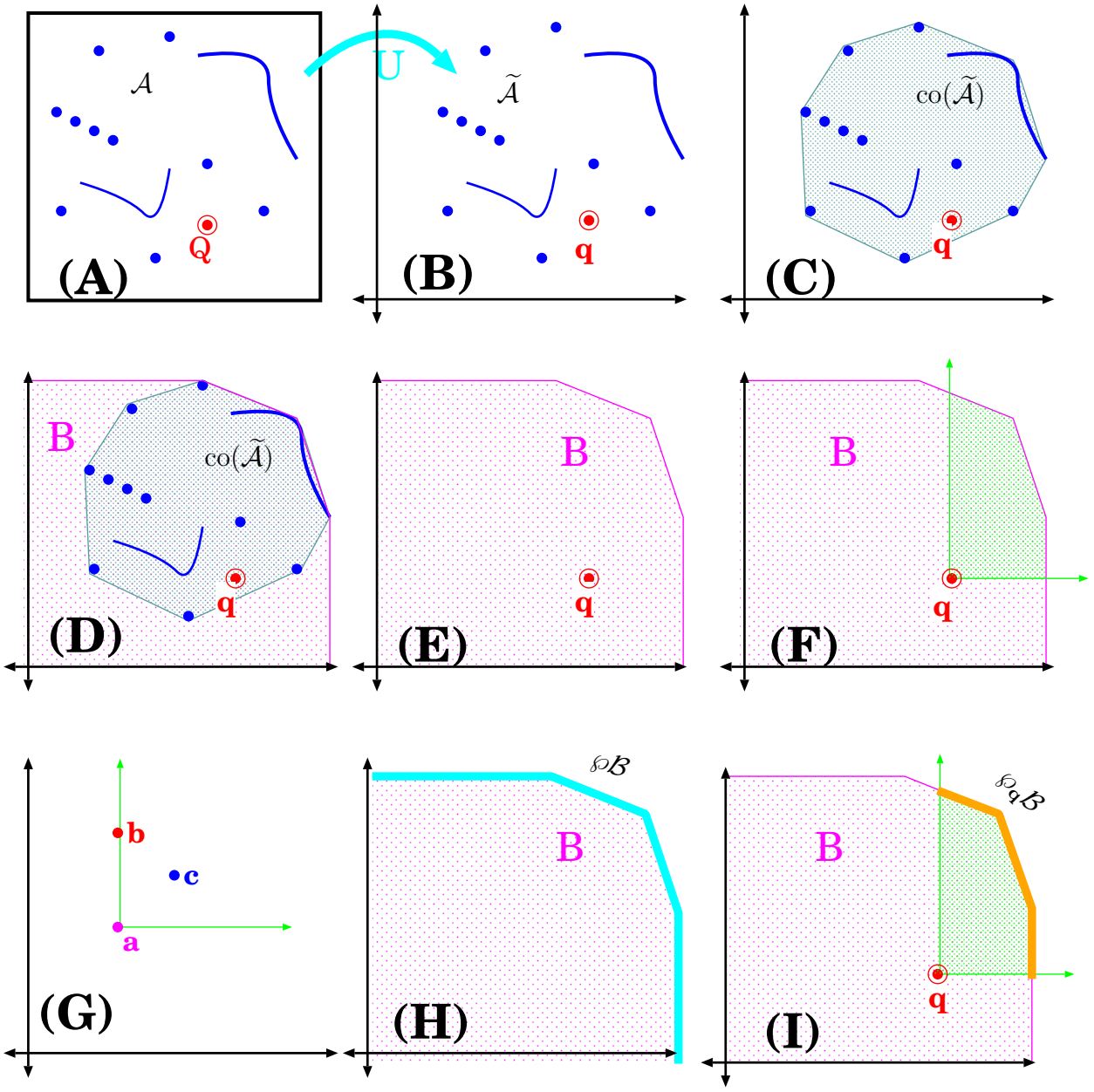


Figure 4.1: The von Neumann-Morgenstern bargaining model. **(A)** $\mathcal{A}$ is an abstract set of 'alternatives', and Q is the 'status quo' **(B)** $\widetilde{\mathcal{A}} \subset \mathbb{R}^2$ is the image of $\mathcal{A}$ under $U : \mathcal{A} \longrightarrow \mathbb{R}^2$, and $\mathbf{q} = U(Q)$.
(C) $\mathrm{co}(\widetilde{\mathcal{A}})$ is the *convex closure* of $\widetilde{\mathcal{A}}$. **(D)** $\mathcal{B}$ is the *comprehensive closure* of $\mathrm{co}(\widetilde{\mathcal{A}})$.
(E) We forget about $\widetilde{\mathcal{A}}$; the bargaining problem is determined by $\mathcal{B}$ and $\mathbf{q}$.
(F) The axiom **(MB)** says any viable bargain must be Pareto-preferred to the status quo. This restricts us to the green-shaded 'northeast' corner.
(G) **b** is Pareto-preferred to **a** because $b_0 \geq a_0$ and $b_1 \geq a_1$, while **c** is *strictly* Pareto prefered to **a** because $c_0 > a_0$ and $c_1 > a_1$. (Note that **b** is *not* strictly Pareto-preferred to **a**, because $b_0 = a_0$). Finally, neither of **b** and **c** is Pareto-prefered to the other (they are *Pareto-incomparable*).
(H) The axiom **(P)** says any viable bargain must be Pareto-optimal. This restricts us to the *Pareto frontier* $\wp\mathcal{B}$. **(I)** Axioms **(MB)** and **(P)** together restrict us to the *negotiating set* $\wp_{\mathbf{q}}\mathcal{B}$.

of *Alternating Offers* games in sections 5F-5G). However, for simplicity of exposition, we will mostly present these results in the context of *bilateral* bargaining, when $|\mathcal{I}| = 2$ (the extensions of these results to $|\mathcal{I}| \geq 3$ appear as exercises throughout the text). Thus, unless otherwise noted, we will set $\mathcal{I} := \{0, 1\}$. Rather than referring to the bargainers as 'Player Zero' and 'Player One', we will give then names; Player Zero is named *Zara*, while Player One is named *Owen* (we presume the mnemonic is obvious).

Let $\mathcal{A}$ be a (possibly infinite) set of alternatives, which Zara and Owen can jointly choose amongst (Figure 4.1A). Each element of $\mathcal{A}$ represents an outcome for Zara and a corresponding outcome for Owen. For example:

- Suppose Zara and Owen are negotiating an exchange of several commodities (say, ale, beer, and cider), then each element of $\mathcal{A}$ allocates a specific amount of ale, beer, and cider to Zara, and a complementary amount of ale, beer, and cider to Owen.
- If Zara is a labour union and Owen is management, then each element of $\mathcal{A}$ corresponds to a labour contract with particular terms concerning wages, benefits, holiday time, job security, etc.
- If Zara and Owen are roommates, then each element of $\mathcal{A}$ corresponds to some agreement about how to split the housework responsibilities (e.g. dishwashing) and how to peacefully coexist (e.g. who gets to play music, entertain guests, or watch television at particular times).

We assume that Zara and Owen have cardinal utility functions, $U_0 : \mathcal{A} \longrightarrow \mathbb{R}$ and $U_1 : \mathcal{A} \longrightarrow \mathbb{R}$. By adding some large enough constant to the utility function of each player, we can assume that $U_0 : \mathcal{A} \longrightarrow \mathbb{R}_{\neq}$ and $U_1 : \mathcal{A} \longrightarrow \mathbb{R}_{\neq}$, where $\mathbb{R}_{\neq} := \{r \in \mathbb{R} \,;\, r \geq 0\}$. Together, these determine a *joint utility function* $U : \mathcal{A} \longrightarrow \mathbb{R}^2_{\neq}$, where $U(A) = \Big(U_0(A),\ U_1(A)\Big)$ for each $A \in \mathcal{A}$. For the purposes of bargaining, the details of each alternative in $\mathcal{A}$ are unimportant; all that is really important is how much utility each alternative has for each bargainer. Thus, we can forget $\mathcal{A}$ and instead consider the image set $\widetilde{\mathcal{A}} = U(\mathcal{A}) \subset \mathbb{R}^2_{\neq}$ (Figure 4.1B). We treat the bargainers as negotiating over elements of $\widetilde{\mathcal{A}}$.

A *convex combination* of elements in $\widetilde{\mathcal{A}}$ is any linear combination:

$$\sum_{j=1}^{J} c_j \mathbf{a}_j$$

where $\mathbf{a}_1, \ldots, \mathbf{a}_J \in \widetilde{\mathcal{A}}$, and where $c_1, \ldots, c_j \in [0, 1]$ are coefficients such that $\sum_{j=1}^{J} c_j = 1$. This convex combination represents the 'average utility' obtained by 'mixing' several alternatives together. There are several ways in which we could 'mix' alternatives:

- Suppose $\mathcal{A}$ represents alternatives which could be shared over time. Then a convex combination represents a 'time-sharing' agreement.

For example, suppose Zara and Owen are roommates, and each element of $\mathcal{A}$ represents a division of household chores. Suppose A and B are two such divisions:

A: Zara washes dishes, Owen washes floors.

B: Zara washes floors, Owen washes dishes.

If $\mathbf{a} = U(A)$ and $\mathbf{b} = U(B)$, then the convex combination $\frac{1}{3}\mathbf{a} + \frac{2}{3}\mathbf{b}$ represents the 'time shared' outcome, 'one third of the time, Zara washes the dishes and Owen washes the floor; two thirds of the time, Zara washes the floor and Owen washes the dishes'.

- Suppose $\mathcal{A}$ is a set of *quantitative* alternatives (e.g. a division of goods or money). Then convex combinations represent compromises obtained by 'splitting the difference' between alternatives in $\mathcal{A}$.

  For example, suppose Zara is a worker and Owen is an employer. Suppose A and B are two possible contracts:

  A: Wage of \$10.00/hour, and 21 days paid vacation

  B: Wage of \$12.00/hour, and only 7 days paid vacation.

  If $\mathbf{a} = U(A)$ and $\mathbf{b} = U(B)$, then the convex combination $\frac{1}{2}\mathbf{a} + \frac{1}{2}\mathbf{b}$ might[1] represent a compromise contract with a wage of \$11.00/hour and 14 days paid vacation.

- If Zara and Owen are willing to gamble over the outcome, then a convex combination represents a lottery. If $\mathbf{a}_j = U(A_j)$ for all $j \in [1..J]$, then the convex combination $\sum_{j=1}^{J} c_j \mathbf{a}_j$ represents the *expected utility* of the lottery where alternative A_j has probability c_j of occurring.

We assume that it is possible to convex-combine the alternatives in $\widetilde{\mathcal{A}}$ in some fashion; thus, it is possible for the negotiators to realize any point in the *convex closure* of $\widetilde{\mathcal{A}}$, which is the set $\mathrm{co}(\widetilde{\mathcal{A}})$ shown in Figure 4.1(C), defined:

$$\mathrm{co}(\widetilde{\mathcal{A}}) \quad := \quad \overline{\left\{ \sum_{j=1}^{J} c_j \mathbf{a}_j \;;\; \mathbf{a}_1, \ldots, \mathbf{a}_J \in \widetilde{\mathcal{A}};\; c_1, \ldots, c_j \in [0,1];\; \sum_{j=1}^{J} c_j \;=\; 1 \right\}}$$

Thus, $\mathrm{co}(\widetilde{\mathcal{A}})$ is the set of all joint utilities which are realizable through some kind of timesharing, compromise, or lottery. Note that, if $\widetilde{\mathcal{A}}$ is bounded, then $\mathrm{co}(\widetilde{\mathcal{A}})$ is a convex, compact subset of of $\mathbb{R}^2_{\swarrow}$. "*Compact*" means that the set $\mathrm{co}(\widetilde{\mathcal{A}})$ is *bounded* (which means there is an upper bound on how happy any feasible bargain could make either player) and *closed* (which means that if a sequence of feasible bargains tends to some limit, then the limit is also a feasible bargain).

[1] I say 'might', because this assumes that the utilities of worker and employer are 'linear' in the variables 'wage' and 'holiday time', and this assumption is generally not true.

Exercise 4.1: Let $\mathcal{C} \subset \mathbb{R}^2$. Show that the following are equivalent:

(a) $\mathcal{C}$ is convex.

(b) If $\mathbf{c}$ and $\mathbf{c}'$ are in $\mathcal{C}$, then $\mathcal{C}$ also contains the entire line segment from $\mathbf{c}$ to $\mathbf{c}'$.

(c) If $\mathbf{c}_1, \ldots, \mathbf{c}_n \in \mathcal{C}$ and $r_1, \ldots, r_n \in [0,1]$ are such that $r_1 + r_2 + \cdots + r_n = 1$, then the point $r_1\mathbf{c}_1 + \cdots + r_n\mathbf{c}_n$ is also in $\mathcal{C}$.

(d) If $\mathbf{c}_1, \ldots, \mathbf{c}_n \in \mathcal{C}$, then $\mathcal{C}$ contains the polygon with vertices at $\mathbf{c}_1, \ldots, \mathbf{c}_n$.

A subset $\mathcal{C} \subset \mathbb{R}^2_{\neq}$ is *comprehensive* if, for all $\mathbf{c} \in \mathcal{C}$ and $\mathbf{b} \in \mathbb{R}^2_{\neq}$, if $b_0 \leq c_0$ and $b_1 \leq c_1$, then $\mathbf{b} \in \mathcal{C}$ also. Geometrically speaking, the set $\mathcal{C}$ is comprehensive if it contains the entire 'cone' which extends out of each point into the southwest quadrant. If we interpret $\mathbf{c}$ as a feasible utility allocation, then this means that it is always possible for the players to deliberately reduce their own utility (say, by burning some money) to move from $\mathbf{c}$ to $\mathbf{b}$; thus, if $\mathbf{c}$ is feasible, then so is $\mathbf{b}$ (this is sometimes described as *free disposal* of utility). Of course, rational people would never actually *do* this, so it will make absolutely no difference to the behaviour of our rational bargainers if we include points like $\mathbf{b}$. For technical reasons, however, it is often convenient to assume that the set of feasible utility allocations is comprehensive. We therefore define the *bargaining set* $\mathcal{B}$ to be the *comprehensive closure* of $\text{co}(\widetilde{\mathcal{A}})$; this is the smallest comprehensive set containing $\text{co}(\widetilde{\mathcal{A}})$, as shown in Figure 4.1(D). Note that $\mathcal{B}$ is also convex and compact. We assume that the bargainers can choose any point in $\mathcal{B}$. This means we can assume the axiom of *Three C's:*

(CCC) *The bargaining set $\mathcal{B}$ is always a convex, compact, comprehensive subset of $\mathbb{R}^2_{\neq}$.*

Next, we assume that each player begins with an initial *endowment* or *bargaining position*; these endowments determine the *status quo* alternative $Q \in \mathcal{A}$. If the players cannot come to a mutually agreeable arrangement, then either one can terminate negotiations and both will receive the 'status quo' outcome. For example

- If Zara and Owen are potential roommates, then Q means they cannot come to a satisfactory cohabitation agreement, and thus, they do not become roommates.
- If a labour union and an employer are negotiating a new contract, then Q means 'no contract'. If the union chooses Q, this is tantamount to a strike; if the employer chooses Q, this is tantamount to a lock-out.

Let $\mathbf{q} = U(Q) \in \mathbb{R}^2_{\neq}$, and suppose $\mathbf{q} = (q_0, q_1)$, where q_0 is the utility of Q for Zara, and q_1 is the utility of Q for Owen. If $\mathbf{b} \in \mathcal{B}$, and $\mathbf{b} = (b_0, b_1)$, then clearly, $\mathbf{b}$ is acceptable to Zara only if $b_0 \geq q_0$. Likewise, $\mathbf{b}$ is acceptable to Owen only if $b_1 \geq q_1$. In other words, a bargain will only occur if it is mutually beneficial. We therefore assume the axiom of *Mutual Benefit*:

(MB) *An element $\mathbf{b} \in \mathcal{B}$ is an acceptable bargain only if $b_0 \geq q_0$ and $b_1 \geq q_1$.* (Figure 4.1F).

(Sometimes this axiom is called *Individual Rationality*, because it means each individual would be rational to choose the bargain over the status quo. Also, this axiom is sometimes called *No Coercion*, because it means that neither part can be 'forced' to accept a bargain inferior to her status quo utility.)

If $\mathbf{a}, \mathbf{b} \in \mathcal{B}$, then we say $\mathbf{b}$ is *Pareto preferred* to $\mathbf{a}$ if $a_0 \leq b_0$ and $a_1 \leq b_1$ —in other words, both Zara and Owen agree that $\mathbf{b}$ is 'no worse' than $\mathbf{a}$ (and it is perhaps better for at least one of them); see Figure 4.1(G). We then write $\mathbf{a} \overset{\wp}{\preceq} \mathbf{b}$. Thus, the axiom **(MB)** can be rephrased:

$\mathbf{b}$ *is an acceptable bargain only if* $\mathbf{q} \overset{\wp}{\preceq} \mathbf{b}$.

If $\mathbf{a}, \mathbf{c} \in \mathcal{B}$, then we say $\mathbf{c}$ is *strictly Pareto preferred* to $\mathbf{a}$ if $a_0 < c_0$ and $a_1 < c_1$ —in other words, both Zara and Owen agree that $\mathbf{c}$ is strictly *better* than $\mathbf{a}$; see Figure 4.1(G). We then write $\mathbf{a} \overset{\wp}{\prec} \mathbf{c}$. Clearly, if $\mathbf{c}$ is strictly Pareto-preferred to $\mathbf{a}$, then Zara and Owen will never choose $\mathbf{a}$ if they could instead choose $\mathbf{c}$. We say that a point $\mathbf{b} \in \mathcal{B}$ is *Pareto optimal* (or *Pareto efficient*) in $\mathcal{B}$ if there exists no $\mathbf{c} \in \mathcal{B}$ with $\mathbf{b} \overset{\wp}{\prec} \mathbf{c}$. This means: if $\mathbf{c}$ is any other point with $c_0 > b_0$, then we must have $c_1 \leq b_1$ (and vice versa). In other words, starting from $\mathbf{b}$, it is not possible to simultaneously make both players better off.

The *Pareto frontier* of $\mathcal{B}$ is the set $\wp\mathcal{B}$ of Pareto-optimal points in $\mathcal{B}$ (Figure 4.1H). If $\mathcal{B}$ is a comprehensive domain in $\mathbb{R}^2_{\nearrow}$, then $\wp\mathcal{B}$ is the 'northeast' frontier of this domain. Clearly, Zara and Owen will only agree to a Pareto-optimal bargain. We therefore assume *Pareto Optimality:*

(P) *An element* $\mathbf{b} \in \mathcal{B}$ *is an acceptable bargain only if* $\mathbf{b} \in \wp\mathcal{B}$ *—ie.* $\mathbf{b}$ *is Pareto-optimal.*

Lemma 4A.1 *Let* $\mathcal{B} \subset \mathbb{R}^2_{\nearrow}$. *Then*

$$\Big(\mathcal{B} \textit{ satisfies } \textbf{(CCC)}\Big) \iff \begin{pmatrix} \wp\mathcal{B} \textit{ is the graph of a continuous, nonincreasing} \\ \textit{function } \Gamma_1 : [0, M_0] \longrightarrow \mathbb{R}_{\nearrow}, \textit{ for some } M_0 > 0. \end{pmatrix}$$

Proof: **Exercise 4.2** □

Interpretation: For any $b_0 \in [0, M_0]$, $\Gamma_1(b_0)$ represents the most utility that Owen can get, given that Zara is getting b_0. Likewise, if $M_1 = \Gamma_1(0)$, and $\Gamma_0 := \overline{b}_1^{-1} : [0, M_1] \longrightarrow [0, M_0]$, then for any $b_1 \in [0, M_1]$, $\Gamma_0(b_1)$ represents the most utility that Zara can get, given that Owen is getting b_1.

Example 4A.2: In a *surplus division problem*, Zara and Owen must agree on how to divide a fixed quantity of some resource. If they cannot come to an agreement, then no one gets anything. (In the standard example, they must divide a dollar between them.)

Suppose there is one unit of resource to be divided. Let $U_0, U_1 : [0, 1] \longrightarrow \mathbb{R}_{\nearrow}$ be two nondecreasing functions, so that $U_0(x)$ is the utility which Zara obtains from quantity x of the resource, and $U_1(x)$ is the quantity which Owen obtains. This results in the feasible set

$$\mathcal{B} \;:=\; \{[U_0(x_0),\; U_1(x_1)] \;;\; x_0, x_1 \in [0, 1] \text{ and } \; 0 \leq x_0 + x_1 \leq 1\}. \tag{4.1}$$

In this case, for any $x \in [0,1]$, we have $\Gamma_1(x) = U_1\left[1 - U_0^{-1}(x)\right]$ and $\Gamma_0(x) = U_0\left[1 - U_1^{-1}(x)\right]$ (**Exercise 4.3**). ?

Thus, $\wp\mathcal{B}$ is the graph of Γ_1. That is: $\wp\mathcal{B} := \{[b_0, \Gamma_1(b_0)] \; ; \; b_0 \in \mathbb{R}_{\not\infty}\}$.

Also, $\mathbf{q} := [U_0(0), \; U_1(0)]$. Thus, $\wp_{\mathbf{q}}\mathcal{B} := \{(b_0, \Gamma_1(b_0)) \; ; \; b_0 \geq U_0(0) \text{ and } \Gamma_1(b_0) \geq U_1(0)\}$. ♢

As we've seen, any bargaining problem can be represented by a pair $(\mathcal{B}, \mathbf{q})$, where $\mathcal{B} \subset \mathbb{R}^2_{\not\infty}$ is some convex, compact, comprehensive *bargaining set*, and $\mathbf{q} \in \mathcal{B}$ is some *status quo* point. We will thus refer to the ordered pair $(\mathcal{B}, \mathbf{q})$ as a *bargaining problem* (Figure 4.1E). Given a bargaining problem $(\mathcal{B}, \mathbf{q})$, the von Neumann-Morgenstern *negotiating set* is the set of bargains satisfying axioms **(MB)** and **(P)**:

$$\wp_{\mathbf{q}}\mathcal{B} \quad := \quad \left\{ \mathbf{b} \in \wp\mathcal{B} \; ; \; \mathbf{b} \overset{\wp}{\succeq} \mathbf{q} \right\} \qquad \text{[see Figure 4.1(I)].}$$

The above reasoning implies that any mutually agreeable outcome will always be an element of $\wp_{\mathbf{q}}\mathcal{B}$. If $\wp_{\mathbf{q}}\mathcal{B}$ contains a single point (namely, $\mathbf{q}$), then this is the unique bargain which is acceptable to both players. However, in general, $\wp_{\mathbf{q}}\mathcal{B}$ contains many points. Clearly, Zara prefers the easternmost point on $\wp_{\mathbf{q}}\mathcal{B}$, while Owen prefers the northernmost point in $\wp_{\mathbf{q}}\mathcal{B}$. These goals are incompatible, and the bargaining problem comes down to finding a 'fair' compromise between these extremes.

Let $\mathfrak{B}$ be the set of all bargaining problems satisfying axiom **(CCC)**. That is,

$$\mathfrak{B} \quad := \quad \left\{ (\mathcal{B}, \mathbf{q}) \; ; \; \mathcal{B} \subset \mathbb{R}^2_{\not\infty} \text{ is convex, compact, and comprehensive, and } \mathbf{q} \in \mathcal{B} \right\}.$$

A *bargaining solution* (or *arbitration scheme*) is a function $\alpha : \mathfrak{B} \longrightarrow \mathbb{R}^2_{\not\infty}$, which takes any bargaining problem $(\mathcal{B}, \mathbf{q})$ as input, and yields, as output, a unique point $\alpha(\mathcal{B}, \mathbf{q}) \in \wp_{\mathbf{q}}\mathcal{B}$. Note that $\alpha(\mathcal{B}, \mathbf{q})$ must satisfy axioms **(MB)** and **(P)**, by definition. Bargaining solutions have two (quite different) interpretations:

Normative: If we are concerned with questions of 'fairness' and 'justice', then $\alpha(\mathcal{B}, \mathbf{q})$ should suggest a 'fair' compromise between the competing claims of the bargainers. In this case, the bargaining solution should satisfy certain axioms encoding our notions of 'fairness', 'rationality', 'consistency', 'impartiality', etc.[2]

Descriptive: If we approach bargaining as a strategic confrontation (i.e. a 'game'), then a $\alpha(\mathcal{B}, \mathbf{q})$ should predict the 'inevitable' outcome of this bargaining game, given the strategic positions of the players. In this case, the bargaining solution should satisfy certain axioms consistent with our assumption that the players are rational utility maximizers.[3]

[2] Do not confuse *normative* with *prescriptive*; see page 89.

[3] The *descriptive* interpretation is also sometimes called *predictive* or *positive*.

Exercise 4.4: Suppose that Zara and Owen are both *strictly risk averse*, (see Exercise 3.10). Show that the functions $\Gamma_1(x)$ and $\Gamma_0(x)$ from Example 4A.2 are both concave. Conclude that, in eqn.(4.1), the bargaining set $\mathcal{B}$ is strictly convex.

Exercise 4.5: Let $(\mathcal{B}, \mathbf{0})$ be any bargaining problem, where $\mathcal{B} \subset \mathbb{R}^2_{\nearrow}$ is a convex, compact, comprehensive set. Show that we can always 'represent' $(\mathcal{B}, \mathbf{0})$ as a surplus-division problem like Example 4A.2. In other words, there exist risk-averse utility functions $U_0, U_1 : \mathbb{R}_{\nearrow} \longrightarrow \mathbb{R}_{\nearrow}$ such that $\Gamma_1(x) = U_1\left[1 - U_0^{-1}(x)\right]$ and $\Gamma_0(x) = U_0\left[1 - U_1^{-1}(x)\right]$.

Exercise 4.6: We say that a point $\mathbf{b} \in \mathcal{B}$ is *strictly* Pareto optimal in $\mathcal{B}$ if there exists no $\mathbf{c} \in \mathcal{B}$ with $\mathbf{b} \overset{\wp}{\preceq} \mathbf{c}$. Thus, if $\mathbf{c}$ is any other point, and $c_0 > b_0$, then we must have $c_1 < b_1$ (and vice versa). In other words, starting from $\mathbf{b}$, it is not possible to make one player better off without making the other player strictly worse off. Clearly, if $\mathbf{b}$ is strictly Pareto optimal, then it is Pareto optimal. However, the converse is generally false.

(a) Find an example of a domain $\mathcal{B}$ and a point $\mathbf{b} \in \mathcal{B}$ which is Pareto optimal but not strictly Pareto optimal.

(b) What conditions must you impose on $\mathcal{B}$ to guarantee that all Pareto optimal points are *strictly* Pareto optimal?]

Exercise 4.7: For simplicity, we have formulated the ideas in this section for *bilateral* bargains (i.e. those involving only two parties). However, all of the ideas in this section can be formulated for *multilateral* bargains (i.e. those involving three or more parties).

(a) Reformulate axioms **(CCC)**, **(MB)** and **(P)** for bargains with three or more parties.

(b) Generalize the statement and proof of Lemma 4A.1 for three or more parties.

(c) Generalize Example 4A.2 to three or more parties, and then restate and solve Exercise 4.5 in this setting.

4B The Nash Solution

Nash [Nas50] argued that any 'reasonable' bargaining outcome should satisfy three axioms; he then showed that these three axioms together determine a unique solution for any bargaining problem.

Invariance under rescaling of utility functions: Recall that, in the von Neumann-Morgenstern definition of cardinal utility, the cardinal utility functions of Zara and Owen are only well-defined up to *rescaling*. Thus, if Zara has utility function $U_0 : \mathcal{A} \longrightarrow \mathbb{R}$, and we define $U_0' : \mathcal{A} \longrightarrow \mathbb{R}$ by $U_0'(A) = k \cdot U_0(A) + j$ for some constants k and j, then U_0' and U_0 are both equally valid as utility functions for Zara (ie. both satisfy the conditions of the von Neumann Morgenstern theorem).

If Zara asserts that her utility function is U_0', then a 'fair' arbitration scheme should produce the same outcome as if Zara had said her utility function was U_0. If the arbitration scheme was sensitive to a rescaling of Zara's utility function, then she could manipulate the outcome by choosing the constants k and j so as to skew results in her favour.

Suppose $U_0' : \mathcal{A} \longrightarrow \mathbb{R}$ and $U_1' : \mathcal{A} \longrightarrow \mathbb{R}$ were rescaled versions of U_0 and U_1, and let $U' : \mathcal{A} \longrightarrow \mathbb{R}^2$ be the resulting joint utility function. Let $\widetilde{\mathcal{A}}' = U'(\mathcal{A})$; let $\mathcal{B}'$ be the convex, comprehensive closure of $\widetilde{\mathcal{A}}'$ and let $\mathbf{q}' = U'(Q)$ be the 'status quo'. We say that the bargaining problem $(\mathcal{B}', \mathbf{q}')$ is a *rescaling* of the bargaining problem $(\mathcal{B}, \mathbf{q})$.

To be precise, suppose $U_0'(A) = k_0 \cdot U_0(A) + j_0$ and $U_1'(A) = k_1 \cdot U_1(A) + j_1$ for some constants $k_0, k_1, j_0, j_1 \in \mathbb{R}$. If we define the *rescaling function* $F : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ by $F(b_0, b_1) = (k_0 b_0 + j_0,\ k_1 b_1 + j_1)$, then $\mathcal{B}' = F(\mathcal{B})$ and $\mathbf{q}' = F(\mathbf{q})$. (**Exercise 4.8**) ?

$(\mathcal{B}', \mathbf{q}')$ and $(\mathcal{B}, \mathbf{q})$ represent the *same* bargaining problem, only with the utility functions for each player 'rescaled' by some amount. A 'fair' arbitration scheme should therefore yield the 'same' outcome for $(\mathcal{B}', \mathbf{q}')$ and $(\mathcal{B}, \mathbf{q})$, after accounting for the rescaling. Thus, for any $\mathbf{b} \in \mathcal{B}$, it is reasonable to assert: "$\mathbf{b}$ is a fair outcome of $(\mathcal{B}, \mathbf{q})$ if and only if $F(\mathbf{b})$ is a fair outcome of $(\mathcal{B}', \mathbf{q}')$." We therefore require the bargaining solution $\alpha : \mathfrak{B} \longrightarrow \mathbb{R}^2_{\nearrow}$ to satisfy axiom of *Rescaling Invariance:*

(RI) (Rescaling Invariance) *Let $(\mathcal{B}, \mathbf{q}) \in \mathfrak{B}$, let $F : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ be an affine 'rescaling' function, and let $F(\mathcal{B}) = \mathcal{B}'$ and $F(\mathbf{q}) = \mathbf{q}'$. Then $\alpha(\mathcal{B}', \mathbf{q}') = F[\alpha(\mathcal{B}, \mathbf{q})]$.*

Symmetry: In a 'fair' arbitration scheme, the two parties should be treated equally. Thus, if Zara and Owen switch places, then the arbitration scheme should respond by switching the outcomes. Let $R : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ be the 'reflection' (or 'role reversal') function $R(x_0, x_1) = (x_1, x_0)$. If $(\mathcal{B}, \mathbf{q})$ is a bargaining problem, then the *reflected* bargaining problem $(R(\mathcal{B}), R(\mathbf{q}))$ is obtained by applying R. That is,

$$R(\mathcal{B}) \ := \ \big\{(b_1, b_0) \in \mathbb{R}^2 \ ; \ (b_0, b_1) \in \mathcal{B}\big\},$$

and, if $\mathbf{q} = (q_0, q_1)$, then $R(\mathbf{q}) = (q_1, q_0)$. We assert: "$\mathbf{b}$ is a fair outcome of $(\mathcal{B}, \mathbf{q})$ if and only if $R(\mathbf{b})$ is a fair outcome of $(R(\mathcal{B}), R(\mathbf{q}))$." Formally, this means α should satisfy the axiom of *Symmetry*

(S) (Symmetry) *Let $(\mathcal{B}, \mathbf{q}) \in \mathfrak{B}$. If $\widehat{\mathcal{B}} := R(\mathcal{B})$ and $\widehat{\mathbf{q}} := R(\mathbf{q})$, then $\alpha(\widehat{\mathcal{B}}, \widehat{\mathbf{q}}) = R[\alpha(\mathcal{B}, \mathbf{q})]$.*

Lemma 4B.1 *Suppose $\mathcal{B}$ is a* symmetric *set* (i.e. $R(\mathcal{B}) = \mathcal{B}$) and $q_0 = q_1$ (ie. the bargainers have identical status quo positions). *If $\mathbf{b}$ is any fair outcome satisfying* **(S)**, *then $b_0 = b_1$* (i.e. the outcome is identical for each bargainer).

Proof: **Exercise 4.9** □

(In fact, Lemma 4B.1 is really the only consequence of **(S)** we will need, so in Nash's original work, he *defined* axiom **(S)** to be the statement of Lemma 4B.1).

Independence of Irrelevant Alternatives: Suppose the bargainers are initially negotiating over some bargaining problem $(\mathcal{B}, \mathbf{q})$, and they agree on outcome $\mathbf{b}$. Suppose that they then discover that, *actually*, their range of alternatives was more restricted than they thought; so that the *real* bargaining set was some subset $\mathcal{B}' \subset \mathcal{B}$. The good news, however, is that $\mathbf{b}$ is *still* an admissible outcome; ie. $\mathbf{b} \in \mathcal{B}'$. Clearly, if $\mathbf{b}$ was a mutually agreeable outcome for the bargaining set $\mathcal{B}$, it should *still* be mutually agreeable for $\mathcal{B}'$.

We can reverse this scenario: suppose the bargainers initially agree on an outcome $\mathbf{b}$ for the bargaining problem $(\mathcal{B}', \mathbf{q})$. Suppose now that their alternatives are *enhanced*, so that the bargaining set is *expanded* to some superset $\mathcal{B} \supset \mathcal{B}'$. The outcome of the new bargaining problem $(\mathcal{B}, \mathbf{q})$ should either *remain* $\mathbf{b}$, or, if it changes, it should change to some previously inaccessible outcome —ie. an element of $\mathcal{B} \setminus \mathcal{B}'$. It certainly makes no sense for the players to change from $\mathbf{b}$ to another element of $\mathcal{B}'$, when confronted with the richer possibilities of $\mathcal{B}$. In other words, "If the bargain $\mathbf{b}$ is a fair outcome of $(\mathcal{B}, \mathbf{q})$, and $\mathbf{b} \in \mathcal{B}' \subset \mathcal{B}$, then $\mathbf{b}$ is also a fair outcome of $(\mathcal{B}', \mathbf{q})$." We thus arrive at the axiom of *Independence of Irrelevant Alternatives*:

(IIA) (Independence of Irrelevant Alternatives) *Let $(\mathcal{B}, \mathbf{q}) \in \mathfrak{B}$ and $(\mathcal{B}', \mathbf{q}) \in \mathfrak{B}$, and suppose $\mathbf{q} \in \mathcal{B}' \subset \mathcal{B}$. If $\alpha(\mathcal{B}, \mathbf{q}) \in \mathcal{B}'$, then $\alpha(\mathcal{B}, \mathbf{q}) = \alpha(\mathcal{B}', \mathbf{q})$.*

The Nash Solution: If $(\mathcal{B}, \mathbf{q})$ is a bargaining problem, then the *Nash product* is the function $N_{\mathbf{q}} : \mathcal{B} \longrightarrow \mathbb{R}$ defined:

$$N_{\mathbf{q}}(b_0, b_1) \quad := \quad (b_0 - q_0) \cdot (b_1 - q_1).$$

In other words, for any $\mathbf{b} \in \mathcal{B}$, we first compute the *net change* in utility from the status quo for each bargainer. Then we multiply these net changes. The *Nash solution* is the (unique) point $\eta(\mathcal{B}, \mathbf{q})$ in $\mathcal{B}$ which maximizes the value of $N_{\mathbf{q}}$.

Theorem 4B.2 (Nash)

(a) *For any bargaining problem $(\mathcal{B}, \mathbf{q}) \in \mathfrak{B}$, the Nash solution $\eta(\mathcal{B}, \mathbf{q})$ is well-defined, because there exists a unique maximizer for $N_{\mathbf{q}}$ in $\wp_{\mathbf{q}}\mathcal{B}$.*

(b) *The Nash bargaining solution η satisfies axioms* **(RI)**, **(S)** *and* **(IIA)**.

(c) *η is the unique bargaining solution satisfying axioms* **(RI)**, **(S)** *and* **(IIA)**.

Proof: **(a)** *Existence:* To see that a Nash solution exists, we must show that the function $N_{\mathbf{q}}$ takes a maximum on $\mathcal{B}$. This follows from the fact that $N_{\mathbf{q}}$ is continuous, and that $\wp_{\mathbf{q}}\mathcal{B}$ is a compact subset of $\mathbb{R}^2$.

(a) *Uniqueness:* We must show that $N_{\mathbf{q}}$ cannot have *two* maxima in $\mathcal{B}$. We'll use two facts:

- $\mathcal{B}$ is a convex set.
- $N_{\mathbf{q}}$ is a *strictly quasiconcave* function. That is, for any $\mathbf{b}_1 \neq \mathbf{b}_2 \in \mathbb{R}^2$, and any $c_1, c_2 \in (0,1)$ such that $c_1 + c_2 = 1$, we have $N_{\mathbf{q}}(c_1\mathbf{b}_1 + c_2\mathbf{b}_2) > \min\{N_{\mathbf{q}}(\mathbf{b}_1), N_{\mathbf{q}}(\mathbf{b}_2)\}$. (**Exercise 4.10**)

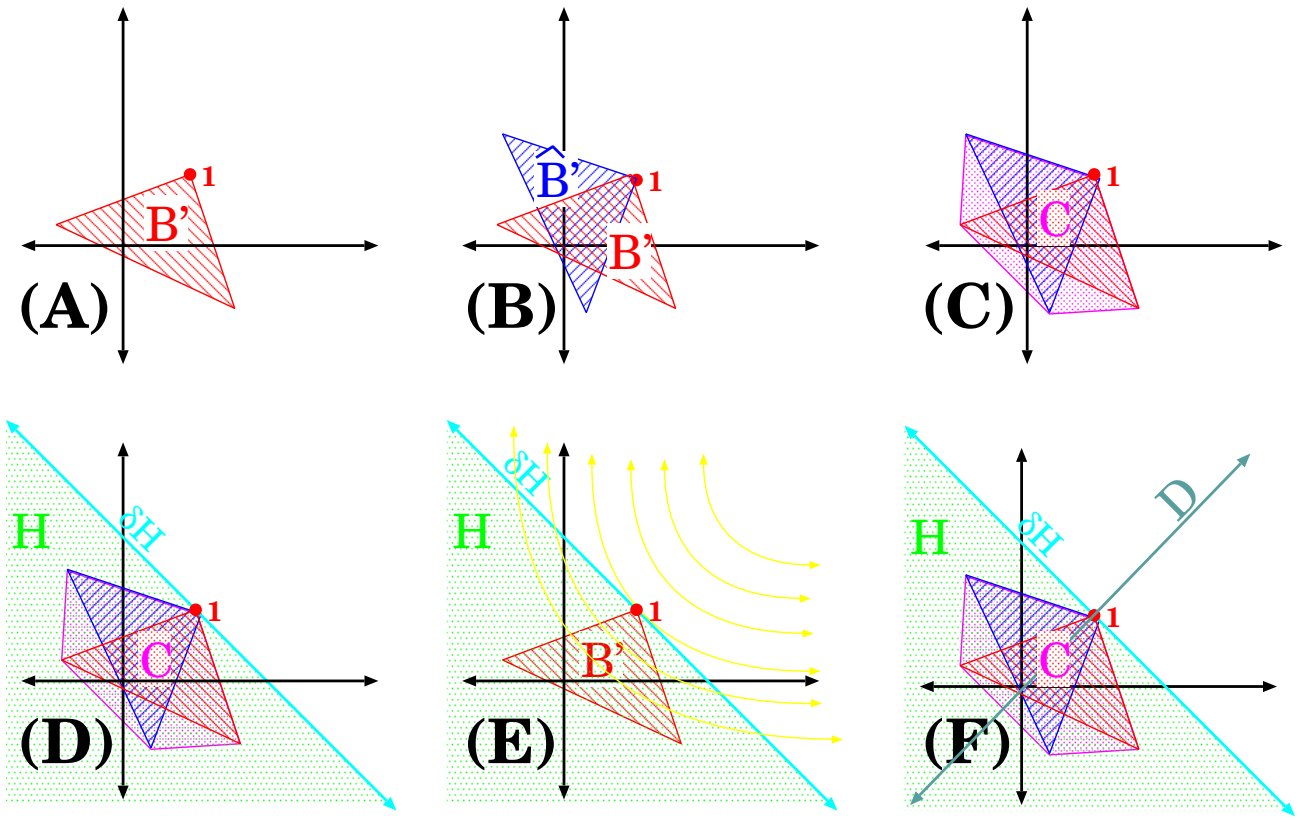


Figure 4.2: The Nash Solution: **(A)** The bargaining set $\mathcal{B}'$. **(B)** $\widehat{\mathcal{B}}' = \big\{(b_1, b_0) \in \mathbb{R}^2 \,;\, (b_0, b_1) \in \mathcal{B}'\big\}$. **(C)** $\mathcal{C}$ is the convex closure of $\mathcal{B}' \cup \widehat{\mathcal{B}}'$. **(D)** $\mathcal{H} = \big\{\mathbf{h} \in \mathbb{R}^2 \,;\, h_0 + h_1 < 2\big\}$. **(E)** The level curve of $\widetilde{U}$ must be tangent to the boundary of $\mathcal{B}'$ at $\mathbf{1}$. **(F)** $\mathcal{D} = \big\{\mathbf{d} \in \mathbb{R}^2 \,;\, d_0 = d_1\big\}$, and $\wp\mathcal{C} \cap \mathcal{D} = \{\mathbf{1}\}$.

Suppose M was the maximal value of $N_{\mathbf{q}}$ in $\mathcal{B}$, and suppose $\mathbf{b}_1$ and $\mathbf{b}_2$ were both maxima for $N_{\mathbf{q}}$, so that $N_{\mathbf{q}}(\mathbf{b}_1) = M = N_{\mathbf{q}}(\mathbf{b}_2)$. Let $\mathbf{b} = \frac{1}{2}\mathbf{b}_1 + \frac{1}{2}\mathbf{b}_2$. Then $\mathbf{b} \in \mathcal{B}$ (because $\mathcal{B}$ is convex), and $N_{\mathbf{q}}(\mathbf{b}) > \min\{M, M\} = M$, (because $N_{\mathbf{q}}$ is strictly quasiconcave), thereby contradicting the maximality of M. By contradiction, the maximum of F in $\mathcal{B}$ must be *unique*.

(b) *The Nash solution satisfies* **(RI)***:* Let $F : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ be a rescaling function defined: $F(b_0, b_1) = (k_0 b_0 + j_0,\ k_1 b_1 + j_1)$. Let $\mathcal{B}' = F(\mathcal{B})$ and let $\mathbf{q}' = F(\mathbf{q})$. Let $N_{\mathbf{q}'} : \mathcal{B}' \longrightarrow \mathbb{R}$ be the Nash function for the bargaining problem $(\mathcal{B}', \mathbf{q}')$.

Claim 1: **(a)** *For any* $\mathbf{b} \in \mathcal{B}$, *if* $\mathbf{b}' = F(\mathbf{b})$, *then* $N_{\mathbf{q}'}(\mathbf{b}') = k_0 k_1 \cdot N_{\mathbf{q}}(\mathbf{b})$.

(b) *Thus,* $\Big(\mathbf{b}$ *is the maximum of* $N_{\mathbf{q}}$ *on* $\mathcal{B}\Big) \iff \Big(\mathbf{b}'$ *is the maximum of* $N_{\mathbf{q}'}$ *on* $\mathcal{B}'\Big)$.

Proof: (a) If $\mathbf{b} = (b_0, b_1)$, then $\mathbf{b}' = (k_0 b_0 + j_0,\ k_1 b_1 + j_1)$. Also, $\mathbf{q}' = (k_0 q_0 + j_0,\ k_1 q_1 + j_1)$. Thus,

$$\begin{aligned} N_{\mathbf{q}'}(\mathbf{b}') &= (b'_0 - q'_0) \cdot (b'_1 - q'_1) \\ &= \Big((k_0 b_0 + j_0) - (k_0 q_0 + j_0)\Big) \cdot \Big((k_1 b_1 + j_1) - (k_1 q_1 + j_1)\Big) \\ &= (k_0 b_0 - k_0 q_0) \cdot (k_1 b_1 - k_1 q_1) \\ &= k_0 \cdot k_1 \cdot (b_0 - q_0) \cdot (b_1 - q_1) \quad = \quad k_0 k_1 \cdot N_{\mathbf{q}}(\mathbf{b}) \end{aligned}$$

Part (b) follows from (a). ◇ Claim 1

(b) *The Nash solution satisfies* **(S)***:* Let $\widehat{\mathcal{B}}$ and $\widehat{\mathbf{q}}$ be as in the definition of **(S)**. Let $N_{\widehat{\mathbf{q}}} : \widehat{\mathcal{B}} \longrightarrow \mathbb{R}$ be the Nash function for the bargaining problem $(\widehat{\mathcal{B}}, \widehat{\mathbf{q}})$.

Claim 2: **(a)** *For any* $\mathbf{b} = (b_0, b_1) \in \mathcal{B}$*, if* $\widehat{\mathbf{b}} = (b_1, b_0) \in \widehat{\mathcal{B}}$*, then* $N_{\widehat{\mathbf{q}}}(\widehat{\mathbf{b}}) = N_{\mathbf{q}}(\mathbf{b})$.

(b) *Thus,* $\Big(\mathbf{b}$ *is the maximum of* $N_{\mathbf{q}}$ *on* $\mathcal{B}\Big) \iff \Big(\widehat{\mathbf{b}}$ *is the maximum of* $N_{\widehat{\mathbf{q}}}$ *on* $\widehat{\mathcal{B}}\Big)$.

Proof: $N_{\widehat{\mathbf{q}}}(\widehat{\mathbf{b}}) = (\widehat{b}_0 - \widehat{q}_0) \cdot (\widehat{b}_1 - \widehat{q}_1) = (b_1 - q_1) \cdot (b_0 - q_0) = (b_0 - q_0) \cdot (b_1 - q_1) = N_{\mathbf{q}}(\mathbf{b})$. ◇ Claim 2

(b) *The Nash arbitration scheme satisfies* **(IIA)***:* **Exercise 4.11** .

(c): Suppose α was some bargaining solution satisfying axioms **(RI)**, **(S)** and **(IIA)**. Let $(\mathcal{B}, \mathbf{q}) \in \mathfrak{B}$ and let $\mathbf{b} := \eta(\mathcal{B}, \mathbf{q})$. We must show that $\alpha(\mathcal{B}, \mathbf{q}) = \mathbf{b}$.

Let $F : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ be the rescaling defined: $F(c_0, c_1) = \left(\frac{c_0 - q_0}{b_0}, \frac{c_1 - q_1}{b_1}\right)$, for any $\mathbf{c} = (c_0, c_1) \in \mathcal{B}$.

Thus, $F(\mathbf{q}) = \mathbf{0} := (0, 0)$ and $F(\mathbf{b}) = \mathbf{1} := (1, 1)$. Let $\mathcal{B}' = F(\mathcal{B})$ (Figure 4.2A). Thus, the axiom **(RI)** implies that

$$\Big(\alpha(\mathcal{B}, \mathbf{q}) = \mathbf{b}\Big) \iff \Big(\alpha(\mathcal{B}', \mathbf{0}) = \mathbf{1}\Big). \tag{4.2}$$

Hence we will prove that $\alpha(\mathcal{B}', \mathbf{0}) = \mathbf{1}$. To do this, let $\widehat{\mathcal{B}'} = \{(b_1, b_0) \in \mathbb{R}^2 \,;\, (b_0, b_1) \in \mathcal{B}'\}$ (Figure 4.2B), and let $\mathcal{C}$ be the convex comprehensive closure of $\mathcal{B}' \cup \widehat{\mathcal{B}'}$ (Figure 4.2C). Then $\mathcal{C}$ is a convex, compact, comprehensive, symmetric set containing $\mathcal{B}'$.

Consider the halfspace $\mathcal{H} = \{\mathbf{h} \in \mathbb{R}^2 \,;\, h_0 + h_1 \leq 2\}$ (Figure 4.2D).

Claim 3: $\mathcal{C} \subset \mathcal{H}$.

Proof: The Nash utility function for $(\mathcal{B}', \mathbf{0})$ is just the function $N_{\mathbf{0}}(b_0, b_1) = b_0 \cdot b_1$. Claim 1(b) implies that $\mathbf{1}$ is the maximum of $N_{\mathbf{0}}$ in $\mathcal{B}'$. Thus, the level curve of $N_{\mathbf{0}}$ must be tangent

to the boundary of $\mathcal{B}'$ at $\mathbf{1}$ (Figure 4.2E). But the level curve of $N_{\mathbf{0}}$ at $\mathbf{1}$ has slope -1 (**Exercise 4.12**); in other words, it is tangent to the line $\partial\mathcal{H} = \{\mathbf{h} \in \mathbb{R}^2 \; ; \; h_1 + h_2 = 2\}$.

It follows that the boundary of $\mathcal{B}'$ is tangent to $\partial\mathcal{H}$ at $\mathbf{1}$. But $\mathcal{B}'$ is convex, so it follows that *all* of $\mathcal{B}'$ must be below $\partial\mathcal{H}$ (**Exercise 4.13**); in other words, $\mathcal{B}' \subset \mathcal{H}$. Thus, $\widehat{\mathcal{B}'} \subset \widehat{\mathcal{H}}$. But $\mathcal{H}$ is symmetric under exchange of coordinates; hence $\widehat{\mathcal{H}} = \mathcal{H}$, so we have $\widehat{\mathcal{B}'} \subset \mathcal{H}$. Thus, $\mathcal{B}' \cup \widehat{\mathcal{B}'} \subset \mathcal{H}$. Since $\mathcal{H}$ is convex, we conclude that the convex closure $\mathcal{C}$ must also be a subset of $\mathcal{H}$. ◇ Claim 3

Let $\wp\mathcal{C}$ be the Pareto frontier of $\mathcal{C}$.

Claim 4: $\mathbf{1} \in \wp\mathcal{C}$.

Proof: We know that $\mathbf{1} \in \mathcal{C}$ because $\mathbf{1} \in \mathcal{B}$. Suppose $\mathbf{c} \in \mathcal{C}$ was strictly Pareto preferred to $\mathbf{1}$. Then $c_0 \geq 1$ and $c_1 \geq 1$, and at least one of these is a strict inequality. Hence $c_0 + c_1 > 1 + 1 = 2$. Hence $\mathbf{c} \notin \mathcal{H}$, contradicting Claim 3. ◇ Claim 4

Claim 5: $\alpha(\mathcal{C}, \mathbf{0}) = \mathbf{1}$.

Proof: Let $\alpha(\mathcal{C}, \mathbf{0}) = \mathbf{c}$. Since $\mathcal{C}$ and $\mathbf{0}$ are both symmetric, the symmetry axiom **(S)** implies that $\mathbf{c}$ must also be symmetric —ie. $c_0 = c_1$. Thus, if $\mathcal{D} = \{\mathbf{d} \in \mathbb{R}^2 \; ; \; d_0 = d_1\}$ is the diagonal line in Figure 4.2(F), then we know $\mathbf{c} \in \mathcal{D}$. However, the Pareto axiom **(P)** also requires that $\mathbf{c}$ be an element of the Pareto frontier $\wp\mathcal{C}$. Thus, $\mathbf{c} \in \wp\mathcal{C} \cap \mathcal{D}$.

Now Claim 4 says $\mathbf{1} \in \wp\mathcal{C}$, and clearly, $\mathbf{1} \in \mathcal{D}$; hence $\mathbf{1} \in \wp\mathcal{C} \cap \mathcal{D}$; However, $\wp\mathcal{C}$ can never be tangent to $\mathcal{D}$ (**Exercise 4.14**), so we know that $\wp\mathcal{C}$ and $\mathcal{D}$ can intersect in at most one place; hence $\wp\mathcal{C} \cap \mathcal{D} = \{\mathbf{1}\}$. Thus, we must have $\mathbf{c} = \mathbf{1}$. ◇ Claim 5

Since $\mathbf{1} \in \mathcal{B}' \subset \mathcal{C}$, it follows from Claim 5 and axiom **(IIA)** that $\alpha(\mathcal{B}', \mathbf{0}) = \mathbf{1}$. Thus, eqn.(4.2) implies that $\alpha(\mathcal{B}, \mathbf{q}) = \mathbf{b}$. □

Discussion: The Nash arbitration scheme is applicable when the axioms **(RA)**, **(S)**, and **(IIA)** are appropriate. However, the Nash scheme has been rejected by some who argue that the axioms **(RA)** and **(S)** are often inappropriate.

For example, the 'Rescaling' axiom **(RA)** explicitly disallows any judgements about the 'intensity' of the utility valuations of the bargainers. In some cases, this may be appropriate, because we have limited information, and because subjective testimonials by the bargainers concerning the 'intensity' of their feelings will be at best imprecise, and at worst, deliberately manipulative.

However, in some scenarios, it is *clear* that one bargainer has much stronger preferences than the other. Suppose that Zara is starving, and trying to beg food from Owen the grocer. We would all agree that Zara's utility valuations regarding food will be much more 'intense' than Owen's. However, the axiom **(RA)** does not allow us to include this information in our model.

The 'Symmetry' axiom **(S)** is also contentious. **(S)** is appropriate when dealing with two individual persons, if we assume that all persons must be treated equally. However, in many bargaining situations, the bargainers are not *people* but *groups*. For example, Zara might represent a union with 5000 workers, while Owen is the manager of a company, representing the interests of 3000 shareholders. It is not clear how we should weigh the 'moral importance' of 5000 workers against 3000 shareholders, but it is clear that an *a priori* insistence on exactly symmetric treatment is simpleminded at best.

To obviate the objections to **(S)**, we can insist that the Nash solution only be applied to bargaining between individuals. A bargaining scenario between two *groups* can then be treated as a *multiparty* bargain between all the individuals comprising these groups. For example, arbitration between a labour union (representing 5000 workers) and management (representing 3000 shareholders) can be treated as a multiparty arbitration involving 8000 individuals. The von Neumann-Morgenstern model and the Nash arbitration scheme generalize to multiparty bargaining in the obvious way, and Nash's Theorem still holds in this context.

To obviate the objections to **(RI)**, we can insist that all individuals must assign utilities to the same collection of alternatives. For example, if starving Zara is begging food from Owen the grocer, the axiom **(S)** is inappropriate, because Zara's alternatives are eating versus starving, while Owen's alternatives are more versus less revenue. To apply **(S)**, we should ask Owen to also assign utility to the (imaginary) scenario where he is a starving person, and we should ask Zara to assign a utility to the (imaginary) scenario where she is a struggling grocer trying to remain in business (even though these imaginary scenarios are not part of the bargaining set). The problem is that we may not be able to obtain accurate utility estimates for such far-fetched imaginary scenarios.

Exercise 4.15: The Nash solution can be generalized to bargains involving three or more parties.

- **(a)** Reformulate axiom **(S)** for three or more parties; you must allow any possible permutation of coordinates. [In this context, **(S)** is often refered to as *Anonymity* **(A)**.]
- **(b)** Reformulate and prove Lemma 4B.1 for three or more parties.
- **(c)** Reformulate axioms **(RI)** and **(IIA)** for three or more parties.
- **(d)** Define the *Nash product* and *Nash solution* for three or more parties.
- **(e)** Reformulate and prove Theorem 4B.2 for three or more parties.

Exercise 4.16: A function $C : \mathbb{R}^2_{\nearrow} \longrightarrow \mathbb{R}$ is *concave* if

$$\text{For any } \mathbf{x} \neq \mathbf{y} \in \mathbb{R}^2_{\nearrow}, \text{ and } r \in (0,1), \qquad C(r\mathbf{x} + (1-r)\mathbf{y}) \quad \geq \quad rC(\mathbf{x}) + (1-r)C(\mathbf{y}).$$

We say C is *strictly concave* if this inequality is strict. We say C is *quasiconcave* if

$$\text{For any } \mathbf{x} \neq \mathbf{y} \in \mathbb{R}^2_{\nearrow}, \text{ and } r \in (0,1), \qquad C(r\mathbf{x} + (1-r)\mathbf{y}) \quad \geq \quad \min\left\{C(\mathbf{x}),\ C(\mathbf{y})\right\}.$$

We say C is *strictly quasiconcave* if this inequality is strict.

For our purposes, C represents some way of combining the utilities of Zara and Owen to get a measurement of "collective good" for society.

(a) If $c_0, c_1 \in \mathbb{R}_{\neq}$ are any constants, show that the function $C(x_0, x_1) = c_0 x_0 + c_1 x_1$ is concave (but not strictly).

(b) Show that any strictly concave function is strictly quasiconcave.

(c) Show that the *Nash product* $N(x_0, x_1) = x_0 \cdot x_1$ is a strictl quasiconcave.

(d) Let $r_0, r_1 \in \mathbb{R}_{\neq}$. Show that the *generalized Nash product* $C(x_0, x_1) = x_0^{r_0} \cdot x_1^{r_1}$ is strictly quasiconcave.

(e) A bargaining solution $\alpha : \mathfrak{B} \longrightarrow \mathbb{R}^2_{\neq}$ is *strictly quasiconcave-optimizing* if there is some strictly quasiconcave functional $C : \mathbb{R}^2_{\neq} \longrightarrow \mathbb{R}$ such that $\alpha(\mathcal{B}, \mathbf{q})$ is the point in $\wp_{\mathbf{q}}\mathcal{B}$ which maximizes C. For example, the Nash bargaining solution is strictly quasiconcave-optimizing, because it maximizes the Nash product $N(x_0, x_1) = x_0 \cdot x_1$.

Show that any strictly quasiconcave-optimizing bargaining solution must satisfy the axiom **(IIA)**.

(f) Suppose α is strictly quasiconcave-optimizing as in (d), and assume that C is differentiable. Recall that the *gradient* of C is defined $\nabla C := (\partial_0 C, \partial_1 C)$. Suppose that the negotiating set $\wp_{\mathbf{q}}\mathcal{B}$ is the graph of some function $\Gamma_1 : \mathbb{R}_{\neq} \longrightarrow \mathbb{R}_{\neq}$, as in Lemma 4A.1. Show that $\alpha(\mathcal{B}, \mathbf{q})$ is the point $(b_0, \Gamma_1(b_0))$ on $\wp_{\mathbf{q}}\mathcal{B}$ where ∇C is *orthogonal* to the vector $(1, \Gamma_1'(b_0))$; that is, where

$$\partial_0 C(b_0, \Gamma_1(b_0)) \quad = \quad -\Gamma_1'(b_0) \cdot \partial_1 C(b_0, \Gamma_1(b_0)).$$

(g) Continuing (e), show that the Nash bargaining solution is the unique point $(b_0, \Gamma_1(b_0))$ such that $\Gamma_1'(b_0) = \dfrac{-\Gamma_1(b_0)}{b_0}$.

Exercise 4.17:(Risk Aversion and the Nash Solution)

Consider the *surplus division* problem of Example 4A.2. For simplicity, suppose Zara and Owen are trying to divide one dollar. Thus, if x_0 is Zara's share of the dollar, then $x_1 := 1 - x_0$ is Owen's share. Assume Zara is risk-neutral, so that her utility function for money is $b_0(x_0) = x_0$ (this is plausible if Zara is quite wealthy, so that one dollar represents a very small fraction of her existing wealth). Assume Owen is risk-averse, with monetary utility function $b_1(x_1) = x_1^\alpha$ for some $\alpha \in (0, 1)$ (this is plausible if Owen is much less wealthy than Zara). Assume the status quo is $\mathbf{0} := (0, 0)$.

(a) Show that the negotiating set is the graph of the function $\Gamma_1(b_0) := (1 - b_0)^\alpha$.

(b) Show that the Nash solution to this bargaining problem allocates $x_0 = \frac{1}{\alpha+1}$ dollars to Zara, and the remaining $x_1 = \frac{\alpha}{\alpha+1}$ dollars to Owen, thereby yielding Zara a utility of $b_0 = \frac{1}{\alpha+1}$ and Owen a utility of $b_1 = \left(\frac{\alpha}{\alpha+1}\right)^\alpha$. [Hint: Use Exercise 4.16(g)]

Figure 4.3(a) shows that that $x_0 \nearrow 1$ and $x_1 \searrow 0$ as $\alpha \searrow 0$. Thus, the more risk averse Owen becomes, the more the Nash solution favours Zara in material terms. This feature perhaps enhances the realism of the Nash solution as a description of real bargaining, while simultaneously diminishing its appeal as a normative ideal of justice.

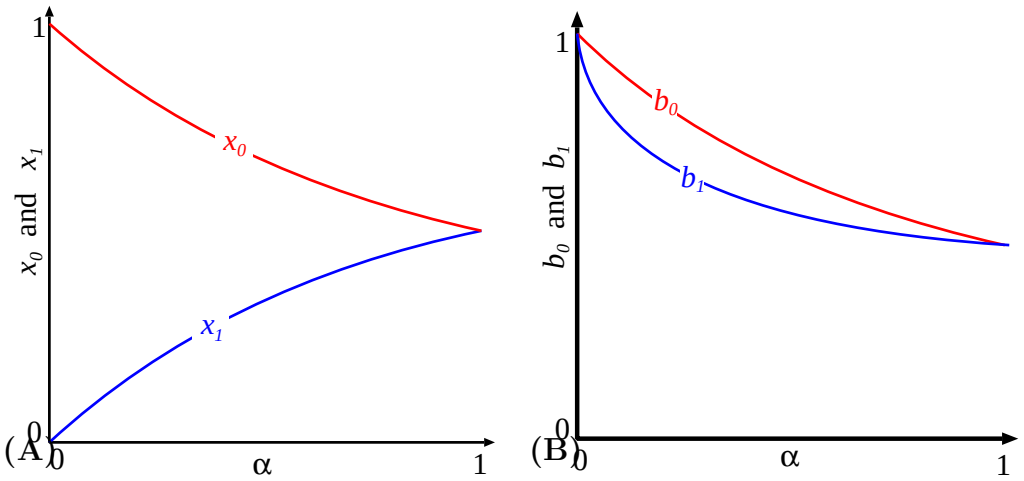


Figure 4.3: The Nash bargaining solution as a function of the risk aversion of Owen (Exercise 4.17). Zara and Owen are dividing a dollar, so $x_0 + x_1 = 1$. Zara has utility function $b_0(x_0) = x_0$ and Owen has utility function $b_1(x_1) = x_1^\alpha$, where $\alpha \in (0, 1)$.
(A) x_0 and x_1 as functions of α. Notice that $x_0 \nearrow 1$ and $x_1 \searrow 0$ as $\alpha \searrow 0$.
(B) b_0 and b_1 as functions of α. Note that $b_1 < b_0$ except when $\alpha = 0$ or $\alpha = 1$.

Figure 4.3(b) shows that $b_1 < b_0$ except when $\alpha = 0$ or $\alpha = 1$. Thus, even in 'utility' terms, the Nash solution favours Zara over Owen; however, as $\alpha \searrow 0$, Owen's utility function becomes so abject that he finds the bargain almost as rewarding as Zara, even though she gets the lion's share of the money.]

4C Hausdorff Continuity

Prerequisites: §4A **Recommended:** §4B

Two similar bargaining problems should result in similar outcomes. In other words, if $\alpha : \mathfrak{B} \longrightarrow \mathbb{R}^2_\nearrow$ is a bargaining solution, and $\mathcal{B}, \mathcal{B}' \subset \mathbb{R}^2$ are two 'similar' sets, with $\mathbf{q} \in \mathcal{B} \cap \mathcal{B}'$, then $\alpha(\mathcal{B}, \mathbf{q})$ should be 'close' to $\alpha(\mathcal{B}', \mathbf{q})$. To make this idea precise, we need a way to measure the 'distance' between two sets. This is the purpose of the *Hausdorff metric*. If $\mathcal{B} \subset \mathbb{R}^2$ is a closed subset, and $x \in \mathbb{R}^2$, then define

$$d(x, \mathcal{B}) \quad := \quad \inf_{b \in \mathcal{B}} d(x, b).$$

Exercise 4.18: (a) Show that $d(x, \mathcal{B}) = 0$ if and only if $x \in \mathcal{B}$.

(b) Show that this is *not* true if $\mathcal{B}$ is not a *closed* subset of $\mathbb{R}^2$.

Next, if $\mathcal{B}, \mathcal{C} \subset \mathbb{R}^2$ are two closed subsets, the *Hausdorff distance* from $\mathcal{B}$ to $\mathcal{C}$ is defined:

$$d_H(\mathcal{B}, \mathcal{C}) \quad := \quad \sup_{b \in \mathcal{B}} d(b, \mathcal{C}) \; + \; \sup_{c \in \mathcal{C}} d(c, \mathcal{B}).$$

Exercise 4.19: Let $\mathfrak{K} = \big\{\mathcal{B} \subset \mathbb{R}^2 \; ; \; \mathcal{B} \text{ compact}\big\}$. Show that d_H is a *metric* on the set $\mathfrak{K}$. That is, for any compact subsets $\mathcal{B}, \mathcal{C}, \mathcal{D} \subset \mathbb{R}^2$:

(a) $d_H(\mathcal{B}, \mathcal{C}) \geq 0$. Furthermore, $d_H(\mathcal{B}, \mathcal{C}) = 0$ if and only if $\mathcal{B} = \mathcal{C}$.

(b) $d_H(\mathcal{B}, \mathcal{C}) = d_H(\mathcal{C}, \mathcal{B})$.

(c) $d_H(sB, \mathcal{D}) \leq d_H(\mathcal{B}, \mathcal{C}) + d_H(\mathcal{C}, \mathcal{D})$.

If $\{\mathcal{B}_n\}_{n=1}^{\infty}$ is a sequence of compact subsets, then we say that $\{\mathcal{B}_n\}_{n=1}^{\infty}$ *Hausdorff-converges* to $\mathcal{B}$, (or that $\mathcal{B}$ is the *Hausdorff limit* of $\{\mathcal{B}_n\}_{n=1}^{\infty}$) if $\lim_{n \to \infty} d_H(\mathcal{B}_n, \mathcal{B}) = 0$.

Exercise 4.20: Show that $\mathcal{B}$ is the Hausdorff limit of the sequence $\{\mathcal{B}_n\}_{n=1}^{\infty}$ if and only if $\mathcal{B}$ is the set of all points $\mathbf{b} = \lim_{n \to \infty} \mathbf{b}_n$, where $\{\mathbf{b}_n\}_{n=1}^{\infty}$ is any sequence such that $\mathbf{b}_n \in \mathcal{B}_n$ for all $n \in \mathbb{N}$.

A bargaining solution $\alpha : \mathfrak{B} \longrightarrow \mathbb{R}^2_{\nearrow}$ satisfies the axiom of *Hausdorff continuity* if the following is true:

(HC) *For any sequence of bargaining problems $\{(\mathcal{B}_n, \mathbf{q}_n)\}_{n=1}^{\infty} \subset \mathfrak{B}$, if $\mathcal{B}$ is the Hausdorff limit of $\{\mathcal{B}_n\}_{n=1}^{\infty}$, and $\mathbf{q} = \lim_{n \to \infty} \mathbf{q}_n$, then $\alpha(\mathcal{B}, \mathbf{q}) = \lim_{n \to \infty} \alpha(\mathcal{B}_n, \mathbf{q}_n)$.*

Lemma 4C.1 *Let $\eta : \mathfrak{B} \longrightarrow \mathbb{R}^2_{\nearrow}$ be the Nash bargaining solution. Then η is Hausdorff-continuous.*

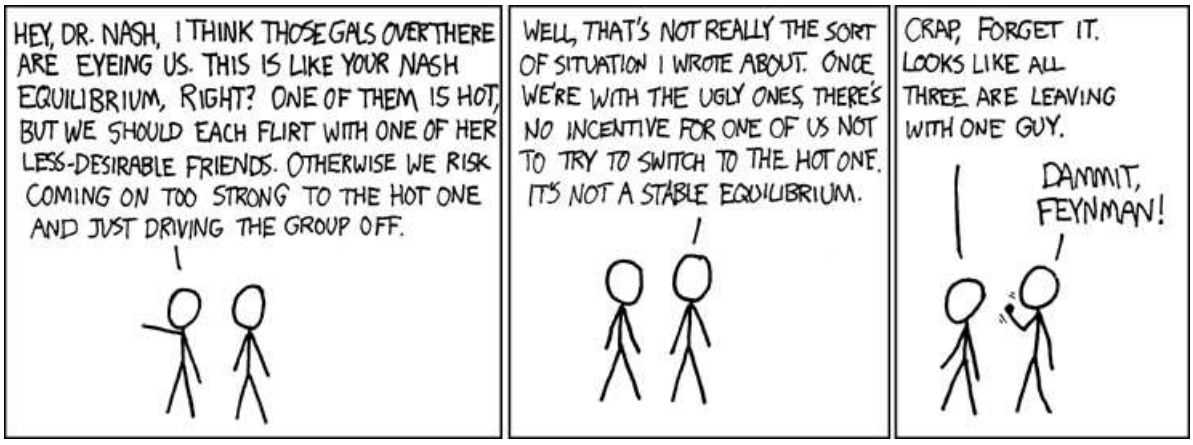
Proof: **Exercise 4.21** . □

Further Reading

Most books on game theory or social choice theory contain at least a chapter on bargaining theory; see for example [Mye91, Chapt.8], [OR94, Chapt.7 & 10], [Bin91], or [LR80, Chapt.6], [Mou84, Chapt.3], [Roe98, Chapt.1]; or [Mue03, Chapt.23]. There are also at least two books dedicated to bargaining theory: [Mut99] and [Nap02]. For a nice and informal introduction to many aspects of bargaining theory, with emphasis on philosophical implications, see [Bin98, Chapt 2 and Appendix 3].

Chapter 5

The Nash Program



from `http://xkcd.com/`

5A Introduction

Prerequisites: §4A **Recommended:** §4B

Game theory is the mathematical analysis of strategic interactions between rational individuals ('players') who each seek to maximize their expected utility ('payoff'). A game-theoretic analysis has two closely related components:

Prescriptive: This component prescribes what each player *should* do in each situation, so as to maximize her payoff.

Descriptive: This component predicts what each player *will* do in each situation (assuming she is rational and seeks to maximize her payoff).

These two components are intertwined: if the players are perfectly rational and can analyse their situation using game theory, then what they *will* do is likely to be what game theory says

they *should* do. Conversely, what you *should* do depends, to some extent, on what you predict other people *will* do. For example, in a Nash equilibrium (see §5B below) you *should* play your equilibrium strategy, because you predict that everyone else *will* play their equilibrium strategy, and you predict that other players *will* play this way because, like you, they deduce that they *should* (because they predict that you *will*, etc.)

Note that *prescriptive* is not the same as *normative* (see page 77). Game theory prescribes what each player 'should' do to maximize her payoff, but the word "should" has no moral content here. Each player has a utility function, which presumably encodes not only her personal tastes, but also her moral values. However, we do not necessarily *endorse* her utility function (and the moral values it encodes). Whether or not we endorse a player's moral values is irrelevant to game theory.

Nash proposed a bargaining solution which is 'fair' with respect to certain criteria [i.e. it satisfies axioms **(P)**, **(MB)**, **(S)**, **(IR)**, and **(IIA)**]. However, this does not imply that people in a real bargaining situation will converge upon this solution. In other words, Nash's solution is *normative*, but not yet either *descriptive* or *prescriptive*. Nash proposed the following problem:

> *Find a realistic description of the bargaining process as a game, such that this game has a unique equilibrium strategy, which we interpret to be the predicted 'outcome' of the bargaining process.*

Obviously, Nash hoped that the predicted outcome would correspond to the Nash bargaining solution of §4B. The search for such a game-theoretic justification of the Nash solution became known as the *Nash program*.

Bargaining vs. Arbitration: To pursue the Nash program, we must make careful and explicit modelling assumptions. First of all, note that the Nash program is concerned with *bargaining*, and not with *arbitration*. In *arbitration*, there is a third party (the 'arbitrator' or 'referee') who helps the bargainers reach a consensus. For example, all of the 'fair division' methods of Chapter IV implicitly rely upon an arbitrator, who implements a certain fair division procedure or enforces the rules of a certain division game.

Indeed, the arbitrator may be authorized to *propose* a certain solution, and then legally empowered to *enforce* this solution if the two bargainers refuse to comply voluntarily. Thus, an arbitrator can focus more on the normative question of what the bargain 'should' be, rather than the descriptive question of what it 'would' be if the bargainers were left to themselves. However, he cannot be totally autocratic; his enforcement powers are probably limited, and if his proposed settlement is too outrageous, then one or both parties may simply walk away from the bargaining process. Even if he is legally empowered, in a modern civilized legal system, the arbitrator's decision will be subject to legal appeal, and if he wishes to maintain his reputation (and keep his job), the arbitrator wants to avoid forcing the bargaining parties to appeal his decision. Because of this, the arbitrator will seek some outcome which would be deemed 'reasonable' or 'fair' by both parties, and the Nash solution is a good candidate for this. Hence, we could propose the Nash solution as an *arbitration scheme*. However, this is not

what the Nash program requires. The Nash program is concerned with what would happen *without* an arbitrator.

This is an important point, because models which predict the Nash solution as the outcome of a bargaining game are sometimes interpreted *normatively*, to mean that the Nash solution is the 'correct' or 'fair' outcome of a bargaining situation. However, in the presence of an arbitrator, other solutions are possible, and there may be valid normative reasons for favouring these other solutions over the Nash solution. Even if the Nash solution is the inevitable outcome of unrefereed bargaining, that doesn't necessarily make it the right thing to do.

Anarchy and coalition-formation: However, there can be no arbitration in situations where there is no 'rule of law', such as negotations between nation-states, negotiations between criminals, or negotiations between individuals in some anarchic Hobbesian 'state of nature'. Even in a modern civilized state, there will be many-player, noncompetitive[1] games (such as a capitalist economy or a legislative assembly) where players may find it mutually advantageous to form *coalitions*, and coordinate their strategies within each coalition. In a game with N players, there are 2^N possible coalitions, each of which may offer some benefit to its members. Each player will presumably join the coalition which offers him the greatest benefit. Thus, to determine which coalitions are likely to form, we must compute the benefit each coalition can offer to each of its members, which means predicting the way in which each coalition will divide the surplus it acquires. This means predicting the outcomes of 2^N separate bargaining problems. Presumably, each player is looking around for his best option, and hence, is simultaneously engaged in quiet and informal bargaining processes with many other players or groups of players. It is unrealistic to suppose that each of these many informal bargaining sessions is governed by an arbitrator (we would need more arbitrators than players), so we must predict what would happen *without* an arbitrator. We could use the Nash solution of of §4B for this purpose, but first we must justify this by explaining why the quiet and informal bargaining sessions would likely yield Nash solutions as outcomes.

Bargaining vs. Haggling: Do the players have perfect knowledge of each other's preferences? If not, then each player might gain advantage by exaggerating his utility function, or misrepresenting his status quo point, or by pretending to be less averse to risk. In this context, Nash distinguished between *bargaining* (where the players have perfect knowledge of each others preferences) and *haggling* (where misrepresenting your preferences becomes possible, and hence is an essential strategy). For example, some of the fair division games of Chapter **??** are ingenious partly because each player has an incentive to be honest about his preferences. (But these games require an arbitrator.) Haggling is much more complex than bargaining: the space of strategies is vast, because it must include all the possible acts of deception. We will

[1]This means that the game is not 'zero-sum'; in other words, *some* players may find cooperation mutually beneficial at least *some* of the time. But of course, the players are still competing with each other.

therefore concentrate on bargaining[2].

A similar issue: do the players have perfect knowledge about the shape of the bargaining set $\mathcal{B}$? Or do they have incomplete information, so that some bargains are 'probably' feasible and others 'probably' aren't? This is not a question of willfull *misrepresentation*, but rather of *ignorance*. Nevertheless, incomplete information can affect the bargaining outcome. In some situations (e.g. the Nash Demand Game, §5C) we will see that the assumption of incomplete information actually makes things simpler.

Arbitration vs. Fair Division: We earlier pointed to the fair division protocols of Chapter IV to illustrate the role of an arbitrator. However, fair division is a much simpler problem than general arbitration. In a fair division problem, we effectively assume that the status quo or worst-case scenario (getting nothing) has utility *zero* for each player, while the best-case scenario (getting the whole 'cake') has utility of *one* for each player. We furthermore implicitly assume that the utilities of different players are *comparable* —indeed, a concept like 'equitable' allocation doesn't even make sense, otherwise. Neither of these assumptions is necessarily appropriate in a general bargaining problem, where one player may have much more to gain or lose than the other, and where interpersonal comparison of utility raises thorny philosophical issues. Nevertheless, if we rescale the player's utilities in a general bargaining problem, so that the worst-case scenario is has utility zero and the best case has utility one for each player, and we then try to divide the surplus 'equitably' on this scale, then we arrive at the Kalai-Smorodinsky solution of §7A.

Time and commitment: Are the players engaged in a long-term process where they can repeatedly propose offers and consider the other player's offers? Or is the bargaining a one-shot, take-or-leave situation? If the bargaining is protracted in time, then how much value does time have to the players? In other words, at what rate do they *discount* predicted future earnings? A closely related question: what is the probability that the bargaining 'breaks down' before an agreement is reached (either because one of the players walks away, or because some exogenous force intervenes)? Also, what powers of *commitment* do the players have? For example, can the players make credible[3] threats? Conversely, can the players make credible promises (i.e. can they engage in 'binding contracts')? One of the key roles of a modern government is the *enforcement* of contracts and such external enforcement is necessary precisely because certain bargains may be unenforceable —and hence, unobtainable —without it.

Implementation and renegotiation: Having selected a bargaining outcome, can the players move *instantaneously* from the status quo to this outcome? Or must the implementation process itself occur over time? The implementation then becomes a continuous path from the

[2]Note that this perhaps greatly limits the applicability of these mathematical models to real life negotations, which are often much closer to 'haggling' than to 'bargaining'.

[3]'Credible' means that Zara can be expected to follow through on her threat, even though it may also be detrimental to herself. Thus, the Owen must take her threat seriously.

status quo to the bargaining outcome. If the players don't trust each other, then this implementation path must be such that neither player is vulnerable to defection by the other player at any position along the path. Also, what stops one of the players from stopping half way through the implementation process, and demanding that the two players *renegotiate* their agreement? (This question of renegotiation is especially salient when nothing binds the two players to their commitments). If renegotiation is possible, then any bargain must be 'renegotiation-proof', meaning that neither player will ever have an incentive to renegotiate half way through the implementation.

Different assumptions are appropriate in different real-world situations. They lead to different mathematical models, which may have different outcomes.

5B Normal-form games and Nash equilibria

Recommended: §3A

A *normal form game* is a game where each player simultaneously chooses a single 'action', without knowing what actions the other players will take. The game then ends immediately, and the payoffs for each player are a function of the actions taken by her and all the other players.

Formally, a normal form game is totally described by the following data:

1. A set $\mathcal{I}$ of *players*.
2. For each $i \in \mathcal{I}$, a set $\mathcal{A}_i$ of possible *actions* (or 'strategies') for player i.
3. For each $i \in \mathcal{I}$, a *payoff function* $u_i : \mathcal{A} \longrightarrow \mathbb{R}$, where $\mathcal{A} := \prod_{i \in \mathcal{I}} \mathcal{A}_i$.

Suppose each player $i \in \mathcal{I}$ chooses action $a_i \in \mathcal{A}_i$, and let $\mathbf{a} := (a_i)_{i \in \mathcal{I}} \in \mathcal{A}$ be the resulting *outcome* of the game. Then the *payoff* for player j will be $u_j(\mathbf{a})$.

A normal-form game involves no probabilistic component[4]. Thus, the players' payoff functions can be interpreted as *ordinal* utilities, not *cardinal* utilities (see §3A for the distinction). In other words, if $\mathbf{a}, \mathbf{a}' \in \mathcal{A}$ are two outcomes, and $u_j(\mathbf{a}) = 5$ while $u_j(\mathbf{a}') = 2.5$, then this means that player j *prefers* outcome $\mathbf{a}$ to outcome $\mathbf{a}'$, but it does *not* necessarily mean that player j likes $\mathbf{a}$ "twice as much" as $\mathbf{a}'$.

Thus, if $\phi : \mathbb{R} \longrightarrow \mathbb{R}$ is any mononotone increasing function, and we define $u'_j := \phi \circ u_j$, then u'_j is *equivalent* to u_j as a description of the preferences (and thus, the strategic behaviour) of player j. In particular, if the set $\mathcal{A}$ is finite, then we can always choose u_j to take integer values; in other words, we can assume that $u_j : \mathcal{A} \longrightarrow \mathbb{Z}$ for all $j \in \mathcal{I}$. This is often a convenient assumption.

Suppose there are two players, so that $\mathcal{I} := \{0, 1\}$. In this case, $\mathcal{A} = \mathcal{A}_0 \times \mathcal{A}_1$, and the two payoff functions $u_0, u_1 : \mathcal{A} \longrightarrow \mathbb{Z}$ can be concisely expressed using a *payoff matrix*, as in the following examples.

[4] Unless we introduce randomized ('mixed') strategies; see Remark 5B.5(a) below.

Example 5B.1: (a) *Bistro or Szechuan:* Zara and Owen have agreed to meet for lunch on Tuesday. They discussed meeting at either *Belle's Bistro* or *Lucky Dragon Szechuan Palace*, but they forgot to decide which one! Now they are separated, and each one must choose to go to one restaurant or the other, and hope that the other person makes the same choice. Thus, $\mathcal{A}_0 = \mathcal{A}_1 = \{B, S\}$, where B means "Go to the Bistro" and S means "Go to Szechuan".

Zara prefers the *Bistro*, but Owen prefers *Szechuan*. However, both would prefer to dine together rather than dine alone. Thus, their payoff matrices are as follows:

Zara's action \ Owen's action	B	S
B	4	2
S	1	3

Zara's payoff matrix

Zara's action \ Owen's action	B	S
B	3	2
S	1	4

Owen's payoff matrix

In other words, Zara orders the game outcomes as follows:

4 (Most preferred) Eat at Bistro with Owen.

3 Eat Szechuan with Owen.

2 Eat at Bistro while Owen eats Szechuan.

1 (Least preferred) Eat Szechuan alone while Owen goes to the Bistro.

Owen's ordering is similar, but with "Szechuan" and "Bistro" reversed. We can combine these two payoff matrices into a single *joint payoff matrix*:

Zara's action \ Owen's action	B	S
B	$_4\backslash^3$	$_2\backslash^2$
S	$_1\backslash^1$	$_3\backslash^4$

Here, in each box '$[_a\backslash^b]$', the payoff for Zara is a and the payoff for Owen is b.

(b) *Stag Hunt:* Zara and Owen are hunting a stag in the forest. If they work together, they can capture the stag and share in the bounty. However, each one also has the option of deserting the hunt to catch a hare instead. The capture of a hare is guaranteed, regardless of what the other player does. A stag is more than four times the size of a hare, so a successful stag hunt would yield more than twice as much food for each hunter, if they evenly divide the meat. However, if either person deserts (to go catch a hare), then the stag will escape. Thus, $\mathcal{A}_0 = \mathcal{A}_1 = \{C, D\}$, where C means "Cooperate in the hunt" and D means "Desert (and catch a hare)". The joint payoff matrix is as follows:

Zara's action \ Owen's action	C	D
C	$_2\backslash^2$	$_0\backslash^1$
D	$_1\backslash^0$	$_1\backslash^1$

(c) *Prisoner's Dilemma:* Zara and Owen are partners in crime who have been arrested and imprisoned in separate cells. The police do not have enough evidence to convict them, so they approach Zara and ask her to testify against Owen. Meanwhile they ask Owen to testify against Zara. They offer each one the following deal:

1. If you testify against your partner, then we will immediately let you go free, and imprison your partner for ten years.
2. However, if your partner testifies against you instead, then he/she will immediately go free, and *you* will go to jail for ten years.
3. If neither one of you testifies, then the trial will probably drag on for a year, but eventually you will probably both be acquitted.
4. If you *both* testify, then we will have enough evidence to convict you *both* for ten years. But we will reduce the sentence to eight years because you cooperated with the investigation.

Thus, in this case, $\mathcal{A}_0 = \mathcal{A}_1 = \{C, D\}$, where C stands for "Conform to the Code of Silence", and D stands for "Defect" (i.e. testify against your buddy). If we measure the payoffs in years, then the joint payoff matrix is as follows:

Zara's action \ Owen's action	C	D
C	$-1 \backslash -1$	$-10 \backslash 0$
D	$0 \backslash -10$	$-8 \backslash -8$

However, if we treat payoffs as ordinal (not cardinal) utilities, then it is equivalent to use the matrix

Zara's action \ Owen's action	C	D
C	$3 \backslash 3$	$1 \backslash 4$
D	$4 \backslash 1$	$2 \backslash 2$

(d) *Chicken*[5]*:* Zara and Owen are racing their Kawasaki motorcycles along a narrow bridge at 180 km/h in opposite directions. Collision is imminent! Each player has a choice to either swerve off the bridge (and land in the water) or keep going straight. Each player would prefer *not* to swerve (to avoid getting wet and wrecking his/her Kawasaki). Each player also feels that, if he/she must get wet, then the other player should also get wet. But if neither player swerves, then they will collide and both die. Thus, $\mathcal{A}_0 = \mathcal{A}_1 = \{S, K\}$, where S means "Swerve" and K means "Keep going". The joint payoff matrix is as follows:

Zara's action \ Owen's action	S	K
S	$3 \backslash 3$	$2 \backslash 4$
K	$4 \backslash 2$	$1 \backslash 1$

[5]Sometimes called *Hawk and Dove*.

(e) *Scissors-Rock-Paper:* In this popular game, $\mathcal{A}_0 = \mathcal{A}_1 = \{S, R, P\}$, where S is "Scissors", R is "Rock", and P is "Paper". The rules are: "Scissors cuts Paper, Paper wraps Rock, and Rock blunts Scissors." Thus, the payoff matrix is as follows:

Zara's action \ Owen's action	S	R	P
S	$0 \backslash 0$	$-1 \backslash 1$	$1 \backslash -1$
R	$1 \backslash -1$	$0 \backslash 0$	$-1 \backslash 1$
P	$-1 \backslash 1$	$1 \backslash -1$	$0 \backslash 0$

♢

Given a normal form game, our goal is either to prescribe what action each player *should* take, or (closely related) to predict which action each player *will* take. In some cases, there is clearly a best strategy for each player. Given an action $a_1 \in \mathcal{A}_1$ for Owen, we say that $b_0 \in \mathcal{A}_0$ is a ***best response*** to a_1 for Zara if for all other $a_0 \in \mathcal{A}_0$, we have $u_0(b_0, a_1) \geq u_0(a_0, a_1)$. In other words, b_0 is Zara's optimal strategy, assuming that Owen has already committed to action a_1.

Likewise, given an action $a_0 \in \mathcal{A}_0$ for Zara, we say that $b_1 \in \mathcal{A}_1$ is a ***best response*** to a_0 for Owen if for all other $a_1 \in \mathcal{A}_1$, we have $u_1(a_0, b_1) \geq u_1(a_0, a_1)$. In other words, b_1 is Owen's optimal strategy, assuming that Zara has already committed to action a_0.

Example 5B.2: (a) Consider *Bistro or Szechuan*. If Owen chooses *Bistro*, then Zara's best response is *Bistro*, and vice versa.

If Owen chooses *Szechuan*, then Zara's best response is *Szechuan*, and vice versa.

(b) Consider *Stag Hunt*. If Owen chooses to *Cooperate*, then Zara's best response is to also *Cooperate*, and vice versa.

If Owen chooses to *Desert*, then Zara's best response is to also *Desert*, and vice versa.

(c) Consider *Chicken*. If Owen chooses to *Keep going*, then Zara's best response is to *Swerve*.

If Owen chooses to *Swerve*, then Zara's best response is to *Keep going*.

If Zara chooses to *Keep going*, then Owen's best response is to *Swerve*.

If Zara chooses to *Swerve*, then Owen's best response is to *Keep going*.

(d) In *Scissors-Rock-Paper*, the best response (for either player) to *Scissors* is *Rock*. The best response to *Rock* is *Paper*, and the best response to *Paper* is *Scissors*. ♢

An action $b_0 \in \mathcal{A}_0$ is ***dominant strategy*** for Zara if b_0 is a best response to *every* possible action in $\mathcal{A}_1$. An action $b_1 \in \mathcal{A}_1$ is ***dominant strategy*** for Owen if b_1 is a best response to *every* possible action in $\mathcal{A}_0$. If one of the players has a unique dominant strategy, then clearly we should prescribe that she use this strategy. Furthermore, if the player is rational, then we can predict with some confidence that she *will* use this strategy.

Example 5B.3: (a) Consider *Prisoner's Dilemma*. The action *Defect* is the unique dominant strategy for each player. Thus, we can predict that each player, if rational, will 'Defect' and betray his/her partner. The outcome will be (D, D), and each person will serve eight years in jail.

(b) In *Bistro or Szechuan*, or *Stag hunt*, or *Chicken* or *Scissors, Rock, Paper*, neither player has a dominant strategy. ♢

If one player (say, Zara) has a dominant strategy b_0, then the other player (Owen) can predict that she will choose b_0. Then Owen's rational reaction is to choose his best response b_1 to b_0. The pair (b_0, b_1) is then called a *dominant strategy equilibrium* for the game. Thus, in a two-player game, if either player has a dominant strategy, then the game has a completely predictable outcome.

However, if neither player has a dominant strategy, then the analysis is more complex. Each player must 'guess' what the other player might do. For example, in *Bistro or Szechuan*, Owen's best response depends on his prediction of what Zara will do, and vice versa. A *Nash equilibrium* is a situation where each player *correctly* predicts the other player's action, so that each player's action is a best response to the action of the other player. Formally, a *Nash equilibrium* is a pair $(b_0, b_1) \in \mathcal{A}_0 \times \mathcal{A}_1$ such that:

- b_0 is a best response for Zara to b_1.
- b_1 is a best response for Owen to b_0.

Example 5B.4: (a) If either player has a *dominant* strategy, then in any Nash equilibrium, she will always play that strategy. For example, in *Prisoner's Dilemma*, the only Nash Equilibrium is the dominant-strategy equilibrium (D, D).

(b) *Bistro or Szechuan* has two Nash equilibria: (B, B) and (S, S).

(c) *Stag hunt* has two Nash equilibria: (C, C) and (D, D).

(d) *Chicken* has two Nash equilibria: (S, K) and (K, S).

(e) *Scissors-Rock-Paper* has *no* Nash equilibrium. ♢

Remarks 5B.5: (a) The Nash equilibrium concept has two major problems, which limit its predictive/prescriptive value:

- Some games (e.g. *Stag Hunt*, *Chicken* and *Bistro or Szechuan*) have multiple Nash equilibria.
- Some games (e.g. *Scissors-Rock-Paper*) have *no* Nash equilibrium.

The first problem is called the problem of *equilibrium selection*. To solve it, we must introduce additional criteria or constraints which select one of the Nash equilibria as being somehow better or more natural than others. There is no generally satisfactory solution to this problem (but see Remark (e) below).

The second problem can be solved by allowing each player to randomly choose an action according to some probability distribution which she specifies; this probability distribution is called her *mixed strategy*. The problem is then to choose a mixed strategy which is a *best response* to your opponent's mixed strategy, meaning that it maximizes your *expected* payoff. (Note that, once we start computing *expected* payoffs, we must treat the payoffs as *cardinal* utilities and not *ordinal* utilities; see §3A).

The famous *Nash Equilibrium Theorem* says that *every* finite, normal-form game has at least one Nash equilibrium, if we allow mixed strategies [OR94, Proposition 33.1]. For example *Scissors-Rock-Paper* has a unique mixed-strategy Nash equilibrium where both players use probability distributions $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. We will not discuss mixed strategies further, because they are not relevant to bargaining theory.

(b) The *Prisoner's Dilemma* is often used as a model of a more general problem called the *Tragedy of the Commons*. This refers to any situation where two or more people share the benefits of some 'common good'. Each person benefits if everyone else 'Cooperates' by contributing to the common good, but each person has an incentive to unilaterally 'Defect' at everyone else's expense. For example, instead of criminals, imagine that Zara and Owen are two countries which must individually decide whether to participate in a multinational agreement to reduce their pollution output, ban the use of land mines, etc.

The 'tragedy' is that each player's dominant strategy is to Defect, so the dominant strategy equilibrium results in the worst-case scenario for everyone. This example highlights the fact that game-theoretic equilibria may possess *predictive* or *prescriptive* content, but not necessarily any *normative* content.

(c) In all the examples we have introduced, both players have the same set of actions. Also, there is usually some kind of symmetry (or antisymmetry) between the payoffs for the two players. This is a coincidence. In general, different players may have different sets of actions; even if they have the same action-sets, there is not necessarily any relationship between their payoffs.

(d) *Scissors-Rock-Paper* is an example of a *purely competitive* game, meaning that the payoffs of one player are ordered inversely to the payoffs of the other player. Such games are sometimes called *zero-sum* games, because we can always choose the (ordinal) payoff functions such that the sum of the payoffs in each box equals zero (as is the case here). (Note, however, that the 'zero-sum' property depends on a particular choice of ordinal payoff functions; a game can be purely competitive without being zero-sum).

(e) At the opposite extreme are *purely cooperative* games, in which the payoffs for the Nash equilibrium boxes are ordered in the *same* way by each player. For example, *Stag Hunt* is a purely cooperative game, because both players prefer equilibrium (C, C) to equilibrium (D, D). (Note that this does *not* mean the players assign the *same* payoff to each equilibrium. Nor does it mean that they order the nonequilibrium boxes in the same way).

The games *Stag Hunt* and *Bistro or Szechuan* have exactly the same strategic structure and exactly the same Nash equilibria. The difference is that *Stag Hunt* is cooperative, whereas *Bistro or Szechuan* is not. Thus, in *Stag Hunt*, the problem of 'equilibrium selection' can be resolved: we simply select the equilibrium (C, C), because it is Pareto-preferred to the equilibrium (D, D). This resolution does not apply to *Bistro or Szechuan*.

Stag Hunt is a model of the problem of *coordination*: the players can both benefit, as long each player trusts the other player to do the right thing and 'Cooperate'. This is different than *Prisoner's Dilemma*, because in *Stag Hunt*, the box (C, C) is a Nash Equilibrium; in other words, each player *will* cooperate as long as she believes that the other player will also cooperate. (In *Prisoner's Dilemma*, the box (C, C) is *not* an equilibrium, because each player will Defect even if she knows the other player intends to Cooperate). The 'Stag Hunt' story comes from a parable by Rousseau, which he used to illustrate the problem of coordination in society and to explain the origin of 'social contracts'; see [Sky03].

(f) In the normal-form games we have considered here, the payoffs from every outcome are deterministic —there is no random component. This is not a good model of games of chance (e.g. card games), where each player's payoff is determined by a random factor in addition to her own strategic choices. However, it is easy to extend the framework developed here to encompass games of chance. We proceed as if each outcome $\mathbf{a} \in \mathcal{A}$ actually defines a *lottery* which awards a random utility to each player according to some probability distribution (which depend on $\mathbf{a}$). Then we treat the payoffs in each box of the matrix as the *expected* utilities of these lotteries. (Note that this requires us to interpret these payoffs as *cardinal*, not *ordinal* utilities).

(g) Normal-form games are unrealistic because we assume that every player has 'perfect knowledge' about the utility function (i.e. preferences) of every other player. However, in many real-life 'games', players are partially or totally ignorant of one another's preferences (indeed, it is often a powerful strategy to systematically misrepresent your preferences, so as to manipulate the other players). This can be formally modeled using a *Bayesian game*. In such a game, each player is randomly assigned a *type* (that is, a particular payoff function), which is known only to her and not to the other players. Each player must then choose a strategy to maximize her payoff, even though she is ignorant of the 'types' of the other players (and hence, cannot necessarily predict their behaviour). The appropriate equilibrium concept in such a game is a *Bayesian Nash equilibrium*. Loosely speaking, each player chooses an optimal strategy for *each* of her possible types, based on assumptions about the strategies which will be employed by all the other players/types, and assumptions about the probability of every type for each player; furthermore, the players' assumptions must all be mutually consistent, as in the case of a Nash equilibrium.

(h) Normal-form games are also unrealistic because the entire game occurs in a single instant, with all players moving simultaneously. This does not represent the unfolding sequence of moves and countermoves which is found in most real-life games. A game which unfolds in time is called an *extensive game*; we will consider these in §5F and §5G.

Exercise 5.1: For each of the following games, determine the following:

(a) What are the dominant strategies (if any)?

(b) What are the Nash equilibria (if any)?

(c) Is the game purely competitive? Purely cooperative? Why or why not?

Zara's action \ Owen's action	A	B
A	$1 \backslash -1$	$-1 \backslash 1$
B	$-1 \backslash 1$	$1 \backslash -1$

Cat & Mouse

Zara's action \ Owen's action	H	T
H	$1 \backslash 1$	$0 \backslash 0$
T	$0 \backslash 0$	$1 \backslash 1$

Matching Pennies

Zara's action \ Owen's action	$L > R$	$L = R$	$L < R$
L	$3 \backslash 1$	$2 \backslash 2$	$1 \backslash 3$
R	$1 \backslash 3$	$2 \backslash 2$	$3 \backslash 1$

I cut, you choose (simultaneously)

Interpretation: *Cat & Mouse:* Zara is a cat and Owen is a mouse. Owen can hide in one of two locations, A or B, and Zara can search one of these locations. If Zara searches where Owen is hiding, she wins and he loses. Otherwise he wins.

Matching Pennies: Zara and Owen can each place a penny on the table with either the head (H) or tail (T) facing up. If the pennies match, they both win; otherwise they both lose.

I cut, you choose (simultaneously): Zara and Owen are dividing a cake. Owen will cut the cake into two pieces, L and R; he can either make the left piece bigger than the right ($L > R$), make them both the same size ($L = R$) or make the left piece smaller than the right ($L < R$). Zara must choose which piece she wants, but she must make this choice *before* she sees how Owen has cut the cake.

Note that this is different than the 'I cut, you choose' cake division game 8C.1 on page 168, because in *that* game, Zara can make her choice *after* Owen cuts the cake. This is the difference between a normal-form game (simultaneous moves) and an *extensive game* (with successive moves).

Exercise 5.2: An action $b_1 \in \mathcal{A}_1$ is a *maximin* strategy for Owen if

$$\forall\, a_1 \in \mathcal{A}_1, \quad \min_{a_0 \in \mathcal{A}_0} u_1(a_0, b_1) \quad \geq \quad \min_{a_0 \in \mathcal{A}_0} u_1(a_0, a_1).$$

Thus, b_1 is the 'risk-averse' strategy which maximizes the worst-case-scenario payoff for Owen no matter what Zara does. This is especially appropriate in a purely competitive game, where Owen can assume that Zara will always try to 'minimize' his payoff, so as to maximize her own. Maximin strategies are defined analogously for Zara.

What are the maximin strategies (if any) for the players in each of the three games above?

Exercise 5.3: The concepts in this section also apply to games with three or more players (although we can no longer easily represent the payoffs using a two-dimensional matrix).

(a) Generalize the definition of *best response* to games with three or more players.

(b) Generalize the definition of *dominant strategy* to games with three or more players.

(c) For any $N \geq 3$, construct an N-player version of *Prisoner's Dilemma*: a game with a single dominant-strategy equilibrium which gives every player a low payoff, and which also has another (non-equilibrium) outcome where every player gets a higher payoff.

(d) Generalize the definition of *Nash equilibrium* to games with three or more players.

(e) Generalize the definition of *purely cooperative* to three or more players.

(f) For any $N \geq 3$, construct an N-player version of *Stag Hunt*: a purely cooperative game with two Nash equilibria, where one equilibrium is Pareto-preferred to the other one.

5C The Nash demand game

Prerequisites: §4B, §5B **Recommended:** §4C, §5A

Let $(\mathcal{B}, \mathbf{0})$ be a bargaining problem with status quo $\mathbf{0} = (0, 0)$. Nash himself proposed a simple model of one-shot bargaining called the *demand game* [Nas50]. The rules are as follows:

- Zara issues a 'demand' for some amount of utility $b_0 \in \mathbb{R}_{\neq}$, while simultaneously, Owen issues a 'demand' for some amount of utility $b_1 \in \mathbb{R}_{\neq}$.
- If the ordered pair (b_0, b_1) is inside the bargaining set $\mathcal{B}$, then each player gets exactly what he/she asked for. If (b_0, b_1) is *not* in $\mathcal{B}$. then each player gets their status quo payoff of zero.

Formally, this is a normal-form game where $\mathcal{A}_0 = \mathcal{A}_1 = \wp_{\mathbf{0}}\mathcal{B}$, and where the payoffs are simply the player's utility functions. The players face conflicting incentives: each one wants to ask for as much as possible, but if s/he asks for too much, then s/he may get nothing.

Notes: (a) Both players must make their demands *simultaneously*, each in ignorance of the other player's offer. If Zara is allowed to announce her demand first, then Owen is placed in a 'take-it-or-leave-it' sitation. This is called the *Ultimatum game*, and clearly gives the Zara an overwhelming advantage. The *Ultimatum game* gives the Zara the power to make a perfectly credible threat, since there is exactly one chance to reach an agreement.

(b) The *Demand Game* is appropriate in one-shot bargaining situations where the players will not have a chance to compromise or make a second offer, either because there is no time, or because neither player can stand to 'lose face' by backing down from his original demand. In most civilized bargaining scenarios, these are obviously not reasonable assumptions.

Lemma 5C.1 *Every point on the negotiation set $\wp_{\mathbf{0}}\mathcal{B}$ is a Nash equilibrium of the Nash Demand Game.*

Proof: Lemma 4A.1 says we can assume that the Pareto frontier $\wp\mathcal{B}$ is the graph of some continuous, nonincreasing function $\Gamma_1 : \mathbb{R}_{\neq} \longrightarrow \mathbb{R}_{\neq}$. In other words, for any $b_0 \in \mathbb{R}_{\neq}$, $\Gamma_1(b_0)$ represents the most utility that Owen can get, given that Zara is getting b_0. Likewise, if $\Gamma_0 := \overline{b}_1^{-1} : \mathbb{R}_{\neq} \longrightarrow \mathbb{R}_{\neq}$, then for any $b_1 \in \mathbb{R}_{\neq}$, $\Gamma_0(b_1)$ represents the most utility that Zara can get, given that Owen is getting b_1. At this point, it suffices to make the following observations:

(a) Suppose Owen believes that Zara will demand b_0. Then his best response is to demand $\Gamma_1(b_0)$.

(b) Suppose Zara believes that Owen will demand b_1. Then her best response is to demand $\Gamma_0(b_1)$.

(c) Thus, for any $(b_0, b_1) \in \wp_{\mathbf{q}}\mathcal{B}$, if each player believes that the other will play (b_0, b_1), then neither player has an incentive to unilaterally deviate from this strategy; hence it is a Nash equilibrium.

Exercise 5.4 Verify assertions (a), (b), and (c). □

Thus, by itself, the Nash Demand Game does not yield any progress in the Nash program. To obviate this difficulty, Nash introduced a 'smoothed' version of the game. In this game, the players have *imperfect knowledge* of the bargaining set $\mathcal{B}$. In other words, there is a function $\rho : \mathbb{R}^2_{\nearrow} \longrightarrow [0, 1]$, so that, for any $(b_0, b_1) \in \mathbb{R}^2_{\nearrow}$, $\rho(b_0, b_1)$ is the *probability* that (b_0, b_1) will actually be a feasible bargain[6]. Now the players' task is to maximize their *expected utility*, according to ρ. For any demand pair $(b_0, b_1) \in \mathbb{R}^2_{\nearrow}$, the expected utility for Zara is $\rho(b_0, b_1) \cdot b_0$, because there is a probability $\rho(b_0, b_1)$ that (b_0, b_1) is feasible (in which case she gets b_0) and a probability of $1 - \rho(b_0, b_1)$ that (b_0, b_1) is not feasible (in which case she gets 0 because we assume the status quo is at $\mathbf{0}$). Likewise the expected utility for Owen is $\rho(u, b_0) \cdot b_1$.

Proposition 5C.2 *For any $\epsilon \in (0, 1]$, let $\mathcal{B}_\epsilon := \big\{(b_0, b_1) \in \mathbb{R}^2_{\nearrow} \; ; \; \rho(b_0, b_1) \geq \epsilon\big\}$. Let $(b_0^\epsilon, b_1^\epsilon) \in \mathcal{B}_\epsilon$ be the Nash Solution to the bargaining problem $(\mathcal{B}_\epsilon, \mathbf{0})$ —in other words, $(b_0^\epsilon, b_1^\epsilon)$ is the point in $\mathcal{B}_\epsilon$ which maximizes the Nash product $b_0 \cdot b_1$.*

The set of Nash equilibria for the Smoothed Demand Game defined by ρ is then the set $\{(b_0^\epsilon, b_1^\epsilon) \; ; \; \epsilon \in (0, 1]\}$.

If we let the 'uncertainty level' of the Smoothed Demand Game tend to zero in an appropriate sense, then this collection of Nash equilibria converges on the Nash bargaining solution for the original bargaining problem. To state this precisely, we use the Hausdorff metric d_H of §4C.

Corollary 5C.3 *Let $\{\rho^n\}_{n=1}^\infty$ be a sequence of functions from $\mathbb{R}^2_{\nearrow}$ to $[0, 1]$. For each $n \in \mathbb{N}$, and each $\epsilon \in (0, 1]$, let*

$$\mathcal{B}^{n,\epsilon} \quad := \quad \big\{(b_0, b_1) \in \mathbb{R}^2_{\nearrow} \; ; \; \rho^n(b_0, b_1) \geq \epsilon\big\}.$$

and let $(b_0^{n,\epsilon}, b_1^{n,\epsilon})$ be the Nash solution for the bargaining problem $(\mathcal{B}^{n,\epsilon}, \mathbf{0})$. Suppose that, for all $\epsilon \in [0, 1]$, $\lim_{n\to\infty} d_H(\mathcal{B}^{n,\epsilon}, \mathcal{B}) = 0$. Then $\lim_{n\to\infty} (b_0^{n,\epsilon}, b_1^{n,\epsilon}) = (b_0, b_1)$, where $(b_0, b_1) = \eta(\mathcal{B}, \mathbf{0})$ is the Nash solution for $(\mathcal{B}, \mathbf{0})$.

Proof: This follows immediately from Proposition 5C.2 and from the fact that the Nash solution is Hausdorff-continuous by Lemma 4C.1. □

[6]Note: ρ is *not* a probability density function.

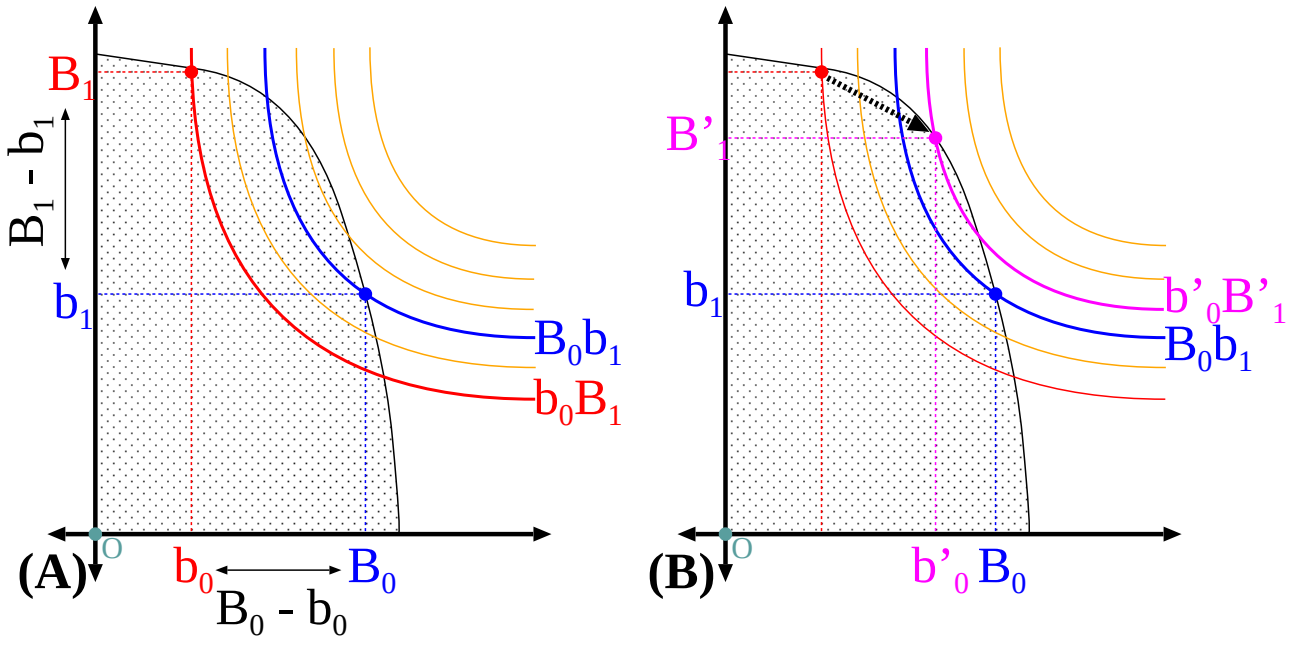


Figure 5.1: The Harsanyi-Zeuthen concession model.

Thus, loosely speaking, we can say the following: In a Nash Demand Game where there is a 'small' amount of uncertainty about which outcomes are feasible, the players will pick an equilibrium strategy which is 'close' to the Nash solution to the underlying bargaining problem. This is not a completely satisfactory solution to the Nash program; the Demand Game is highly restrictive and artificial in its assumptions, and it only yields the Nash solution in a somewhat technical and convoluted way.

5D The Harsanyi-Zeuthen concession model

Prerequisites: §4B **Recommended:** §5A

Let $(\mathcal{B}, \mathbf{0})$ be a bargaining problem with status quo $\mathbf{0} = (0, 0)$. Zeuthen [Zeu30] proposed a bargaining model in 1930, long before von Neumann, Morgenstern, or Nash. Zeuthen's model was later examined by Harsanyi [Har56], [LR80, §6.7, p.135]. As in Lemma 4A.1, we suppose that the negotiating set $\wp_{\mathbf{q}}\mathcal{B}$ is the graph of some continuous, nonincreasing function $\Gamma_1 : \mathbb{R}_{\neq} \longrightarrow \mathbb{R}_{\neq}$. Thus, for any $b_0 \in \mathbb{R}_{\neq}$, $\Gamma_1(b_0)$ is the most that Owen can get, given that Zara is getting b_0. Likewise, if $\Gamma_0 := \overline{b}_1^{-1} : \mathbb{R}_{\neq} \longrightarrow \mathbb{R}_{\neq}$, then for any $b_1 \in \mathbb{R}_{\neq}$, $\Gamma_0(b_1)$ is the most that Zara can get, given that Owen is getting b_1. The Harsanyi-Zeuthen model is then the following caricature of two people haggling over a price in a bazaar:

1. Each player makes an initial demand. Let's say Zara demands B_0 and Owen demands B_1. (Figure 5.1A)

2. If (B_0, B_1) is not feasible (which is likely), then one of the players must make a concession. To predict who will concede first, let $b_1 := \Gamma_1(B_0)$ and $b_0 := \Gamma_0(B_1)$. We presume that $b_0 < B_0$ and $b_1 < B_1$. Then we predict

   - Owen will concede if $\dfrac{B_1 - b_1}{B_1} < \dfrac{B_0 - b_0}{B_0}$.
   - Zara will concede if $\dfrac{B_1 - b_1}{B_1} > \dfrac{B_0 - b_0}{B_0}$.

   Exercise 5.5 Check that $\dfrac{B_1 - b_1}{B_1} < \dfrac{B_0 - b_0}{B_0}$ if and only if $B_1 b_0 < b_1 B_0$.

   In other words, Owen concedes if and only if the Nash product of his proposal is smaller than the Nash product of Zara's proposal, (and vice versa.)

   Suppose Owen concedes. Then he will make an offer $B_1' < B_1$ such that, if $b_0' = \Gamma_0(B_1')$, then $B_1' b_0' \geq b_1 B_0$, as shown in Figure 5.1B (otherwise he would just have to immediately concede again).

3. Next, Zara must concede, by making an offer $B_0' < B_0$ such that, if $b_1' = \Gamma_1(B_0')$, then $b_1' B_0' \geq B_1' b_0'$.

The players then make alternating concessions, leading to a series of offers such that

$$B_1 > B_1' > B_1'' > \cdots > b_1'' > b_1' > b_1 \quad \text{and} \quad B_0 > B_0' > B_0'' > \cdots > b_0'' > b_0' > b_0,$$

while the Nash product of these offers steadily increases; for example, if Owen concedes first, we get the sequence:

$$B_1 b_0 < b_1 B_0 < B_1' b_0' < b_1' B_0' < B_1'' b_0'' < b_1'' B_0'' < \cdots$$

Define $b_0^1 := b_0'$, $b_0^2 := b_0''$, $b_0^3 := b_0'''$, etc., and similarly define B_0^n, b_1^n, and B_1^n. If $b_0^* := \lim_{n\to\infty} B_0^n = \lim_{n\to\infty} b_0^n$ and $b_1^* := \lim_{n\to\infty} B_1^n = \lim_{n\to\infty} b_1^n$, then it follows that (b_0, b_1) maximizes the value of the Nash product $b_0 b_1$. In other words, (b_0^*, b_1^*) is the *Nash solution* to the bargaining problem.

It remains to justify Zeuthen's "concession" rule. To do this, recall that neither player wants the negotations to break down, because then each one ends up with their status quo payoff of zero. Thus, Owen will concede if he thinks that Zara is more likely to terminate negotiations than to conceede herself (and vice versa). Hence, each player tries to estimate the probability that the other player will terminate negotiations. Suppose Zara believes that Owen is unwilling to concede, and is intransigent in his current demand of B_1. Then her choice is either to terminate negotiations, or to accept a payoff of $b_0 = \Gamma_0(B_1)$, in which case her concession relative to her current demand is $\frac{B_0 - b_0}{B_0}$. Likewise, if Owen believes Zara to be intransigent, then he can either terminate negotiations or take a relative loss of $\frac{B_1 - b_1}{B_1}$. The player who is *more* likely to terminate negotiations is the one facing the bigger relative concession. In other words, if $\frac{B_0 - b_0}{B_0} > \frac{B_1 - b_1}{B_1}$, then Zara is more likely to terminate negotiations. Knowing

this, it is in Owen's best interest to concede, so that the negotiations continue. Likewise, if if $\frac{B_0-b_0}{B_0} < \frac{B_1-b_1}{B_1}$, then Owen is more likely to walk away, so it is Zara who should concede.

This justification for Zeuthen's concession rule is merely a piece of *ad hoc* psychological reasoning; it does *not* establish that the concessions represent equilibrium strategies in some bargaining game. Indeed, the Zeuthen-Harsanyi model is *not* a bargaining game; it is merely a quasideterministic bargaining *model*. Thus, it does not constitute a very satisfactory solution to the Nash program.

5E Discount Factors

Prerequisites: §4B **Recommended:** §5A

In §5F we will introduce a fairly complex and realistic model of bargaining as a 'game' of alternating offers, which was developed by Ariel Rubinstein and Ingolf Ståhl. A key ingredient of the Rubinstein-Ståhl model is that both Zara and Owen place more value on obtaining a particular bargain *right now* than they do on obtaining the same bargain in the future. In other words, they have what economists call a *discount factor*. This is a value $\delta \in (0, 1)$ which represents the rate at which the *present* utility of some anticipated future gain decays, as we postpone the anticipated arrival of this gain further into the future. For example, suppose that Owen wants an apple, and that, getting this apple *right now* has a utility of 1 for him. If δ represents his 24-hour discount factor, then the anticipation of getting the same apple *tomorrow* has a utility of δ for him right now, and the anticipation of getting the same apple the day after tomorrow will have a utility of δ^2 right now, and so on. Since $\delta < 1$, we see that $\lim_{n\to\infty} \delta^n = 0$, which means that Owen places almost no value on the anticipation of events happening in the far future. The smaller δ is, the more rapidly Owen discounts future gains —i.e. the more 'impatient' he is. Conversely, the closer δ gets to 1, the more patient Owen becomes and the more he values long-term gains.

All dynamical economic models include a discount factor. This explains, for example, why money is always lent with a nonzero interest rate: possessing \$100 today is worth more to Owen than lending it to Zara and getting it back tomorrow. Indeed, if having \$100 *today* is worth roughly the same to him as the anticipation of having \$110 *tomorrow*, then Zara must pay him \$10 of interest to make it worth his while to lend the \$100 to her. Conversely, suppose that having \$100 *today* is worth roughly the same to Zara as the anticipation of having \$110 *tomorrow*; then she is willing to borrow the money *today* with a 10% interest rate. Thus, the *interest rate* of a money market is closely related to the discount factors of the creditors and debtors in that market.

There are several ways to explain Owen's discount factor. The simplest explanation is that Owen is simply impatient, and prefers instant gratification. However, this 'psychological' explanation seems inconsistent with our idealization of Owen as a rational maximizer. Four other explanations are more consistent with rationality:

Immediate necessity: If Owen is starving, then he requires food immediately. The promise of

100 apples tomorrow is worth less to him than one apple today, because if he doesn't eat today he will die. Of course, most people in industrialized countries are not faced with this prospect of imminent starvation. However, on a longer time-scale, we all have basic ongoing expenses that must be paid. For example, no one can afford to be without an income indefinately, and this knowledge most certainly causes people to 'discount' anticipated future earnings relative to present income security. For example, no one would lock *all* of her money into long-term investments, even if these investments offered considerable long-term gain.

Physical decay: If Owen desires physical objects, then these objects will generally decay in value over time. For example, the apple will eventually become rotten. If Owen desires a piece of technology (say, a laptop computer), then over time that technology will become obsolete. If Owen desires a particular quantity of currency, then this currency will be worth less in the future because of inflation. Hence Owen prefers to get what he wants now.

Uncertainty about the future: Owen is unwilling to delay his gratification because there is a small chance that the world will change unpredictably and he will be unable to enjoy this anticipated future gratification. For example, he might die. In moneylending, the interest rate of a loan (the financial equivalent of a discount factor) is strongly correlated with the creditor's uncertainty about the prospect of ever getting his money back, either because the debtor seems likely to default, or because the broader political or economic situation is unstable.

In the context of bargaining with Zara, Owen will be unwilling to delay in reaching a settlement, because he is aware of that outside factors may randomly terminate their negotiations before a settlement can be reached. For example, Zara might suddenly discover a better opportunity elsewhere, and walk away from the bargaining table. Or suppose Zara and Owen represent two firms negotiating a collaborative venture to exploit some business opportunity. While they are negotiating, the market conditions may change; for example, a third party might arrive and exploit the business opportunity before they do, or an unanticipated new technology may render it worthless. We can model these random external forces by saying that, during each round of negotiations where an agreement is *not* reached, there is a probability of $(1-\delta)$ that the negotiations will randomly terminate, and a probability of δ that they will continue. Hence, any utility gains which Owen could expect to make in bargaining tomorrow (or the next day, etc.) must be multiplied by δ (or δ^2, etc.) to compute their *expected* utility for him today.

Time is money: Any time spent on negotiations is time which is *not* being spent on more productive activities, or on leisure activity. Hence, both parties will prefer to conclude the agreement sooner rather than later. (Note that the 'time is money' mentality is perhaps better modelled by subtracting a fixed amount c during each time period, rather than multiplying by δ. Hence a payoff of B_0 at time n is given the value $B_0 - cn$, rather than the value $\delta^n B_0$. We will return to this later).

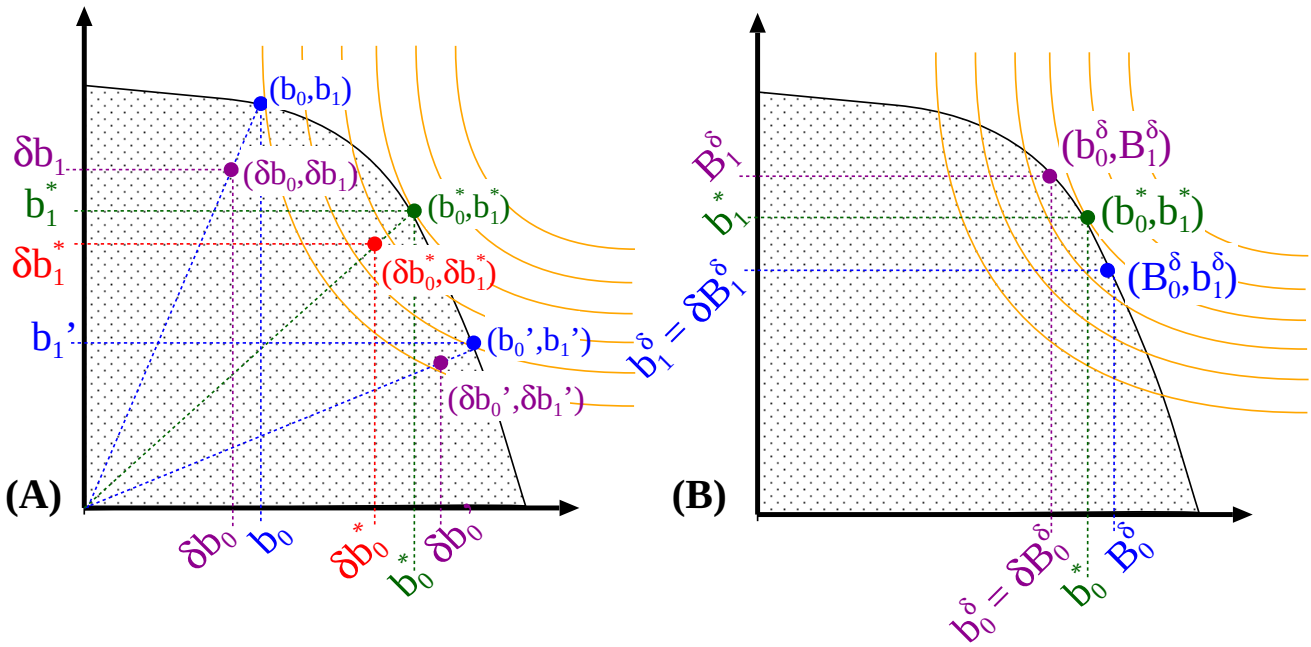


Figure 5.2: **(A)** Let (b_0^*, b_1^*) be the Nash solution, and let (b_0, b_1) and (b_0', b_1') be two other points on the Pareto frontier. As stated in Lemma 5E.1(a), if $\delta b_1 > b_1^*$, then $\delta b_0^* \geq b_0$. Likewise if $\delta b_0' > b_0^*$, then $\delta b_1^* \geq b_1'$. **(B)** As in Lemma 5E.1(b), $b_0^\delta \leq b_0^* \leq B_0^\delta$ while $b_1^\delta \leq b_1^* \leq B_1^\delta$, and $b_0^\delta = \delta \cdot B_0^\delta$ while $b_1^\delta = \delta \cdot B_1^\delta$.

The next result characterizes the Nash bargaining solution in terms of discount factors.

Lemma 5E.1 *Let $(\mathcal{B}, \mathbf{0})$ be a bargaining problem (with status quo $\mathbf{0}$). Let $\wp\mathcal{B}$ be the Pareto frontier of $\mathcal{B}$.*

(a) *The Nash solution of the bargaining problem $(\mathcal{B}, \mathbf{0})$ is the unique point (b_0^*, b_1^*) in $\wp\mathcal{B}$ such that, for any $\delta \in (0,1)$, and any other point $(b_0, b_1) \in \wp\mathcal{B}$,*

$$\Big(\delta b_1 > b_1^*\Big) \Longrightarrow \Big(\delta b_0^* \geq b_0\Big), \quad \textit{and} \quad \Big(\delta b_0 > b_0^*\Big) \Longrightarrow \Big(\delta b_1^* \geq b_1\Big). \quad \text{[Figure 5.2(A)]}$$

(b) *For any $\delta \in (0,1)$, there exist unique points (B_0^δ, b_1^δ) and (b_0^δ, B_1^δ) in $\wp\mathcal{B}$, such that $b_0^\delta \leq b_0^* \leq B_0^\delta$ and $b_1^\delta \leq b_1^* \leq B_1^\delta$, and such that*

$$b_0^\delta = \delta \cdot B_0^\delta \quad \textit{and} \quad b_1^\delta = \delta \cdot B_1^\delta. \qquad \text{[Figure 5.2(B)]}$$

(c) $\lim_{\delta \nearrow 1} (B_0^\delta, b_1^\delta) = (b_0^*, b_1^*) = \lim_{\delta \nearrow 1} (b_0^\delta, B_1^\delta)$.

Heuristic Interpretation of Lemma 5E.1(a): Suppose that Zara has proposed the Nash solution (b_0^*, b_1^*) to Owen, but Owen prefers the bargain (b_0, b_1) because $b_1 > b_1^*$. If $\delta b_1 > b_1^*$, then Owen will be intransigent in his demand, if he believes that this intransigence will eventually lead to outcome (b_0, b_1) during the next round of bargaining, because the (discounted) utility δb_1 of this future outcome is *greater* for him than the utility b_1^* of accepting the Nash solution (b_0^*, b_1^*) right now.

If $\delta b_0^* < b_0$, then Zara will capitulate to Owen's demand, because if she defies him and insists on (b_0^*, b_1^*), then she will obtain (at best) the bargain (b_0^*, b_1^*) during the next round of bargaining, which yields her a (discounted) utility of at most δb_0^*, which is *less* for her than the utility of b_0 she would get right now if she capitulated.

Thus, in order for the Nash solution to be robust against Owen's intransigence, it must be the case that $\delta b_0^* \geq b_0$ for any δ and (b_0, b_1) such that $\delta b_1 > b_1^*$. By reversing the roles of Zara and Owen, we see that it must also be the case that $\delta b_1^* \geq b_1$ for any δ and (b_0, b_1) such that $\delta b_0 > b_0^*$ (otherwise (b_0^*, b_1^*) is susceptible to Zara's intransigence).

Heuristic Interpretation of Lemma 5E.1(b): (B_0^δ, b_1^δ) and (b_0^δ, B_1^δ) represent 'bottom line' equilibrium bargaining positions for Zara and Owen, respectively. If Owen were to offer (b_0^δ, B_1^δ), then Zara would be indifferent between accepting his offer right now or holding out for the possibility of (B_0^δ, b_1^δ) in the future (discounted by δ). If he were to make any offer (b_0', B_1') such that $b_0' > b_0^\delta$ and $B_1' \leq B_1^\delta$, then Zara would certainly accept his offer right now, rather hold out for (B_0^δ, b_1^δ) in the future. Likewise if Zara were to offer (B_0^δ, b_1^δ), then Owen would be indifferent between accepting her offer right now or holding out for the possibility of (b_0^δ, B_1^δ) in the future (discounted by δ). If she were to make any offer (B_0', b_1') such that $B_0' \leq B_0^\delta$ and $b_1' \geq b_1^\delta$, then Owen would certainly accept her offer right now, rather hold out for (b_0^δ, B_1^δ) in the future.

Heuristic Interpretation of Lemma 5E.1(c): When δ is much less than 1, both Zara and Owen are highly impatient, and willing to make significant compromises to reach an agreeement quickly. However, as $\delta \nearrow 1$, they both become more and more patient, and are less willing to compromise. Their bottom line bargaining positions (B_0^δ, b_1^δ) and (b_0^δ, B_1^δ) each become more demanding —in other words, B_0^δ and B_1^δ both increase, while b_1^δ and b_0^δ both decrease. Eventually the positions (B_0^δ, b_1^δ) and (b_0^δ, B_1^δ) close in on (b_0^*, b_1^*) from opposite sides.

Proof of Lemma 5E.1: (a) **Exercise 5.6** (Hint: This follows from the fact that (b_0^*, b_1^*) is the unique maximum in $\wp\mathcal{B}$ of the Nash product $b_0 b_1$.)

(b) Let M_0 be the maximum utility for Zara of any element in $\mathcal{B}$, and let M_1 be the maximum utility for Owen. As described by Lemma 4A.1, suppose the Pareto frontier of $\mathcal{B}$ is the graph of the monotone decreasing function $\Gamma_1 : [0, M_0] \longrightarrow [0, M_1]$. Thus, for any $b_0 \in \mathbb{R}_{\neq}$, $\Gamma_1(b_0)$ is the most utility that Owen can get, given that Zara is getting b_0. Let $\Gamma_0 := \overline{b_1}^{-1} : [0, M_1] \longrightarrow [0, M_0]$; then for any $b_1 \in \mathbb{R}_{\neq}$, $\Gamma_0(b_1)$ is the most utility that Zara can get, given that Owen is getting b_1. To prove **(b)**, we must find values B_0 and B_1

such that $\Gamma_1(B_0) =: b_1 = \delta B_1$ and $\Gamma_0(B_1) =: b_0 = \delta B_0$. To do this, we define a function $\Phi_\delta : [0, M_0] \longrightarrow [0, M_0]$ by

$$\Phi_\delta(b_0) \;:=\; \delta \cdot \Gamma_0 \Big[\delta \cdot \Gamma_1(b_0)\Big].$$

Claim 1: *Φ_δ has a fixed point —i.e. there is some $b_0^\delta \in [0, M_0]$ such that $\Phi_\delta(b_0^\delta) = b_0^\delta$.*

Proof: **Exercise 5.7** (Hint: First explain why you can assume without loss of generality that $M_0 = 1 = M_1$. Then use the Contraction Mapping Theorem). ◇ Claim 1

Now, let $B_0^\delta := \dfrac{b_0^\delta}{\delta}$, let $B_1^\delta := \Gamma_1(b_0^\delta)$, and let $b_1^\delta := \delta B_1^\delta$. Then

$$\Gamma_0(b_1^\delta) \;=\; \Gamma_0(\delta \cdot B_1^\delta) \;=\; \frac{\delta}{\delta}\Gamma_0\Big[\delta \cdot \Gamma_1(b_0^\delta)\Big] \;=\; \frac{1}{\delta}\Phi_\delta(b_0^\delta) \;=\; \frac{b_0^\delta}{\delta} \;=\; B_0^\delta.$$

and conversely, $B_1^\delta = \Gamma_1(b_0^\delta)$. Thus, (B_0^δ, b_1^δ) and (b_1^δ, B_0^δ) are the pair we seek.

The proof that $b_0^\delta \le b_0^* \le B_0^\delta$ and $b_1^\delta \le b_1^* \le B_1^\delta$ is **Exercise 5.8** (Hint: use part **(a)**).

(c) For any $\delta < 1$, we have $\frac{b_0^\delta}{B_0^\delta} = \delta$, because $b_0^\delta = \delta B_0^\delta$ by definition. Thus, $\lim\limits_{\delta \nearrow 1} \dfrac{b_0^\delta}{B_0^\delta} = \lim\limits_{\delta \nearrow 1} \delta = 1$, which means that $\lim\limits_{\delta \nearrow 1} b_0^\delta = \lim\limits_{\delta \nearrow 1} B_0^\delta$. From **(b)** we know that $b_0^\delta \le b_0^* \le B_0^\delta$; hence we conclude that $\lim\limits_{\delta \nearrow 1} b_0^\delta = \lim\limits_{\delta \nearrow 1} B_0^\delta = b_0^*$.

By an identical argument, we get $\lim\limits_{\delta \nearrow 1} b_1^\delta = \lim\limits_{\delta \nearrow 1} B_1^\delta = b_1^*$. □

5F The Rubinstein-Ståhl *Alternating Offers* model

Prerequisites: §5E **Recommended:** §5A, §5B

Vulcans never bluff. —Spock, *Star Trek*, "The Doomsday Machine"

In 1974, Ingolf Ståhl introduced a model of bargaining as a 'game' where two players alternate in making 'offers' to one another [Stå72]. In 1982, Ariel Rubinstein showed that, under reasonable assumptions about the players' attitudes towards time, each player in Ståhl's bargaining game had a unique optimum strategy, so that the game had a unique outcome: the bargain that would be reached by perfectly rational playars [Rub82]. In 1986, Ken Binmore showed that this outcome converged to the Nash bargaining solution, thereby realizing the objective of the Nash program [Bin87, BRW86].

In this section, we will introduce the notion of an extensive game, define the Rubinstein-Ståhl model, and informally define the key concept of *subgame perfect equilibrium*, so that

we can precisely state Rubinstein's result (Theorem 5F.1). In §5G, we will rigorously prove Theorem 5F.1, after formally developing some background theory about extensive games and subgame perfect equilibria. Thus, the present section is intended to be accessible to a general audience, whereas §5G will require some amount of mathematical sophistication (or at least, stamina).

Let $(\mathcal{B}, \mathbf{0})$ be a bargaining problem with status quo $\mathbf{0}$. Assume some discount factor $\delta < 1$. The Rubinstein-Ståhl *Alternating Offers* game proceeds in stages, where at each stage, each player has the chance to make an offer and to consider the offer of the other player. We describe these stages inductively:

Stage 1: **(a)** Zara makes an initial offer $(B_0^1, b_1^1) \in \mathcal{B}$.

(b) Owen either accepts (B_0^1, b_1^1) (and the game ends), or he rejects it and makes a counter-offer, $(b_0^1, B_1^1) \in \mathcal{B}$.

Stage n: (*Assume Owen has just made an offer* $(b_0^{n-1}, B_1^{n-1}) \in \mathcal{B}$)

(a) Zara either accepts (b_0^{n-1}, B_1^{n-1}), or she rejects it and makes a counter-offer, $(B_0^n, b_1^n) \in \mathcal{B}$.

If Zara accepts the offer, then the game ends, and she receives a (discounted) payoff of $\delta^{2n} b_0^n$, and Owen receives a (discounted) payoff of $\delta^{2n} B_1^n$.

(b) Owen either accepts (B_0^n, b_1^n) (and the game ends), or he rejects it and makes a counter-offer, $(b_0^n, B_1^n) \in \mathcal{B}$.

If Owen accepts the offer, then the game ends, and he receives a (discounted) payoff of $\delta^{2n+1} b_1^n$, and Zara receives a (discounted) payoff of $\delta^{2n+1} B_0^n$.

Stage ∞: If the game does not end after finitely many stages, then both players get a payoff of zero.

Notes: (a) The game is exactly the same at every stage. In particular, the player's offers during the first $(n-1)$ rounds of play place no constraints on their bids at round n. Thus, for example, Zara's demand of B_0^n need bear no relation to her previous demand of B_0^{n-1}, or to Owen's offer of b_0^n.

(b) If $\delta = 1$, then the players can go on bargaining forever without penalty; there is no incentive for either one to compromise. It is the assumption that $\delta < 1$ which creates an incentive for the players to compromise and reach an agreement.

(c) Rather than directly discounting the nth stage payoffs by a factor of δ^n, we could let be payoffs remain *non*-discounted, but instead assume that, every stage, with probability $(1-\delta)$, some exogenous random event ends the bargaining prematurely. This would have exactly the same effect, because the probability that the nth stage even happens is then δ^n, so any payoff anticipated in the nth round must be multiplied by δ^n to obtain its expected utility. ♢

An *extensive game* is a game where the players alternate moves over time, so that the structure of this game can be described by an infinite 'tree', where each 'branch' of this tree

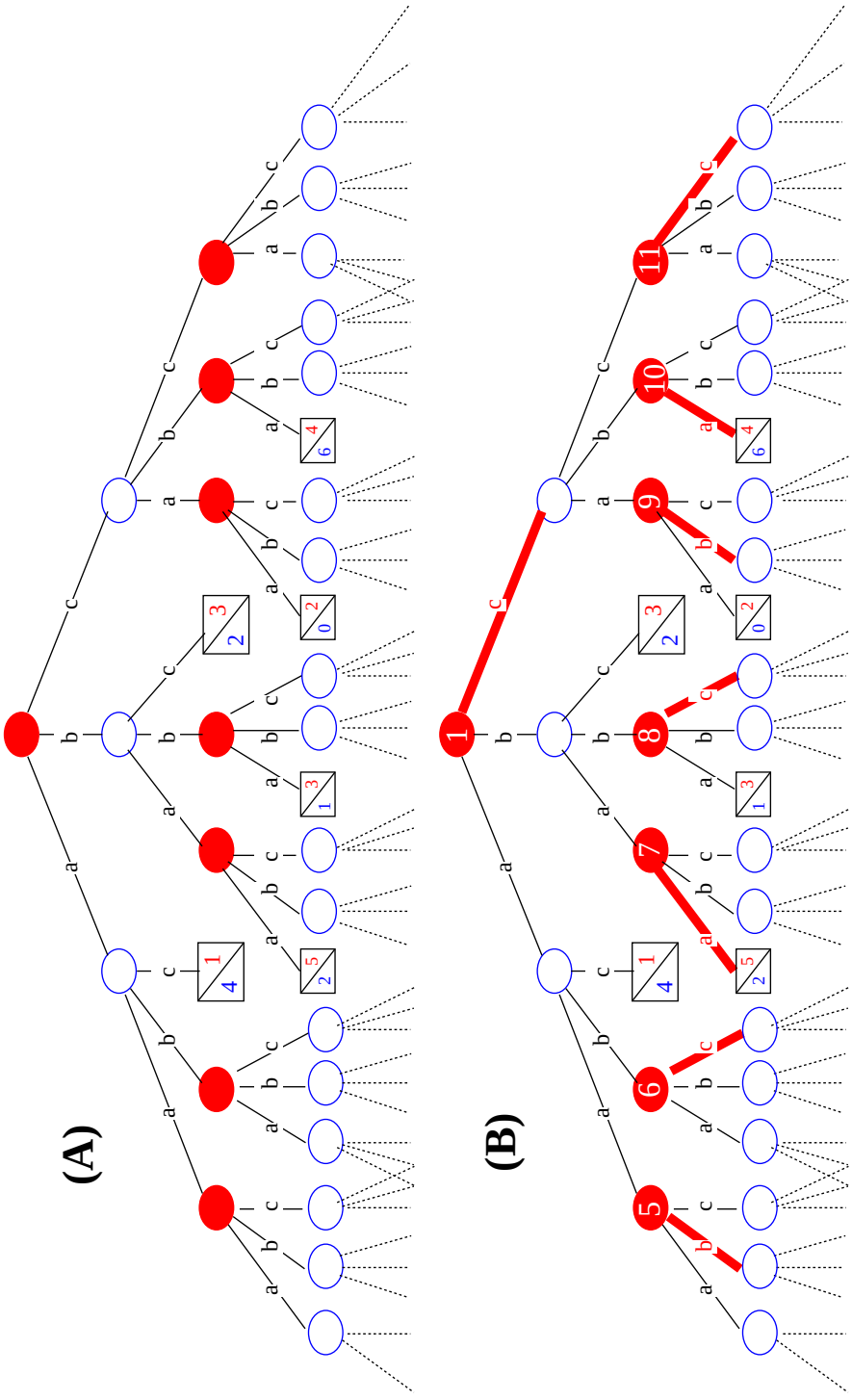


Figure 5.3: **(A)** The tree representation of an extensive game (not *Alternating Offers*). Each solid circle is a game state when it is Zara's turn to move, while each hollow circle is a game state when it is Owen's turn to move. The boxes represent terminal states, and the two numbers are the payoffs for the two players. Each labelled edge represents is a transition from one gamestate to another resulting from a particular move.
(B) A strategy σ_0 for Zara assigns a move to each of her gamestates. This strategy could be encoded, "In gamestate 1, do c. In state 5, do b. In state 6, do c. In state 7, do a,..." etc.

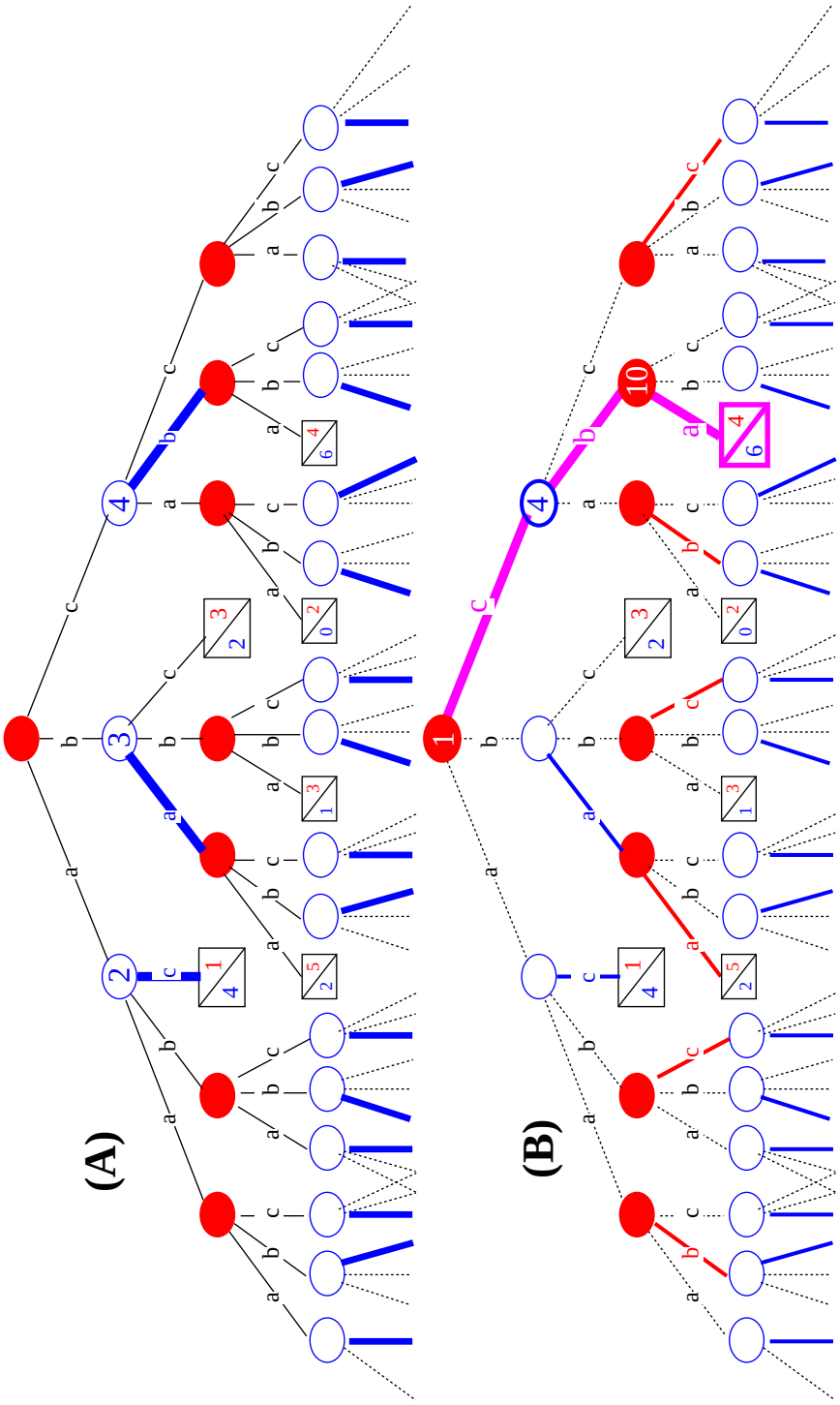


Figure 5.4: **(A)** A strategy σ_1 for Owen. "In gamestate 2, do c; In state 3, do a; ...." etc.
(B) Given the strategies σ_0 and σ_1, we can predict how the entire game will unfold, how it will end, and the ultimate payoffs for the players. In this case, the game unfolds: $1 \xrightarrow{c} 4 \xrightarrow{b} 10 \xrightarrow{a} \boxed{6\backslash 4}$. The game ends, Owen gets a payoff of 6, and Zara gets 4.

corresponds to some history of prior moves, as in Figure 5.3(A). Thus, *Alternating offers* is an extensive game, albeit with an uncountably infinite game tree (see Figure 5.6(A) for a crude sketch). Furthermore, *Alternating offers* is a game of *perfect information* because everything about each player's past moves is known to the other player (although this information actually has no relevance in the *Alternating offers* game). At the nth stage, if it is Zara's turn to make an offer, then this history consists of a sequence of prior offers and counteroffers like this:

$$[(B_0^1, b_1^1);\ (b_0^1, B_1^1);\quad (B_0^2, b_1^2);\ (b_0^2, B_1^2);\quad \ldots\ldots\ldots;\quad (B_0^{n-1}, b_1^{n-1});\ (b_0^{n-1}, B_1^{n-1})].$$

If it is Owen's turn, then the history looks like this:

$$[(B_0^1, b_1^1);\ (b_0^1, B_1^1);\quad (B_0^2, b_1^2);\ (b_0^2, B_1^2);\quad \ldots\ldots\ldots;\quad (B_0^{n-1}, b_1^{n-1});\ (b_0^{n-1}, B_1^{n-1});\quad (B_0^n, b_1^n)].$$

The *subgame* starting from such history is just game which begins at "**Stage** n" as described above (with either Owen or Zara moving first, as appropriate); see Figure 5.5(A) for an example. Notice that (unlike most extensive games, such as the one in Figure 5.5(A)), the *Alternative offers* game tree is "self-similar", in the sense that each subgame is clearly isomorphic to the game as a whole (except for the discounting of payoffs). This is a result of the assumption that past offers have no strategic effect on future offers.

Strategies and subgame-perfect equilibria: In an extensive game, a *strategy* for Zara is a rule σ_0 which specifies exactly which move she should make at each possible branch of the game tree, as in Figure 5.3(B). Likewise, a *strategy* for Owen is a rule σ_1 specifying his moves, as in Figure 5.4(A). Given any pair of strategies (σ_0, σ_1), we can predict exactly how the game will unfold, and thus predict the ultimate payoff for each player, as shown in Figure 5.4(B). Let $\overline{U}_0(\sigma_0, \sigma_1)$ be the (discounted) ultimate payoff for Zara, and $\overline{U}_1(\sigma_0, \sigma_1)$ be the (discounted) ultimate payoff for Owen.

A pair of strategies (σ_0^*, σ_1^*) forms a *Nash equilibrium* if, for any other strategy $\sigma_0 \neq \sigma_0^*$, we have $\overline{U}_0(\sigma_0^*, \sigma_1^*) \geq \overline{U}_0(\sigma_0, \sigma_1^*)$, and likewise, for any other strategy $\sigma_1 \neq \sigma_1^*$, we have $\overline{U}_0(\sigma_0^*, \sigma_1^*) \geq \overline{U}_0(\sigma_0^*, \sigma_1)$. In other words, if Zara believes that Owen will use strategy σ_1^*, then her *best reply* is σ_0^*; conversely, if Owen believes that Zara will use strategy σ_0^*, then his *best reply* is σ_1^*. Hence, if each player believes that the other is playing their half of the Nash equilibrium, then neither player has any incentive to unilaterally deviate.

The concept of Nash equilibrium is fairly satisfactory for games in normal form (i.e. 'one-shot' games described by a payoff matrix; see §5B), but it is inadequate for extensive games. To see why, observe that, so long as 'everything goes as planned', Zara will always make moves according to the rule specified by σ_0^*, which means that only a very small part of the tree of possible game histories will ever be realized. Thus, to be a 'best reply' to σ_0^*, Owen's strategy σ_1^* only has to suggest optimal responses in those game branches which could ever be reached via σ_0^*; in every other possible game branch, σ_1^* is free to make wildly suboptimal responses, because *these situations will never occur*. In particular, this means that Owen can 'bluff', by making threats which would be very suboptimal for him if he ever had to carry them out. If he encodes these threats in σ_1^*, then Zara's best response strategy σ_0^* will explicitly avoid doing

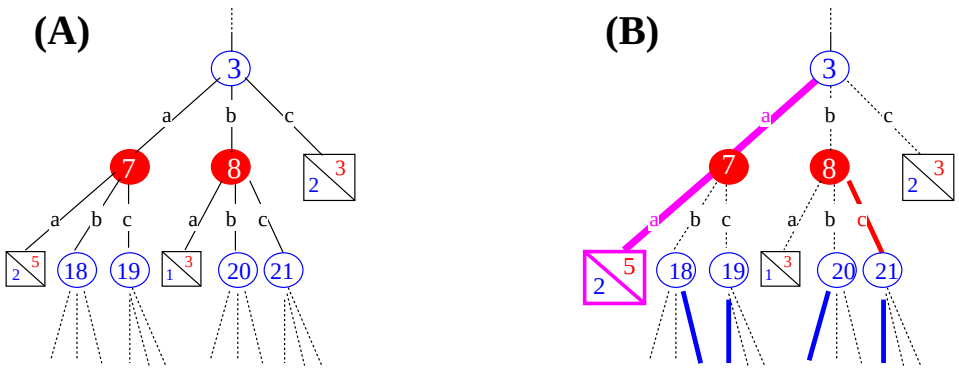


Figure 5.5: **(A)** The subgame beginning at node 3 of the game tree from Figure 5.3(A). **(B)** The strategies σ_0 and σ_1 in Figures 5.3(B) and 5.4(A) restrict to strategies on this subgame, which in turn determine the unfolding and the eventual payoff of this subgame *even though this subgame will never occur according to the strategy pair in the Figure 5.4(B).*

anything which causes Owen to carry out one of his threats. But this means that Owen can be confident he will never have to execute these threats, so he can put extreme threats into σ_1^* (which would be harmful even to himself) so as to put more pressure on Zara so that her best response strategy σ_0^* gives him more of what he wants. If (counterfactually) he was ever in a part of the game tree where he had to follow through on his σ_1^* threats, he would actually prefer *not* to, but as long as Zara uses σ_0^*, this will never happen, so he can make his σ_1^* threats as wild as he likes.

For example, in the context of the bargaining game, suppose $\epsilon > 0$ is very small, and let $B_1^\epsilon = \Gamma_1(\epsilon)$, so that (ϵ, B_1^ϵ) is just about Owen's best possible outcome (and just about Zara's worst). Then his strategy σ_1^ϵ might be the rule, "Always demand the bargain (ϵ, B_1^ϵ), and accept nothing less, no matter what Zara offers." If Owen adopts this strategy, then Zara's best response to σ_1^ϵ is the strategy σ_0^ϵ defined by "Always offer Owen (ϵ, B_1^ϵ), and always accept this offer." Then $(\sigma_0^\epsilon, \sigma_1^\epsilon)$ is a Nash equilibrium. Clearly, this is true for any $\epsilon > 0$. Also, by symmetric reasoning, it is a Nash equilibrium for Zara to always demand an exorbitant amount, and for Owen to always accept it. Hence, the Nash equilibrium concept fails to identify any particular bargain as the likely outcome of the game.

Furthermore, if Zara were to perform a more sophisticated analysis, she would see that these threats are hollow, because Owen himself would be unwilling to follow through on them (regardless of the fact that his σ_1^* strategy commits him to do so). Would he *really* just keep insisting on (ϵ, B_1^ϵ), no matter what? Suppose she committed herself to the strategy σ_0^n, where she always demands the Nash solution (b_0^*, b_1^*). If Owen keeps insisting on (ϵ, B_1^ϵ) for eternity, then he will get nothing; this is as irrational for him as it is for her. In other words, Owen's strategy looks impressive, but it commits him to *playing irrationally in each subgame.* Humans are not automata; Owen cannot commit himself irrevocably to playing a suicidal strategy as if he was a ballistic missile. At each subgame, he will have a chance to reconsider his strategy, and a truly rational strategy would be one where he wouldn't change his mind if he had the

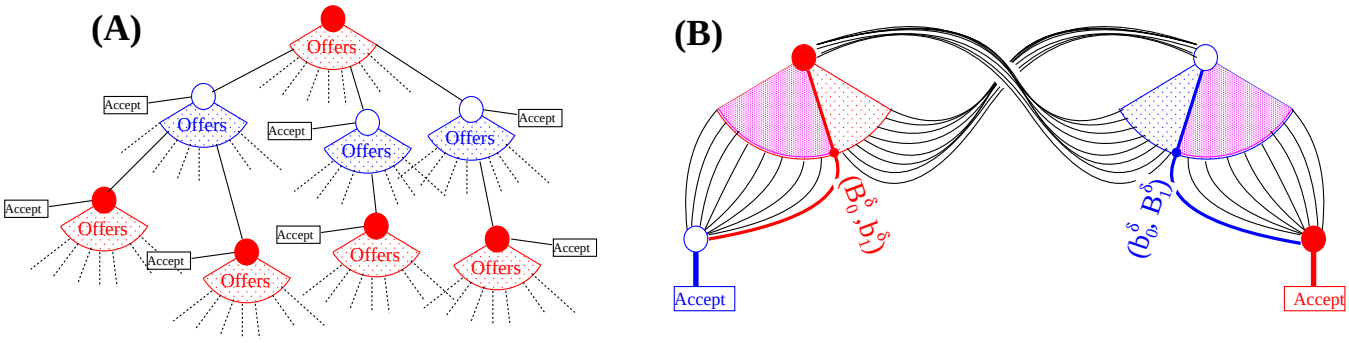


Figure 5.6: **(A)** A tiny fragment of the (uncountably infinite) game tree for the *Alternating Offers* game. **(B)** Rubinstein's subgame perfect equilibrium for *Alternating Offers*. Zara offers (B_0^δ, b_1^δ), and will accept no less than (b_0^δ, B_1^δ). Owen offers (b_0^δ, B_1^δ), and will accept no less than (B_0^δ, b_1^δ).

chance to reconsider —in other words, a strategy which recommends the optimal choice in each subgame, *even subgames which Owen thinks would never happen if Zara makes the best reply to his strategy.* In particular, a perfectly rational strategy would not make threats which Owen would be unwilling to execute —because Zara could *predict* that he would not execute these threats, and therefore they would not be credible. Like Spock, a perfectly rational strategy would never bluff.[7]

To formalize this reasoning, notice that the strategies σ_0 and σ_1 implicitly define strategies for Zara and Owen on all subgames. In any subgame, the strategy pair (σ_0, σ_1) then determines the outcome and payoff which would obtained *in that subgame*, as shown in Figure 5.5(B). Notice that this analysis is meaningful even for subgames that *would never occur* if the players had employed the strategies σ_0 and σ_1 from the very beginning of the original game. We say that the strategy-pair (σ_0^*, σ_1^*) is a ***subgame perfect equilibrium*** (SPE) if (σ_0^*, σ_1^*) is a Nash equilibrium for the whole game, and furthermore, given any initial history **h** (even one which could never happen if the players used the strategies σ_0^* and σ_1^*), the strategy pair (σ_0^*, σ_1^*) is *also* a Nash equilibrium for the subgame starting at **h**. The SPE concept was introduced by Selten [Sel65], and is generally agreed to be the 'right' definition of equlibrium for perfect-information extensive games like *Alternating offers.*

Theorem 5F.1 (Rubinstein, 1982)
Let $\delta < 1$, and let (B_0^δ, b_1^δ) and (b_0^δ, B_1^δ) be as in Lemma 5E.1(b).

(a) *In the alternating offers game, there is a unique subgame-perfect equilibrium, where the players have the following strategies* [shown in Figure 5.6(B)]:

[7]Note that this argument depends on critically on the following three assumptions: (1) Zara is perfectly rational. (2) Zara believes Owen to be perfectly rational. (3) Zara has perfect information about Owen's utility function (and hence, she can predict that he would be unwilling to execute certain threats). In real life, assumption (3) is usually false (and often (1) and (2) are false as well, unless perhaps Zara and Owen are Vulcans), so that bluffing may often be an excellent strategy for Owen. Again, this is the difference between 'bargaining' and 'haggling'.

- *Zara will always offer* (B_0^δ, b_1^δ). *She will accept any offer* (b'_0, B'_1) *by Owen such that* $b'_0 \geq b_0^\delta$. *She will reject any offer* (b'_0, B'_1) *such that* $b'_0 < b_0^\delta$.
- *Owen will always offer* (b_0^δ, B_1^δ). *He will accept any offer* (B'_0, b'_1) *by Zara such that* $b'_1 \geq b_1^\delta$. *He will reject any offer* (B'_0, b'_1) *such that* $b'_1 < b_1^\delta$.

(b) *As a consequence, the bargaining game will terminate immediately. Zara will offer* (B_0^δ, b_1^δ) *and Owen will accept, so that* (B_0^δ, b_1^δ) *will be the bargaining outcome.*

(c) *In the limit as* $\delta \nearrow 1$, *the bargaining outcome* (B_0^δ, b_1^δ) *converges to the Nash solution.*

The proof of Theorem 5F.1 will be the subject of section 5G below.

Interpretation of "$\delta \nearrow 1$": When $\delta = 1$, there is no incentive for cooperation, and the *Alternating offers* game has an infinity of subgame perfect equilibria describing various 'intransigent' behaviours by the players. So what are we to make of the fact that the unique subgame perfect equilibria $(\sigma_0^\delta, \sigma_1^\delta)$ converges to the Nash solution as $\delta \nearrow 1$? What does "$\delta \nearrow 1$" mean? Does it mean that we are imagining the limiting scenario where Owen and Zara become infinitely patient immortals, living in a riskless and eternal universe, bargaining over invincible assets which will never decay in value?

A better interpretation is to imagine the players as having a *continuous time discount rate* $\frac{1}{\alpha} > 0$, so, for any $t \in \mathbb{R}_{\neq}$, their discount factor at time t is $\delta_t := e^{-t/\alpha}$. Imagine that each 'stage' of the bargaining game takes some fixed time duration $\tau > 0$. Then the discount factor after a single stage (i.e. at time $t = \tau$) will be $\delta_\tau := e^{-\tau/\alpha}$. The discount factor after n stages (i.e. at time $t = n\tau$) will be $e^{-n\tau/\alpha} = \delta_\tau^n$. Hence, δ_τ is determined by the continuous time rate α, but also by the timespan τ. Now, τ reflects the 'efficiency' of the bargaining process; it is the amount of time it takes for each player to consider the other player's offer and/or make a counteroffer. If Zara and Owen are experienced bargainers who know exactly what they want and don't want to waste time, then they will exchange offers and counteroffers as fast as they can think and speak (visualize the trading floor at a stock exchange). Thus, the limit as $\delta_\tau \nearrow 1$ is really the limit as $\tau \searrow 0$, which is the limiting case of extremely efficient trade with very rapid exchange of offers.

5G Proof of Rubinstein's Theorem

Prerequisites: §5F **Recommended:** §5B

The purpose of this section is to develop sufficient background in the theory of extensive games to provide a rigorous proof of Rubinstein's Theorem 5F.1. Thus, the material is somewhat more formal and mathematically sophisticated than the informal treatment of §5F.

In a *stationary symmetric extensive game with perfect information*, we first define a set $\mathcal{A}$ of *nonterminal actions* and another set $\mathcal{A}^\dagger$ *terminal actions*, which are available to each player during each round of play, and whose execution is visible to the other player(s). The game

is *symmetric* because all players have exactly the same set of actions; the game is *stationary* because this set of actions remains the same during each round of play; the game has *perfect information* because each player knows exactly what both she and the other player(s) did during all previous rounds of play.

Suppose there are two players, Zara and Owen. Let $\mathcal{H}_0$ be the set of all finite sequences of nonterminal actions of the form

$$(\mathbf{a}^0, \mathbf{a}^1;\ \mathbf{a}^2, \mathbf{a}^3;\ \mathbf{a}^4, \mathbf{a}^5; \ldots;\ \mathbf{a}^{2T}),$$

where $\mathbf{a}^t \in \mathcal{A}$ for all $t \in [0...2T]$. Such a sequence represents a finite 'history' of possible game play, where Zara begins with nonterminal action $\mathbf{a}^0$, then Owen counters with $\mathbf{a}^1$, then Zara counters with $\mathbf{a}^2$ and so on, with Zara making all *even*-numbered moves and Owen making all *odd*-numbered moves, until the last nonterminal action (by Zara) is $\mathbf{a}^{2T}$. Let $\mathcal{H}_1$ be the set of all finite sequences of nonterminal actions ending with a nonterminal action by Owen; that is, all sequences of the form:

$$(\mathbf{a}^0, \mathbf{a}^1;\ \mathbf{a}^2, \mathbf{a}^3;\ \mathbf{a}^4, \mathbf{a}^5; \ldots;\ \mathbf{a}^{2T}, \mathbf{a}^{2T+1}).$$

Let $\mathcal{H}_0^\dagger$ be the set of all finite *terminating* sequences of the form

$$(\mathbf{a}^0, \mathbf{a}^1;\ \mathbf{a}^2, \mathbf{a}^3;\ \mathbf{a}^4, \mathbf{a}^5; \ldots;\ \mathbf{a}^{2T-2}, \mathbf{a}^{2T-1};\ \mathbf{a}^{2T}),$$

where $\mathbf{a}^t \in \mathcal{A}$ for all $t \in [0...2T)$, but $\mathbf{a}^{2T} \in \mathcal{A}^\dagger$. This represents a finite game history ending in a terminal move $\mathbf{a}^{2T}$ by Zara (at which point the game ends). Likewise, let $\mathcal{H}_1^\dagger$ be the set of all finite terminating sequences of the form

$$(\mathbf{a}^0, \mathbf{a}^1;\ \mathbf{a}^2, \mathbf{a}^3;\ \mathbf{a}^4, \mathbf{a}^5; \ldots;\ \mathbf{a}^{2T-2}, \mathbf{a}^{2T-1};\ \mathbf{a}^{2T}, \mathbf{a}^{2T+1}),$$

where $\mathbf{a}^t \in \mathcal{A}$ for all $t \in [0...2T]$, but $\mathbf{a}^{2T+1} \in \mathcal{A}^\dagger$. This represents a finite game history ending in a terminal move $\mathbf{a}^{2T+1}$ by Owen (at which point the game ends). Finally, let $\mathcal{H}^\infty$ be the set of all infinite sequences

$$(\mathbf{a}^0, \mathbf{a}^1;\ \mathbf{a}^2, \mathbf{a}^3;\ \mathbf{a}^4, \mathbf{a}^5; \ldots;\ \mathbf{a}^{2T}, \mathbf{a}^{2T+1}; \ldots\ldots\ldots)$$

(representing games which never terminate). Let $\overline{\mathcal{H}} := \mathcal{H}_0^\dagger \sqcup \mathcal{H}_1^\dagger \sqcup \mathcal{H}^\infty$; this is the set of all possible 'complete histories' for the game. The game description is then completed by a pair of *payoff functions* $U_0, U_1 : \overline{\mathcal{H}} \longrightarrow \mathbb{R}$. Thus, for any finite terminating history $\mathbf{h} \in \mathcal{H}_0^\dagger \sqcup \mathcal{H}_1^\dagger$ or infinite history $\mathbf{h} \in \mathcal{H}^\infty$, $U_0(\mathbf{h})$ is the ultimate payoff utility for Zara and $U_1(\mathbf{h})$ is the ultimate payoff utility for Owen when the game is over. Each player strives to maximize her utility. Formally, the game is entirely encoded by the data structure $(\mathcal{A}, \mathcal{A}^\dagger, U_0, U_1)$.

Example 5G.1: Let $(\mathcal{B}, \mathbf{0})$ be a bargaining problem (i.e. $\mathcal{B}$ is a convex, compact, comprehensive subset of $\mathbb{R}^2_\nearrow$, and $\mathbf{0} \in \mathcal{B}$). In the Rubinstein-Ståhl *Alternating Offers* game, the set $\mathcal{A}$ of *nonterminal* actions is just the negotiating set $\wp_{\mathbf{q}}\mathcal{B}$ of the bargaining set $\mathcal{B}$, because at each

stage of the game, each player's 'action' is to propose a point on $\wp_{\mathbf{q}}\mathcal{B}$ to the other player. The set $\mathcal{A}^\dagger$ of *terminal actions* contains a single element, Accept; if Owen 'accepts' Zara's offer, the game ends; otherwise Owen must make a counteroffer (in $\mathcal{A}$), and the game continues.

Let $\delta \in (0,1)$ be a *discount factor*. If $\mathbf{h} = (\mathbf{a}^0, \mathbf{a}^1; \ldots; \mathbf{a}^{2t}, \text{Accept}) \in \mathcal{H}_1^\dagger$ is a terminal history ending with Owen's acceptance of Zara's offer $\mathbf{a}^{2t} = (A_0, a_1) \in \wp_{\mathbf{q}}\mathcal{B}$, then

$$U_0(\mathbf{h}) \quad := \quad \delta^{2t} A_0 \quad \text{and} \quad U_1(\mathbf{h}) \quad := \quad \delta^{2t} a_1.$$

If $\mathbf{h} = (\mathbf{a}^0, \mathbf{a}^1; \ldots; \mathbf{a}^{2t}, \mathbf{a}^{2t+1}; \text{Accept}) \in \mathcal{H}_0^\dagger$ is a terminal history ending with Zara's acceptance of Owen's offer $\mathbf{a}^{2t+1} = (a_0, A_1) \in \wp_{\mathbf{q}}\mathcal{B}$, then

$$U_0(\mathbf{h}) \quad := \quad \delta^{2t+1} a_0 \quad \text{and} \quad U_1(\mathbf{h}) \quad := \quad \delta^{2t+1} A_1.$$

In other words, both players receive a whatever utility they get from the last bargain proposed by either player (the one which is 'accepted' by the other player), but multiplied by a discount factor of δ^t if $t+1$ offers have so far been made (and t of these have been rejected). If $\mathbf{h} \in \mathcal{H}^\infty$, then $U_0(\mathbf{h}) = 0 = U_1(\mathbf{h})$ (i.e. negotiating until eternity yields zero payoff for both players). Note:

- The payoffs only depend upon the terminal actions of the players; they do not depend on the entire past history of play.
- $\lim_{t\to\infty} \delta^t = 0$, so the players' payoffs decay exponentially over time.

Both of these are special properties of the *Alternating Offers* game. The exponential decay provides some incentive to reach an agreement quickly. The 'ahistorical' property of the payoffs means that our analysis of strategies at time T is identical to our analysis of the strategies at time 0 (except for the exponential discount). This greatly simplifies the analysis of the game. ♢

Suppose $(\mathcal{A}, \mathcal{A}^\dagger, U_0, U_1)$ is a symmetric, stationary, two-person extensive game with perfect information. A *strategy* for Zara is a function $\sigma_0 : \mathcal{H}_1 \longrightarrow \mathcal{A} \sqcup \mathcal{A}^\dagger$; this dictates an action for Zara in response to each possible game history ending in a nonterminal action by Owen. Likewise, a *strategy* for Owen is a function $\sigma_1 : \mathcal{H}_0 \longrightarrow \mathcal{A} \sqcup \mathcal{A}^\dagger$; this dictates an action for Owen in response to each possible game history ending in a nonterminal action by Zara.

For simplicity, we assume $\mathcal{H}_0$ and $\mathcal{H}_1$ each includes the 'empty' sequence $\emptyset$. Thus, if Zara moves first, then her 'opening move' will be the action $a_0^0 := \sigma_0(\emptyset)$; if Owen moves first, then his 'opening move' will be the action $a_1^0 := \sigma_1(\emptyset)$.

Example 5G.2: Consider the *Alternating Offers* game from Example 5G.1. Let (B_0^δ, b_1^δ) and (b_0^δ, B_1^δ) be as in Lemma 5E.1(b) on page 107, and consider the strategy $\sigma_0 : \mathcal{H}_1 \longrightarrow \mathcal{A} \sqcup \mathcal{A}^\dagger$ for Zara defined as follows: $\sigma_0(\emptyset) = (B_0^\delta, b_1^\delta) \in \wp_{\mathbf{q}}\mathcal{B}$, and for any $\mathbf{h} := (\mathbf{a}^0, \mathbf{a}^1; \mathbf{a}^2, \mathbf{a}^3; \ldots; \mathbf{a}^{2T}, \mathbf{a}^{2T+1}) \in \mathcal{H}_1$, if $\mathbf{a}^{2T+1} := (a_0, A_1) \in \wp_{\mathbf{q}}\mathcal{B}$, then

$$\sigma_0(\mathbf{h}) \quad = \quad \begin{cases} \text{Accept} & \text{if} \quad a_0 \geq b_0^\delta; \\ (B_0^\delta, b_1^\delta) & \text{if} \quad a_0 < b_0^\delta. \end{cases}$$

Likewise, we define the strategy $\sigma_1 : \mathcal{H}_1 \longrightarrow \mathcal{A} \sqcup \mathcal{A}^\dagger$ for Owen as follows: $\sigma_1(\emptyset) = (b_0^\delta, B_1^\delta) \in \wp_{\mathbf{q}}\mathcal{B}$, and for any $\mathbf{h} := (\mathbf{a}^0, \mathbf{a}^1;\ \mathbf{a}^2, \mathbf{a}^3; \ldots;\ \mathbf{a}^{2T}) \in \mathcal{H}_0$, if $\mathbf{a}^{2T} := (A_0, a_1) \in \wp_{\mathbf{q}}\mathcal{B}$, then

$$\sigma_1(\mathbf{h}) \quad = \quad \begin{cases} \text{Accept} & \text{if} \quad a_1 \geq b_1^\delta; \\ (b_0^\delta, B_1^\delta) & \text{if} \quad a_1 < b_1^\delta. \end{cases}$$

(Note that, during each round of play, each player's response only depends on the *immediately prior* action of the other player. This is not typical; for a typical strategy, each player's response depends upon the entire prior history of actions by herself and the other player.) ♢

Suppose Zara moves first. The game play then proceeds as follows:

Time 0: Zara begins with $\mathbf{a}^0 := \sigma_0(\emptyset) \in \mathcal{A}$.

Time 1: Owen responds with $\mathbf{a}^1 := \sigma_1(\mathbf{a}^0) \in \mathcal{A}$.

Time 2: Zara responds with $\mathbf{a}^2 := \sigma_0(\mathbf{a}^0, \mathbf{a}^1) \in \mathcal{A}$.

Time 3: Owen responds with $\mathbf{a}^3 := \sigma_1(\mathbf{a}^0, \mathbf{a}^1;\ \mathbf{a}^2) \in \mathcal{A}$.

Time 4: Zara responds with $\mathbf{a}^4 := \sigma_0(\mathbf{a}^0, \mathbf{a}^1;\ \mathbf{a}^2, \mathbf{a}^3) \in \mathcal{A}$.

Time 5: Owen responds with $\mathbf{a}^5 := \sigma_1(\mathbf{a}^0, \mathbf{a}^1;\ \mathbf{a}^2, \mathbf{a}^3;\ \mathbf{a}^4) \in \mathcal{A}$.

and so on, either until an infinite sequence of moves and countermoves has transpired, or until one player or the other responds with a terminal move in $\mathcal{A}^\dagger$, at which point the game ends and both players collect their payoffs. Thus, given two strategies σ_0 and σ_1, we can completely predict how the gameplay will unfold. In other words, σ_0 and σ_1 determine a unique *induced game history* $H(\sigma_0, \sigma_1) = (\mathbf{a}^0, \mathbf{a}^1;\ \mathbf{a}^2, \mathbf{a}^3;\ \mathbf{a}^4, \mathbf{a}^5; \ldots) \in \overline{\mathcal{H}}$, defined as above. The *induced payoffs* for each player are then defined:

$$\overline{U}_0(\sigma_0, \sigma_1) \quad := \quad U_0[H(\sigma_0, \sigma_1)] \quad \text{and} \quad \overline{U}_1(\sigma_0, \sigma_1) \quad := \quad U_1[H(\sigma_0, \sigma_1)].$$

Thus, if $\mathcal{S}_0$ and $\mathcal{S}_1$ are the space of all possible extensive strategies for Zara and Owen, then we get a pair of *induced payoff functions* $\overline{U}_0, \overline{U}_1 : \mathcal{S}_0 \times \mathcal{S}_1 \longrightarrow \mathbb{R}$. In this sense, any extensive game can be represented as a *normal-form* game, where each player's single 'move' in the normal game is commit to an entire extensive strategy for the original game.

(We can imagine that, instead of playing the game herself, each player programs an 'automaton' with a complete specification of an extensive strategy, and then launches this automaton to play on her behalf; thus, her only 'move' is to choose which strategy to program into her automaton at time zero. Why would the player launch automata to play on her behalf? Because a perfectly rational, hyperintelligent player could figure out, in advance, what the best possible response was to every possible game situation. Once she has performed these computations, it is redundant to actually play the game herself; she could program the automaton to play exactly as she would. Of course, this is totally impractical in real life.)

Example 5G.3: Let σ_0^* and σ_1^* be the strategies from Example 5G.2. Then

$$\mathcal{H}(\sigma_0^*, \sigma_1^*) \quad = \quad \left((B_0^\delta, b_1^\delta), \text{Accept}\right).$$

In other words, Zara opens by offering the bargain (B_0^δ, b_1^δ), and Owen immediately accepts. Thus, $\overline{U}_0(\sigma_0^*, \sigma_1^*) = B_0^\delta$ and $\overline{U}_1(\sigma_0^*, \sigma_1^*) = b_1^\delta$. ♢

A pair $(\sigma_0^*, \sigma_1^*) \in \mathcal{S}_0 \times \mathcal{S}_1$ of extensive strategies is a *Nash Equilibrium* if:

- σ_0^* is a *best response* to σ_1^*. In other words, for any other $\sigma_0 \in \mathcal{S}_0$, $\overline{U}_0(\sigma_0^*, \sigma_1^*) \geq \overline{U}_0(\sigma_0, \sigma_1^*)$. (Thus, if Zara assumes that Owen will play σ_1^*, then σ_0^* is her optimal strategy, given this assumption).

- σ_1^* is a *best response* to σ_0^*. In other words, for any other $\sigma_1 \in \mathcal{S}_1$, $\overline{U}_1(\sigma_0^*, \sigma_1^*) \geq \overline{U}_1(\sigma_0^*, \sigma_1)$. (Thus, if Owen assumes that Zara will play σ_0^*, then σ_1^* is his optimal strategy, given this assumption).

Example 5G.4: Let σ_0^* and σ_1^* be the strategies from Example 5G.2. Then (σ_0^*, σ_1^*) is a Nash equilibrium. To see this, let $\sigma_0 \in \mathcal{S}_0$ be some other strategy for Zara; we must show that $\overline{U}_0(\sigma_0, \sigma_1^*) \leq B_0^\delta$ (because Example 5G.3 shows that $\overline{U}_0(\sigma_0^*, \sigma_1^*) = B_0^\delta$).

Suppose $H(\sigma_0, \sigma_1^*) = \mathbf{h}$. If $\mathbf{h}$ ends in round t with Zara accepting an offer made by Owen, then $U_0(\mathbf{h}) = \delta^t b_0^\delta \leq b_0^\delta < B_0^\delta$, because strategy σ_1^* *always* offers (b_0^δ, B_1^δ). If $\mathbf{h}$ ends in round t with Owen accepting an offer (a_0, a_1) made by Zara, then

$$U_0(\mathbf{h}) \quad = \quad \delta^t a_0 \quad \leq \quad \delta^t B_0^\delta \quad \leq \quad B_0^\delta,$$

with equality if and only if $t = 0$. This is because σ_1^* *only* accepts (a_0, a_1) if $a_1 \geq b_1^\delta$, which means that $a_0 \leq B_0^\delta$.

Thus, $\overline{U}_0(\sigma_0, \sigma_1^*) \leq B_0^\delta$, with equality if and only if σ_0 offers (B_0^δ, b_1^δ) during the first round (which σ_1 immediately accepts). Thus, σ_0^* is a best response to σ_1^*.

By exactly symmetric reasoning, σ_1^* is a best response to σ_0^*. Thus, (σ_0^*, σ_1^*) is a Nash equilibrium. ♢

Neither player has an incentive to unilaterally deviate from her half of the Nash equilibrium strategy pair, as long as she believes that the other player also will not deviate. In this sense, a Nash equilibrium is 'stable'. However, a Nash equilibrium may commit the players to choosing 'suboptimal' strategies in various subgames.

Example 5G.5: Consider the *Alternating Offers* game from Example 5G.1. Let $\epsilon > 0$, and let $A_0 := \Gamma_0(\epsilon) = \max\{b_0 > 0 \,;\, (b_0, \epsilon) \in \mathcal{B}\}$. Thus, the bargain $(A_0, \epsilon) \in \wp_{\mathbf{q}}\mathcal{B}$ is just about the best for Zara, and just about the worst for Owen. Consider the 'intransigent' strategy

$\sigma_0^\epsilon : \mathcal{H}_1 \longrightarrow \mathcal{A} \sqcup \mathcal{A}^\dagger$ for Zara defined as follows: $\sigma_0^\epsilon(\emptyset) = (A_0, \epsilon) \in \wp_{\mathbf{q}}\mathcal{B}$, and for any $\mathbf{h} := (\mathbf{a}^0, \mathbf{a}^1;\ \mathbf{a}^2, \mathbf{a}^3; \ldots;\ \mathbf{a}^{2T}, \mathbf{a}^{2T+1}) \in \mathcal{H}_1$, if $\mathbf{a}^{2T+1} := (a_0, A_1) \in \wp_{\mathbf{q}}\mathcal{B}$, then

$$\sigma_0^\epsilon(\mathbf{h}) \quad = \quad \begin{cases} \text{Accept} & \text{if} \quad a_0 \geq A_0; \\ (A_0, \epsilon) & \text{if} \quad a_0 < A_0. \end{cases}$$

This is a strategy where Zara makes an unreasonable demand and sticks to it, no matter what. Owen's best response, in this situation, is to capitulate (he gains nothing by defying her intransigence). Thus, he might use the 'capitulation' strategy $\sigma_1^\epsilon : \mathcal{H}_1 \longrightarrow \mathcal{A} \sqcup \mathcal{A}^\dagger$ defined as follows: $\sigma_0^\epsilon(\emptyset) = (A_0, \epsilon) \in \wp_{\mathbf{q}}\mathcal{B}$, and for any $\mathbf{h} := (\mathbf{a}^0, \mathbf{a}^1;\ \mathbf{a}^2, \mathbf{a}^3; \ldots;\ \mathbf{a}^{2T}) \in \mathcal{H}_0$, if $\mathbf{a}^{2T} := (a_0, a_1) \in \wp_{\mathbf{q}}\mathcal{B}$, then

$$\sigma_1^\epsilon(\mathbf{h}) \quad = \quad \begin{cases} \text{Accept} & \text{if} \quad a_1 \geq \epsilon; \\ (A_0, \epsilon) & \text{if} \quad a_1 < \epsilon. \end{cases}$$

Exercise 5.9 (a) Show that σ_1^ϵ is a best response to σ_0^ϵ.

(b) Show that σ_0^ϵ is a best response to σ_1^ϵ.

Thus, $(\sigma_0^\epsilon, \sigma_1^\epsilon)$ is a Nash Equilibrium, for any choice of $\epsilon \in (0, M_0)$ (where $M_0 = \Gamma_1(0)$).

Heuristically speaking, in this equilibrium, Zara threatens to behave in an intransigent fashion, and Owen, believing this threat, capitulates to her demands. However, if Owen had *not* capitulated, then Zara's intransigence would hurt her as well as him. For example, suppose that, instead of adopting the strategy σ_1^ϵ, Owen deploys the strategy σ_1^* from Example 5G.4 (and assume $\epsilon < b_1^\delta$). Then

$$H(\sigma_0^\epsilon, \sigma_1^*) \quad = \quad \big[(A_0, \epsilon),\, (a_0^\delta, B_1^\delta);\ \ (A_0, \epsilon),\, (b_0^\delta, B_1^\delta);\ \ (A_0, \epsilon),\, (b_0^\delta, B_1^\delta); \ldots\ldots\big]$$

After four or five iterations of this cycle of rejectionism, it should become apparent to Zara that her intransigent strategy is not having the desired effect. Meanwhile, the payoff of any possible outcome is exponentially decaying to zero. Will she *really* continue to behave in such an intransigent way? Surely it would be rational to change her strategy.

The strategy σ_0^ϵ is based on a threat. Zara presumably believes that Owen will be so impressed by this threat that he would never choose σ_1^*, and so the aforementioned 'rejectionist' history will never unfold. But suppose he *did* chose σ_1^*; then it would *irrational* for Zara to persist in a strategy which will ultimately yield a payoff of zero to *both* players. Perceiving that Owen has chosen σ_1^*, a rational player would abort her original strategy σ_0^ϵ, and try to salvage the situation by adopting a more suitable strategy (say, σ_0^*).

Thus, if Zara is perfectly rational, then Owen can predict that she will *not* actually persist in the intransigent strategy σ_0^ϵ, if he refuses to capitulate. In other words, she is *bluffing*. Being perfectly rational himself, Owen *knows* she is bluffing, and will respond accordingly. Zara's threat lacks credibility. Thus, the strategy pair $(\sigma_0^\epsilon, \sigma_1^\epsilon)$, although it is a Nash equilibrium, is *not* a realistic description of the behaviour of rational agents, because: (1) A rational player would modify her strategy in response to unforeseen behaviour by other player, and (2) The other player knows this (and, being rational, will optimally exploit this knowledge). ♢

The solution to the quandary of Example 5G.5 lies in the concept of a *subgame perfect equilibrium*. Given a strategy pair $(\sigma_0, \sigma_1) \in \mathcal{S}_0 \times \mathcal{S}_1$ and any finite, nonterminal history $\mathbf{h} \in \mathcal{H}_0 \sqcup \mathcal{H}_1$, we can extrapolate the unfolding game beyond this point, just as before. Suppose $\mathbf{h} \in \mathcal{H}_1$, and let $\mathbf{h} = (\mathbf{a}^0, \mathbf{a}^1;\ \ldots;\ \mathbf{a}^{2t}, \mathbf{a}^{2t+1})$; then the future unfolds as follows:

Time $(2t+2)$**:** Zara responds with $\mathbf{a}^{2t+2} := \sigma_0(\mathbf{a}^0, \mathbf{a}^1;\ \ldots;\ \mathbf{a}^{2t}, \mathbf{a}^{2t+1}) \in \mathcal{A}$.

Time $(2t+3)$**:** Owen responds with $\mathbf{a}^{2t+3} := \sigma_1(\mathbf{a}^0, \mathbf{a}^1;\ \ldots;\ \mathbf{a}^{2t}, \mathbf{a}^{2t+1};\ \mathbf{a}^{2t+2}) \in \mathcal{A}$.

Time $(2t+4)$**:** Zara responds with $\mathbf{a}^{2t+4} := \sigma_0(\mathbf{a}^0, \mathbf{a}^1;\ \ldots;\ \mathbf{a}^{2t}, \mathbf{a}^{2t+1};\ \mathbf{a}^{2t+2}, \mathbf{a}2t+3) \in \mathcal{A}$.

Time $(2t+5)$**:** Owen responds with $\mathbf{a}^{2t+5} := \sigma_1(\mathbf{a}^0, \mathbf{a}^1;\ \ldots;\ \mathbf{a}^{2t}, \mathbf{a}^{2t+1};\ \mathbf{a}^{2t+2}, \mathbf{a}^{2t+3};\ \mathbf{a}^{2t+4}) \in \mathcal{A}$.

and so on, either until an infinite sequence of moves and countermoves has transpired, or until one player or the other responds with a terminal move in $\mathcal{A}^\dagger$, at which point the game ends and both players collect their payoffs. We can likewise extrapolate the future of any history $\mathbf{h} \in \mathcal{H}_0$. In other words, any $\mathbf{h} \in \mathcal{H}_0 \sqcup \mathcal{H}_1$, and any strategies σ_0 and σ_1 determined a unique *induced subgame history* $H(\mathbf{h}; \sigma_0, \sigma_1) = (\mathbf{a}^0, \mathbf{a}^1;\ \mathbf{a}^2, \mathbf{a}^3;\ \mathbf{a}^4, \mathbf{a}^5; \ldots) \in \overline{\mathcal{H}}$, defined as above. The *induced subgame payoffs* for each player are then defined:

$$\overline{U}_0(\mathbf{h}; \sigma_0, \sigma_1) \quad := \quad U_0[H(\mathbf{h}; \sigma_0, \sigma_1)] \quad \text{and} \quad \overline{U}_1(\mathbf{h}; \sigma_0, \sigma_1) \quad := \quad U_1[H(\mathbf{h}; \sigma_0, \sigma_1)].$$

The strategy pair (σ_0^*, σ_1^*) is a *subgame perfect equilibrium* (SPE) if

- For any finite nonterminal history $\mathbf{h} \in \mathcal{H}_0 \sqcup \mathcal{H}_1$, the strategy is σ_0^* is a *best response* to σ_1^*, *given* history $\mathbf{h}$. That is, for any other $\sigma_0 \in \mathcal{S}_0$, $\overline{U}_0(\mathbf{h}; \sigma_0^*, \sigma_1^*) \geq \overline{U}_0(\mathbf{h}; \sigma_0, \sigma_1^*)$.
- For any finite nonterminal history $\mathbf{h} \in \mathcal{H}_0 \sqcup \mathcal{H}_1$, the strategy is σ_1^* is a *best response* to σ_0^*, *given* history $\mathbf{h}$. That is, for any other $\sigma_1 \in \mathcal{S}_1$, $\overline{U}_1(\mathbf{h}; \sigma_0^*, \sigma_1^*) \geq \overline{U}_1(\mathbf{h}; \sigma_0^*, \sigma_1)$.

In other words, the pair (σ_0^*, σ_1^*) is not only a Nash equilibrium for the original game, but furthermore, for any finite nonterminal history $\mathbf{h} \in \mathcal{H}_0 \sqcup \mathcal{H}_1$, the pair (σ_0^*, σ_1^*) is *also* a Nash equilibrium for the 'subgame' beginning with history $\mathbf{h}$.

Example 5G.6: Let σ_0^* and σ_1^* be the strategies from Example 5G.2. Then (σ_0^*, σ_1^*) is a subgame perfect equilibrium. ♢

Exercise 5.10 Prove that (σ_0^*, σ_1^*) is an SPE. Proceed as follows:

Let $\mathbf{h} \in \mathcal{H}_0 \sqcup \mathcal{H}_1$ be a history of length t. We must show that (σ_0^*, σ_1^*) is a Nash equilibrium (i.e. each strategy is a 'best response' to the other one) in the subgame beginning with history $\mathbf{h}$. So, let $\sigma_0 \in \mathcal{S}_0$ be any other strategy. We must show that $\overline{U}_0(\mathbf{h}; \sigma_0, \sigma_1^*) \leq \overline{U}_0(\mathbf{h}; \sigma_0^*, \sigma_1^*)$. There are two cases: either $\mathbf{h} \in \mathcal{H}_1$, or $\mathbf{h} \in \mathcal{H}_0$.

Case I: $\mathbf{h} \in \mathcal{H}_1$ (*i.e. Owen moved last, and Veroniqe moves next.*)

Suppose Owen's last offer was (a_0, A_1). There are now two subcases: either σ_0 accepts (a_0, A_1), or σ_0 rejects (a_0, A_1) and makes a counteroffer

Case I(a) Suppose σ_0 accepts (a_0, A_1). Show that $\overline{U}_0(\mathbf{h}; \sigma_0, \sigma_1^*) = \delta^{t-1} a_0 \leq \overline{U}_0(\mathbf{h}; \sigma_0^*, \sigma_1^*)$ (with equality if and only if $a_0 \geq b_0^\delta$).

Case I(b) Suppose σ_0 rejects (a_0, A_1). Show that $\overline{U}_0(\mathbf{h}; \sigma_0, \sigma_1^*) \leq \delta^t B_0^\delta \leq \overline{U}_0(\mathbf{h}; \sigma_0^*, \sigma_1^*)$.

Case II: $\mathbf{h} \in \mathcal{H}_0$ (*i.e. Zara moved last, and Owen moves next*). Reduce this to Case I.

Conclude that in either of Cases I or II, σ_0^* is a best response to σ_1^* in the subgame determined by $\mathbf{h}$. By reversing the roles of Zara and Owen in the previous argument, we see that σ_1^* is a best response to σ_0^* in any subgame determined by any history $\mathbf{h}$. Conclude that (σ_0^*, σ_1^*) is an SPE. ⌞ ⌟

Example 5G.7: Let σ_0^ϵ and σ_1^ϵ be the strategies from Example 5G.5. Then $(\sigma_0^\epsilon, \sigma_1^\epsilon)$ is *not* a subgame perfect equilibrium. (**Exercise 5.11** Prove this. *Hint:* Consider a subgame where Owen has just made an offer (a_0, A_1), where $\delta A_0 < a_0 < A_0$. Strategy σ_0^ϵ says that Zara should reject (a_0, A_1), and make counteroffer (A_0, ϵ). Owen will then accept this counteroffer (according to σ_1^ϵ). However, is σ_0^ϵ really Zara's best response?) ♢

To prove Rubinstein's Theorem 5F.1, it thus remains to prove the following

Proposition 5G.8 *Let $(\mathcal{B}, \mathbf{0})$ be a bargaining problem and let $\delta > 0$. Let σ_0^* and σ_1^* be the strategies from Example 5G.2. Then (σ_0^*, σ_1^*) is the only subgame perfect equilibrium for the Rubinstein-Ståhl Alternating Offers game.*

Proof: Let $\mathsf{SPE}(0) := \{(\sigma_0, \sigma_1)\ ;\ (\sigma_0, \sigma_1)$ is an SPE in *Alternating Offer*$\}$. Then let

$$B_0^+ := \sup_{(\sigma_0,\sigma_1)\in\mathsf{SPE}(0)} \overline{U}_0(\sigma_0, \sigma_1) \qquad \text{and} \qquad B_0^- := \inf_{(\sigma_0,\sigma_1)\in\mathsf{SPE}(0)} \overline{U}_0(\sigma_0, \sigma_1)$$

be Zara's best and worst possible payoffs in any subgame perfect equilibrium. Observe that, if $\mathbf{h} \in \mathcal{H}_1$ is any initial history of length t, then

$$\delta^t B_0^+ = \sup_{(\sigma_0,\sigma_1)\in\mathsf{SPE}(0)} \overline{U}_0(\mathbf{h}; \sigma_0, \sigma_1), \tag{5.1}$$

$$\text{and} \quad \delta^t B_0^- = \inf_{(\sigma_0,\sigma_1)\in\mathsf{SPE}(0)} \overline{U}_0(\mathbf{h}; \sigma_0, \sigma_1). \tag{5.2}$$

(This is because any subgame beginning at time t is isomorphic to the original game of *Alternating Offers*, except that the payoffs have been discounted by δ^t.)

Next, consider the 'modified' Alternating Offers game game where *Owen* moves first, and let $\mathsf{SPE}(1)$ be the set of subgame perfect equilibria in this game. Let

$$b_0^+ := \sup_{(\sigma_0,\sigma_1)\in\mathsf{SPE}(1)} \overline{U}_0(\sigma_0, \sigma_1) \qquad \text{and} \qquad b_0^- := \inf_{(\sigma_0,\sigma_1)\in\mathsf{SPE}(1)} \overline{U}_0(\sigma_0, \sigma_1)$$

be Zara's best and worst possible payoffs in any subgame perfect equilibrium where Owen moves first. Again observe: if $\mathbf{h} \in \mathcal{H}_0$ is any initial history of length t, then

$$\delta^t b_0^+ = \sup_{(\sigma_0,\sigma_1)\in \mathsf{SPE}(1)} \overline{U}_0(\mathbf{h};\sigma_0,\sigma_1), \tag{5.3}$$

$$\text{and} \quad \delta^t b_0^- = \inf_{(\sigma_0,\sigma_1)\in \mathsf{SPE}(1)} \overline{U}_0(\mathbf{h};\sigma_0,\sigma_1). \tag{5.4}$$

Claim 1: $b_0^+ = \delta B_0^+$.

Proof:

Claim 1.1: $b_0^+ \leq \delta B_0^+$.

Proof: Suppose Owen makes an initial offer (a_0, A_1) and Zara rejects this offer. Equation (5.1) implies that the best possible payoff she can obtain in the resulting subgame (where she will move first) is $\delta \cdot B_0^+$. Thus, if her strategy is subgame-perfect, then she *must* accept Owen's initial offer (a_0, A_1) if $a_0 \geq \delta \cdot B_0^+$.
Since Owen knows this, in a subgame-perfect strategy he may offer $a_0 = \delta B_0^+$, but will *never* offer more than this. Thus, any SPE (σ_0, σ_1) will always yield a payoff for Zara of at most δB_0^+, if Owen moves first. ▽ Claim 1.1

Claim 1.2: $b_0^+ \geq \delta B_0^+$.

Proof: Suppose Owen makes an initial offer (a_0, A_1), and let $\mathbf{h} = [(a_0, A_1)] \in \mathcal{H}_1$. For any $\epsilon > 0$, equation (5.1) implies that there is an SPE (σ_0, σ_1) such that $\overline{U}_0(\mathbf{h};\sigma_0,\sigma_1) > \delta B_0^+ - \epsilon$. Thus, in this SPE, Zara would never accept Owen's initial offer (a_0, A_1) unless $a_0 \geq \delta B_0^+ - \epsilon$. Thus, $\overline{U}_0(\sigma_0,\sigma_1) \geq \delta B_0^+ - \epsilon$ in the modified game where Owen moves first, which means that $b_0^+ \geq \delta B_0^+ - \epsilon$. Since this holds for all ϵ, we conclude that $b_0^+ \geq \delta B_0^+$. ▽ Claim 1.2

Claims 1.1 and 1.2 together imply that $b_0^+ = \delta B_0^+$. ◇ Claim 1

Claim 2: $b_0^- = \delta B_0^-$.

Proof:

Claim 2.1: $b_0^- \geq \delta B_0^-$.

Proof: Suppose Owen makes an initial offer (a_0, A_1) and Zara rejects this offer. Equation (5.2) implies that the *worst* possible payoff she can obtain in the resulting subgame (where she will move first) is δB_0^-. Thus, a subgame-perfect strategy for Zara will reject the proposal (a_0, A_1) if $a_0 < \delta B_0^-$ (she may accept it if $a_0 \geq \delta B_0^-$). Thus, any SPE will always guarantee a payoff for Zara of at least δB_0^- if Owen plays first; in other words, $b_0^- \geq \delta B_0^-$. ▽ Claim 2.1

Claim 2.2: $b_0^- \leq \delta B_0^-$.

Proof: Suppose Owen makes an initial offer (a_0, A_1). Let $\mathbf{h} = [(a_0, A_1)] \in \mathcal{H}_1$. For any $\epsilon > 0$, equation (5.2) implies that there is an SPE (σ_0, σ_1) such that $\overline{U}_0(\mathbf{h},\sigma_0,\sigma_1) < \delta B_0^- + \epsilon$.

In this SPE, Zara *must* accept his initial offer (a_0, A_1), as long as $a_0 \geq \delta B_0^- + \epsilon$. But Owen knows this, and σ_1 is a best response to σ_0, so σ_1 may offer $a_0 = \delta B_0^- + \epsilon$, but will *never* offer more than this.

Thus, $\overline{U}_0(\sigma_0, \sigma_1) \leq \delta B_0^- + \epsilon$, in the modified game where Owen moves first, which means that $b_0^- \leq \delta B_0^- + \epsilon$. Since this holds for all ϵ, we conclude that $b_0^- \leq \delta B_0^-$. ▽ Claim 2.2

Claims 2.1 and 2.2 together imply that $b_0^- = \delta B_0^-$. ◇ Claim 2

Next, let

$$B_1^+ := \sup_{(\sigma_0,\sigma_1)\in \mathsf{SPE}(1)} \overline{U}_1(\sigma_0,\sigma_1), \qquad B_1^- := \inf_{(\sigma_0,\sigma_1)\in \mathsf{SPE}(1)} \overline{U}_1(\sigma_0,\sigma_1),$$
$$b_1^+ := \sup_{(\sigma_0,\sigma_1)\in \mathsf{SPE}(0)} \overline{U}_1(\sigma_0,\sigma_1), \quad \text{and} \quad b_1^- := \inf_{(\sigma_0,\sigma_1)\in \mathsf{SPE}(0)} \overline{U}_1(\sigma_0,\sigma_1)$$

be Owen's best/worst payoffs in any SPE of the games where either he or Zara plays first.

Claim 3: $b_1^+ = \delta B_1^+$ *and* $b_1^- = \delta B_1^-$.

Proof: **Exercise 5.12** Argue analogously to Claims 1 and 2. ◇ Claim 3

Let $\Gamma_1 : \mathbb{R}_{\neq} \longrightarrow \mathbb{R}_{\neq}$ be as in Lemma 4A.1 on page 76.

Claim 4: $B_1^+ = \Gamma_1(b_0^-)$, $B_1^- = \Gamma_1(b_0^+)$, $b_1^+ = \Gamma_1(B_0^-)$, *and* $b_1^- = \Gamma_1(B_0^+)$.

Proof: **Exercise 5.13** Hint: the *best* payoff for Owen in a game where he plays first corresponds to the *worst* payoff for Zara in such a game, and vice versa. ◇ Claim 4

Let (B_0^δ, b_1^δ) and (b_0^δ, B_1^δ) be as in Lemma 5E.1(b) on page 107.

Claim 5: $(B_0^+, b_1^-) = (B_0^\delta, b_1^\delta)$ *and* $(b_0^+, B_1^-) = (b_0^\delta, B_1^\delta)$.

Proof: Claim 4 implies that the points (B_0^+, b_1^-) and (b_0^+, B_1^-) are on the Pareto frontier $\wp\mathcal{B}$. But Claims 1 and 3 say $b_0^+ = \delta B_0^+$ and $b_1^- = \delta B_1^-$, so (B_0^+, b_1^-) and (b_0^+, B_1^-) satisfy the defining equations of (B_0^δ, b_1^δ) and (b_0^δ, B_1^δ). But Lemma 5E.1(b) says that (B_0^δ, b_1^δ) and (b_0^δ, B_1^δ) are the *only* pair of points on $\wp_{\mathbf{q}}\mathcal{B}$ satisfying these equations. Thus, we must have $(B_0^+, b_1^-) = (B_0^\delta, b_1^\delta)$ and $(b_0^+, B_1^-) = (b_0^\delta, B_1^\delta)$. ◇ Claim 5

Claim 6: $(B_0^-, b_1^+) = (B_0^\delta, b_1^\delta)$ *and* $(b_0^-, B_1^+) = (b_0^\delta, B_1^\delta)$.

Proof: **Exercise 5.14** Argue analogously to Claim 5. ◇ Claim 6

Thus, if (σ_0, σ_1) is *any* subgame perfect equilibrium, and Zara plays first, then

$$B_0^\delta \underset{(*)}{=} B_0^- \leq \overline{U}_0(\sigma_0,\sigma_1) \leq B_0^+ \underset{(\dagger)}{=} B_0^\delta, \quad \text{which means} \quad \overline{U}_0(\sigma_0,\sigma_1) = B_0^\delta,$$
$$\text{and } b_1^\delta \underset{(\dagger)}{=} b_1^- \leq \overline{U}_1(\sigma_0,\sigma_1) \leq b_1^+ \underset{(*)}{=} b_1^\delta, \quad \text{which means} \quad \overline{U}_1(\sigma_0,\sigma_1) = b_1^\delta,$$

Here $(*)$ is by Claim 6, $(\dagger)$ is by Claim 5, and all the '$\leq$' follow from the definitions of $B_0^\pm$ and $b_1^\pm$. Similarly, if Owen plays first, then the same argument implies that

$$\overline{U}_0(\sigma_0,\sigma_1) = b_0^\delta \quad \text{and} \quad \overline{U}_1(\sigma_0,\sigma_1) = B_1^\delta.$$

Likewise, for any initial history $\mathbf{h} \in \mathcal{H}_1$ of length t, Claims 5 and 6, along with equations (5.1) and (5.2), imply that

$$\overline{U}_0(\mathbf{h}; \sigma_0, \sigma_1) \;=\; \delta^t\, B_0^\delta \quad \text{and} \quad \overline{U}_1(\mathbf{h}; \sigma_0, \sigma_1) \;=\; \delta^t\, b_1^\delta. \tag{5.5}$$

whereas, for any $\mathbf{h} \in \mathcal{H}_0$ of length t, Claims 5 and 6, along with equations (5.3) and (5.4), imply that

$$\overline{U}_0(\mathbf{h}; \sigma_0, \sigma_1) \;=\; \delta^t\, b_0^\delta \quad \text{and} \quad \overline{U}_1(\mathbf{h}; \sigma_0, \sigma_1) \;=\; \delta^t\, B_1^\delta. \tag{5.6}$$

However, the *only* subgame perfect equilibrium (σ_0, σ_1) which satisfies equations (5.5) and (5.6) for every $\mathbf{h} \in \mathcal{H}_0 \sqcup \mathcal{H}_1$ is the SPE (σ_0^*, σ_1^*) from Example 5G.2 (**Exercise 5.15** Check this). Thus, (σ_0^*, σ_1^*) is the only subgame perfect equilibrium for the Alternating Offers game. □

For other proofs of Rubinstein's theorem, please see [OR94, Prop. 122.1, p.122], [Mye91, Thm. 8.3, p.395], [Mut99, Prop. 3.3, p.62] or [Bin98, §1.7.1].

Asymmetric Bargaining The original *Alternating Offers* game assumes that both players discount future utility with the same factor δ. This is not realistic, because one of the players may be more "impatient" than the other. For example, one interpretation of the discount factor δ is that $\delta = 1 - \beta$, where β is the probability of a breakdown in negotiations due to a random exogenous event (and hence, δU represents the *expected utility* of an anticipated future bargain yielding utility U, because there is only probability δ that this anticipated future will ever transpire). However, the different players may have different estimates of β, and hence obtain different values for δ.

In the *Asymmetric Alternating Offers* game, we give both the players the same set of non-terminal and terminal actions (i.e. $\mathcal{A} = \wp_{\mathbf{q}}\mathcal{B}$ and $\mathcal{A}^\dagger = \{\text{Accept}\}$), but we introduce *two* discount factors $\delta_0, \delta_1 \in (0,1)$, and define the payoff functions U_0 and U_1 as follows: If $\mathbf{h} = (\mathbf{a}^0, \mathbf{a}^1; \ldots;\ \mathbf{a}^{2t}, \text{Accept}) \in \mathcal{H}_1^\dagger$ is a terminal history ending with Owen's acceptance of Zara's offer $\mathbf{a}^{2t} = (A_0, a_1) \in \wp_{\mathbf{q}}\mathcal{B}$, then

$$U_0(\mathbf{h}) \quad := \quad \delta_0^{2t}\, A_0 \quad \text{and} \quad U_1(\mathbf{h}) \quad := \quad \delta_1^{2t}\, a_1.$$

If $\mathbf{h} = (\mathbf{a}^0, \mathbf{a}^1; \ldots;\ \mathbf{a}^{2t}, \mathbf{a}^{2t+1};\ \text{Accept}) \in \mathcal{H}_0^\dagger$ is a terminal history ending with Zara's acceptance of Owen's offer $\mathbf{a}^{2t+1} = (a_0, A_1) \in \wp_{\mathbf{q}}\mathcal{B}$, then

$$U_0(\mathbf{h}) \quad := \quad \delta_0^{2t+1}\, a_0 \quad \text{and} \quad U_1(\mathbf{h}) \quad := \quad \delta_1^{2t+1}\, A_1.$$

Theorem 5G.9 *Let $(\mathcal{B}, \mathbf{0})$ be a bargaining problem, and let $\boldsymbol{\delta} = (\delta_0, \delta_1)$, where $\delta_0, \delta_1 \in (0,1)$.*

(a) *There exist unique $b_0^{\boldsymbol{\delta}} \leq B_0^{\boldsymbol{\delta}}$ and $B_1^{\boldsymbol{\delta}} \geq b_1^{\boldsymbol{\delta}}$ such that $(B_0^{\boldsymbol{\delta}}, b_1^{\boldsymbol{\delta}})$ and $(b_0^{\boldsymbol{\delta}}, B_1^{\boldsymbol{\delta}})$ are in $\wp_{\mathbf{q}}\mathcal{B}$, and satisfy the equations*

$$b_0^{\boldsymbol{\delta}} \quad = \quad \delta_0 \cdot B_0^{\boldsymbol{\delta}} \quad \textit{and} \quad b_1^{\boldsymbol{\delta}} \quad = \quad \delta_1 \cdot B_1^{\boldsymbol{\delta}}.$$

(b) *Define strategies* $\sigma_0^* : \mathcal{H}_1 \longrightarrow \mathcal{A} \sqcup \mathcal{A}^\dagger$ *and* $\sigma_1^* : \mathcal{H}_1 \longrightarrow \mathcal{A} \sqcup \mathcal{A}^\dagger$ *as follows:* $\sigma_0^*(\emptyset) = (B_0^{\boldsymbol{\delta}}, b_1^{\boldsymbol{\delta}})$ *and* $\sigma_1^*(\emptyset) = (b_0^{\boldsymbol{\delta}}, B_1^{\boldsymbol{\delta}})$. *For any* $\mathbf{h} \in \mathcal{H}_1$ *ending in an offer* $(a_0, A_1) \in \wp_{\mathbf{q}}\mathcal{B}$ *by Owen,*

$$\sigma_0^*(\mathbf{h}) \quad = \quad \begin{cases} \text{Accept} & \text{if} \quad a_0 \geq b_0^{\boldsymbol{\delta}}; \\ (B_0^{\boldsymbol{\delta}}, b_1^{\boldsymbol{\delta}}) & \text{if} \quad a_0 < b_0^{\boldsymbol{\delta}}. \end{cases}$$

For any $\mathbf{h} \in \mathcal{H}_0$ *ending in an offer* $(A_0, a_1) \in \wp_{\mathbf{q}}\mathcal{B}$ *by Zara,*

$$\sigma_1^*(\mathbf{h}) \quad = \quad \begin{cases} \text{Accept} & \text{if} \quad a_1 \geq b_1^{\boldsymbol{\delta}}; \\ (b_0^{\boldsymbol{\delta}}, B_1^{\boldsymbol{\delta}}) & \text{if} \quad a_1 < b_1^{\boldsymbol{\delta}}. \end{cases}$$

Then (σ_0^*, σ_1^*) *is the unique subgame perfect equilibrium for the Asymmetric Alternating Offers game defined by* $(\mathcal{B}, \mathbf{0})$ *and* $\boldsymbol{\delta}$. *In this equilibrium,* $\overline{U}_0(\sigma_0^*, \sigma_1^*) = B_0^{\boldsymbol{\delta}}$ *and* $\overline{U}_1(\sigma_0^*, \sigma_1^*) = b_1^{\boldsymbol{\delta}}$.

Proof: **Exercise 5.16** (a) Prove part (a). *Hint:* Imitate the proof of Lemma 5E.1(b) on page 107; use the Contraction Mapping Theorem.

(b) Prove part (b). *Hint:* Imitate the proof strategies of Example 5G.6 and Proposition 5G.8. □

Let $\boldsymbol{\alpha} := (\alpha_0, \alpha_1)$, for some $\alpha_0, \alpha_1 > 0$. For any bargaining problem $(\mathcal{B}, \mathbf{q})$, the *generalized Nash bargaining solution* $\eta_{\boldsymbol{\alpha}}(\mathcal{B}, \mathbf{q})$ is the unique point $(b_0, b_1) \in \wp_{\mathbf{q}}\mathcal{B}$ which maximizes the *generalized Nash product* $N_{\mathbf{q}}^{\boldsymbol{\alpha}}(b_0, b_1) = (b_0 - q_0)^{\alpha_0}(b_1 - q_1)^{\alpha_1}$. (For example, if $\mathbf{q} = \mathbf{0}$, then $N_{\mathbf{0}}^{\boldsymbol{\alpha}}(b_0, b_1) = b_0^{\alpha_0} b_1^{\alpha_1}$.)

Exercise 5.17 (a) Let $\|\boldsymbol{\alpha}\|_1 := \alpha_0 + \alpha_1$. Define $\beta_0 := \alpha_0/\|\boldsymbol{\alpha}\|_1$ and $\beta_1 := \alpha_1/\|\boldsymbol{\alpha}\|_1$, so that $\|\boldsymbol{\beta}\|_1 = 1$. Show that $\eta_{\boldsymbol{\beta}} = \eta_{\boldsymbol{\alpha}}$. (Thus, when analyzing the generalized Nash solution, we can assume without loss of generality that $\|\boldsymbol{\alpha}\|_1 = 1$.)

(b) If $\alpha_0 = \alpha_1$, show that $\eta_{\boldsymbol{\alpha}}$ is just the classical Nash bargaining solution.

(c) For any $\alpha_0, \alpha_1 > 0$, show that $\eta_{\boldsymbol{\alpha}}$ satisfies axioms **(RI)** and **(IIA)**.

(d) However $\eta_{\boldsymbol{\alpha}}$ satisfies **(S)** only if $\alpha_0 = \alpha_1$.

(e) Let $(\mathcal{B}, \mathbf{0})$ be a symmetric bargaining problem and let $\eta_{\boldsymbol{\alpha}}(\mathcal{B}, \mathbf{0}) = (b_0, b_1)$. If $\alpha_0 > \alpha_1$, show that $b_0 \geq b_1$.

One way to interpret α_0 and α_1 is as 'exponential discount rates', so that a delay of duration $\tau \in \mathbb{R}_{\neq}$ induces a discount factor $\delta_0 = e^{-\tau/\alpha_0}$ for Zara and a discount factor $\delta_1 = e^{-\tau/\alpha_1}$ for Owen. If we let $\tau \searrow 0$, then both δ_0 and δ_1 tend to one (i.e. neither player discounts the immediate future very much), but δ_0 and δ_1 tend to one at different rates. If $\alpha_1 < \alpha_0$, then $\tau/\alpha_1 > \tau/\alpha_0$, so that $\delta_1 < \delta_0$; this means that Owen is more 'impatient' than Zara, and his impatience weakens his bargaining position in the Alternating Offers game. This interpretation of Exercise 5.17(c), is justified by the next result, due to Ken Binmore:

Proposition 5G.10 *$(\mathcal{B}, \mathbf{0})$ be a bargaining problem, let $\alpha_0, \alpha_1 > 0$, and let $(b_0^{\boldsymbol{\alpha}}, b_1^{\boldsymbol{\alpha}}) = \eta_{\boldsymbol{\alpha}}(\mathcal{B}, \mathbf{0})$ be the generalized Nash solution. For any $\tau > 0$, let $\boldsymbol{\delta}(\tau) = (\delta_0, \delta_1)$, where $\delta_0 := e^{-\tau/\alpha_0}$ and $\delta_1 := e^{-\tau/\alpha_1}$, and then let $b_0^{\boldsymbol{\delta}(\tau)} < B_0^{\boldsymbol{\delta}(\tau)}$ and $b_1^{\boldsymbol{\delta}(\tau)} < B_1^{\boldsymbol{\delta}(\tau)}$ be as in Theorem 5G.9(a). Then*

$$\lim_{\tau \searrow 0} (B_0^{\boldsymbol{\delta}(\tau)}, b_1^{\boldsymbol{\delta}(\tau)}) \quad = \quad (b_0^{\boldsymbol{\alpha}}, b_1^{\boldsymbol{\alpha}}) \quad = \quad \lim_{\tau \searrow 0} (b_0^{\boldsymbol{\delta}(\tau)}, B_1^{\boldsymbol{\delta}(\tau)}).$$

Proof: **Exercise 5.18** Hint: Mimic the proof of Lemma 5E.1(c) on page 107. □

In Proposition 5G.10, we can interpret τ as 'reaction time' —i.e the minimum delay between an offer by one player and the counteroffer by the other player. Proposition 5G.10 says that, if this reaction time becomes very small, then the payoffs of the Asymmetric Alternating Offers game converge to the generalized Nash Solution which favours the 'more patient' player. See [Bin98, §1.7.1] for an excellent discussion of this asymmetric bargaining model. See [Mye91, Thm 8.3] for a very similar model, where δ_0 and is interpreted as the probability (in Zara's perception) that Owen will walk away from the bargaining table each time one of his offers is refused; likewise δ_1 is Owen's subjective probability that Zara will walk away from the bargaining table when her offer is refused. Myerson derives a result very similar to Theorem 5G.9.

Other variations: Rubinstein's original alternating offers model has since been modified for enhanced realism in several ways, most of which lead to results similar to Theorems 5F.1 and 5G.9.

(a) *Payoff flows; Breakdown vs. Deadlock:* In our bargaining models, we have assumed that the players get their entire payoff if and when a bargain is concluded, and get the status quo otherwise. In real life, however, the parties are often trying to *renegotiate* an existing and ongoing economic relationship (think of negotiations between a labour union and management). In these situations, the parties are not negotiating over a share of a fixed payoff (which they get on completion of the bargain), but rather, negotiating over a share of an ongoing ***payoff flow***, some of which arrives during each time unit. We can imagine that the *status quo* payoff flows continue even while the bargaining is under way. In this situation, the bargaining can have three outcomes:

- *Success:* The parties eventually move from the status quo payoff flow to some mutually agreeable payoff flow on the Pareto-frontier of the bargaining set.
- *Deadlock:* The parties continue negotating indefinitely, and receive the status quo payoff flow forever.
- *Breakdown:* The parties discontinue their relationship (e.g. strike, lockout, bankruptcy, etc.), and each party takes its best outside option (which is presumably a payoff flow less than the status quo).

Under these hypotheses, Binmore, Shaked and Sutton have proved a generalization of Rubinstein's Theorems 5F.1 and 5G.9; see [Bin98, Appendix C].

(b) *Linear discounting* (or, 'Time is money'): Suppose that, instead of the exponential discounting described above, the players discount future earnings linearly. In other words, there are constants $c_0, c_1 > 0$ so that Zara assigns utility $-c_0$ to each lost time unit, and Owen assigns it utility $-c_1$. Thus a bargain (b_0, b_1) reached at time n is only worth $b_0 - c_0 n$ to Zara and is worth $b_1 - c_0 n$ to Owen. In this case, if $c_0 < c_1$, then there is a unique subgame perfect equilibrium where Zara gets everything and Owen gets nothing (and vice versa). If $c_0 = c_1$, then there are many subgame perfect equilibria. See [OR94, Exercise 125.2].

(c) *Discrete bargaining set* (or 'Money comes in whole numbers of cents'): Suppose that we replace $\mathcal{B}$ with some large but discrete set of points. (Note that this violates the convexity assumption, and implicitly assumes that lotteries over outcomes are not possible). Then van Damme, Selten, and Winters have shown that *any* point in $\mathcal{B}$ is the outcome of some subgame-perfect equilibrium [vDSW90]; see also [Mye91, §8.9, p.403].

(d) *Incomplete information:* Theorems 5F.1 and 5G.9 only apply to bargaining games where Zara has *complete information* about Owen's preferences (and vice versa) so that she can correctly identify when Owen is making a credible threat and when he is simply bluffing. In games of incomplete information, successful bluffing is possible, and this can change the bargaining outcome [Mye91, §8.8, p.399].

(e) *Bargains with three or more players:* The Nash solution (and the other bargaining solutions discussed in Chapters 6 and 7) have obvious extensions to bargaining problems with three or more players, but Rubinstein's Theorems 5F.1 and 5G.9 do not. Although there is a three-person version of the Rubinstein-Ståhl game of *Alternating Offers*, it does *not* yield the Nash solution, but instead has infinitely many subgame perfect equilibria [OR94, §7.4.5, p.130].

(f) *Ordinal utility and ordinal discounting:* For those who are uncomfortable with the 'strong rationality' assumptions behind von Neumann and Morgenstern's theory of cardinal utility functions, or who don't believe that people necessarily discount future utility exponentially in time, Rubinstein has studied the *Alternating Offers* game under much weaker assumptions. Instead of a cardinal utility function and a discount factor, each player merely has a *preference ordering* over the space of possible bargains, which extends to a preference ordering over the space of all possible future bargains, with 'discounting' expressed by the fact that each player always prefers a given bargain now to the same bargain in the future. Under certain apparently weak assumptions about the 'consistency' or 'rationality' of these preference orderings, Rubinstein proves a theorem essentially identical to Theorem 5F.1(a). I say *apparently* weak because a key step in Rubinstein's argument is the use of a result by himself and Fishburn [FR82] which says that such a preference

ordering can always be *represented* using a cardinal utility function and an exponential discount factor. In other words, just as von Neumann and Morgenstern's 'revealed preference' Theorem 3A.1 says that rational people will behave 'as if' maximizing the expected value of some cardinal utility function (whether they are conscious of this or not), the Rubinstein-Fishburn theorem says that rational people playing an extensive game will behave 'as if' maximizing the (future-discounted) value of some cardinal utility function (whether they are conscious of this or not). See [OR94, §7.2, p.118].

Further reading

For more information on the *Alternating Offers* model of §5F and §5G and its various extensions, see [Mut99]. See [Nap02] for more about *Alternating Offers*, as well as the Zeuthen model of §5D, and other classic bargaining models by Edgeworth and Hicks, and also for a very interesting 'evolutionary' analysis of the 'ultimatum game' (see p.101). Finally, for a very different application of game theory to negotiation (almost disjoint from what we've covered here), see [Bra90].

Chapter 6

Interpersonal Comparison Models

The Nash bargaining solution (see §4B) deliberately avoids the contentious issue of interpersonal utility comparisons (see §3B); indeed it effectively legislates interpersonal comparisons to be meaningless, by imposing the axiom **(RI)**. However, in some situations, there may be some reasonable criteria by which such comparisons could be made, and **(RI)** is inappropriate.

Recall that Theorem 3A.1 (page 58) only defines the von Neumann-Morgenstern cardinal utility function U_0 of Zara up to affine transformation. Thus, if $c_0 > 0$ is some constant, then c_0U_0 is also a valid cardinal utility function for Zara. Likewise, if U_1 is a vNM cardinal utility function for Owen, and $c_1 > 0$, then c_1U_1 is also a vNM cardinal utility function for Owen.

Suppose, for the sake of argument, that we could find some *calibration constants* $c_0, c_1 > 0$ such that c_0U_0 and c_1U_1 were 'comparable', meaning that increasing c_0U_0 by one unit (i.e. increasing U_0 by $1/c_0$ units) was somehow 'ethically equivalent' to increasing c_1U_1 by one unit (i.e. increasing U_1 by $1/c_1$ units). Then it would be coherent to interpret the utilitarian sum $c_0U_0 + c_1U_1$ as the 'total happiness' of Zara and Owen as a group; thus it would be a coherent social goal to try to maximize this sum. Likewise, if $c_0U_0 < c_1U_1$, then it would be coherent to interpret this to mean that Zara was (in some objective sense) 'less happy' than Owen; hence it would be a coherent social goal to try to 'equalize' the happiness of both players by by setting $c_0U_0 = c_1U_1$.

6A The Utilitarian Solution

Prerequisites: §4A, §4A

The utilitarian bargaining solution is the solution that would probably have been proposed by 19th century utilitarian thinkers such as Jeremy Bentham or John Stuart Mill. To be precise, we fix a pair of calibration constants $\mathbf{c} = (c_0, c_1) \in [0, \infty)$. The **c**-*utilitarian bargaining solution* is the point $\mu_{\mathbf{c}}(\mathcal{B}, \mathbf{q}) := (b_0, b_1)$ in the negotiating set $\wp_{\mathbf{q}}\mathcal{B}$ which maximizes the **c**-*utilitarian sum* $U_{\mathbf{c}}(b_0, b_1) := c_0b_0 + c_1b_1$. In other words, $\mu_{\mathbf{c}}$ maximizes the 'total utility' of the society, assuming that $\frac{1}{c_0}$ units of Zara's utility are calibrated to be 'equivalent' to $\frac{1}{c_1}$ units of Owen's utility. Geometrically speaking, we can draw parallel lines of the form $\mathbf{L}_r := \{(b_0, b_1) \; ; \; c_0b_0 + c_1b_1 = r\}$

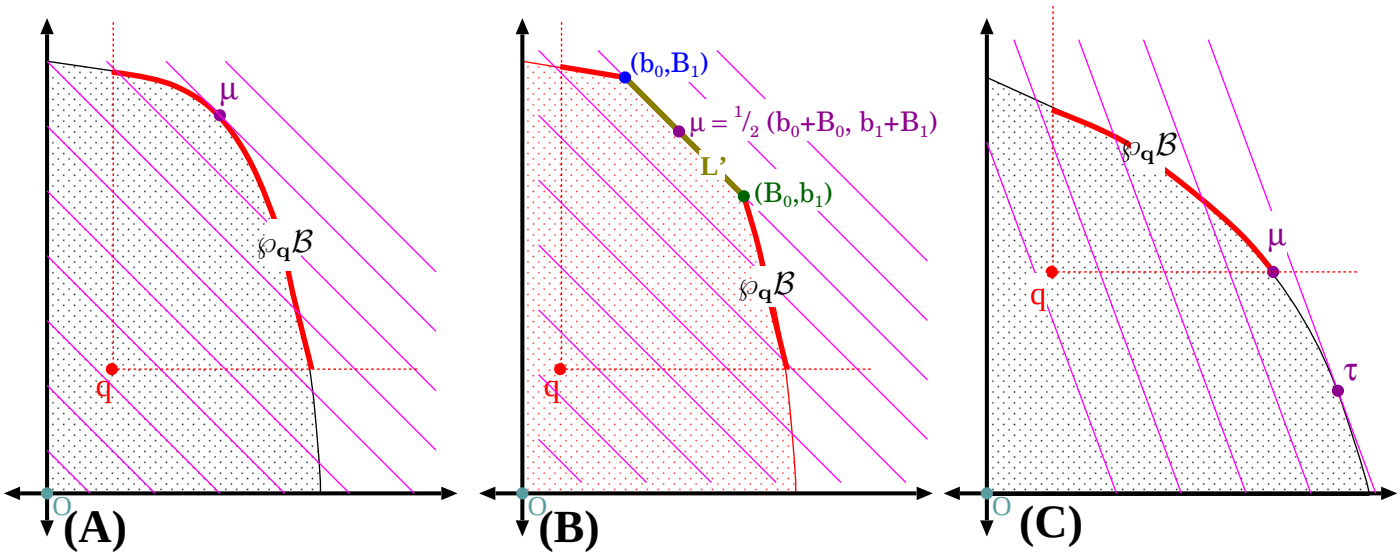


Figure 6.1: The utilitarian solution to the bargaining problem.

for each $r \in \mathbb{R}$. Then $\mu_{\mathbf{c}}$ is the point where some line $\mathbf{L}_r$ is *tangent* to the negotiating set $\wp_{\mathbf{q}}\mathcal{B}$, as shown in Figure 6.1(A).

Notes: (a) If $\mathcal{B}$ is *not* strictly convex, then the negotiating set $\wp_{\mathbf{q}}\mathcal{B}$ may contain a line segment $\mathbf{L}'$ between two ponts (b_0, B_1) and (B_0, b_1). If aforementioned lines $\mathbf{L}_r$ are parallel to $\mathbf{L}'$ [i.e. if $c_0(B_0 - b_0) = c_1(B_1 - b_1)$], then every point on $\mathbf{L}'$ will be a tangency point between $\wp_{\mathbf{q}}\mathcal{B}$ and some $\mathbf{L}_r$; see Figure 6.1(B). Thus, that the utilitarian solution is indifferent between all the pointson $\mathbf{L}'$ [although clearly Zara will favour (B_0, b_1) and Owen will favour (b_0, B_1)]. In this case, for the sake of specifying a unique solution, we could define the utilitarian solution to be the midpoint $\frac{1}{2}(b_0 + B_0,\ b_1 + B_1)$.

(b) The utilitarian solution depends upon the slopes of the lines $\mathbf{L}_r$, which depends upon the calibration constants c_0 and c_1. Indeed, if $\mathcal{B}$ is strictly convex, then, with suitable a calibration constant $\mathbf{c}$, we can make $\mu_{\mathbf{c}}$ any point on $\wp_{\mathbf{q}}\mathcal{B}$ (**Exercise 6.1** Check this).

(c) In an extreme situation, the tangency point τ between the line $\mathbf{L}_r$ and the Pareto frontier $\wp\mathcal{B}$ may lie *outside* the negotiating set $\wp_{\mathbf{q}}\mathcal{B}$ —i.e. it may be below or left of the the status quo point $\mathbf{q}$, as in Figure 6.1(C). Since we assume that neither player will accept less than his/her status quo payoff, we must then define $\mu_{\mathbf{c}}$ to be the extreme point of the negotiating set $\wp_{\mathbf{q}}\mathcal{B}$ that lies closest to τ.

(c) The utilitarian solution is totally independent of the choice of zero point $\mathbf{0}$ for the utility functions of Zara and Owen. It also depends only weakly on the status quo point $\mathbf{q}$ (i.e. only to the extent that $\mu_{\mathbf{c}}$ must be Pareto-prefered to $\mathbf{q}$). In this sense $\mu_{\mathbf{c}}$ is 'egalitarian' in that it cares nothing for vested interests or existing endowments of wealth. The utilitarian solution is highly 'inegalitarian', however, in the sense that the player who has the most to gain from the bargain is the one who will gain the most, as the next example shows.

Example 6A.1: (a) Suppose Zara has \$16 and Owen has \$25, and they are bargaining over how to divide an additional \$5. We assume both players have identical utility functions for money, and we assume that these utility functions are *concave*, which means that each additional cent is worth more for the poorer Zara than it is for the richer Owen (See Exercise 3.10 on page 60). It follows that greater 'total utility' will be generated by giving Zara all \$5 than would be generated by splitting the money in any other way; hence the utilitarian solution will award all of the money to Zara. (Whether or not this is 'fair' depends on your politics.)

(b) If the players were bargaining over a larger sum (say, \$20), then the solution would award money to Zara up until they were equal, and then equally split the remaining money (because each additional cent would be worth slightly more to whoever had one less cent at that instant). Thus, Zara would get \$14.50, while Owen would get \$5.50, and both players would then end up with \$30.50. ♢

In other contexts, the utilitarian solution is 'inegalitarian' in that it favours the more 'efficient' or 'productive' player, or at least, the one who will benefit most from the resource in question:

Example 6A.2: Suppose that Zara and Owen are bargaining over how to divide one acre of farmland they wish to cultivate. Owen can expect to produce \$1 of crop from this acre, but Zara is a more efficient farmer, and expects to product \$4 of crop. Assume both players have the utility function $U(x) = \sqrt{x}$ for money, and (unlike the previous example) assume that both have the same initial endowment (say, both have \$0). Thus, if a_0 acres are given to Zara, and $a_1 = 1 - a_0$ acres are given to Owen, then Zara will produce $\$4a_0$ worth of crop, yielding a utility of $U(4a_0) = \sqrt{4a_0} = 2\sqrt{a_0}$, whereas Owen will produce $\$(1 - a_0)$ worth of crop, yielding a utility of $U(1 - a_0) = \sqrt{1 - a_0}$. The total utility of the allocation will then be $\overline{U}(a_0) = 2\sqrt{a_0} + \sqrt{1 - a_0}$. It is a simple exercise in calculus to see that $\overline{U}(a_0)$ is maximized when $a_0 = \frac{4}{5}$. With this allocation, Zara produces $\$4 \times \frac{4}{5} = \$3.20$, yielding her a utility of $\sqrt{3.20} \approx 1.789$, whereas Owen produces $\$1 \times \frac{1}{5} = \$0.20$, yielding him a utility of $\sqrt{0.2} \approx 0.447$. Note: we used the function $U(x) = \sqrt{x}$ only to make the computations easy. A similar phenomenon occurs with any concave (i.e. 'risk-averse') utility function; see Exercise 3.10 on page 60. ♢

Exercise 6.2: Let $\mathbf{C} := \left\{(c_0, c_1) \in \mathbb{R}^2_{\not\succ} \; ; \; c_0 + c_1 = 1\right\}$. If $\mathbf{c} \in \mathbb{R}^2_{\not\succ}$, and $\widetilde{\mathbf{c}} := \mathbf{c}/(c_0 + c_1)$, then $\widetilde{\mathbf{c}} \in \mathbf{C}$. Show that $\mu_{\mathbf{c}} = \mu_{\widetilde{\mathbf{c}}}$. (Thus, we can assume without loss of generality that the calibration constant $\mathbf{c}$ is an element of $\mathbf{C}$).

We will now show that the utilitarian solution has a natural axiomatic characterization. Let $(\mathcal{B}_0, \mathbf{q}_0)$ and $(\mathcal{B}_1, \mathbf{q}_1)$ be two bargaining problems. For any $r \in [0, 1]$, let $\mathbf{q}_r := (1 - r)\mathbf{q}_0 + r\mathbf{q}_1$, and let

$$\mathcal{B}_r \quad := \quad \{(1 - r)\mathbf{b}_0 + r\mathbf{b}_1 \; ; \; \mathbf{b}_0 \in \mathcal{B}_0 \text{ and } \mathbf{b}_1 \in \mathcal{B}_1\}.$$

Exercise 6.3 (a) Recall that $\mathcal{B}_0$ and $\mathcal{B}_1$ are convex; show that $\mathcal{B}_r$ is also convex.
(b) Suppose $\mathcal{B}_0$ and $\mathcal{B}_1$ are *strictly* convex; show that $\mathcal{B}_r$ is also strictly convex.
(c) Show that $\mathbf{q}_r \in \mathcal{B}_r$. Conclude that $(\mathcal{B}_r, \mathbf{q}_r)$ is a bargaining problem. Furthermore, if $(\mathcal{B}_0, \mathbf{q}_0) \in \mathfrak{B}^*$ and $(\mathcal{B}_1, \mathbf{q}_1) \in \mathfrak{B}^*$, then $(\mathcal{B}_r, \mathbf{q}_r) \in \mathfrak{B}^*$ also.

Intuitively, $(\mathcal{B}_r, \mathbf{q}_r)$ represents the set of *expected payoffs* of a 'random' bargaining problem. The feasible set and status quo of this random bargaining problem are currently unknown; with probability r it will turn out to be the problem $(\mathcal{B}_1, \mathbf{q}_1)$, whereas with probability $(1-r)$ it will turn out to be the problem $(\mathcal{B}_0, \mathbf{q}_0)$.

Example 6A.3: Suppose Zara and Owen are potential business partners who are negotiating how to split the anticipated profits of some joint business venture. They must agree on a binding contract *now*, but the outcome of their business venture will not be known for another year (because it perhaps depends upon future market prices which are presently unknown). Indeed, even their respective payoffs under the status quo (i.e. no cooperation) are unknown, one year in advance. In this case, $(\mathcal{B}_1, \mathbf{q}_1)$ and $(\mathcal{B}_0, \mathbf{q}_0)$ represent the feasible sets and status quo payoffs in two possible futures, which will occur with probability r and $(1-r)$ respectively. But Zara and Owen must reach an agreement now, so they are really facing the bargaining problem $(\mathcal{B}_r, \mathbf{q}_r)$. ♢

A bargaining solution $\alpha : \mathfrak{B}^* \longrightarrow \mathbb{R}^2_\nearrow$ is *linear* if, for any $(\mathcal{B}_0, \mathbf{q}_0)$ and $(\mathcal{B}_1, \mathbf{q}_1)$ in $\mathfrak{B}^*$, and any $r \in [0, 1]$,

$$\alpha(\mathcal{B}_r, \mathbf{q}_r) \quad = \quad r\,\alpha(\mathcal{B}_1, \mathbf{q}_1) \;+\; (1-r)\,\alpha(\mathcal{B}_0, \mathbf{q}_0).$$

In terms of Example 6A.3, this means that it doesn't matter whether Zara and Owen...

(1) ...negotiate an agreement *now* for the bargaining problem $(\mathcal{B}_r, \mathbf{q}_r)$;

or

(2) ...wait one year to see what the future holds, and then end up either negotating over the bargaining problem $(\mathcal{B}_1, \mathbf{q}_1)$ [with probability r] or negotating over the bargaining problem $(\mathcal{B}_0, \mathbf{q}_0)$ [with probability $1-r$].

Options (1) and (2) will yield the same expected utility for each player. Myerson calls this the 'No Timing' property. The 'No Timing' property is important; if (1) and (2) did *not* yield the same expected utilities, then one player would have an incentive to push for a final settlement *now*, whereas the other player would have an incentive to delay a final settlement until after one year has passed. This difference in agendas could create a conflict which scuttles the whole agreement.

Myerson has shown that the *only* linear bargaining solutions are the utilitarian ones:

Theorem 6A.4 [Mye81]

(a) *For any* $\mathbf{c} \in \mathbb{R}^2_{\nearrow}$, *the* $\mathbf{c}$*-utilitarian bargaining solution* $\mu_{\mathbf{c}}$ *is linear.*

(b) *If* $\alpha : \mathfrak{B}^* \longrightarrow \mathbb{R}^2_{\nearrow}$ *is any linear bargaining solution, then* $\alpha = \mu_{\mathbf{c}}$ *for some* $\mathbf{c} \in \mathbb{R}^2_{\nearrow}$.

Proof: (a) **Exercise 6.4**

(b) Let $\alpha : \mathfrak{B}^* \longrightarrow \mathbb{R}^2_{\nearrow}$ be a linear bargaining solution; we must find some $\mathbf{c} \in \mathbb{R}^2_{\nearrow}$ such that $\alpha(\mathcal{B}, \mathbf{q}) = \mu_{\mathbf{c}}(\mathcal{B}, \mathbf{q})$ for every $(\mathcal{B}, \mathbf{q}) \in \mathfrak{B}^*$. Note that it suffices to look for $\mathbf{c}$ in the line segment $\mathbf{C} := \left\{(c_0, c_1) \in \mathbb{R}^2_{\nearrow} \; ; \; c_0 + c_1 = 1\right\}$ (by Exercise 6.2 above). Thus, we seek some $\mathbf{c} \in \mathbf{C}$ so that $\alpha(\mathcal{B}, \mathbf{q})$ always maximizes the value of the linear functional $U_{\mathbf{c}}(\mathbf{x}) := c_0 x_0 + c_1 x_1$ over the negotiating set of $(\mathcal{B}, \mathbf{q})$. That is:

$$\forall\ (\mathcal{B}, \mathbf{q}) \in \mathfrak{B}^*, \qquad U_{\mathbf{c}}[\alpha(\mathcal{B}, \mathbf{q})] \quad = \quad \max_{\mathbf{b} \in \wp_{\mathbf{q}}\mathcal{B}} U_{\mathbf{c}}(\mathbf{b}). \tag{6.1}$$

Now, for any $(\mathcal{B}, \mathbf{q}) \in \mathfrak{B}^*$, let

$$\mathbf{O}_{(\mathcal{B},\mathbf{q})} \quad := \quad \left\{ \mathbf{c} \in \mathbf{C} \; ; \; \max_{\mathbf{b} \in \wp_{\mathbf{q}}\mathcal{B}} U_{\mathbf{c}}(\mathbf{b}) \; > \; U_{\mathbf{c}}[\alpha(\mathcal{B}, \mathbf{q})] \right\} \quad \subseteq \quad \mathbf{C}. \tag{6.2}$$

Thus, if $\mathbf{c} \in \mathbf{O}_{(\mathcal{B},\mathbf{q})}$ for some $(\mathcal{B}, \mathbf{q}) \in \mathfrak{B}^*$, then α cannot possibly be $\mu_{\mathbf{c}}$. Let

$$\mathbf{O} \quad := \quad \bigcup_{(\mathcal{B},\mathbf{q}) \in \mathfrak{B}^*} \mathbf{O}_{(\mathcal{B},\mathbf{q})} \quad \subseteq \quad \mathbf{C}.$$

Thus, if $\mathbf{c} \in \mathbf{O}$, then α can't be $\mu_{\mathbf{c}}$. On the other hand, if $\mathbf{c} \in \mathbf{C} \setminus \mathbf{O}$, then this means $\mathbf{c} \notin \mathbf{O}_{(\mathcal{B},\mathbf{q})}$ for *any* $(\mathcal{B}, \mathbf{q}) \in \mathfrak{B}^*$, which means that $\mathbf{c}$ satisfies eqn.(6.1) —hence $\alpha = \mu_{\mathbf{c}}$.

Thus, it suffices to show that $\mathbf{C} \setminus \mathbf{O}$ is nonempty, which means showing that $\mathbf{O} \subsetneq \mathbf{C}$.

Suppose, by contradiction, that $\mathbf{C} = \mathbf{O}$.

Claim 1: *For any* $(\mathcal{B}, \mathbf{q}) \in \mathfrak{B}^*$, *the set* $\mathbf{O}_{(\mathcal{B},\mathbf{q})}$ *is an open subset of* $\mathbf{C}$.

Proof: **Exercise 6.5** ◇ Claim 1

Thus, the collection $\left\{\mathbf{O}_{(\mathcal{B},\mathbf{q})} \; ; \; (\mathcal{B}, \mathbf{q}) \in \mathfrak{B}^*\right\}$ is an open cover of $\mathbf{C}$. But $\mathbf{C}$ is a compact set (because it is closed and bounded), so the open cover $\left\{\mathbf{O}_{(\mathcal{B},\mathbf{q})} \; ; \; (\mathcal{B}, \mathbf{q}) \in \mathfrak{B}^*\right\}$ has a finite subcover —in other words, there exist some $(\mathcal{B}_1, \mathbf{q}_1)$, $(\mathcal{B}_2, \mathbf{q}_2)$, $\ldots$, $(\mathcal{B}_J, \mathbf{q}_J)$ in $\mathfrak{B}^*$ such that

$$\mathbf{C} \quad = \quad \mathbf{O}_{(\mathcal{B}_1,\mathbf{q}_1)} \cup \mathbf{O}_{(\mathcal{B}_2,\mathbf{q}_2)} \cup \cdots \cup \mathbf{O}_{(\mathcal{B}_J,\mathbf{q}_J)}. \tag{6.3}$$

Now, let $\mathcal{B} := \frac{1}{J}\sum_{j=1}^{J} \mathcal{B}_j$ and let $\mathbf{q} := \frac{1}{J}\sum_{j=1}^{J} \mathbf{q}_j$. Let $\mathbf{a} := \alpha(\mathcal{B}, \mathbf{q})$; then the linearity of α implies that

$$\mathbf{a} \quad = \quad \frac{1}{J}\sum_{j=1}^{J} \mathbf{a}_j, \qquad \text{where } \mathbf{a}_j = \alpha(\mathcal{B}_j, \mathbf{q}_j), \text{ for all } j \in [1...J]. \tag{6.4}$$

Claim 2: *There exists some* $\mathbf{c} \in \mathbf{C}$ *such that* $U_{\mathbf{c}}(\mathbf{a}) = \max_{\mathbf{b} \in \wp_{\mathbf{q}}\mathcal{B}} U_{\mathbf{c}}(\mathbf{b})$.

Proof: **Exercise 6.6** *Hint:* This is because $\mathbf{a} \in \wp_{\mathbf{q}}\mathcal{B}$, and $\mathcal{B}$ is convex. ◇ Claim 2

But eqn.(6.3) implies that there is some $i \in [1...J]$ such that $\mathbf{c} \in \mathbf{O}_{(\mathcal{B}_i, \mathbf{q}_i)}$. Thus, there is some $\mathbf{b}_i \in \wp_{\mathbf{q}_i}\mathcal{B}_i$ such that

$$U_{\mathbf{c}}(\mathbf{b}_i) \quad > \quad U_{\mathbf{c}}(\mathbf{a}_i), \tag{6.5}$$

by definition of $\mathbf{O}_{(\mathcal{B}_i, \mathbf{q}_i)}$ [see eqn.(6.2)]. Now, define $\mathbf{b} := \frac{1}{J}\left(\mathbf{b}_i + \sum_{i \neq j=1}^{J} \mathbf{a}_j\right)$. Then $\mathbf{b} \in \wp_{\mathbf{q}}\mathcal{B}$, and

$$\begin{aligned} U_{\mathbf{c}}(\mathbf{b}) &\underset{(\mathrm{L})}{=} \frac{1}{J}\left[U_{\mathbf{c}}(\mathbf{b}_i) + \sum_{i \neq j=1}^{J} U_{\mathbf{c}}(\mathbf{a}_j)\right] \underset{(*)}{>} \frac{1}{J}\left[U_{\mathbf{c}}(\mathbf{a}_i) + \sum_{i \neq j=1}^{J} U_{\mathbf{c}}(\mathbf{a}_j)\right] \\ &\underset{(\mathrm{L})}{=} U_{\mathbf{c}}\left[\frac{1}{J}\sum_{j=1}^{J} \mathbf{a}_j\right] \underset{(\dagger)}{=} U_{\mathbf{c}}(\mathbf{a}), \end{aligned}$$

where **(L)** is because $U_{\mathbf{c}}$ is a linear function, $(*)$ is by inequality (6.5), and $(\dagger)$ is by equation (6.4).

Thus, $U_{\mathbf{c}}(\mathbf{b}) > U_{\mathbf{c}}(\mathbf{a})$. But this contradicts the conclusion of Claim 2. This contradiction means that $\mathbf{O} \subsetneq \mathbf{C}$, which means that there is some $\mathbf{c} \in \mathbf{C} \setminus \mathbf{O}$, which means that α is the $\mathbf{c}$-utilitarian solution, as desired. □

Exercise 6.7: Generalize the utilitarian bargaining solution to bargains involving three or more players.

Exercise 6.8: Generalize the statement and proof of Theorem 6A.4 to bargains involving three or more players.

Exercise 6.9: Suppose Zara and Owen are bargaining over how to divide some quantity M of money (see Example 4A.2), and that they begin with status quo endowment (q_0, q_1). Let $\mathbf{c} = (1, 1)$.

(a) Suppose Zara and Owen have identical, strictly concave utility functions $U_0 = U_1 = U$. Assume $q_0 \leq q_1$ (i.e. Zara is poorer). Show that:

(i) If $q_1 - q_0 > M$, then the utilitarian solution $\mu_{\mathbf{c}}$ awards all the money to Zara.

(ii) If $q_1 - q_0 < M$, then the utilitarian solution $\mu_{\mathbf{c}}$ awards the first $q_1 - q_0$ dollars to Zara, and then splits the remaining $M - q_1 + q_0$ dollars evenly between them.

(b) Say that Owen is *more acquisitive* if $U_1'(x) \geq U_0'(x)$ for all $x \in \mathbb{R}_{\not\prec}$. Heuristically, this means that Owen can get 'more marginal happiness' out of each additional dollar than Zara can. Suppose they begin *the same* initial wealth; i.e. $q_0 = q_1$. Show that the Utilitarian solution $\mu_{\mathbf{c}}$ gives *all* the money to Owen.

(Where would you locate Utilitarianism on the 'left vs. right' political spectrum?)

Exercise 6.10: (Triage)

During a major natural disaster, or in a battlefield hospital, the inundation of injured patients often totally overwhelms the available medical resources. In this situation, some hospitals implement a policy of *triage*, whereby the incoming wounded are divided into three categories:

1. Slightly injured patients, who will probably survive without any medical intervention.
2. Severely injured patients, who will likely survive if and only if they receive immediate and major medical attention.
3. Critically or fatally injured patients, who probably will not survive even if they receive major medical attention.

Category 2 patients receive the vast majority of medical resources, e.g. surgery, drugs, antibiotics, etc. Category 1 patients are given a place to sit and perhaps something to eat or drink. If any medical personnel are available after dealing with Category 2, then they might administer minor first-aid (e.g. bandages, splints and disinfectants). Category 3 patients are made as comfortable as possible, given a lot of morphine, and left to die.

Explain the *triage* policy in terms of the utilitarian bargaining solution (where the incoming wounded are the 'bargainers').

6B The Proportional Solution

Prerequisites: §4A

The proportional solution stipulates that both players should gain 'the same amount' when we moving from the status quo the outcome of the bargain. To be precise, fix a pair of calibration constants $\mathbf{c} = (c_0, c_1)$. The **c-*proportional bargaining solution*** is the function $\rho_{\mathbf{c}} : \mathfrak{B} \longrightarrow \mathbb{R}^2_{\not\prec}$ so that, for any bargaining problem $(\mathcal{B}, \mathbf{q})$, $\rho_{\mathbf{c}}(\mathcal{B}, \mathbf{q})$ is the unique point (r_0, r_1) in the negotiating set $\wp_{\mathbf{q}}\mathcal{B}$ such that $c_0(r_0 - q_0) = c_1(r_1 - q_1)$.

Geometrically speaking, let $\mathbf{L}$ be the line through the point $\mathbf{q}$ parallel to the vector $\mathbf{p} := (c_0^{-1}, c_1^{-1})$ (or equivalently, the line of slope c_1/c_0). Then $\rho_{\mathbf{c}}(\mathcal{B}, \mathbf{q})$ is the (unique) point where $\mathbf{L}$ intersects $\wp_{\mathbf{q}}\mathcal{B}$; see Figure 6.2(P). In other words, $\rho_{\mathbf{c}}(\mathcal{B}, \mathbf{q}) = \mathbf{q} + \overline{r}\mathbf{p}$, where $\overline{r} := \max\{r > 0 \; ; \; \mathbf{q} + r\mathbf{p} \in \mathcal{B}\}$.

If we use the calibration $c_0 = c_1 = 1$, then $\mathbf{L}$ is at 45^o, and then another way to define $\rho_{\mathbf{c}}(\mathcal{B}, \mathbf{q})$ is to draw the largest possible square inside of $\mathcal{B}$ whose bottom left corner is at the status quo $\mathbf{q}$; then the top right corner touches $\wp_{\mathbf{q}}\mathcal{B}$, and this point is $\rho_{\mathbf{c}}(\mathcal{B}, \mathbf{q})$.

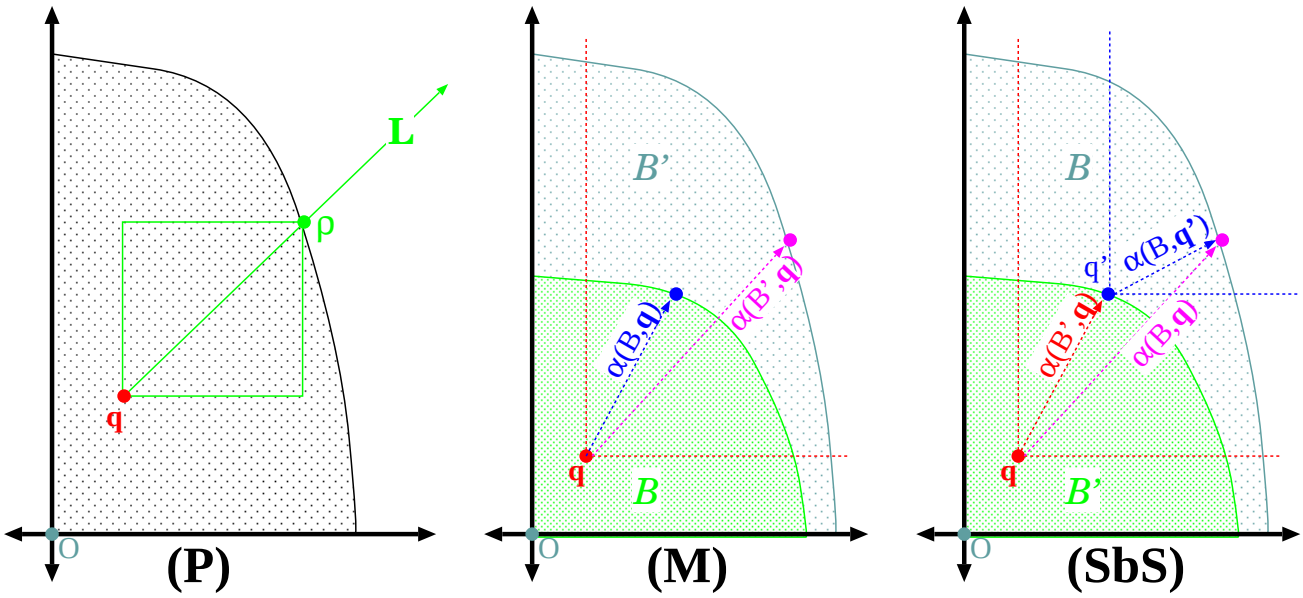


Figure 6.2: **(P)** The proportional bargaining solution. **(M)** The monotonicity axiom. **(SbS)** The 'step-by-step' axiom.

Note that the proportional solution is 'egalitarian' in an opposite sense to the utilitarian solution. The proportional solution explicitly does *not* favour the player who stands to gain the most from the bargaining problem, since it awards both the same amount. However, in some sense, the proportional solution perpetuates the status quo, as the next example shows:

Example 6B.1: Suppose Zara has \$16 and Owen has \$25, and they are bargaining over how to divide an additional \$20 [as in Example 6A.1(b)]. For simplicity, assume both players gain a utility of $U(x) = \sqrt{x}$ from possessing x dollars. Hence Owen's current utility is $U(25) = \sqrt{25} = 5$, whereas Zara's current utility is $U(16) = 4$. The proportional solution will then award \$9 to Zara (bringing her wealth up to \$25 and her utility up to $U(25) = 5$), whereas it will award \$11 to Owen (bringing his wealth up to \$36, and his utility up to $U(36) = 6$). In this way, both players gain exactly one unit of utility, and thus, this is the proportional bargaining solution —i.e. the unique solution which gives them equal gains of utility. However, notice that the richer player (Owen) actually got *more* of the money. (Whether or not this is 'fair' depends on your politics).

Again we used the function $U(x) = \sqrt{x}$ only to make the computations easy; a similar phenomenon occurs with any con utility function. ♢

Exercise 6.11: Suppose Zara and Owen are bargaining over how to divide some quantity M of money (see Example 4A.2), and that they begin with status quo endowment (q_0, q_1). Let $\mathbf{c} = (1, 1)$.

(a) Suppose Zara and Owen have identical, strictly concave utility functions $U_0 = U_1 = U$. Assume $q_0 \leq q_1$ (i.e. Zara is poorer). Show that the proportional solution $\rho_{\mathbf{c}}$ gives *less* money to Zara than it gives to Owen.

(b) Say that Owen is *more acquisitive* if $U_1'(x) \geq U_0'(x)$ for all $x \in \mathbb{R}_{\not}$. (Heuristically, this means that Owen can get 'more marginal happiness' out of an additional dollar than Zara can.) Suppose they begin *the same* initial wealth; i.e. $q_0 = q_1$. Show that the proportional solution $\rho_{\mathbf{c}}$ gives *less* money to Owen than it gives to Zara.

(Where would you locate the proportional solution on the "left vs. right" political spectrum? Compare this to the outcome of Exercise 6.9)

Ehud Kalai has shown that, like the Nash solution (§4B) and the utilitarian solution (§6A), the proportional solution can be characterized as the bargaining solution satisfying certain axioms of 'rationality', 'consistency' and/or 'fairness' [Kal77]. To see this, let $\mathfrak{B}$ be the set of all *bargaining problems* —i.e. ordered pairs $(\mathcal{B}, \mathbf{q})$, where $\mathcal{B} \subseteq \mathbb{R}^2_{\not}$ is a convex, compact, comprehensive subset, and $\mathbf{q} \in \mathcal{B}$ is some 'status quo' point. Recall that a *bargaining solution* is a function $\alpha : \mathfrak{B} \longrightarrow \mathbb{R}^2_{\not}$ satisfying the axioms **(MB)** and **(P)** —in other words, $\alpha(\mathcal{B}, \mathbf{q})$ is always a point on the negotiating set $\wp_{\mathbf{q}}\mathcal{B}$. We might also stipulate the following assumptions:

(H) (Homogeneity) *Let $(\mathcal{B}, \mathbf{q})$ be a bargaining problem, and let $r > 0$. Let $r\mathcal{B} := \{r\mathbf{b} \; ; \; \mathbf{b} \in \mathcal{B}\}$. If $\alpha(\mathcal{B}, \mathbf{q}) = \mathbf{b}$, then $\alpha(r\mathcal{B}, r\mathbf{q}) = r\mathbf{b}$.*

(T) (Translation invariance)[1] *Let $(\mathcal{B}, \mathbf{q})$ be a bargaining problem. Let $\mathcal{B}_0 := \{\mathbf{b} - \mathbf{q} \; ; \; \mathbf{b} \in \mathcal{B}\}$. If $\alpha(\mathcal{B}, \mathbf{q}) = \mathbf{b}$, then $\alpha(\mathcal{B}_0, \mathbf{0}) = \mathbf{b} - \mathbf{q}$.*

Observe that **(H)** is a weakened form of Nash's *Rescaling Invariance* axiom **(RI)** from §4B. Axiom **(RI)** allowed us to rescale the two player's utilties using *different* linear transformations, whereas **(H)** requires us to apply the *same* transformation to each axis. (This makes sense because we are assuming some standard for interpersonal comparison of utility, and this would be meaningless if we could independently rescale the axes). Axiom **(T)** allows us to move the status quo point to $\mathbf{0}$; again this is a special case of axiom **(RI)**.

(M) (Monotonicity) *Suppose $\mathbf{q} \in \mathcal{B} \subset \mathcal{B}' \subset \mathbb{R}^2_{\not}$. Then $\alpha(\mathcal{B}, \mathbf{q}) \overset{\wp}{\preceq} \alpha(\mathcal{B}', \mathbf{q})$.* [Figure 6.2(M)].

Normatively speaking, axiom **(M)** says, "If the set $\mathcal{B}$ of feasible bargains expands to a larger set $\mathcal{B}'$, then each player should do *at least* as well when bargaining over $\mathcal{B}'$ as he did when bargaining over $\mathcal{B}$." Descriptively speaking, we might interpret axiom **(M)** as follows: suppose the expansion of $\mathcal{B}$ to $\mathcal{B}'$ is not possible without the consent/cooperation of both players (for example, $\mathcal{B}'$ might be result of some collaboration between them). If either player expects to

[1] Kalai does not explicitly include the axiom **(T)** in his original formulation. However, the fact that he *defines* bargaining problems to always have the status quo at $\mathbf{0}$ means that he is implicitly assuming something like **(T)**.

get *less* in $\mathcal{B}'$ than he did in $\mathcal{B}$, then he will simply withhold his cooperation, and $\mathcal{B}'$ will never happen. Conversely, if $\mathcal{B}'$ *does* happen, it is only because both players expect to do at least as well in $\mathcal{B}'$ as in $\mathcal{B}$.

Kalai's other axiom is based on a rough model of a standard real-life negotiation strategy of breaking a big negotiation problem into separate sub-problems, and resolving these independently. Such 'step-by-step' negotiation reduces the cognitive complexity of a complex, multifaceted bargaining problem to manageable levels, and can also diffuse emotionally volatile situations by isolating contentious issues. Given a bargaining problem $(\mathcal{B}, \mathbf{q})$, we can represent a two-stage negotation as follows: At the first stage, we isolate some subset $\mathcal{B}' \subset \mathcal{B}$ (describing the range of agreements which can be reached at the first stage). The players then reach an agreement $\mathbf{q}'$ to the bargaining problem $(\mathcal{B}', \mathbf{q})$. At the second stage, we then resolve the remaining issues by solving the problem $(\mathcal{B}, \mathbf{q}')$; see Figure 6.2(SbS).

However, 'step by step negotiation' cannot happen without the consent of both players, and Owen will not agree if he thinks he will lose more in step-by-step bargaining than in a one-shot bargaining session. Conversely, Zara will *insist* on step-by-step bargaining if she thinks she will gain more. Thus, the possibility of step-by-step negotiation *itself* can create an impasse, unless we stipulate that it will have no effect on the utility of either player. Step-by-step negotation must merely be a device for dialogue, with no substantive impact on the bargaining outcome. This is the content of the next axiom:

(SbS) (Step-by-step negotiation) *Suppose* $\mathbf{q} \in \mathcal{B}' \subset \mathcal{B} \subset \mathbb{R}^2_{\not\prime}$. *Let* $\mathbf{q}' = \alpha(\mathcal{B}', \mathbf{q})$. *Then* $\alpha(\mathcal{B}, \mathbf{q}') = \alpha(\mathcal{B}, \mathbf{q})$. [Figure 6.2(SbS)].

Finally, we say that a bargaining solution is *proportional* if there exists some calibration $\mathbf{c} = (c_0, c_1)$ such that $\alpha = \rho_{\mathbf{c}}$ —i.e. for every bargaining problem $(\mathcal{B}, \mathbf{q}) \in \mathfrak{B}$, the point $\alpha(\mathcal{B}, \mathbf{q})$ is the $\mathbf{c}$-proportional solution $\rho_{\mathbf{c}}(\mathcal{B}, \mathbf{q})$.

Theorem 6B.2 (E. Kalai, 1977) *Let* $\alpha : \mathfrak{B} \longrightarrow \mathbb{R}^2_{\not\prime}$ *be a bargaining solution satisfying axioms* **(H)** *and* **(T)**. *The following are equivalent:*

(M) α *satisfies the monotonicity axiom.*

(SbS) α *satisfies step-by-step axiom.*

(P) α *is a proportional bargaining solution.*

Proof: "$(\mathbf{P}) \Longrightarrow (\mathbf{SbS})$" is **Exercise 6.12**. "$(\mathbf{SbS}) \Longrightarrow (\mathbf{M})$" is **Exercise 6.13**.

"$(\mathbf{M}) \Longrightarrow (\mathbf{P})$": By axiom **(T)**, we can assume that $\mathbf{q} = 0$. As shown in Figure 6.3(A), let $\Delta := \big\{(x_0, x_1) \in \mathbb{R}^2_{\not\prime} \; ; \; x_0 + x_1 \leq 1\big\}$, and let $\mathbf{p} := \alpha(\Delta, \mathbf{0})$. If $\mathbf{p} = (p_0, p_1)$, then let $\mathbf{c} := (p_0^{-1}, p_1^{-1})$; we will show that α must be the $\mathbf{c}$-proportional bargaining solution $\rho_{\mathbf{c}}$. In other words, for any bargaining problem $(\mathcal{B}, \mathbf{0})$, we must will show that $\alpha(\mathcal{B}, \mathbf{0}) = \overline{r}\mathbf{p}$, where $\overline{r} := \max\{r > 0 \; ; \; r\mathbf{p} \in \mathcal{B}\}$

As shown in Figure 6.3(A), let $\square_\epsilon \subset \Delta$ be the trapezoidal region with vertices at $(0, 0)$, $(p_0 + \epsilon, 0)$, $(0, p_1 + \epsilon)$ and $\mathbf{p}$, where $\epsilon > 0$ is very small. Thus, $\square_\epsilon$ is essentially a rectangle, but

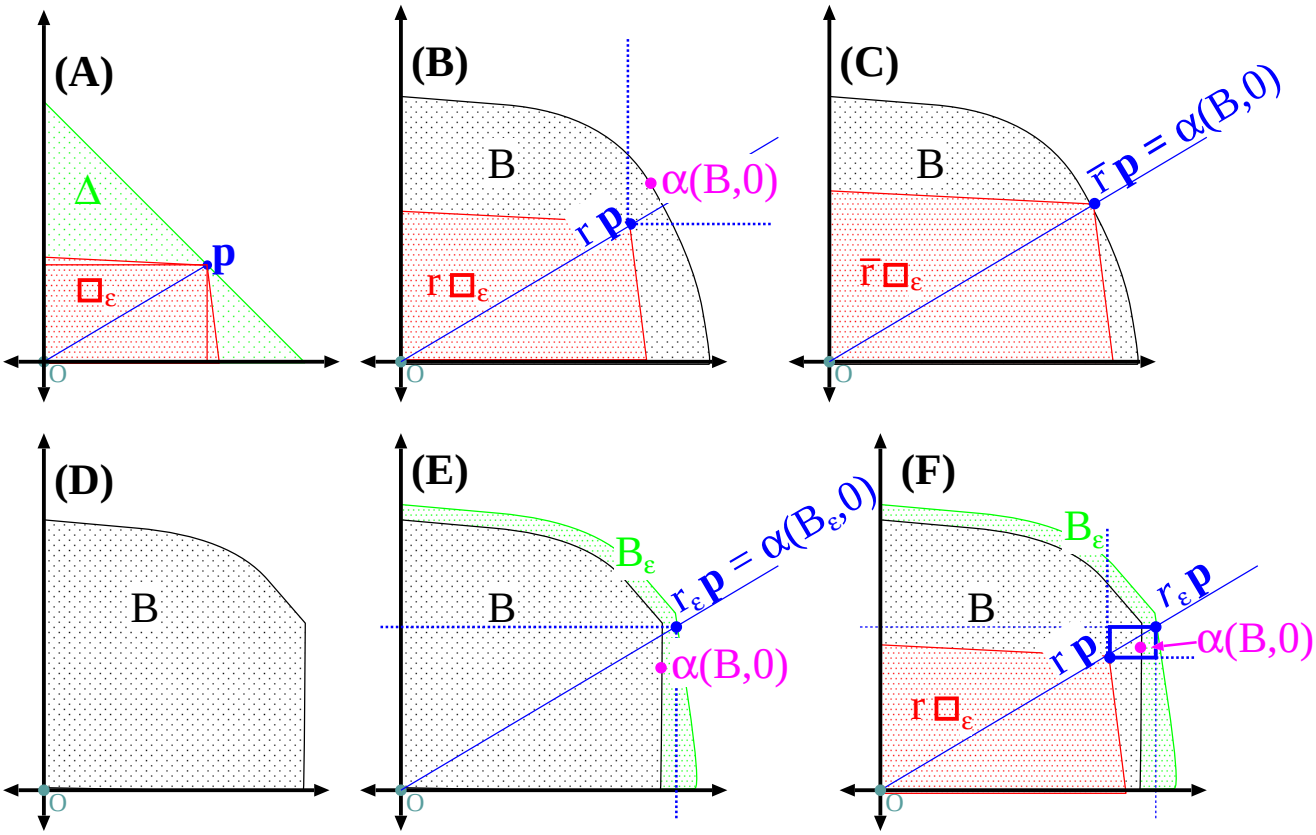


Figure 6.3: The proof of Theorem 6B.2.

we tilt the top and right sides very slightly, so that $\mathbf{p}$ is *not* Pareto-preferred to any point on these sides.

Claim 1: $\alpha(\square_\epsilon, \mathbf{0}) = \mathbf{p}$.

Proof: Let $\mathbf{b} := \alpha(\square_\epsilon, \mathbf{0})$. Then axiom **(P)** says $\mathbf{b}$ is on the Pareto frontier of $\square_\epsilon$, which is its top and right sides. But $\square_\epsilon \subset \Delta$, so axiom **(M)** says that $\mathbf{b} \overset{\wp}{\preceq} \mathbf{p}$. But by construction, the *only* point $\mathbf{b}$ on the Pareto frontier of $\square_\epsilon$ such that $\mathbf{b} \overset{\wp}{\preceq} \mathbf{p}$ is $\mathbf{b} = \mathbf{p}$ itself. ◇ Claim 1

For any $r > 0$, let $r\square_\epsilon := \{r\mathbf{b}\ ;\ \mathbf{b} \in \square_\epsilon\}$.

Claim 2: *Let* $r > 0$. *If* $r\square_\epsilon \subseteq \mathcal{B}$, *then* $r\mathbf{p} \overset{\wp}{\preceq} \alpha(\mathcal{B}, \mathbf{0})$. [Figure 6.3(B)]

Proof: $r\mathbf{p} \;=\; r\alpha(\square_\epsilon, \mathbf{0}) \underset{(*)}{=} \alpha(r\square_\epsilon, \mathbf{0}) \;\overset{\wp}{\preceq}\; \alpha(\mathcal{B}, \mathbf{0})$.

Here, $(*)$ is by axiom **(H)**, and " $\overset{\wp}{\preceq}$ " is by axiom **(M)**, because $r\square_\epsilon \subseteq \mathcal{B}$. ◇ Claim 2

Let $\overline{r} := \max\{r > 0 \; ; \; r\mathbf{p} \in \mathcal{B}\}$ (this maximum exists because $\mathcal{B}$ is compact). Thus, $\overline{r}\mathbf{p}$ is on the Pareto frontier of $\mathcal{B}$.

Claim 3: $\overline{r}\mathbf{p} \stackrel{\wp}{\preceq} \alpha(\mathcal{B}, \mathbf{0})$.

Proof: For any $r < \overline{r}$, we can find some $\epsilon > 0$ small enough that $r\square_\epsilon \subset \mathcal{B}$. Then Claim 2 says that $r\mathbf{p} \stackrel{\wp}{\preceq} \alpha(\mathcal{B}, \mathbf{0})$. Taking the limit as $r \nearrow \overline{r}$, we conclude that $\overline{r}\mathbf{p} \stackrel{\wp}{\preceq} \alpha(\mathcal{B}, \mathbf{0})$. $\diamond$ Claim 3

A point $\mathbf{b} \in \mathcal{B}$ is *strictly Pareto optimal* if $\mathbf{b}$ is Pareto optimal and furthermore, there is no other point $\mathbf{b}' \in \mathcal{B}$ (besides $\mathbf{b}$ itself) with $\mathbf{b} \stackrel{\wp}{\preceq} \mathbf{b}'$. Typically, every point on the Pareto frontier $\wp\mathcal{B}$ is strictly Pareto optimal. The only exceptions occur when $\wp\mathcal{B}$ contains a perfectly vertical (or horizontal) line segment; in this case, a point $\mathbf{b}$ near the bottom (or right end) of this line sigment will be Pareto optimal, but $\mathbf{b}$ will not be *strictly* Pareto optimal because $\mathbf{b}$ will be Pareto inferior to a point $\mathbf{b}'$ at the top (or left end) of the line segment. (Usually, feasible sets do not contain vertical or horizontal line segments, so usually, Pareto optimality is equivalent to strict Pareto optimality).

Claim 4: *If $\overline{r}\mathbf{p}$ is strictly Pareto-optimal in $\mathcal{B}$, then $\alpha(\mathcal{B}, \mathbf{0}) = \overline{r}\mathbf{p}$.* [Figure 6.3(C)]

Proof: Claim 3 says that $\overline{r}\mathbf{p} \stackrel{\wp}{\preceq} \alpha(\mathcal{B}, \mathbf{0})$, and of course $\alpha(\mathcal{B}, \mathbf{0}) \in \mathcal{B}$. But if $\overline{r}\mathbf{p}$ is strictly Pareto optimal, this means $\alpha(\mathcal{B}, \mathbf{0}) = \overline{r}\mathbf{p}$. $\diamond$ Claim 4

Thus, Claim 4 proves the result for almost all feasible sets, except for 'pathological' cases containing strictly vertical or horizontal edges on their Pareto frontiers. Suppose $\mathcal{B}$ was such a set, shown in Figure 6.3(D).

Claim 5: $\alpha(\mathcal{B}, \mathbf{0}) \stackrel{\wp}{\preceq} \overline{r}\mathbf{p}$.

Proof: Let $\epsilon > 0$, and let $\mathcal{B}_\epsilon \supset \mathcal{B}$ be a slight perturbation of $\mathcal{B}$ where we tilt the edges slightly outwards, as in Figure 6.3(E). Then Claim 4 implies that $\alpha(\mathcal{B}_\epsilon, \mathbf{0}) = r_\epsilon\mathbf{p}$, where $r_\epsilon := \max\{r > 0 \; ; \; r\mathbf{p} \in \mathcal{B}_\epsilon\}$. But $\mathcal{B} \subset \mathcal{B}_\epsilon$, so **(M)** says that $\alpha(\mathcal{B}, \mathbf{0}) \stackrel{\wp}{\preceq} r_\epsilon\mathbf{p}$. Now let $\epsilon \searrow 0$, so that $r_\epsilon \searrow \overline{r}$. $\diamond$ Claim 5

Thus, as suggested in Figure 6.3(F), Claims 3 and 5 together mean $\overline{r}\mathbf{p} \stackrel{\wp}{\preceq} \alpha(\mathcal{B}, \mathbf{0}) \stackrel{\wp}{\preceq} \overline{r}\mathbf{p}$, which forces $\alpha(\mathcal{B}, \mathbf{0}) = \overline{r}\mathbf{p}$.

Thus, for any bargaining set $\mathcal{B}$, we conclude that $\alpha(\mathcal{B}, \mathbf{0}) = \overline{r}\mathbf{p}$, which is the unique intersection point of the line $\mathbb{R}\mathbf{p}$ with the frontier of $\mathcal{B}$, which is $\rho_{\mathbf{c}}(\mathcal{B}, \mathbf{0})$ by definition. $\square$

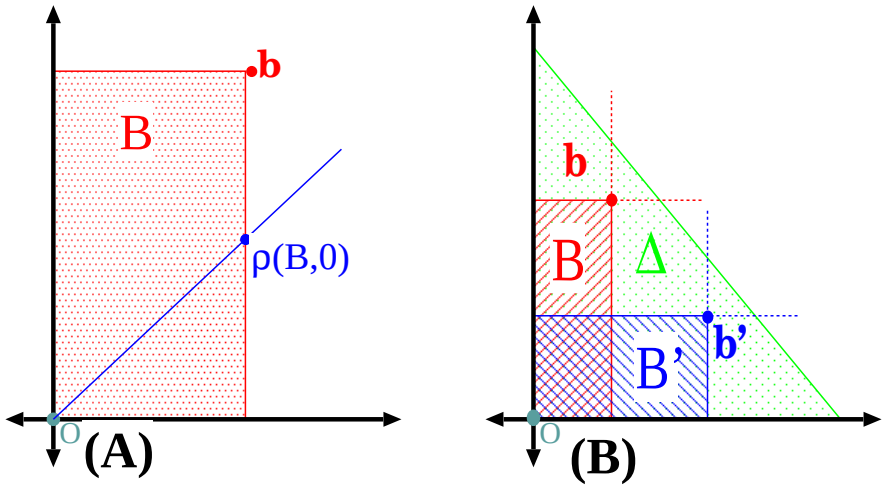


Figure 6.4: **(A)** A suboptimal outcome of the proportional bargaining solution.
(B) Why this is inevitable in any bargaining solution satisfying monotonicity axiom **(M)**.

Remarks: (a) Kalai has also proved that the proportional bargaining solution is the only solution satisfying axioms **(H)** and **(T)**, a strengthened form of **(M)**, Nash's axiom **(IIA)**, and the Hausdorff continuity axiom **(HC)** of §4C; see [Kal77, Theorem 3].

(b) Myerson [Mye77] weakened Kalai's conditions even further by considering bargaining problems with *ordinal* utility (i.e. preference orderings). He showed that a bargaining solution satisfies **(P)**, **(MB)**, **(SbS)**, and an extremely weak homogeneity axiom if and only if it is a proportional bargaining system with respect to some cardinal utility representation of the players' preference orderings.

Exercise 6.14: (a) Generalize the definition of the proportional bargaining solution to bargains with $N \geq 3$ players.

(b) Generalize the axioms **(H)**, **(T)**, **(M)**, and **(SbS)** to bargains with $N \geq 3$ players.

(c) Generalize the statement and proof of Theorem 6B.2 to bargains with $N \geq 3$ players.

Exercise 6.15: Consider the medical *triage* scenario from Exercise 6.9. How would the 'proportional bargaining solution' allocate medical resources amongst the patients in Categories 1, 2, and 3 during a natural disaster?

Exercise 6.16: The *Monotonicity* Axiom **(M)** seems quite similar to the *Individual Monotonicity* Axiom **(IM)** satisfied by the Kalai-Smorodinsky bargaining solution of §7A. However, **(M)** is actually a more restrictive axiom. To see this, construct an example where the Kalai-Smorodinsky solution violates axiom **(M)**.

The Lexmin solution: The 'exceptional' case which required Claim 5 of Theorem 6B.2 suggests a slightly pathological outcome of the proportional bargaining solution. Suppose the bargaining set $\mathcal{B}$ is a rectangle, as in Figure 6.4(A). Clearly, the 'best' bargaining outcome is the corner $\mathbf{b}$, as this simultaneously maximizes utility for both players. However, in general the proportional solution $\rho(\mathcal{B}, \mathbf{0})$ will be a point somewhere along one of the sides of the rectangle, which needlessly robs one of the players of some of his possible utility.

Because of this, one proposed modification to the proportional solution is the *lexmin* solution, defined as follows:

- First, compute the proportional solution $\mathbf{p}$. If $\mathbf{p}$ is strictly Pareto optimal, then stop.
- If $\mathbf{p}$ is not strictly Pareto optimal, then find the (unique) strictly Pareto optimal point $\mathbf{p}'$ which is Pareto-prefered to $\mathbf{p}$. Choose $\mathbf{p}'$ as the bargaining solution.

For example, in Figure 6.4(A), the lexmin solution would select the point $\mathbf{b}$.

However, that the lexmin solution violates axiom **(M)**. Indeed there is *no* bargaining solution which can satisfy axiom **(M)** and which will also choose $\mathbf{b}_0$ as the bargaining outcome in Figure 6.4(A). To see this, consider the two rectangles $\mathcal{B}$ and $\mathcal{B}'$ in Figure 6.4(B). Suppose α was a bargaining solution such that $\alpha(\mathcal{B}, \mathbf{0}) = \mathbf{b}$ and $\alpha(\mathcal{B}, \mathbf{0}) = \mathbf{b}'$, as seems obviously fair. Note that $\mathcal{B} \subset \Delta$ and $\mathcal{B}' \subset \Delta$. Hence, axiom **(M)** says that $\alpha(\Delta, \mathbf{0})$ must be a point which is Pareto-prefered to both $\mathbf{b}$ and $\mathbf{b}'$. But there is *no such point* on the boundary of Δ. Contradiction. This example is adapted from Luce and Raiffa [LR80, §6.6, p.134].

The Egalitarian or Maximin Solution: The egalitarian bargaining solution is quite similar to the proportional solution, in that both players are intended to gain the 'same' amount of utility from the bargain. Now, however, these gains are measured relative to some 'absolute zero' point on their utility scales, *not* relative to the status quo point $\mathbf{q}$. To be precise, we fix some calibration constants $\mathbf{c} = (c_0, c_1) \in \mathbb{R}^2_{\nearrow}$, We then define $\epsilon_{\mathbf{c}}(\mathcal{B}, \mathbf{q})$ to be the unique point (b_0, b_1) on the Pareto frontier of $\mathcal{B}$ such that $c_0 b_0 = c_1 b_1$. In extreme situations, this point may actually be below or to the left of the status quo $\mathbf{q}$, in which case it is inadmissible as a bargain. In this case, we simply choose the extreme point of the negotiating set $\wp_{\mathbf{q}}\mathcal{B}$ which is closest to the egalitarian point.

The egalitarian solution is often called the *maximin solution*, because it maximizes the *minimum* of the utilities for all players involved in the bargain (which usually means equalizing all the players' utilities). The egalitarian solution combines the egalitarian properties of the proportional and utilitarian solutions. Like the proportional solution, the egalitarian solution explicitly does *not* favour the player who stands to gain the most from the bargaining problem, since it awards both the same amount. However, like the utilitarian solution, the egalitarian solution usually obliterates the status quo.

Example 6B.3: Suppose Zara has \$16 and Owen has \$25, and they are bargaining over how to divide an additional \$$X$ (as in Example 6A.1 and Example 6B.1). Assume both players have the same (concave) utility function for money. If $X \leq 9$, then the entire amount will

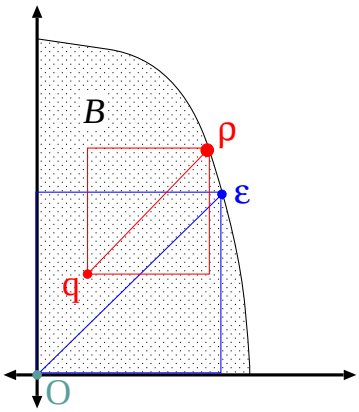


Figure 6.5: The egalitarian bargaining solution ϵ versus the proportional solution ρ.

be awarded to Zara. However, if $X > 9$, then first \$9 will be given to Zara, and then the remaining money will be split equally between Zara and Owen, so they both end up with the same total. Note that, in this situation, the outcome is identical to the utilitarian bargain of Example 6A.1(b). This is because we assumed that utility is concave in money (a standard assumption of risk-aversion), so that the 'poorer' person always enjoys money more. Also, we assumed that both players enjoy money identically. ♢

However, whereas the utilitarian solution favours the more 'efficient' or 'productive' player, a egalitarian solution favours the *less* efficient player:

Example 6B.4: Suppose that Zara and Owen are bargaining over how to divide one acre of farmland, as in Example 6A.2. Again, Zara can produce \$4 of crop with this acre, whereas Owen can only produce \$1. Assume both begin with zero dollars, and assume both have the identical utility function for money (it doesn't matter what this function is). Then the egalitarian solution will award $\frac{1}{5}$ of the land to Zara, and $\frac{4}{5}$ of the land to Owen, so that each of them produces \$0.80 worth of crop. ♢

Remark: (a) In the literature, the egalitarian and proportional solutions are often treated as the same. I distinguish between them because the egalitarian solution compares possible bargains to the absolute zero point (which is maybe 'fairer'), whereas proportional solution compares possible bargains to the existing status quo (which is arguably more realistic).

(b) The egalitarian solution postulates the existence of an 'absolute zero' on the player's utility scales, which raises interesting philosophical questions. In some contexts, there is a natural definition of absolute zero. For example, in a strictly financial bargaining problem (say, between two large companies, or between a labour union and the management of a firm), if we assume that utility is a linear function of monetary net worth, then 'absolute zero' corresponds to zero net worth (i.e. assets are exactly balanced by liabilities). In terms of the utility of a person's life-plan, absolute zero might correspond to death. (Note that, in both examples, it is possible to have utilities much less than zero.)

Exercise 6.17: Generalize the egalitarian solution to bargains with $N \geq 3$ players.

Exercise 6.18: The proportional and egalitarian solutions both endeavour to 'equalize' the utilities for all players. The difference is the reference point: the proportional solution measures utilities relative to the status quo point $\mathbf{q}$, whereas in the egalitarian solution measures them relative to some 'absolute zero'; this makes a difference in our definition of 'equality'.

Now consider the utilitarian solution of §6A. Does it make a difference what we use as our 'reference' utility? Why or why not?

6C Solution Syzygy

Prerequisites: §4B, §6A, §6B

At this point we have three different bargaining solutions (shown in Figure 6.6), each with some plausible justification. We now have two problems:

1. In general, these three solutions will be distinct, and it isn't clear which is the 'right' one to use.

2. Two of the solutions (namely ρ and μ) depend upon an arbitrary calibration constant $\mathbf{c}$ which defines our interpersonal comparison of utility. It isn't clear what the 'right' value of $\mathbf{c}$ is.

The good news is that, to some extent, problem #2 can be used to obviate problem #1. Since the constant $\mathbf{c}$ is arbitrary, we can manipulate $\mathbf{c}$ so that ρ and μ will align with each other and with η. A value of $\mathbf{c}$ which achieves such an alignment is perhaps the 'right' value to use.

Theorem 6C.1 (Yaari, 1981) *Let $\mathbf{q} \subset \mathcal{B} \subset \mathbb{R}^2_{\nearrow}$, and suppose $\mathcal{B}$ is strictly convex —i.e. the Pareto frontier of $\mathcal{B}$ contains no straight line segments. Let $\mathbf{n} := \eta(\mathcal{B}, \mathbf{q})$ be the Nash solution to the bargaining problem $(\mathcal{B}, \mathbf{q})$. Fix calibration a constant $\mathbf{c} \in \mathbb{R}^2_{\nearrow}$, and let $\mathbf{u} := \mu_{\mathbf{c}}(\mathcal{B}, \mathbf{q})$ and $\mathbf{p} := \rho_{\mathbf{c}}(\mathcal{B}, \mathbf{q})$ be the $\mathbf{c}$-utilitarian and $\mathbf{c}$-proportional solutions. Then:*

(a) $\Big(\mathbf{p} = \mathbf{u}\Big) \iff \Big(\mathbf{p} = \mathbf{n}\Big) \iff \Big(\mathbf{u} = \mathbf{n}\Big)$.

(b) *There is a unique value of $\mathbf{c} \in \mathbb{R}^2_{\nearrow}$ satisfying the equations of part* **(a)**, *and with $c_0 + c_1 = 1$.*

Proof: For any fixed calibration constant $\mathbf{c} \in \mathbb{R}^2_{\nearrow}$, the proportional bargaining solution $\rho_{\mathbf{c}}$ satisfies the translation invariance axiom **(T)**, by Proposition 6B.2. The Nash solution η also satisfies **(T)** [indeed it satisfies the much stronger rescaling axiom **(RI)**]. Finally, the utilitarian solution $\mu_{\mathbf{c}}$ also satisfies **(T)** (**Exercise 6.19** Verify this sentence). Thus, we can assume without loss of generality that we have translated $\mathcal{B}$ so that $\mathbf{q} = 0$.

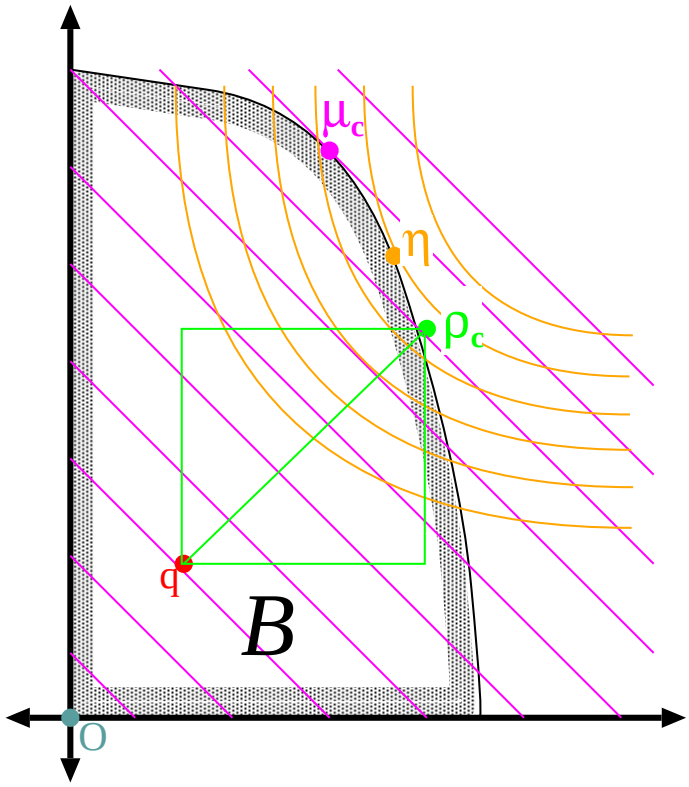


Figure 6.6: A comparison of the classic utilitarian solution ($\mu_{\mathbf{c}}$), Nash solution (η) and proportional solution ($\rho_{\mathbf{c}}$) to the bargaining problem. Note that the exact positions of all three solutions depend on the shape of the bargaining set $\mathcal{B}$ and the location of the status quo $\mathbf{q}$. Also, the positions of $\mu_{\mathbf{c}}$ and $\rho_{\mathbf{c}}$ depend on the calibration constant $\mathbf{c}$. Hence the three points can have almost any geometric relationship; the picture here is just one possibility.

If $F : \mathbb{R}^2 \longrightarrow \mathbb{R}$ is a smooth function, recall that the *gradient vector field* of F is the (vector-valued) function $\nabla F : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ defined by $\nabla F(\mathbf{x}) = (\partial_1 F(\mathbf{x}), \partial_2 F(\mathbf{x}))$ for all $\mathbf{x} \in \mathbb{R}^2$ (here, $\partial_k F := \frac{\partial F}{\partial x_k}$). If $\mathcal{B}$ is is some compact domain with smooth boundary $\partial\mathcal{B}$, and $\mathbf{b} \in \partial\mathcal{B}$, then $\mathbf{b}$ is a maximum or minimum point of F if and only if $\nabla F(\mathbf{b})$ is orthogonal to $\partial\mathcal{B}$ at $\mathbf{b}$.

We review the following facts about the three bargaining solutions:

Claim 1: *Let $\wp\mathcal{B}$ be the Pareto frontier of $\mathcal{B}$.*

(a) $\mathbf{p} = (p_0, p_1)$ *is the unique point on* $\wp\mathcal{B}$ *such that* $c_0 p_0 = c_1 p_1$.

(b) *There is some* $\overline{r} > 0$ *such that* $\mathbf{p} = \overline{r}(c_1, c_0)$ [note the coordinate reversal].

(c) *Let* $U_{\mathbf{c}}(x_0, x_1) = c_0 x_0 + c_1 x_1$ *be the* $\mathbf{c}$*-weighted utilitarian sum. Then* $\nabla U_{\mathbf{c}}(x_0, x_1) = (c_0, c_1)$ *for all* $(x_0, x_1) \in \mathbb{R}^2$.

(d) *Let* $N(x_0, x_1) = x_0 x_1$ *be the Nash product. Then* $\nabla N(x_0, x_1) = (x_1, x_0)$ *for all* $(x_0, x_1) \in \mathbb{R}^2$ [again, note the coordinate reversal].

(e) $\mathbf{u}$ *is the unique point where* $\wp\mathcal{B}$ *is orthogonal to* (c_0, c_1).

(f) $\mathbf{n} = (n_0, n_1)$ *is the unique point where* $\wp\mathcal{B}$ *is orthogonal to* $\nabla N(n_0, n_1) = (n_1, n_0)$.

Proof: (a) and (b) are true by definition of the proportional solution (see the first paragraph of §6B). (c) and (d) are elementary computations. (e) and (f) follow from (c) and (d), because $\mathbf{u}$ and $\mathbf{n}$ are the maximizers of $U_{\mathbf{c}}$ and N, respectively. In both (e) and (f), the *uniqueness* of the orthogonal point follows from the assumption that $\mathcal{B}$ is strictly convex. ◇ Claim 1

Claim 2: $\Big(\mathbf{n} = \mathbf{u}\Big) \iff \Big(\mathbf{n} = \mathbf{p}\Big)$.

Proof: $\Big(\mathbf{n} = \mathbf{u}\Big) \underset{(*)}{\Longleftrightarrow} \Big(\wp\mathcal{B}$ is orthogonal to both (c_0, c_1) and (n_1, n_0) at the point $\mathbf{n}\Big) \iff \Big($The vectors (c_0, c_1) and (n_1, n_0) are parallel$\Big) \iff \Big((c_0, c_1) = k\,(n_1, n_0)$, for some $k \in \mathbb{R}\Big)$ $\iff \Big(c_0/n_1 = k = c_1/n_0\Big) \iff \Big(c_0 n_0 = c_1 n_1\Big) \underset{(\dagger)}{\Longleftrightarrow} \Big(\mathbf{n} = \mathbf{p}\Big)$. Here, $(*)$ is by Claim 1(e,f), while $(\dagger)$ is by Claim 1(a). ◇ Claim 2

Claim 3: $\Big(\mathbf{p} = \mathbf{u}\Big) \Longrightarrow \Big(\mathbf{n} = \mathbf{u}\Big)$.

Proof: Suppose $\mathbf{p} = \mathbf{u}$. Then

$$\nabla N(\mathbf{u}) \underset{(*)}{=} (u_1, u_0) \underset{(\dagger)}{=} (p_1, p_0) \underset{(\diamond)}{=} \overline{r}(c_0, c_1) \underset{(\ddagger)}{=} \overline{r}\nabla U_{\mathbf{c}}(\mathbf{u}).$$

Here, $(*)$ is by Claim 1(d); $(\dagger)$ is because $\mathbf{p} = \mathbf{u}$ by hypothesis; $(\diamond)$ is by Claim 1(b) [note the double coordinate reversal]; and $(\ddagger)$ is by Claim 1(c).

But $\nabla U_{\mathbf{c}}(\mathbf{u})$ is orthogonal to $\wp\mathcal{B}$ at $\mathbf{u}$ [by Claim 1(e)]. Thus, $\nabla N(\mathbf{u})$ is also orthogonal to $\wp\mathcal{B}$ at $\mathbf{u}$. But that means $\mathbf{u} = \mathbf{n}$, by Claim 1(f). ◇ Claim 3

Claim 4: $\Big(\mathbf{u} = \mathbf{n}\Big) \Longrightarrow \Big(\mathbf{u} = \mathbf{p}\Big)$.

Proof: If $\mathbf{u} = \mathbf{n}$, then $\mathbf{n} = \mathbf{p}$ by Claim 2. But then $\mathbf{u} = \mathbf{n} = \mathbf{p}$. ◇ Claim 4

Claims 2, 3, and 4 complete the circle of implications and establish part **(a)**.

Part **(b)** is **Exercise 6.20**. □

The unique calibration constant $\mathbf{c}$ in Theorem 6C.1(b) which causes $\rho_{\mathbf{c}}(\mathcal{B}, \mathbf{q}) = \mu_{\mathbf{c}}(\mathcal{B}, \mathbf{q}) = \eta(\mathcal{B}, \mathbf{q})$ will be called the *Yaari calibration* for $(\mathcal{B}, \mathbf{q})$. Note that different bargaining problems will generally have *different* Yaari calibrations.

Exercise 6.21: Generalize the statement and proof of Theorem 6C.1 to bargains involving $N \geq 3$ players.]

Exercise 6.22: For any exponent $p \in (-\infty, 1]$ with $p \neq 0$, define $F_p : \mathbb{R}^2_{\not\nearrow} \longrightarrow \mathbb{R}_{\not\nearrow}$ by

$$F_p(u_0, u_1) \quad := \quad (u_0^p + u_1^p)^{1/p}.$$

Let $\alpha_p : \mathfrak{B} \longrightarrow \mathbb{R}^2_{\not\nearrow}$ be the bargaining solution such that $\alpha_p(\mathcal{B}, \mathbf{q})$ is the point (b_0, b_1) maximizing the value of $F_p(b_0 - q_0, b_1 - q_1)$ in $\wp_{\mathbf{q}}\mathcal{B}$.

For example, if $p = 1$, then $F_1(u_0, u_1) = u_0 + u_1$ is just the standard utilitarian sum; thus, α_1 is just the utilitarian solution μ_1 with calibration constant $(1, 1)$.

(a) Show that α_p is a bargaining solution for any $p \in (-\infty, 1]$ with $p \neq 0$. Show that α_p satisfies axioms **(S)** [*Symmetry*], **(T)** [*Translation invariance*], **(H)** [*Homogeneity*], and **(IIA)** [*Independence of Irrelevant Alternatives*], but not axiom **(RI)** [*Rescaling Invariance*] (see the beginnings of §4B and §6B for the definitions of these axioms).

(b) Show that $\lim_{p \to 0} \alpha_p(\mathcal{B}, \mathbf{q}) = \eta(\mathcal{B}, \mathbf{q})$, the Nash solution. [*Hint:* maximizing the Nash product $(b_0 - q_0) \cdot (b_1 - q_1)$ is equivalent to maximizing $\log(b_0 - q_0) + \log(b_1 - q_1)$.]

(c) Show that $\lim_{p \to -\infty} \alpha_p(\mathcal{B}, \mathbf{q}) = \rho(\mathcal{B}, \mathbf{q})$, the proportional solution. [*Hint:* $F_{-p}(b_0 - q_0, b_1 - q_1) = \left(\left\| \left(\frac{1}{b_0 - q_0}, \frac{1}{b_1 - q_1} \right) \right\|_p \right)^{-1}$, where $\|\bullet\|_p$ denotes the ℓ^p norm on $\mathbb{R}^2$. So maximizing $F_{-p}(b_0 - q_0, b_1 - q_1)$ is equivalent to minimizing $\left\| \frac{1}{b_0 - q_0}, \frac{1}{b_1 - q_1} \right\|_p$. But $\lim_{p \to \infty} \|\mathbf{x}\|_p = \|\mathbf{x}\|_\infty$ for any $\mathbf{x} \in \mathbb{R}^2$.]

Thus, the family of bargaining solutions $\{\alpha_p \, ; \, 0 < p < 1\}$ provides a continuous 'interpolation' between the Nash solution and the utilitarian solution. Likewise, the family of bargaining solutions $\{\alpha_p \, ; \, -\infty < p < 0\}$ provides a continuous 'interpolation' between the proportional solution and the Nash solution. The Nash solution itself can be seen as a sort of 'compromise' between total utilitarianism at one extreme, and total egalitarianism at the other.

6D Contractarian Political Philosophy

Prerequisites: §4B, §7A, §6A, §6B, §6.11 **Recommended:** 6C

Contractarian political philosophy originated in the works of Thomas Hobbes, John Locke, and Jean-Jacques Rousseau, and had two (slightly different) goals:

1. To explain the historical emergence of complex societies and political structures. In this form, contractarian thought postulates an anarchic 'state of nature' which existed in the primordial past. In this state of nature, it is conjectured that primitive humans negotiated amongst themselves, and, as free and rational individuals, agreed to create and support a particular political regime for their own mutual benefit.

2. To promote a particular (possibly utopian) social or political order. In this form, contractarian thought argues that, right now, free and rational individuals *would* agree to

> create and support a particular political regime for their own mutual benefit, *if* they had the chance to freely negotiate amongst themselves (which they perhaps do not, because they are constrained by an oppressive status quo).

One must be careful the distinguish between these two claims (and some philosophers do not). The fact that a particular social contract *was* agreed to at some point in the remote past (Goal #1) does not necessarily mean that this contract *would* be agreed to now (Goal #2), and vice versa.

Indeed, goal #1 is essentially a kind of speculative paleoanthropology, and has been pretty thoroughly discredited. It is now recognized that primitive societies (and even modern societies) are really better thought of as *organisms* which evolved by natural selection, rather than as *artifacts* that were consciously created by primordial rational beings in some primordial negotiation. In other words, societies exhibit particular kinds of social and political structures because these structures help the society *survive*; societies with different structures did not survive, and so we don't see them. No one consciously designed or chose or negotiated these sociopolitical structures. Instead, sociopolitical structures emerged in the kin-groups of early human hunter-gatherers (or perhaps even earlier, in our primate ancestors) through a process of random 'mutation' and selection (aided perhaps by the occassional innovations of intelligent humans), and this same process causes these structures to become more sophisticated over time[2]. This recognition has led political scientists like Robert Axelrod [Axe85, Axe97] and philosophers like Brian Skyrms [Sky96, Sky03] to apply *evolutionary game theory* to explain the emergence and persistence of social structures. See also [Rid98].

Contractarian political philosophy still pursues goal #2. In its modern guise, contractarian thought proposes to discover the nature of justice and the ideal political regime through a kind of thought experiment, where we imagine the sort of society which people *would* choose if given the choice. The resulting 'social contract' is presumably obtained through some kind of negotiation which reconciles the competing interests or conflicting moral values of different participants. Thus, bargaining theory is relevant to contractarian political philosophy in at least two ways:

- To predict the social contract that 'would' be reached by freely negotiating rational agents (say, using something like the Rubinstein *Alternating Offers* model of §5F).

- To characterize the social contract that 'should' be reached, according to certain criteria of 'fairness' (such as the axioms used to characterize the Nash, utilitarian, or proportional solutions).

[2]Note that we are speaking here of *social* evolution, not *biological* evolution. The information structures which evolve here are not *genes* encoded in DNA, but are instead *social conventions* like cultural norms, mythologies, ideologies, and habits of thought and behaviour, encoded (perhaps unconsciously) in the minds of the participants. David Lewis has proposed that many social conventions can themselves be seen as Nash equilibria in some 'social interaction game', such that no participant can profit by unilaterally defecting, so long as everyone else conforms to the convention [Lew69].

Harsanyi: For example, in 1955, John Harsanyi proposed that we address social justice issues from the perspective of an 'Ideal Spectator', a hypothetical being who has an equal probability of becoming any person in society [Har55a]. In keeping with the von Neumann-Morgenstern interpretation of utility, the Ideal Spectator will therefore attempt to maximize her expected utility, which will simply be the average utility of all people in the society (since she has an equal probability of becoming any particular person). Thus, the outcome of any social contract negotiation should be the *Utilitarian* bargaining solution of §6A.

Harsanyi claims that interpersonal utility comparison is possible because there are already certain more or less accepted conventions whereby we compare utility between individuals. Indeed, he argues that we make such interpersonal comparisons all the time. He gives the following example: suppose you have bought an expensive theatre ticket, but then discover you cannot attend the performance. Assume that you cannot or will not resell the ticket; then your best choice is to give it to a friend as a gift. You must decide which of your friends would most enjoy the gift of a free ticket to the show. In doing this, you are implicitly making comparisons between the utility functions of your friends.

Rawls: The next major contribution to contractarian political philosophy was John Rawls' *A Theory of Justice* [Raw71]. Rawls also assumed the possibility of interpersonal comparison of utility; he justifies this by claiming that personal utilities can be compared via a set of more or less quantifiable 'primary goods', including physical wealth, 'the social basis of self-respect', and 'the powers and prerogatives of office'.

Rawls rejects the 'Ideal Spectator' approach as impractical. The Ideal Spectator is presumed to have perfect and intimate knowledge of the values, desires, and circumstances of every person, and to able to synthesize this vast quantity of information into some calculation of 'average utility'. But such an omniscient and impartial Spectator does not exist, and we (ignorant and biased as we are) couldn't imagine what the Ideal Spectator would think even if she did exist. Instead of imagining the thought process of an omniscient being, Rawls proposes the opposite thought experiment: we must imagine that people are forced to choose their ideal society from behind a 'veil of ignorance', where they don't know what their own role in that society will be.

Rawls then argued that rational people located behind the veil of ignorance (a place he calls the *original position*) would be highly risk-averse, and would rationally chose the sociopolitical order which best treated the least fortunate member of society. In other words, it would maximise the minimal utility amongst all bargainers; hence it would be the *lexmin* solution described on page 144, which is a slight enhancement of the *proportional* solution of §6B. Rawls describes this as the *difference principle*: inequalities should not be tolerated unless removing them would actually *worsen* the situation of those worst off in society.

For Rawls, the 'original position' is not a place of negotiation, because, behind the 'veil of ignorance', all members of a particular community would have the same moral values, and thus, stripped of their own identities, they would all choose the same social contract. Rawls proposes the 'original position' as a thought experiment to test of the logical consistency of a particular conception of justice: if the original position, applied to our conception of justice,

produces outcomes which are contrary to this conception of justice, then we must revise our concept until these contradictions are eliminated —in Rawl's words, until we obtain a *reflective equilibrium*. He accepts that people in different communities, with different fundamental moral values, may converge upon different reflective equilibria.

This approach seems to limit Rawl's procedure to 'morally homogeneous' communities, whereas modern multicultural societies in the age of globalization clearly are not homogeneous. However, we could perhaps apply the Rawlsian method to such a society by assuming that, behind the veil of ignorance, we also forget our own moral values; hence we must choose a social contract which maximizes the minimum utility of any person with any value system likely to exist in that society.

Gauthier: Rawls' *Theory of Justice* is considered one of the seminal works of political philosophy in the twentieth century. However, it has also been criticised, both for Rawls' (arguably naïve) approach to interpersonal utility comparison, and for his insistence on the egalitarian bargaining solution. In *Morals by Agreement* [Gau86] David Gauthier develops a contractarian theory which avoids these problems by instead proposing that the negotiators use the *Kalai-Smorodinsky* solution of §7A. Gauthier describes the Kalai-Smorodinsky solution as the 'minimax relative concession principle'. The idea is that each participant identifies the most they could potentially gain from the bargain, but then, as a reasonable person, they acknowledge that they must concede some of this potential gain. Gauthier proposes that everyone should concede the same amount, relative to their maximum gain. We could justify this principle *normatively* by observing that it *minimizes* the maximum concession required by anyone (hence 'minimax relative concession'). We could also justify it *pragmatically* by observing that neither party will agree to the bargain if they feel that they are conceding more of their potential gain than the other party.

Gauthier proposes his theory not only to justify a particular political organization for society, but also to rationally justify moral behaviour for each rational individual. The idea is that each of us, separately, can work through the ramifications of this theory and rationally decide that it is in our own best interest to behave 'cooperatively' with other people (e.g. to not defect in Prisoner's Dilemma situations).

Binmore: Recently a contractarian theory has been developed by Ken Binmore in his two-volume *tour de force* entitled *Game theory and the Social Contract* [Bin93, Bin98]. Binmore's argument is subtle and complex, and draws upon a plethora of philosophy, economics, and mathematics. The first difference between Binmore and prior contractarians like Gauthier or Rawls is that Binmore insists that we cannot identify the status quo point $\mathbf{q}$ with some Hobbesian 'state of nature'. We must use the present state of the world as the status quo. To see this, recall that the 'original position' is not a place we can really go, it is merely a *philosophical device*, which real people in the real world can employ to resolve their disputes. But real people will not employ this device if it threatens to take away their status quo privileges and prerogatives. Hence, any bargains reached using the 'original position' device will only be

politically viable if they are Pareto-prefered to the real status quo in the real world. This is not a moral *endorsement* of the status quo; it is just a recognition of political reality.

Also, unlike Rawls, Binmore does not postulate that we will all have the same moral values and therefore achieve instant moral consensus behind the 'veil of ignorance'. Even if they are ignorant of their own exact circumstances, people will still have different values, and so negotiation will still be necessary to obtain a mutually agreeable social contract. Binmore is a political realist; he believes that political philosophy is only useful if it proposes a sociopolitical order which is *stable* (i.e. a Nash equilibrium in some suitably conceived 'game' of human interactions). Thus, he is not interested in utopian descriptions of what sort of agreements 'should' be reached behind the veil of ignorance, but rather, in what sort of agreements 'would' be reached in a real bargaining procedure.

Binmore then analyses what sort of bargaining procedure people might use when employing the 'original position' device. In the presence of a 'philosopher king' (a kind of perfectly incorruptable police officer who will not impose his own decision, but will enforce any decision the players make), the players can make *binding commitments*, and Binmore then argues that they will then select the utilitarian solution (§6A). In real life, however, there are no philosopher kings, and commitments cannot be enforced; players will only abide by a previously negotiated commitment if they feel that they could not do better by *renegotiating* it. In this environment, Binmore claims that the players will chose the proportional solution (§6B). Indeed, Binmore speculates that, over our long evolutionary history of resolving food-sharing disagreements etc. in primitive hunter-gatherer societies, we may well have evolved an *instinct* to apply the Rawlsian 'original position' and the proportional solution to the small problems of social coordination we face in our daily lives —hence the instinctive appeal which these ideas have as a basis for a political philosophy.

However, the proportional and utilitarian bargaining solutions both require a method for interpersonal comparison of utility (which Binmore calls our *empathic preferences*). The calibration constants which we use for these interpersonal comparisons reflect our social judgement of the 'worthiness' of different individuals. However, our notions of 'worthiness' themselves evolve over time. Binmore argues from a game-theoretic perspective that, over time, our interpersonal utility calibration will evolve towards the Yaari calibration of Theorem 6C.1. Thus, eventually, we will actually resolve social contract problems using the *Nash* solution of §5, but our interpersonal utility comparisons will be such that we *perceive ourselves* as using the proportional or utilitarian solutions. This convergence explains why utilitarians like Harsanyi and egalitarians like Rawls often end up recommending similar reforms in practice.

Thus, Binmore concludes that, in the long run, the contractarian approach inevitably degenerates into using the Nash solution, which arguable has no moral content. However, Binmore emphasizes that we *perceive* ourselves to be applying Rawlsian egalitarianism, unconsciously using the Yaari calibration determined by the current bargaining problem. If the bargaining problem changes (for example, if the set $\mathcal{B}$ of feasible outcomes expands because of new technologies or new opportunities), then we will initially apply the *proportional* solution to the new bargaining problem, but using the Yaari calibration from the *old* problem. Hence, in the short term at least, the Rawlsian contractarian framework will be meaningful and relevant. (In the

long term, we will converge to the Yaari calibration for the new bargaining problem, and then we will be back at the Nash solution).

Like David Hume (whom he much admires), Binmore regards real-world socipolitical structures as pragmatic systems which evolved and persist because they *work*, not because they instantiate some ideal of 'justice'. Indeed, our concept of 'the common good' (i.e. the intuition behind utilitarian moral philosophies) and our concept of personal 'righteousness' or 'moral responsibility' (i.e. the intuition behind deontological moral philosophies) are invented *ex post facto* as rationalizations of whatever sociopolitical order our society has evolved. According to Binmore, this is just one of many fallacious 'causal reversals' which cloud our reasoning about sociopolitical phenomena:

> ...An equitable compromise does not assign more to Eve than to Adam because she is more worthy. She is *deemed* to be more worthy because the concept of equity generated by social evolution... assigns her more than Adam. Societies do not become dysfunctional because the old virtues are abandoned. The old virtues cease to be honoured because the social contract has shifted. We do not punish people because they are morally responsible for their actions. We *say* they are morally responsible because our social contract requires that they be punished. We are not unpredictable because we have free will. We *say* that we have free will because we are not always predictable. A society does not choose a social contract because it promotes the common good. Our definition of the common good *rationalizes* our choice of social contract.
>
> [Bin98, §A.1.1, pp.512-513, emphasis mine]

Further reading

See [Mou84, Chapt.3], [Mye91, Chapt.8] and [OR94, Chapt.7] for other introductions to axiomatic characterizations of various bargaining models, including utilitarianism and the proportional solution. Another excellent reference is [Roe98]; chapter 1 contains axiomatic characterizations of various bargaining and social choice solutions, and the remaining chapters explore ramifications to political philosophy. Also, [Nap02] explores applications of bargaining theory to contractarian political philosophy.

Chapter 7

Renormalized Solutions

7A Kalai & Smorodinsky's Relative Egalitarianism

Prerequisites: §4A **Recommended:** §6B

The Nash Solution (§4B) to the bargaining problem has been criticised as sometimes producing unfair outcomes. Nevertheless, it is attractive because of its axiomatic characterization. The utilitarian solution (§6A) and the proportional solution (§6B) also have nice axiomatic characterizations, but unfortunately they require interpersonal comparisons of utility, which can be quite problematic (§3B). In this section we describe another bargaining solution, due to Ehud Kalai and Meir Smorodinsky [KS75]. Like the Nash solution of §4B, Kalai and Smorodinsky treat the two players equally [axiom **(S)**] and regard interpersonal utility comparisons as meaningless [axiom **(IR)**]. However, they replace the axiom **(IIA)** with a "monotonicity" axiom, which superficially has the same flavour, but produces unexpectedly different outcomes.

Consider a bargaining problem with feasible set $\mathcal{B} \subset \mathbb{R}^2_{\nearrow}$ and status quo point $\mathbf{q} \in \mathcal{B}$. To compute the Kalai-Smorodinsky solution, we first define

$$M_0 := \max\left\{ b_0 \; ; \; \mathbf{b} = (b_0, b_1) \in \mathcal{B} \text{ and } \mathbf{b} \overset{\wp}{\succeq} \mathbf{q} \right\}$$

$$\text{and} \qquad M_1 := \max\left\{ b_1 \; ; \; \mathbf{b} = (b_0, b_1) \in \mathcal{B} \text{ and } \mathbf{b} \overset{\wp}{\succeq} \mathbf{q} \right\}.$$

(Recall: "$\mathbf{b} \overset{\wp}{\succeq} \mathbf{q}$" means $b_0 \geq q_0$ and $b_1 \geq q_1$). The point $\mathbf{M} := (M_0, M_1) \in \mathbb{R}^2_{\nearrow}$ is called the *utopian solution*, and represents the idyllic scenario where both bargainers simultaneously get everything they want. Of course, $\mathbf{M}$ is usually *not* in the feasible set $\mathcal{B}$; see Figure 7.1(A).

Next, we draw a line $\mathbf{L}$ from $\mathbf{q}$ to $\mathbf{M}$. The line $\mathbf{L}$ intersects the negotiating set $\wp_{\mathbf{q}}\mathcal{B}$ at a unique point $\kappa = \kappa(\mathcal{B}, \mathbf{q})$. The point κ is the *Kalai-Smorodinsky solution* (also called the *relative egalitarian* solution): it is the solution where both players gain an equal fraction of the maximum amount they *could* gain. In other words $\kappa(\mathcal{B}, \mathbf{q})$ is the unique point $(k_0, k_1) \in \wp_{\mathbf{q}}\mathcal{B}$ such that

$$\frac{k_0 - q_0}{M_0 - q_0} \quad = \quad \frac{k_1 - q_1}{M_1 - q_1}.$$

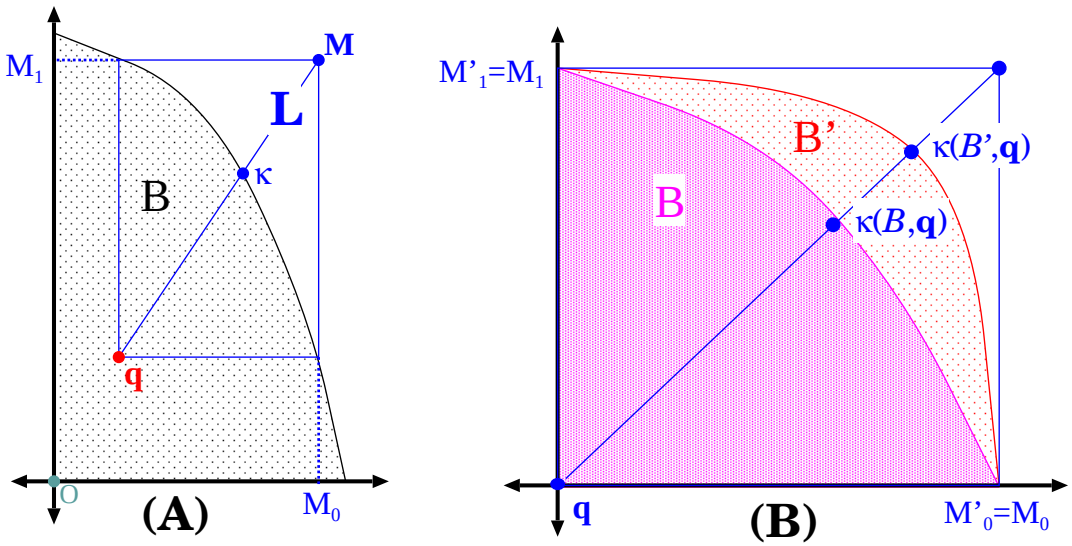


Figure 7.1: **(A)** M_0 and M_1 are the ideal points for each bargainer. **(B)** The Individual Monotonicity Axiom.

Kalai and Smorodinsky motivate this bargaining solution by noting that Nash's axiom **(IIA)** sometimes produces a counterintive result, where one player ends up with a *worse* bargain when the bargaining set is expanded in a manner which seems to favour him. They therefore propose to replace **(IIA)** with the following axiom:

(IM) (Individual Monotonicity) *Suppose* $\mathbf{q} \in \mathcal{B} \subset \mathcal{B}' \subset \mathbb{R}^2_{\nearrow}$, *and* $M_0 = M_0'$ *and* $M_1 = M_1'$. *Then* $\alpha(\mathcal{B}, \mathbf{q}) \overset{\wp}{\preceq} \alpha(\mathcal{B}', \mathbf{q})$ [see Figure 7.1(B)].

Exercise 7.1: Construct an example to show that the Nash bargaining solution of §4B does *not* satisfy **(IM)**.

For convenience, we also restate the other bargaining axioms we will need:

(RI) (Rescaling Invariance) *Suppose* $F : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ *is an affine 'rescaling' function, and let* $F(\mathcal{B}) = \mathcal{B}'$ *and* $F(\mathbf{q}) = \mathbf{q}'$. *Then* $\alpha(\mathcal{B}', \mathbf{q}') = F\left[\alpha(\mathcal{B}, \mathbf{q})\right]$.

(S) (Symmetry) *Let* $\mathcal{B}' = \{(b_1, b_0)\ ;\ (b_0, b_1) \in \mathcal{B}\}$. *If* $\mathbf{q} = (q_0, q_1)$, *then let* $\mathbf{q}' := (q_1, q_0)$. *If* $\alpha(\mathcal{B}, \mathbf{q}) = (b_0, b_1)$, *then* $\alpha(\mathcal{B}', \mathbf{q}') = (b_1, b_0)$.

Theorem 7A.1 (Kalai & Smorodinsky, 1975)
κ is the unique bargaining solution satisfying axioms **(S)**, **(RI)**, *and* **(IM)**.

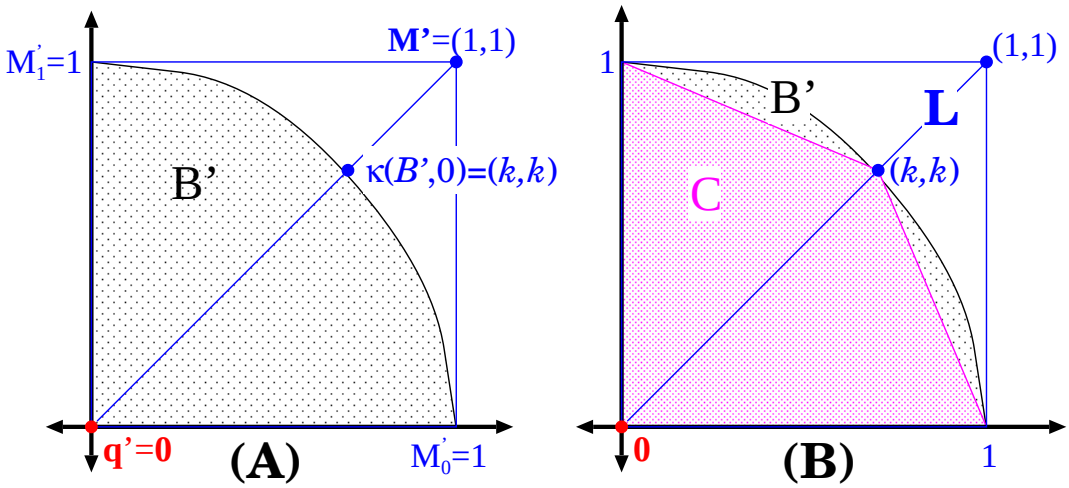


Figure 7.2: The Kalai-Smorodinsky solution.

Proof: **Exercise 7.2** Check that κ satisfies these three axioms.

Now, suppose $\alpha : \mathfrak{B} \longrightarrow \mathbb{R}^2_{\not\nearrow}$ is a bargaining solution satisfying these three axioms; we will show that $\alpha = \kappa$.

Let $(\mathcal{B}, \mathbf{q})$ be any bargaining problem. First, we rescale $(\mathcal{B}, \mathbf{q})$ to a new bargaining problem $(\mathcal{B}', \mathbf{0})$, where $\mathbf{0} = (0, 0)$ and $(M'_0, M'_1) = (1, 1)$, as shown in Figure 7.2(A). Formally, let $F : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ be the rescaling function

$$F(b_0, b_1) \quad := \quad \left(\frac{b_0 - q_0}{M_0 - q_0}, \ \frac{b_1 - q_1}{M_1 - q_1} \right).$$

Let $\mathcal{B}' := F(\mathcal{B})$ and note that $F(\mathbf{q}) = \mathbf{0}$. Thus, the axiom **(RI)** implies that $\alpha(\mathcal{B}', \mathbf{0}) = F[\alpha(\mathcal{B}, \mathbf{q})]$. Since we know that κ is also rescaling invariant, we also have $\kappa(\mathcal{B}', \mathbf{0}) = F[\kappa(\mathcal{B}, \mathbf{q})]$. Thus, it suffices to show that $\alpha(\mathcal{B}', \mathbf{0}) = \kappa(\mathcal{B}', \mathbf{0})$.

First, note that $\kappa(\mathcal{B}', \mathbf{0}) = (k, k)$ for some $k \in [0, 1]$, because $\kappa(\mathcal{B}', \mathbf{0})$ lies on the 45^o line $\mathbf{L}$ from $\mathbf{0}$ to $(1, 1)$. Now, let $\mathcal{C}$ be the kite-shaped region in Figure 7.2(B), which is the convex set with vertices at $\mathbf{0}$, $(1, 0)$, $(0, 1)$, and (k, k). Clearly, $(\mathcal{C}, \mathbf{0})$ is symmetric under reflection across $\mathbf{L}$; hence, axiom **(S)** implies that $\alpha(\mathcal{C}, \mathbf{0})$ must lie on the line $\mathbf{L}$. Meanwhile axiom **(P)** says that $\alpha(\mathcal{C}, \mathbf{0})$ must lie on the Pareto frontier of $\mathcal{C}$. Thus, $\alpha(\mathcal{C}, \mathbf{0}) = (k, k)$, because (k, k) is the unique intersection point of $\mathbf{L}$ with the Pareto frontier of $\mathcal{C}$.

Now, $\mathcal{C} \subseteq \mathcal{B}'$, so axiom **(IM)** implies that $(k, k) = \alpha(\mathcal{C}, \mathbf{0}) \leq \alpha(\mathcal{B}', \mathbf{0})$. Meanwhile axiom **(P)** says that $\alpha(\mathcal{B}', \mathbf{0})$ must lie on the Pareto frontier of $\mathcal{B}'$. But (k, k) itself is the only point on the Pareto frontier of $\mathcal{B}$ which satisfies these two constraints. Thus, $\alpha(\mathcal{B}', \mathbf{0}) = (k, k)$. In other words, $\alpha(\mathcal{B}', \mathbf{0}) = \kappa(\mathcal{B}', \mathbf{0})$. □

Remark: There is a close relationship between the Kalai-Smorodinsky solution and the 'equitable' cake division algorithms of §11B. Indeed, axiom **(RI)** essentially says that we can represent the original bargaining problem by one where the characters divide metaphorical 'cake' which has a total utility of 1 for each of them; the Kalai-Smorodinsky solution then tells us to cut the cake 'equitably', so that Owen's assessment of his portion is the same as Zara's assesment of *her* portion.

Of course, for a two-person cake division, equitable partition is fairly trivial (e.g. 'I cut, you choose'); the real interest is *multiperson* cake divisions. From this point of view, any of the (multiperson) fair division procedures of Chapter IV can be translated into a (multiperson) bargaining solution, with similar properties, as long as we feel justified in first rescaling the player's utilities so that the status quo is at $\mathbf{0}$ and each player's maximum utility gain is 1. (In some settings, such a rescaling may seem quite inappropriate).

One might ask how an exactly algorithm to cut cakes can be applied to bargaining problems like contract negotiations between a labour union and management. To take a simple example, suppose that there are several issues (say, wages, benefits, holiday time, working conditions, etc.), and on each issue there is a 'union position' and a 'management position'. The appropriate version of 'I cut, you choose' might be as follows: One of the two parties (selected at random, say, the union), draws up a contract C, which takes either the 'union position' or the 'management position' on each issue. Let $\overline{C}$ be the 'opposite' contract, which takes the *opposite* stance on each issue. Establishing this dichotomy between C and $\overline{C}$ corresponds to 'cutting the cake'. The other party (say, the management) then chooses between C and $\overline{C}$.

This procedure has a fatal flaw, however: if the parties know in advance that a dispute-resolution protocol like this will be used, then each party will wildly exaggerate its position on one or more issues. For example, the union might demand six month paid vacations, so that the management will feel forced to put this demand into the contract C, and *all* of the other union demands into contract $\overline{C}$ when 'cutting the cake'. The union will then just pick $\overline{C}$ and thereby get 90% of what it wants.

⌜ ⌝

Exercise 7.3: (a) Generalize the Kalai-Smorodinsky solution to bargains with $N \geq 3$ players.

(b) Generalize axioms **(S)**, **(RI)**, and **(IM)** to bargains with $N \geq 3$ players.

(c) Generalize the statement and proof of Theorem 7A.1 to bargains with $N \geq 3$ players.]

Exercise 7.4: (Risk Aversion and the Kalai-Smorodinsky Solution)

Consider the *surplus division* problem of Example 4A.2. For simplicity, suppose Zara and Owen are trying to divide one dollar. Thus, if x_0 is Zara's share of the dollar, then $x_1 := 1 - x_0$ is Owen's share. Assume Zara is risk-neutral, so that her utility function for money is $b_0(x_0) = x_0$ (this is plausible if Zara is quite wealthy, so that one dollar represents a very small fraction of her existing wealth). Assume Owen is risk-averse, with monetary utility function $b_1(x_1) = x_1^\alpha$ for some $\alpha \in (0, 1)$ (this is plausible if Owen is much less wealthy than Zara). Assume the status quo is $\mathbf{0} = (0, 0)$.

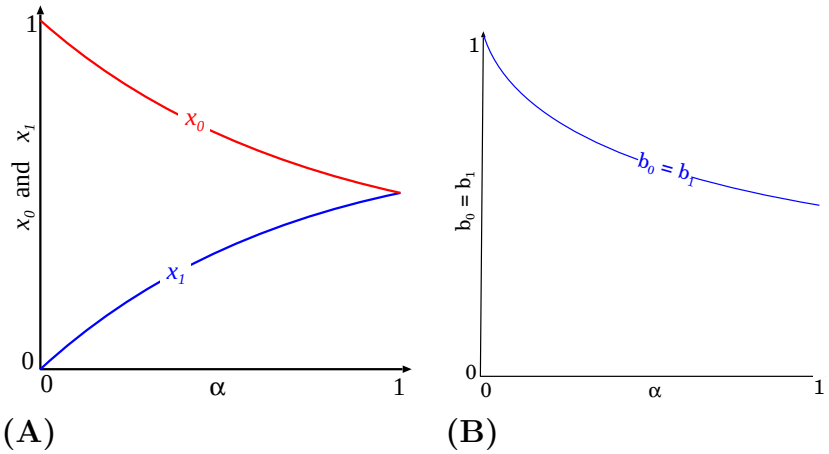


Figure 7.3: The Kalai-Smorodinsky bargaining solution as a function of the risk aversion of Owen (Exercise 7.1). Zara and Owen are dividing a dollar, so $x_0 + x_1 = 1$. Zara has utility function $b_0(x_0) = x_0$ and Owen has utility function $b_1(x_1) = x_1^\alpha$, where $\alpha \in (0, 1)$.
(A) x_0 and x_1 as functions of α. Notice that $x_0 \nearrow 1$ and $x_1 \searrow 0$ as $\alpha \searrow 0$.
(B) $b_0 = b_1$ as functions of α (by definition, $b_0 = b_1$ in the Kalai-Smorodinsky solution). Notice that $b_0 = b_1 \nearrow 1$ as $\alpha \searrow 0$.

(a) Define $f_\alpha(u) := u^{1/\alpha} + u - 1$. Show that the Kalai-Smorodinsky solution to this bargaining problem awards a utility of u to both players, where u is the unique solution in $[0, 1]$ of the equation $f(u) = 0$.

(b) Conclude that the Kalai-Smorodinsky solution awards $x_0 = u$ dollars to Zara, and $x_1 = 1 - u$ dollars to Owen.

(c) In particular, if $\alpha = \frac{1}{2}$, then $x_0 = u = \frac{1}{2}(\sqrt{5} - 1) \approx 0.6180...$, while $x_1 = \frac{1}{2}(3 - \sqrt{5}) \approx 0.3820...$

Figure 7.3(a) shows that that $x_0 \nearrow 1$ and $x_1 \searrow 0$ as $\alpha \searrow 0$. Thus, the more risk averse Owen becomes, the more the Kalai-Smorodinsky solution favours Zara in material terms. This feature perhaps enhances the realism of the solution as a description of real bargaining, while simultaneously diminishing its appeal as a normative ideal of justice. However, by definition, $b_1 = b_0$ in the Kalai-Smorodinsky solution, and indeed, Figure 7.3(b) shows that $b_1 = b_0 \nearrow 1$ as $\alpha \searrow 0$. Hence, in utility terms, neither party is favoured. Compare this with the conclusion of Exercise 4.17 on page 85.]

Moulin's arbitration scheme: Hervé Moulin has proposed an implementation of the Kalai-Smorodinsky solution which does not resort to metaphors about cake cutting [Mou84]. However, like the 'fair division games' of Chapter IV, it is an example of an *implementation procedure*; that is, a game whose rules are designed so that, if each player uses their optimum strategy, then the outcome is the Kalai-Smorodinsky solution. Moulin's game works as follows:

1. Zara names a probability $p_0 \in [0,1]$, while simultaneously, Owen names a probability $p_1 \in [0,1]$.

2. The person who names the higher probability (let's say, Zara) wins the right to propose an agreement $\mathbf{b}_0 \in \mathcal{B}$.

3. Owen can either accept or refuse the proposal $\mathbf{b}_0 \in \mathcal{B}$.

   (a) If he accepts $\mathbf{b}_0$, then an arbitrator sets up a lottery which implements $\mathbf{b}_0$ with probability p_0 and the status quo $\mathbf{q}$ with probability $1 - p_0$. The game ends.

   (b) If Owen refuses $\mathbf{b}_0$, he can make a counterproposal $\mathbf{b}_1 \in \mathcal{B}$.

4. Zara can either accept or refuse $\mathbf{b}_1$.

   (a) If Zara accepts $\mathbf{b}_1$, then then the arbitrator organizes a lottery which implements $\mathbf{b}_1$ with probability p_0 (Note: *not* p_1) and implements the status quo $\mathbf{q}$ with probability $1 - p_0$.

   (b) If Zara refuses, then the arbitrator implements $\mathbf{q}$ for certain.

Note that Moulin's game is *not* intended as a realistic model of bargaining (i.e. it is not an attempt to realize the 'Nash program' of Chapter 5 for the Kalai-Smorodinsky solution). For one thing, Moulin's game requires the intervention of a referee (the arbitrator), whereas real bargaining should not. Instead, we should see Moulin's game as a kind of *arbitration scheme.* Assume that the two players have agreed (in principle) that the Kalai-Smorodinsky bargaining solution should be used, but they are not sure how to identify this solution in practice (partly because they do not trust each other). They can call upon an arbitrator to preside over the Moulin game, and thereby reach an agreement.

7B Relative Utilitarianism

Uzi Segal "All Dictatorships are equally bad" –axiomatic characterization of relative utilitarian bargaining solution.

Dhillon-Mertens, etc.

Dhillon –strong pareto

Karni —impartiality.

Part IV

Fair Division

Chapter 8

Partitions, Procedures, and Games

A compromise is the art of dividing a cake in such a way that everyone believes he has the biggest piece. —Ludwig Erhard (1897-1977)

Prerequisites: None.

Suppose two or more people are dividing a cake. What is a procedure we can use to guarantee that each individual gets a 'fair' portion? We assume that the participants are selfish and do not trust one another, and perhaps even dislike one another. Nevertheless, we seek a procedure, which the people can execute themselves (i.e. without the intervention of some arbiter), which ensures that each individual will come away feeling that he has a fair share. Think of this procedure as a 'game' such that, if each individual plays 'rationally', then all people will be ensured a fair outcome.

Example 8.1: I cut, you choose

The classic cake division procedure for two people is well-known. If Owen and Twyla are trying to split a cake, then one of them (say, Owen) divides the cake into two parts, and the other (Twyla) chooses which half she likes better. Owen then takes the remaining half. Owen therefore has an incentive to make the division as even as possible, to avoid the risk of getting a smaller piece.

To see this, note that Owen's worst-case scenario is that Twyla takes the larger portion, thereby leaving him with the smaller portion. In the language of game theory, the smaller portion is his *minimum payoff*. Hence, he seeks a *maximin* strategy: the cake division which *maximizes* the *minimum* portion which he could receive. Clearly, his unique maximin strategy is to make sure the cake is divided into two exactly equal portions.

Since Twyla always picks what she perceives as the larger portion, she will always perceive the outcome as fair. If Owen plays rationally (i.e. according to his maximin strategy), then he will also see the outcome as 'fair' no matter which piece Twyla chooses, because he has ensured that both portions are equal (in his perception). ♢

The beauty of this procedure is that it does not require any 'referee' or 'arbiter' to decide the fair outcome; the fair outcome arises naturally from the rational choices of the players. Can this elegant solution to the 'cake-cutting problem' be generalized to three or more people? The problem becomes more complicated if the players actually have different preferences (e.g. Owen likes the orange cream part more, but Twyla likes chocolate more) or if the cake has 'indivisible' components (i.e. nuts or cherries which cannot easily be divided).

Of course, our real goal is not to prevent fist-fights at birthday parties. 'Cake-cutting' is a metaphor for a lot of contentious real-life problems, including:

- Resolving a border dispute between two or more warring states.
- Dividing an inheritance amongst squabbling heirs.
- Splitting the property in a divorce settlement.
- Allocating important government positions amongst the various factions in a coalition government.
- Defining 'fair use' of intrinsically common property (e.g. resource rights in international waters).

Fair division procedures can be used to divide up 'bads' as well as 'goods'. In this case, each participant seeks to *minimize* their portion, rather than *maximizing* it. Some examples include:

- Partitioning chores amongst the members of a household.
- Allocating military responsibilities to different member states in an alliance.

Fair division is generally more complicated than the simple and elegant 'I cut, you choose' algorithm for two individuals, because of the following factors:

- There are generally more than two participants.
- The participants may have different preferences (e.g. Owen likes orange cream, Twyla likes chocolate), or at the very least, different perceptions of the situation. Hence, what looks like a 'fair' division to one individual may appear 'unfair' to the other. To mathematically represent this, we endow each individual with a *utility measure*, which encodes how *he* values different parts of the cake. See §8.
- In dividing a piece of physical territory, there are military and economic reasons why the portions should be *connected*. Thus, we cannot achieve 'fairness' by giving each party a chopped up collection of tiny bits. See §9E.
- The participants may be actively hostile and distrusting of one another. Thus, each of four participants may not only demand that he receives *at least* one quarter, but he may also require that (in his perception) no *other* participant receives *more* than he does. We call this an *envy-free* partition. See §11A.

- Ideally, we'd like a partition which maximizes the happiness of the participants. For example, if Owen likes orange cream and Twyla likes chocolate, it makes sense to give him more of the orange cream, and her more of the chocolate, even if this does *not* result in a strict 50/50 division of the cake. See §10 and §10D.
- Some components are indivisible. For example, an inheritance may involve single, high-value items (e.g. a car, a painting) which cannot easily be split or shared amongst two heirs. See §11C.2.
- Some participants may be 'entitled' to a larger share than others. For example, in dividing an inheritance, the spouse of the deceased may be entitled to one half the estate, while each of the three children is entitled to only one sixth. See §11C.1.
- If the participants have knowledge of one another's preferences, they can cooperate to maximize their common well-being. However, one individual can also use this knowledge to *manipulate* the procedure, obtaining a disproportionately large share at someone else's expense. See §11C.4.

8A Utility Measures

Let $\mathbf{X}$ be a set which represents the cake (or the inheritance, or the disputed territory, etc.). A *portion* is some subset $\mathbf{P} \subset \mathbf{X}$. Each individual assigns some utility $\mu(\mathbf{P})$ to the portion $\mathbf{P}$. This defines a function μ from the collection of all subsets of $\mathbf{X}$ to the set of real numbers. We assume that μ satisfies the following axioms:

(U0) $\mu[\emptyset] = 0$. In other words, the value of an empty portion is zero.

(U1) $\mu[\mathbf{X}] = 1$. The value of the entire cake is one.

(UA) For any disjoint subsets $\mathbf{P}, \mathbf{Q} \subset \mathbf{X}$, $\mu[\mathbf{P} \sqcup \mathbf{Q}] = \mu[\mathbf{P}] + \mu[\mathbf{Q}]$. (We say that μ is *additive*.)

More generally, in some procedures involving an sequence of approximations, we require:

(UA$^\infty$) For any infinite sequence of disjoint subsets $\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3, \ldots \subset \mathbf{X}$,

$$\mu[\mathbf{P}_! \sqcup \mathbf{P}_2 \sqcup \mathbf{P}_3 \sqcup \cdots] \quad = \quad \mu[\mathbf{P}_1] \; + \; \mu[\mathbf{P}_2] \; + \; \mu[\mathbf{P}_2] \; + \; \cdots$$

(We say that μ is *sigma-additive*.)

A function μ satisfying properties **(U0)**, **(U1)**, and **(UA)** is called a *utility measure*[1]. We assume that seeks a piece $\mathbf{P}$ which maximizes $\mu(\mathbf{P})$.

[1]Actually μ is a special case of a mathematical object called a *signed measure*. Signed measures cannot, in general, be well-defined on *every* subset of $\mathbf{X}$. Instead, we must choose a collection of 'measurable' subsets called a *sigma-algebra*, which is closed under countable intersections and unions. This is a technicality which we will neglect in this discussion.

Remark on Additivity: The axiom **(UA)** assumes that utilities are *additive*, but in many real situations, this is not true, because it neglects the phenomena of *complementarity* and *substitutability* between various parts of $\mathbf{X}$. For example suppose $\mathbf{X}$ is a collection of food, and $\mathcal{I}$ is a bunch of hungry individuals, who seek to divide the food in a manner such that each individual gets a decent meal. Suppose that $\mathbf{P}, \mathbf{Q}, \mathbf{R}, \mathbf{S} \subset \mathbf{X}$ are four disjoint subsets of $\mathbf{X}$. $\mathbf{P}$ is a piece of *pumpernickel* bread, $\mathbf{Q}$ is *quark* cheese, $\mathbf{R}$ is *rye* bread, and $\mathbf{S}$ is *salami*. For a well-balanced meal, each individual wants to make a sandwich with some bread and some filling. Clearly, if I already have **P**umpernickel bread, then what I want next is either **Q**uark or **S**alami. I *don't* want **R**ye bread. Conversely, if I have the **S**alami, then I want bread, not cheese.

In economics jargon, the items $\mathbf{P}$ and $\mathbf{R}$ are **substitutes**; they are both bread, and thus, if you have one, then you no longer desire the other. In other words, the utility of having both $\mathbf{P}$ and $\mathbf{R}$ is *less* than the sum of their separate utilities:

$$\mu[\mathbf{P} \sqcup \mathbf{R}] \quad < \quad \mu[\mathbf{P}] + \mu[\mathbf{R}].$$

On the other hand, the items $\mathbf{P}$ and $\mathbf{Q}$ are *complements*. By themselves, neither is worth very much (a piece of bread by itself is a pretty poor lunch). But together, they make a tasty sandwich. Thus, the value of the combination $\mathbf{P} \sqcup \mathbf{Q}$ is *greater* than the sum of the separate parts:

$$\mu[\mathbf{P} \sqcup \mathbf{Q}] \quad > \quad \mu[\mathbf{P}] + \mu[\mathbf{Q}].$$

There are other more complex ways in which different subsets of $\mathbf{X}$ can combine to have nonadditive utility. For example, suppose $\mathbf{X}$ is a disputed territory and $\mathcal{I}$ is a collection of hostile military powers trying to divide $\mathbf{X}$ between them. Clearly, a *connected* piece of territory is much more valuable to any party than several *disconnected* components. So, suppose $\mathbf{P} \subset \mathbf{X}$ is a disconnected piece of territory, and $\mathbf{Q} \subset \mathbf{X}$ is another disconnected piece of territory, but the combination $\mathbf{P} \sqcup \mathbf{Q}$ is connected (e.g. $\mathbf{Q}$ forms a 'bridge' between the two parts of $\mathbf{P}$). Then from a strategic point of view, the unified territory $\mathbf{P} \sqcup \mathbf{Q}$ is worth much more than the sum of the two separate pieces. (See §9E for a discussion of connectivity in partitions).

Notwithstanding this objection, we will keep axiom **(UA)** because it is a good approximation in many situations, and because it would be too complicated to mathematically represent complementarity and substitutability[2].

8B Partition Procedures

Let $\mathcal{I} = \{1, \ldots, I\}$ be a set of I individuals dividing $\mathbf{X}$; each individual i has a utility measure μ_i. Instead of referring to these people with the conventional (but boring) terminology of 'Player One', 'Player Two', etc., we will give them names. Player One will always be called

[2]It is perhaps possible to mathematically represent complementarity and substitutability using a measure μ defined on $\mathbf{X} \times \mathbf{X}$ (so that $\mu[\mathbf{P} \times \mathbf{Q}] > 0$ if $\mathbf{P}$ and $\mathbf{Q}$ are complementary, and $\mu[\mathbf{P} \times \mathbf{Q}] < 0$ if $\mathbf{P}$ and $\mathbf{Q}$ are substitutes. We will not pursue this here.

Owen; Player Two will always be called *Twyla*; Player Three will always be called *Trey*; Player Four will always be called *Ford*, and so on. We presume that the mnemonic is obvious.

A *partition* is a collection of disjoint portions $\mathcal{P} = \{\mathbf{P}_1, \ldots, \mathbf{P}_I\}$ so that $\mathbf{X} = \mathbf{P}_1 \sqcup \cdots \sqcup \mathbf{P}_I$. We assume portion $\mathbf{P}_i$ goes to individual i; this individual then assigns the partition a utility of $\mu_i(\mathbf{P}_i)$. Clearly, each individual i seeks a partition $\mathcal{P}$ which maximizes the value of $\mu_i(\mathbf{P}_i)$.

A *partition procedure* is some function which takes information about the utility measures of the various parties, and yields a partition which (we hope) will satisfy each party. Formally, let $\mathfrak{M}$ be the set of all utility measures on $\mathbf{X}$. Then a complete description of the preferences of all parties is given by an I-tuple $(\mu_1, \mu_2, \ldots, \mu_I) \in \mathfrak{M}^I = \underbrace{\mathfrak{M} \times \mathfrak{M} \times \cdots \times \mathfrak{M}}_{I}$. Let $\mathfrak{P}_I$ be the set of all possible partitions of $\mathbf{X}$ into I portions. Then an *I-partition procedure* is a function $\Pi : \mathfrak{M}^I \longrightarrow \mathfrak{P}_I$.

Partition procedures involve 'dividing the value' of the cake, which requires that the value be divisible. Indivisible components of value are called *atoms*, and present obstructions to partition. To be precise, if μ is a utility measure on $\mathbf{X}$, then an *atom* of μ is a point $x \in \mathbf{X}$ such that $\mu\{x\} > 0$. Intuitively, an atom represents a valuable but indivisible item, a 'diamond in the cake' (Akin [Aki95]). We say μ is *nonatomic* if it has no atoms. Failing that, we say that μ is *at most $\frac{1}{I}$ atomic* if the total mass of all atoms of μ is less than $\frac{1}{I}$. In other words, there is a subset $\mathbf{Y} \subset \mathbf{X}$ which contains *no* atoms, such that $\mu[\mathbf{Y}] > 1 - \frac{1}{I}$. The consequence is that any portion of size $\frac{1}{I}$ cannot be *entirely* made of atoms, and hence, is divisible.

Procedure 8B.1: I cut, you choose

Let $\mathbf{X} = [0, 1]$ be the unit interval (representing a one-dimensional cake). Let μ_1 and μ_2 be utility measures on $\mathbf{X}$. Assume μ_1 is at most $\frac{1}{2}$ atomic.

(1) Let $r \in [0, 1]$ be such that $\mu_1[0, r) = \frac{1}{2} = \mu_1[r, 1]$ (i.e. Owen cuts the cake into two pieces which he perceives have equal size; this is possible because μ_1 is at most $\frac{1}{2}$ atomic)

(2a) If $\mu_2[0, r) \geq \mu_2[r, 1]$, then define $\mathbf{P}_2 = [0, r)$ and $\mathbf{P}_1 = [r, 1]$. (If Twyla thinks that $[0, r)$ is larger, then she takes this piece, and Owen takes the other one).

(2b) Otherwise, if $\mu_2[0, r) < \mu_2[r, 1]$, then define $\mathbf{P}_1 = [0, r)$ and $\mathbf{P}_2 = [r, 1]$. (If Twyla thinks that $[r, 1]$ is larger, then she takes *this* piece, and Owen takes the other one).

Now let $\mathcal{P} = \{\mathbf{P}_1, \mathbf{P}_2\}$. ———————————

8C Partition Games

In general, of course, we do *not* have complete information about every individual's preferences. Instead, each individual provides a small amount of information (e.g. by proposing a certain portion as 'fair' or by rejecting a proposed portion as 'too small'). We must use this limited information to *interpolate* his true desires. Also, in general, we cannot assume that an 'objective arbiter' will be present to implement the partition procedure; it must be something which the

participants can do themselves, even if they are *not* friends and do *not* trust one another. Thus, for practical purposes, we seek not a procedure, but a *game*, so that, if all participants play 'rationally', then the outcome will be *as if* we had implemented some partition procedure.

An I-person *partition game* is a structure $\Gamma := (\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_I;\ \gamma)$, where $\mathcal{S}_i$ is some set of 'strategies' (i.e. 'moves') for player i, and where $\gamma : \mathcal{S}_1 \times \mathcal{S}_2 \times \cdots \times \mathcal{S}_I \longrightarrow \mathfrak{P}_I$. An I-tuple of strategies $\mathbf{s} := (s_1, \ldots, s_I)$ is called a *play* of the game (representing a specific choice of strategy by each player). Thus, γ is a function which takes any play $\mathbf{s}$, and produces an **outcome** $\gamma(\mathbf{s}) = \mathcal{P} \in \mathfrak{P}_I$ which is an I-partition of $\mathbf{X}$. Each player might perform a different 'role' in the game, and thus, different players may have different strategy-sets to choose from. We assume each player picks the strategy which he believes will maximize his portion.

Game 8C.1: I cut, you choose

Let $\mathbf{X} = [0, 1]$ be the unit interval (representing a one-dimensional cake). Let μ_1 and μ_2 be two utility measures on $\mathbf{X}$ (with μ_1 being at most $\frac{1}{2}$ atomic). We define the sequential game Γ as follows:

1. First, Owen chooses a number $r \in [0, 1]$ (i.e. Owen 'cuts the cake'.)
2. Next, Twyla choses between partitions **(a)** and **(b)**:
   (a) $\mathbf{P}_1 = [0, r)$ and $\mathbf{P}_2 = [r, 1]$.
   (b) $\mathbf{P}_2 = [0, r)$ and $\mathbf{P}_1 = [r, 1]$.

Thus, Owen's strategy is a point $r \in [0, 1]$, so we can define $\mathcal{S}_1 = [0, 1]$. Twyla's strategy is to then choose either the left portion or the right portion, so we'll say $\mathcal{S}_2 = \{L, R\}$. The function $\gamma : \mathcal{S}_1 \times \mathcal{S}_2 \longrightarrow \mathfrak{P}_2$ is then defined by $\gamma(r, s) = \{\mathbf{P}_1, \mathbf{P}_2\}$, where

$$\mathbf{P}_1 = [0, r) \text{ and } \mathbf{P}_2 = [r, 1] \text{ if } s = L,$$
$$\text{and } \mathbf{P}_2 = [0, r) \text{ and } \mathbf{P}_1 = [r, 1] \text{ if } s = R.$$

Suppose $\mathbf{s} \in \mathcal{S}_1 \times \cdots \times \mathcal{S}_I$ is play of game Γ and $\gamma(\mathbf{s}) = \mathcal{P} = \{\mathbf{P}_1, \ldots, \mathbf{P}_I\}$ is the resulting partition. We define $\mu_i(\mathbf{s}) := \mu_i(\mathbf{P}_i)$; this is called the Γ-*payoff* for player i in the play $\mathbf{s}$.

A *dominant* strategy for Owen is a strategy $s_1^* \in \mathcal{S}_1$ which yields a maximal payoff for Owen, no matter *what* the other players do. Formally, we say that $s_1^* \in \mathcal{S}_1$ is *dominant* if, for *any* counterstrategies $s_2 \in \mathcal{S}_2, \ldots, s_I \in \mathcal{S}_I$, the strategy s_1^* is best for Owen: for any other $s_1 \in \mathcal{S}_1$,

$$\mu_1(s_1^*, s_2, \ldots, s_I) \quad \geq \quad \mu_1(s_1, s_2, \ldots, s_I).$$

Clearly, it is irrational for Owen to choose anything but a dominant strategy, if he has one. Of course, in general, Owen may not have a dominant strategy. In this case, Owen can evaluate the worth of any strategy $s_1 \in \mathcal{S}_1$ by considering its 'worst case scenario'. The *minimum payoff* for s_1 is defined:

$$\underline{\mu}_1(s_1) \quad = \quad \min_{s_2 \in \mathcal{S}_2, \ldots, s_I \in \mathcal{S}_I} \mu_1(s_1, s_2, \ldots, s_I).$$

In other words, $\underline{\mu}_1(s_1)$ is the *worst* payoff which Owen can expect from s_1 under *any* circumstances. A *maximin strategy* is a strategy $s_1^\dagger \in \mathcal{S}_1$ which *maximizes* his worst-case scenario payoff:

$$\underline{\mu}_1(s_1^\dagger) \quad = \quad \max_{s_1 \in \mathcal{S}_1} \underline{\mu}_1(s_1).$$

The value of $\underline{\mu}_1(s_1^\dagger)$ is then the *maximin payoff* for Owen. This is the *worst* he ever expect to do in the game, if he plays according to his maximin strategy.

If $s_1^* \in \mathcal{S}_1$ is dominant for Owen, then s_1^* is automatically a maximin strategy (**Exercise 8.1**). However, a maximin strategy may exist even when a dominant strategy does not. We deem it irrational for Owen to chose anything but a maximin strategy, if one exists.

'I cut, you choose' is a *sequential* game, meaning that the players play one at a time, in numerical order. First Owen plays, then Twyla plays, and so on. Thus, Twyla *knows* the strategy of Owen, and thus, she can choose dominant/maximin strategies *given* this information. More generally, in a sequential game, player i already *knows* the strategies of players $1, \ldots, i-1$, and thus, he can choose dominant/maximin strategies *given* this information.

Example 8C.2: I cut, you choose (maximin strategies)

In Game 8C.1, Owen has no idea which piece Twyla will think is better (he has no idea what her utility measure is). However, he doesn't want to *risk* getting a small piece. Hence, to *maximize* the utility of the worst-case scenario, his *maximin* strategy is to chose r so that $\mu_1[0, r) = \frac{1}{2} = \mu_1[r, 1]$. In other words, he effectively implements step **(1)**. of the 'I cut, you choose' *procedure* (Procedure 8B.1).

Twyla plays after Owen, and *given* a strategy $r \in \mathcal{S}_1$ by Owen, her dominant strategy is clearly to pick the piece she thinks is better. Thus, she will effectively implement steps **(2a)** and **(2b)** of the 'I cut, you choose' *procedure* (Procedure 8B.1). ♢

The outcome of a game appears unpredictable. However, we can often predict the outcome if we make four assumptions about the 'psychology' of the players:

(Ψ1) Each player has **complete self-awareness** about his own preferences. In other words, he has 'perfect knowledge' of his own utility measure.

(Ψ2) Each player has **complete ignorance** of other players' preferences. Thus, he cannot in any way predict or anticipate their behaviour.

(Ψ3) Each player is **rational**, in the sense that he carefully considers all of his strategies and the possible counterstrategies of other players. For each possible play $\mathbf{s} \in \mathcal{S}_1 \times \cdots \times \mathcal{S}_I$, he determines what his payoff would be. He thereby determines his *minimum payoff* for each of his possible strategies, and thereby determines his *maximin strategy*.

(Ψ4) Each player is **conservative** (or **risk-averse**), in the sense that he wants to minimize individualal risk. He will *not* choose 'risky' strategies which threaten low minimum payoffs (even if they also tempt him with high *maximum* payoffs). Instead, he will 'play safe' and choose the strategy with the best minimum payoff: his *maximin* strategy.

Being *psychological* assertions, we can provide no mathematical justification for these axioms[3]. Nevertheless, we must postulate **(Ψ1)**-**(Ψ4)** to develop a predictive theory of partition games.

We can translate a partition *game* into a partition *procedure* if, using axioms **(Ψ1)**-**(Ψ4)**, we can *predict* that the game players will act *as if* they were executing that procedure. To be precise, suppose Π is an I-partition procedure, and Γ is an I-person partition game. We say that Γ *yields* Π if:

1. Each player in Γ has a unique pure maximin strategy, and
2. If all players play their maximin strategies, then the outcome of Γ will be the same partition as produced by Π.

Thus, Example 8C.2 shows that the 'I cut, you choose' *game* (Game 8C.1) yields the 'I cut, you choose' *procedure* (Procedure 8B.1).

[3]Indeed, axioms **(Ψ1)**-**(Ψ4)** are questionable on purely psychological grounds. Nevertheless, we can argue that, even if they are false, these axioms are at least 'reasonable approximations' which are 'good enough' for practical purposes. See [LR80, Chapt. 1 & 2] for more discussion of this.

Chapter 9

Proportional Partitions

9A Introduction

Prerequisites: §8

We say that a partition $\mathcal{P}$ is *proportional* if $\mu_i(\mathbf{P}_i) \geq \frac{1}{I}$ for all $i \in [1..I]$. For example, if $\mathcal{I} = \{1, 2\}$, then the partition $\mathcal{P} = \{\mathbf{P}_1, \mathbf{P}_2\}$ is proportional if $\mu_1(\mathbf{P}_1) \geq \frac{1}{2}$ and also $\mu_2(\mathbf{P}_2) \geq \frac{1}{2}$. In other words, each individual feels that (in their estimation), they received at *least* half the value of the cake. A partition procedure is *proportional* if it *always* produces proportional partitions.

Example 9A.1: 'I cut, you choose' is proportional.

Recall the 'I cut, you choose' procedure (Procedure 8B.1). Notice that $\mu_2(\mathbf{P}_2) \geq \frac{1}{2}$ by definition (steps **(2a)** and **(2b)**) and $\mu_1(\mathbf{P}_1) = \frac{1}{2}$ by step **(1)**. Thus, $\mathcal{P}$ will be a proportional partition. ♢

9B Banach and Knaster's 'Last Diminisher' game

Prerequisites: §9A

Is there a proportional partition procedure for more than two players? Yes.

Procedure 9B.1: (Banach & Knaster) [Kna46, Ste48a, Ste48b]

Let $\mathbf{X} = [0, 1]$ be the unit interval. Suppose $\mathcal{I} = \{1, \ldots, I\}$, and, for each $i \in [1..I]$, let i have a utility measure μ_i that is at most $\frac{1}{I}$ atomic. The Banach-Knaster procedure is defined recursively as follows:

1. If $\mathcal{I} = \{1, 2\}$ has only two players, then play the 'I cut, you choose' game (Game 8C.1)

2. Suppose $\mathcal{I}$ has $I \geq 3$ players. For each $i \in [1..I]$, let r_i be the largest value such that $\mu_i[1, r_i) = \frac{1}{I}$. In other words, $[1, r_i)$ is the largest piece of cake that i thinks is worth $\frac{1}{I}$ of the entire cake.

Claim 1: *Such an r_i exists.*

Proof: Let $R_i \in [0,1]$ be the smallest value such that $\mu_i[1, R_i) \geq \frac{1}{I}$. If $\mu_i[1, R_i) = \frac{1}{I}$, then we are done.

If not, then $\mu_i[1, R_i) > \frac{1}{I}$. Recall that μ_i is at most $\frac{1}{I}$ atomic. Thus $[0, R_i]$ cannot be entirely atomic. We can assume that any atoms in $[0, R_i]$ are clustered near 0, and not near R_i. (This can be achieved if necessary, but cutting $[0, R_i]$ into pieces and reordering them.). Hence we can assume that there is some $S_i < R_i$ so that $\mu_i[0, S_i] < \frac{1}{I}$, and so that all the atoms in $[0, R_i]$ are actually in $[0, S_i]$. Thus, $(S_i, R_i]$ contains *no* atoms. Now define $f : [S_i, R_i] \longrightarrow \mathbb{R}$ by $f(r) = \mu_i[0, r]$. Then f is continuous on $[S_i, R_i]$ (because there are no atoms). But $f(S_i) < \frac{1}{I} < f(R_i)$; hence the Intermediate Value Theorem yields some $r_i \in (S_i, R_i)$ with $f(r_i) = \frac{1}{I}$. ◇ Claim 1

Let i be the player with the smallest value of r_i (if two players are tied, then choose the smaller value of i). We define $\mathbf{P}_i = [0, r_i)$. Observe that $\mu_i(\mathbf{P}_i) = \frac{1}{I}$. (In other words, player i thinks she got $\frac{1}{I}$ of the cake.) Let $\mathbf{X}_1 = [r_i, 1]$ (i.e. $\mathbf{X}_1$ is the remaining cake).

Claim 2: *For every $j \neq i$, $\mu_j[\mathbf{X}_1] \geq \frac{I-1}{I}$.*

Proof: By hypothesis, $r_i \leq r_j$. Thus, $\mu_j[1, r_i) \leq \mu_j[1, r_j) = \frac{1}{I}$. Thus,

$$\mu_j[r_i, 1] \underset{\text{(UA)}}{=\!=} 1 - \mu_j[1, r_i) \quad \geq \quad 1 - \frac{1}{I} \quad = \quad \frac{I-1}{I}.$$

Here, **(UA)** is by axiom **(UA)** on page 165. ◇ Claim 2

Thus, each of the remaining players thinks that at least $\frac{I-1}{I}$ of the cake remains to be divided.

3. Now let $\mathcal{I}_1 = \mathcal{I} \setminus \{i\}$ (the remaining players). We apply the Banach-Knaster procedure recursively to divide $\mathbf{X}_1$ into $I-1$ slices such that each of the players in $j \in \mathcal{I}_1$ thinks she got a portion $\mathbf{P}_j$ such that

$$\mu_j[\mathbf{P}_j] \underset{\text{(S2)}}{\geq} \frac{1}{I-1} \cdot \mu_j[\mathbf{X}_1] \underset{\text{(C1)}}{\geq} \left(\frac{1}{I-1}\right) \cdot \left(\frac{I-1}{I}\right) = \frac{1}{I}. \tag{9.1}$$

where **(S2)** follows from step 2 of the procedure, and **(C1)** follows from Claim 1. ———

The Banach-Knaster partition is proportional, because of equation (9.1). Is there a game which yields the Banach-Knaster procedure? Yes.

Game 9B.2: 'Last Diminisher' (Banach & Knaster)

Let $\mathbf{X} = [0,1]$ be the unit interval, and let $\mathcal{I} = \{1, \ldots, I\}$.

1. Owen cuts a portion from the cake. In other words, Owen chooses some $r_1 \in [0,1]$ (the position of the 'cut').
2. Twyla then has the option (but is not obliged) to 'trim' this portion; i.e. to cut off a small slice and return it to the original cake. In other words, Twyla chooses some $r_2 \in [0,1]$; if $r_2 < r_1$, then Twyla is 'trims' the portion; if $r_2 \geq r_1$, then she leaves it alone.
3. Trey then has the option (but is not obliged) to 'trim' this new portion; i.e. to cut off a small slice and return it to the original cake. In other words, Trey chooses some $r_3 \in [0,1]$; if $r_3 < \min\{r_1, r_2\}$, then Trey 'trims' the portion; otherwise he leaves it alone.

4. The portion passes by each successive player in turn. Each has the option of trimming off a further slice
5. Once all I players have inspected the portion, the 'Last Diminisher' is the *last player who trimmed the portion* (or Owen, if no one else touched it).

   The 'Last Diminisher' receives this portion as *her* portion, and leaves the game.
6. The remaining $(I-1)$ players then repeat the game to divide up the remaining cake.

(Observe that, if there are only two players, then the 'Last Diminisher' game is equivalent to 'I cut, you choose').

Strictly speaking, the 'Last Diminisher' game is not a partition game. Instead of yielding an entire partition all at once, this game consists of a sequence of *apportionment games*, each of which yields a single portion for a single player. We need some machinery to make this precise.

Apportionment games: let $\mathcal{B}$ be the set of all subsets of $\mathbf{X}$ which *could* be a portion for some player[1]. An I-player *apportionment game* is a structure $\Gamma_I := (\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_I;\ \gamma)$, where $\mathcal{S}_i$ is some set of 'strategies' (i.e. 'moves') for player i, and where $\gamma : \mathcal{S}_1 \times \mathcal{S}_2 \times \cdots \times \mathcal{S}_I \longrightarrow \mathcal{B} \times [1..I]$.

Let $\mathbf{s} := (s_1, \ldots, s_I)$ be an I-tuple of strategies, and suppose $\gamma(\mathbf{s}) = (\mathbf{P}, i)$ for some $\mathbf{P} \subset \mathbf{X}$ and $i \in [1..I]$. This means that player i gets the portion $\mathbf{P}$, and the other $(I-1)$ players are left to divide the remaining cake $\mathbf{X} \setminus \mathbf{P}$.

Apportionment cascades: To complete the description, we must describe how the remaining players divide the remaining cake. An *apportionment cascade* is a sequence of apportionment games $\mathbf{\Gamma} := (\Gamma_I, \Gamma_{I-1}, \ldots, \Gamma_3, \Gamma_2)$, each of which awards a slice to *one* player, who then leaves the game (technically Γ_2 is a true *partition* game, since it only has two players, and implicitly assigns a portions to each of them). At the end of this sequence, each player has a portion, so, taken as a totality, the apportionment cascade $\mathbf{\Gamma}$ yields a partition game.

Payoffs and Strategies: To evaluate strategies in the apportionment cascade $\mathbf{\Gamma}$, we must define the $\mathbf{\Gamma}$-payoffs for each players. We do this inductively. First, we define the Γ_2-payoffs by treating it as standard partition game. We can then compute the maximin strategies and maximin payoffs for Γ_2.

Next, we move onto Γ_3. Clearly, if player i receives the portion $\mathbf{P}$ in the game Γ_3, then her Γ_3-payoff is just $\mu_i(\mathbf{P})$. We define the Γ_3-payoffs of all other players to be their *maximin payoffs* in the game Γ_2. Having defined the Γ_3-payoffs for all players, we can then compute the maximin strategies and maximin payoffs for Γ_3.

Inductively, assume we've computed the maximin payoffs for Γ_{I-1}. Consider Γ_I. Clearly, if player i receives the portion $\mathbf{P}$ in the game Γ_I, then her Γ_I-payoff is just $\mu_i(\mathbf{P})$. We define the Γ_I-payoffs of all other players to be their *maximin payoffs* in the game Γ_{I-1}.

[1]Technically, $\mathcal{B}$ is a sigma-algebra of *measurable* subsets, but we're neglecting this technicality

Proposition 9B.3 *If all players use their maximin strategies, then the Last Diminisher game yields the Banach-Knaster procedure.*

Proof: Let Γ_I be the 'Last Diminisher' game with I players.

Claim 1: *Owen's maximin Γ_I-strategy is to cut portion which he believes is exactly $\frac{1}{I}$ of the whole cake. In other words, Owen will choose $r_1 \in [0,1]$ so that $\mu_1[0,r_1] = \frac{1}{I}$.*

His maximin Γ_I-payoff with this strategy is $\frac{1}{I}$.

Proof: (by induction on I)

Base case ($I = 2$): In this case, we're playing 'I cut, you choose', and Example 8C.2 shows that Owen's maximin strategy is to choose r_1 so that $\mu_1[0,r_1] = \frac{1}{2}$. Regardless of how Twyla plays, we know that Owen will end up with a portion $\mathbf{P}_1$ such that $\mu_1[\mathbf{P}_1] = \frac{1}{2}$.

Induction: Suppose the claim is true for $I-1$ players. We first consider three strategies for Owen

`Strategy I:` (*Owen chooses r_1 so that $\mu_1[0,r_1] = \frac{1}{I}$.*)

In this case, either Owen gets this portion (a payoff of $\frac{1}{I}$), or someone else gets gets a 'trimmed' version of it. If someone else gets a trimmed version, then this recipient got a piece *smaller* than $\frac{1}{I}$. Hence, *more* than $\frac{I-1}{I}$ cake remains to be divided in Γ_{I-1}. In other words, $\mu_1[\mathbf{X}_1] > \frac{I-1}{I}$.

Now, Owen enters game Γ_{I-1} with $(I-1)$ other players. By induction hypothesis, Owen's maximin Γ_{I-1}-payoff will be:

$$\frac{1}{I-1}\,\mu_1[\mathbf{X}_1] \quad > \quad \left(\frac{1}{I-1}\right)\cdot\left(\frac{I-1}{I}\right) \quad = \quad \frac{1}{I}.$$

Thus, Owen's minimum Γ_I-payoff under `Strategy I` is $\frac{1}{I}$.

`Strategy II:` (*Owen chooses r_1 so that $\mu_1[0,r_1] > \frac{1}{I}$.*)

In this case, Owen runs the risk that someone else will receive this 'oversized' portion. But then the recipient gets *more* than $\frac{1}{I}$ of the cake, which means that *less* than $\frac{I-1}{I}$ cake remains during the next round of play, which will be played amongst $I-1$ players. In other words, $\mu_1[\mathbf{X}_1] < \frac{I-1}{I}$.

Now, Owen enters game Γ_{I-1}. By induction hypothesis, his maximin Γ_{I-1}-payoff will be:

$$\frac{1}{I-1}\,\mu_1[\mathbf{X}_1] \quad < \quad \left(\frac{1}{I-1}\right)\cdot\left(\frac{I-1}{I}\right) \quad = \quad \frac{1}{I}.$$

Thus, Owen's minimum Γ_I-payoff under `Strategy II` is less than $\frac{1}{I}$.

`Strategy III:` (*Owen chooses r_1 so that $\mu_1[0,r_1] < \frac{1}{I}$*)

In this case, Owen runs the risk of getting this 'undersized' portion $[0,r_1]$ if no one else trims it. Hence his minimum Γ_I-payoff is less than $\frac{1}{I}$.

Clearly, `Strategy I` yields the *best* minimum payoff, so this will be Owen's maximin strategy. ◇ `Claim 1`

Claim 2: *For any $i > 1$, Player i's maximin Γ_I-strategy is to trim the portion if and only if she thinks it is too large, and if so, to trim the portion until she believes it is exactly $\frac{1}{I}$ of the whole cake. In other words, i will choose $r_i \in [0,1]$ so that $\mu_1[0, r_i] = \frac{1}{I}$.*

Her maximin Γ_I-payoff with this strategy is $\frac{1}{I}$.

Proof: **Exercise 9.1** The proof is by induction on I, and is similar to Claim 1; the only difference is that, if $I = 2$, then Twyla takes the role of the 'chooser' in 'I cut, you choose.' ◇ Claim 2

Thus, assuming each player follows their maximin strategy, we can redescribe the Last Diminisher Game as follows:

1. Owen cuts from the cake a portion he believes to be of size $\frac{1}{I}$. In other words, Owen chooses some $r_1 \in [0,1]$ so that $\mu_1[0, r_1] = \frac{1}{I}$.
2. If Twyla thinks that this portion is too large, then she can 'trim' off a small slice (which is returned to the rest of the cake) so that she believes the new trimmed piece is exactly $\frac{1}{I}$ of the whole cake. If Twyla thinks the portion is not too large (or possibly to small), then she leaves it alone.

   That is, Twyla chooses some $r_2 \in [0,1]$ so that $\mu_2[0, r_2] = \frac{1}{I}$. If $r_2 < r_1$ she 'trims' the cake; otherwise she leaves it alone.
3. If Trey thinks that this new portion is too large, then he can 'trim' off a small slice (which is returned to the rest of the cake) so that he believes the new trimmed piece is exactly $\frac{1}{I}$ of the whole cake. If Trey thinks the portion is not too large (or possibly to small), then he leaves it alone.

   That is, Trey chooses some $r_3 \in [0,1]$ so that $\mu_3[0, r_3] = \frac{1}{I}$. If $r_3 < \min\{r_1, r_2\}$, then Trey 'trims' the cake; otherwise he leaves it alone.
4. The portion passes by each successive player in turn. Each has the option of trimming off a further slice, if she thinks the portion is too large.
5. The last individual to trim the cake (i.e. the player i with the smallest value of r_i) is the individual who gets this portion. That is, $\mathbf{P}_i = [1, r_i)$.
6. The rest of the players then play the game with the remaining cake.

Notice how, once we describe the player's maximin strategies, it is clear that the Last Diminisher Game yields the Banach-Knaster Procedure. □

9C The Latecomer Problem: Fink's 'Lone Chooser' game

Prerequisites: §8, §9A

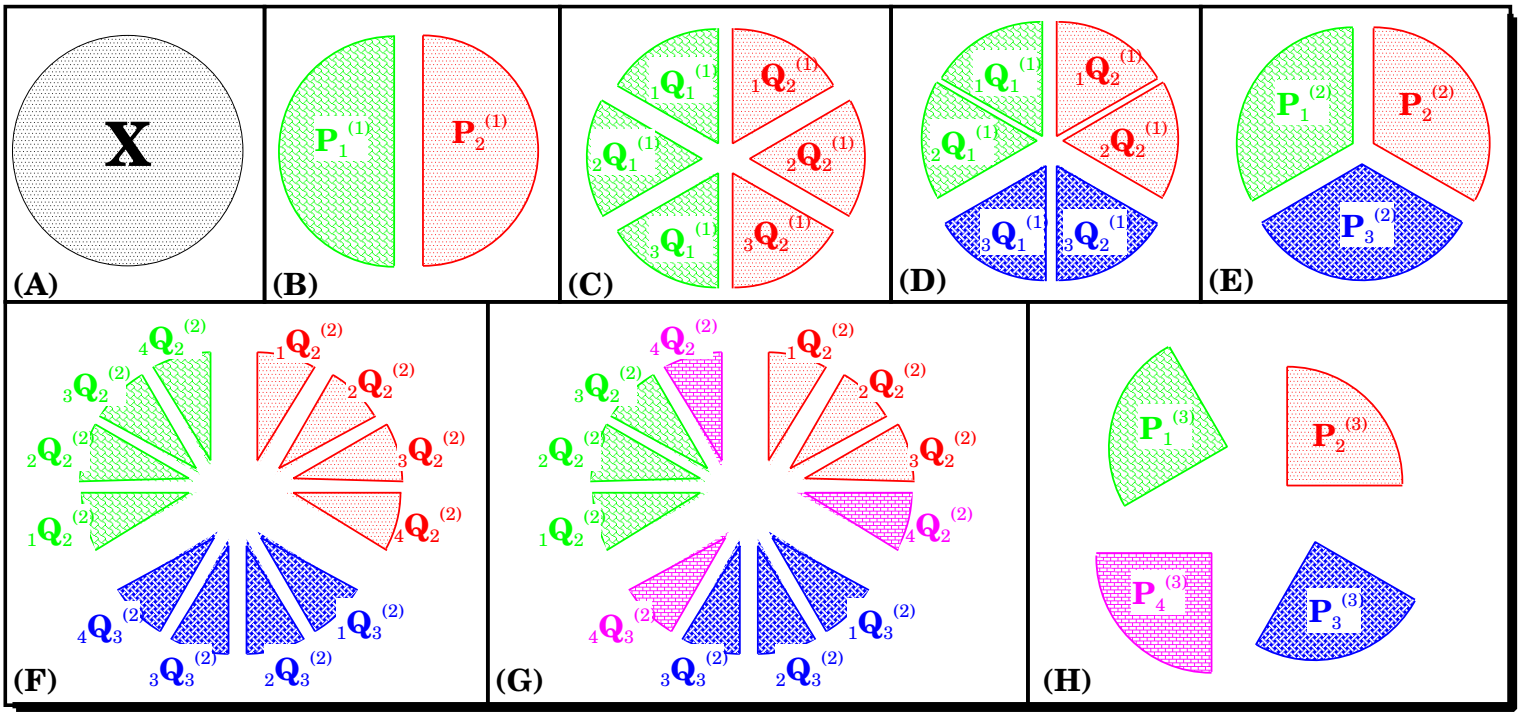


Figure 9.1: The Fink 'Lone Chooser' game.

What if three people have just finished dividing a cake into three fair portions, and suddenly a fourth individual shows up and wants a fair share? In 1964 A.M. Fink [Fin64] devised a proportional partition game which can easily accommodate 'late comers'. We will describe the game rules in normal font, and the maximin strategy in *italics*; the rules and the maximin strategies together yield the partition procedure.

Game 9C.1: 'Lone Chooser' (A.M. Fink)

Suppose $\mathcal{I} = \{1, 2, \ldots, I\}$. Refer to Figure 9.1 for each step of the algorithm:

(A) We begin with a cake $\mathbf{X}$. It is convenient (but not essential) to imagine $\mathbf{X}$ as a disk.

(B) Owen and Twyla use 'I cut, you choose' to split the cake into $\mathbf{P}_1^{(1)}$ and $\mathbf{P}_2^{(1)}$.
Maximin strategy yields $\mu_1[\mathbf{P}_1^{(1)}] = \frac{1}{2}$ *and* $\mu_2[\mathbf{P}_2^{(1)}] \geq \frac{1}{2}$.

(C) Owen trisects his portion $\mathbf{P}_1^{(1)}$ into three equal parts ${}_1\mathbf{Q}_1^{(1)}$, ${}_2\mathbf{Q}_1^{(1)}$, and ${}_3\mathbf{Q}_1^{(1)}$.
Maximin strategy: $\mu_1[{}_1\mathbf{Q}_1^{(1)}] = \mu_1[{}_2\mathbf{Q}_1^{(1)}] = \mu_1[{}_3\mathbf{Q}_1^{(1)}] = \frac{1}{6}$.
Likewise Twyla trisects her portion $\mathbf{P}_2^{(1)}$ into three equal parts ${}_1\mathbf{Q}_2^{(1)}$, ${}_2\mathbf{Q}_2^{(1)}$, and ${}_3\mathbf{Q}_2^{(1)}$.
Maximin strategy: $\mu_2[{}_1\mathbf{Q}_2^{(1)}] = \mu_2[{}_2\mathbf{Q}_2^{(1)}] = \mu_2[{}_3\mathbf{Q}_2^{(1)}] = \frac{1}{3}\mu_2[\mathbf{P}_2^{(1)}] \geq \frac{1}{6}$.

(D) Trey chooses one of ${}_1\mathbf{Q}_1^{(1)}$, ${}_2\mathbf{Q}_1^{(1)}$, or ${}_3\mathbf{Q}_1^{(1)}$, and one of ${}_1\mathbf{Q}_2^{(1)}$, ${}_2\mathbf{Q}_2^{(1)}$, or ${}_3\mathbf{Q}_2^{(1)}$.
Maximin strategy: Choose ${}_j\mathbf{Q}_1^{(1)}$, *and* ${}_k\mathbf{Q}_2^{(1)}$ *so as to maximize* $\mu_3[{}_j\mathbf{Q}_1^{(1)} \sqcup {}_k\mathbf{Q}_2^{(1)}]$.

(E) Assume without loss of generality that Trey chooses ${}_3\mathbf{Q}_1^{(1)}$, and ${}_3\mathbf{Q}_2^{(1)}$. At this point,

- Owen has $\mathbf{P}_1^{(2)} = {}_1\mathbf{Q}_1^{(1)} \sqcup {}_2\mathbf{Q}_1^{(1)}$.
- Twyla has $\mathbf{P}_2^{(2)} = {}_1\mathbf{Q}_2^{(1)} \sqcup {}_2\mathbf{Q}_2^{(1)}$.
- Trey has $\mathbf{P}_3^{(2)} = {}_3\mathbf{Q}_1^{(1)} \sqcup {}_3\mathbf{Q}_2^{(1)}$.

Maximin outcome: $\mu_1[\mathbf{P}_1^{(2)}] = \frac{1}{3}$, $\mu_2[\mathbf{P}_2^{(2)}] \geq \frac{1}{3}$, *and* $\mu_3[\mathbf{P}_3^{(2)}] \geq \frac{1}{3}$. *Thus, each player has (in her estimation) at least one third of the cake.*

Now we introduce Ford.

(F) For each $i \in [1..3]$, player i quadrisects his/her portion $\mathbf{P}_i^{(2)}$ into four equal parts ${}_1\mathbf{Q}_i^{(2)}, \ldots, {}_4\mathbf{Q}_i^{(2)}$.
Maximin strategy: $\mu_i[{}_1\mathbf{Q}_i^{(2)}] = \cdots = \mu_1[{}_4\mathbf{Q}_i^{(2)}] = \frac{1}{4}\mu_i[\mathbf{P}_i^{(2)}]$.

(G) Ford then chooses one of ${}_1\mathbf{Q}_1^{(2)}, \ldots, {}_4\mathbf{Q}_1^{(2)}$, one of ${}_1\mathbf{Q}_2^{(2)}, \ldots, {}_4\mathbf{Q}_2^{(2)}$, and one of ${}_1\mathbf{Q}_3^{(2)}, \ldots, {}_4\mathbf{Q}_3^{(2)}$.
Maximin strategy: For each $i \in [1..3]$, *choose* ${}_{j_i}\mathbf{Q}_i^{(2)}$ *so as to maximize* $\mu_4[{}_{j_i}\mathbf{Q}_i^{(2)}]$.

(H) Assume without loss of generality that Ford chooses ${}_4\mathbf{Q}_1^{(2)}$, ${}_4\mathbf{Q}_2^{(2)}$, and ${}_4\mathbf{Q}_3^{(2)}$. At this point,

- Owen has $\mathbf{P}_1^{(3)} = {}_1\mathbf{Q}_1^{(2)} \sqcup {}_2\mathbf{Q}_1^{(2)} \sqcup {}_3\mathbf{Q}_1^{(2)}$.
- Twyla has $\mathbf{P}_2^{(3)} = {}_1\mathbf{Q}_2^{(2)} \sqcup {}_2\mathbf{Q}_2^{(2)} \sqcup {}_3\mathbf{Q}_2^{(2)}$.
- Trey has $\mathbf{P}_3^{(3)} = {}_3\mathbf{Q}_3^{(2)} \sqcup {}_3\mathbf{Q}_3^{(2)} \sqcup {}_3\mathbf{Q}_3^{(2)}$.
- Ford has $\mathbf{P}_4^{(3)} = {}_4\mathbf{Q}_1^{(2)} \sqcup {}_4\mathbf{Q}_2^{(2)} \sqcup {}_4\mathbf{Q}_3^{(2)}$.

Maximin outcome: $\mu_1[\mathbf{P}_1^{(3)}] = \frac{1}{4}$, $\mu_2[\mathbf{P}_2^{(2)}] \geq \frac{1}{4}$, $\mu_3[\mathbf{P}_3^{(2)}] \geq \frac{1}{4}$, *and* $\mu_4[\mathbf{P}_4^{(2)}] \geq \frac{1}{4}$. *Thus, each player has (in his/her estimation) at least one fourth of the cake.*

A fifth player is dealt with similarly. Proceed inductively. ———————

Exercise 9.2 Verify that the maximin strategies and payoffs for 'Lone Chooser' are as described in above.

Exercise 9.3 One issue with partitioning games is the number of cuts.

1. Show that the I-player Banach-Knaster 'Last Diminisher' game requires *at most* $\frac{I(I-1)}{2}$ cuts.
2. Show that Fink's 'Lone Chooser' game for I players *always* requires $I!$ cuts.

Further reading: The problem of fairly dividing a cake amongst three people was first considered by Steinhaus in 1943, who developed a three-individual game called 'Lone Divider' [Ste48a, Ste48b, Kuh]. Steinhaus was unable to extend her solution to more than three players; this problem was solved by her students, Stefan Banach and Bronislaw Knaster, with their 'Last Diminisher' game [Kna46, Ste48a, Ste48b]. Woodall [Woo86] has modified Fink's scheme so that each of I players is guaranteed *strictly more* than $\frac{1}{I}$ of the cake (in her estimation). Austin [Aus82] has modified Fink's scheme so that each player thinks she gets *exactly* $\frac{1}{I}$ of the cake. A thorough discussion of all these proportional partition procedures is given in Chapters 2 and 3 of Brams and Taylor [BT96].

9D Symmetry: Dubins and Spanier's 'Moving Knife' game

Prerequisites: §9B

One objection to 'I cut, you choose' (Game 8C.1) is that the 'chooser' player has a clear advantage. If 'cutter' plays his maximin strategy, he will be guaranteed *exactly* half the cake (in his perception). However, 'chooser' will be guaranteed *at least* half the cake (in her perception), and possibly much more, if her perception differs sufficiently from that of 'cutter'.

The same objection applies to the Banach-Knaster 'Last Diminisher' game (Game 9B.2), because 'Last Diminisher' reduces to 'I cut, you choose' once the number of players is reduced to two. For example, if five people play 'Last Diminisher', then the individual who ends up being 'chooser' will again be guaranteed *at least* one fifth of the cake, which gives her an advantage over not only 'cutter', but also the three 'diminishers' who have already left the game, each of whom received *exactly* one fifth (in their estimations).

To remove this asymmetry, Dubins and Spanier proposed a 'symmetric' form of the Banach-Knaster procedure, where no specific player gets to be 'chooser', or is forced into the role of 'cutter'. The Dubins-Spanier procedure looks very similar to Banach-Knaster, except that we do *not* resort to 'I cut, you choose' in the base case.

Procedure 9D.1: Dubins & Spanier [DS61]

Again suppose $\mathbf{X} = [0, 1]$ is the unit interval. Suppose $\mathcal{I} = \{1, \ldots, I\}$, and let i have nonatomic utility measure μ_i. The Dubins-Spanier procedure is defined recursively as follows:

1. Suppose $I \geq 2$. For each $i \in [1..I]$, let r_i be the largest value such that $\mu_i[1, r_i) = \frac{1}{I}$. In other words, $[1, r_i)$ is the largest piece of cake that i thinks is $\frac{1}{I}$ of the entire cake (such an r_i exists because μ_i is nonatomic).

Let i be the player with the smallest value of r_i (if two players are tied, then choose the smaller value of i). We define $\mathbf{P}_i = [0, r_i)$. Observe that $\mu_i(\mathbf{P}_i) = \frac{1}{I}$. (In other words, player i thinks she got $\frac{1}{I}$ of the cake.) Let $\mathbf{X}_1 = [r_i, 1]$ (i.e. $\mathbf{X}_1$ is the remaining cake).

As in the Banach-Knaster Procedure (Procedure 9B.1), we have:

Claim 1: *For every* $j \neq i$, $\mu_j[\mathbf{X}_1] \geq \frac{I-1}{I}$.

Thus, each of the remaining players thinks that at least $\frac{I-1}{I}$ of the cake remains to be divided.

2. Now let $\mathcal{I}_1 = \mathcal{I} \setminus \{i\}$ (the remaining players). We apply the Dubins-Spanier procedure recursively to divide $\mathbf{X}_1$ into $I - 1$ slices such that each of the players in $j \in \mathcal{I}_1$ thinks she got a portion $\mathbf{P}_j$ such that

$$\mu_j[\mathbf{P}_j] \underset{\text{(S1)}}{\geq} \frac{1}{I-1} \cdot \mu_j[\mathbf{X}_1] \underset{\text{(C1)}}{\geq} \left(\frac{1}{I-1}\right) \cdot \left(\frac{I-1}{I}\right) = \frac{1}{I}.$$

where **(S1)** follows from step 1 of the procedure, and **(C1)** follows from Claim 1. ∎

In what sense is the Dubins-Spanier procedure symmetric? Let $\sigma : \mathcal{I} \longrightarrow \mathcal{I}$ be a permutation. If $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_I) \in \mathfrak{M}^I$ is an I-tuple of utility measures, let $\sigma(\boldsymbol{\mu}) = (\mu_{\sigma(1)}, \ldots, \mu_{\sigma(I)})$. (Heuristically speaking, the players are rearranged in a different order). We say that a partition procedure Π is *symmetric* if the following is true: Let $\boldsymbol{\mu} \in \mathfrak{M}^I$ and let $\Pi(\boldsymbol{\mu}) = \mathcal{P}$. Let $\sigma : \mathcal{I} \longrightarrow \mathcal{I}$ be any permutation, and let $\Pi(\sigma(\boldsymbol{\mu})) = \mathcal{Q}$. Then for all $i \in [1..I]$, $\mu_i(\mathbf{P}_i) = \mu_{\sigma(i)}(\mathbf{Q}_{\sigma(i)})$. In other words, if we reorder the players and then apply the procedure, then each player receives a portion which she thinks is the same *size* (although it might not actually be the same *portion*) as she received before reordering.

Lemma 9D.2 *The Dubins-Spanier procedure is symmetric*

Proof: **Exercise 9.4** □

In this sense, the Dubins-Spanier procedure is more 'fair' than Banach-Knaster, because the 'last' player has no advantage over the 'first' player. Is there a partition game which yields the Dubins-Spanier procedure? Yes.

Game 9D.3: 'Moving Knife' (Dubins & Spanier)

Suppose $\mathbf{X} = [0, 1]$ is the unit interval.

1. A 'referee' (who could be one of the players) takes a knife, and places it at the left end of the cake (i.e. at 0).

2. The referee then very slowly moves the knife rightwards across the cake. (from 0 to 1).

3. At any moment, any player can say 'cut'. Say Owen says 'cut', then he receives the piece to the left of the knife and exits the game (i.e. if the knife is at $k \in [0, 1]$, then Owen gets the portion $[0, k]$.).

4. The rest of the players then continue playing with the remaining cake. ———————

Notice that 'Moving Knife' is an *apportionment game*, like 'Last Diminisher'. Thus, to create a real *partitioning game*, we actually need to arrange a sequence of 'Moving Knife' games in an *apportionment cascade*. However, 'Moving Knife' is different from 'Last Diminisher' in one important respect. 'Last Diminisher' was a *sequential* game, where the players played one at a time. In contrast, 'Moving Knife' is a *simultaneous* game, where all players must choose and execute their strategies *simultaneously*, each in ignorance of the choices made by the others. In 'Moving Knife', each player's 'strategy' is her choice of the moment when she will say 'cut' (assuming no one else has said it first).

In a simultaneous game, *all* players are in the same strategic position that Owen had in a sequential game. We can thus define 'dominant' and 'maximin' strategies for all players of a simultaneous game in a fashion analogous to the dominant and maximin strategies for Owen in a sequential game (which simplifies analysis).

Recall: in an apportionment game Γ_I, the Γ_I-payoff of the 'winner' i of portion $\mathbf{P}$ is the value of $\mu_i(\mathbf{P})$, while the Γ_I-payoffs of all the *other* players are their maximin payoffs in Γ_{I-1}.

Proposition 9D.4 *The Moving Knife game yields the Dubins-Spanier procedure.*

Proof: Let Γ_I be the 'Moving Knife' game with I players. It suffices to show:

Claim 1: *For any player i, let $r_i \in [0,1]$ be the largest value such that $\mu_i[1, r_i) \leq \frac{1}{I}$. Then i's maximin Γ_I-strategy is to say 'cut' exactly when the knife is at r_i (unless some other player says 'cut' first).*

Her maximin Γ_I-payoff under this strategy is $\frac{1}{I}$.

Proof:

We proceed by induction on I. The logic is very similar to the proof of Proposition 9B.3 (concerning the Banach-Knaster game). Hence, here we will be more sketchy.

Base Case: $(I = 2)$ **Exercise 9.5**.

Induction: If i says 'cut' *before* the knife reaches r_i, then she will end up with a piece which is she *definitely* thinks is less than $\frac{1}{I}$.

If i waits until *after* the knife reaches r_i, then another player might say 'cut' first. Then this *other* player will then get a portion which i believes is *more* than $\frac{1}{I}$ of the cake. Then i enters game Γ_{I-1} with $(I-1)$ other players. By induction, her maximin payoff in Γ_{I-1} now $\frac{1}{I-1}$ of the remaining cake, which she believes is strictly *less* than $\frac{I-1}{I}$. Thus, her maximin Γ_{I-1}-payoff is *less* than $\left(\frac{1}{I-1}\right)\left(\frac{I-1}{I}\right) = \frac{1}{I}$. Hence, her minimum Γ_I-payoff is less than $\frac{1}{I}$.

If i says cut *exactly* at r_i, then she gets a piece of size $\frac{1}{I}$, so her minimum payoff is exactly $\frac{1}{I}$. This strategy yields a *higher* minimum payoff then either of the other two strategies, so this is i's *maximin* strategy in Γ_I. ◇ Claim 1

It follows from Claim 1 that we can expect the player with the smallest value of r_i to be 'honest', and say 'cut' exactly when the knife is at r_i. This player receives the portion $[0, r_i)$, and we continue playing. Thus, step **1** of the Dubins-Spanier procedure is implemented.

The rest of the players now play the game Γ_{I-1}; this implements step **2**. □

If $I = 2$, then the Dubins-Spanier game is called *Austin's single moving knife procedure*, and works as follows: the referee moves the knife across the cake until either of the two players says 'cut'. The player who calls 'cut' gets the left-hand portion, and the other player gets the right-hand portion. By the **Base Case** of Claim 1 above, we see that this game yields a proportional partition.

9E Connectivity: Hill and Beck's Fair Border Procedure

Prerequisites: §8, §9A; Basic Topology (see appendix)

The partitioning procedures of Banach-Knaster (§9B) and Fink (§9C) made no reference to the *topology* of the portions. We *represented* the 'cake' $\mathbf{X}$ with the unit interval $[0,1]$, but the two procedures don't really depend on the linear topology of $[0,1]$, and are equally applicable to subsets of $\mathbb{R}^I$ (or even abstract measure spaces). This is 'topological insensitivity' is acceptable and even helpful, if the real 'cake' does not have any topological structure (e.g. the 'cake' is a set of items to be divided in an inheritance). If the cake is a dessert, these procedures may leave each player with a plate full of chopped up bits, which is suboptimal but at least edible. However, if the 'cake' is a physical piece of land (e.g. a contested property in an inheritance; a disputed territory in a military conflict), then disconnected portions will be unacceptable.

The Dubins-Spanier 'Moving Knife' procedure (§9D) *does* use the topology of $[0,1]$, and guarantees each player a connected subinterval as a portion. The Dubins-Spanier procedure can easily be adapted to higher dimensional spaces such as a two-dimensional territory (we can imagine a metaphorical knife sweeping slowly across a map of the territory until some player says 'stop'). However, the resulting portions may not be connected; if the territory has a complicated boundary (e.g. a coastline), then the 'cuts' made by the knife may intersect this boundary in several places, leaving each player with several disconnected components.

There are economic, logistical, and strategic reasons for desiring *connected* territory. To ensure future peaceful coexistence, we need a procedure which yields a proportional partition with connected portions. In 1983, Theodore P. Hill [Hil83] proved that such a partition existed, but he gave no construction procedure. In 1987, Anatole Beck [Bec87] provided a partition procedure which yields a stronger version of Hill's theorem. The Hill-Beck solution involves some basic planar topology, which is reviewed in the *Appendix on Topology* at the end of this section (page 185).

Theorem 9E.1 (Hill & Beck)

Let $\mathbf{X} \subset \mathbb{R}^2$ be an open connected region in the plane (the disputed territory). *Let $\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_I \subset \partial\mathbf{X}$ be disjoint connected subsets of the boundary such that*

$$\partial\mathbf{X} \quad = \quad \mathbf{B}_1 \sqcup \mathbf{B}_2 \sqcup \cdots \sqcup \mathbf{B}_I \qquad \text{(Figure 9.2A)}$$

($\mathbf{B}_i$ is the 'border' between $\mathbf{X}$ and some adjacent country i which wants a piece of $\mathbf{X}$).

Let $\mu_1, \ldots, \mu_I$ be nonatomic utility measures on $\mathbf{X}$. Then there exists a proportional partition $\mathcal{P} = \{\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_I\}$ of $\mathbf{X}$ such that (as shown in Figure 9.2B), for each $i \in [1..I]$

- $\mathbf{P}_i$ *is connected*
- $\mathbf{B}_i \subset \mathbf{P}_i$ (i.e. $\mathbf{P}_i$ is connected to the country i).
- $\mu_i[\mathbf{P}_i] \ \geq \ \frac{1}{I}$. □

Beck provides a partition game which yields the partition described in this theorem. The partition game is a cascade of apportionment games, each of which gives a portion of territory

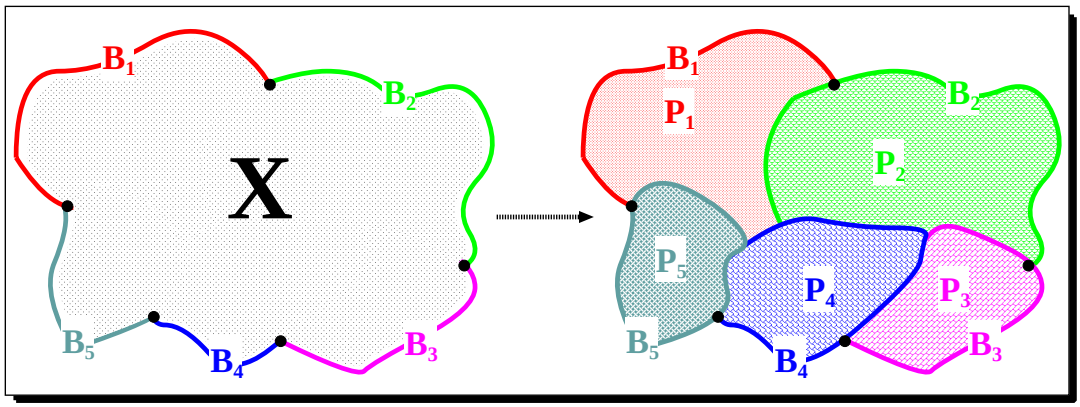


Figure 9.2: The Hill-Beck Theorem

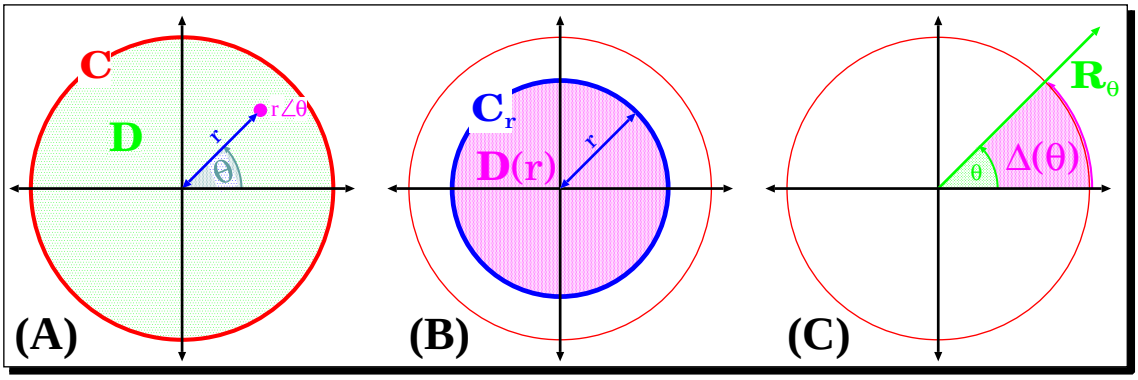


Figure 9.3: The unit disk **D** and subsets.

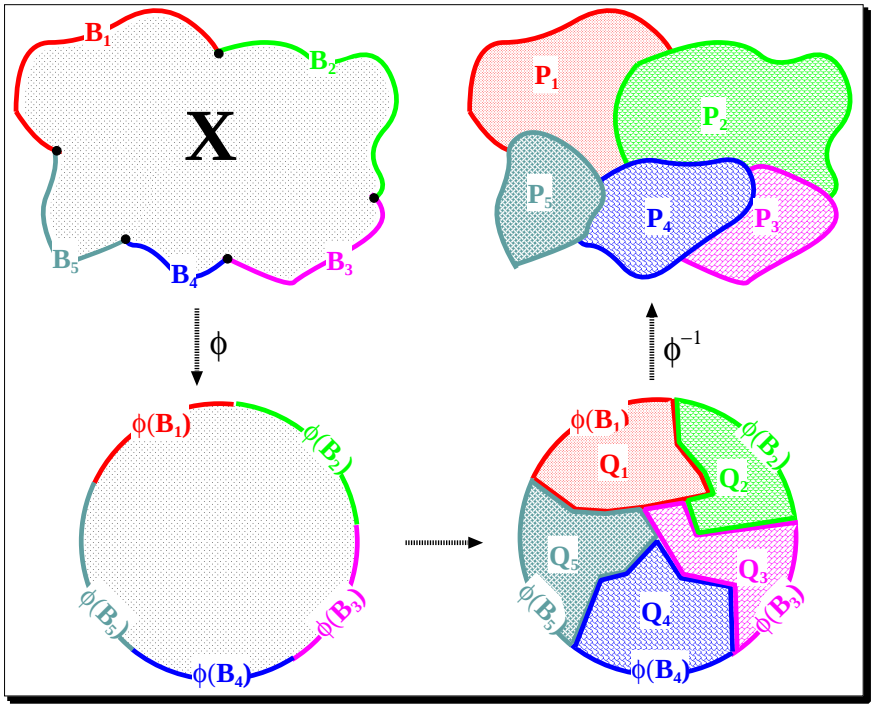


Figure 9.4: Representing the partition problem on the unit disk.

to one player, who then leaves the game. First we need some notation.

$$\begin{aligned}
\text{Let } \mathbf{D} &:= \left\{\mathbf{x} \in \mathbb{R}^2 \,;\, |\mathbf{x}| \leq 1\right\} && \text{be the unit disk. (Fig.9.3A)}\\
\text{Let } \mathbf{C} &:= \left\{\mathbf{x} \in \mathbb{R}^2 \,;\, |\mathbf{x}| = 1\right\} && \text{be the unit circle. (Fig.9.3A)}\\
\text{For each } r \in [0,1], \text{ let } \mathbf{D}(r) &:= \{\mathbf{x} \in \mathbf{D} \,;\, |\mathbf{x}| \leq r\} && \text{be the disk of radius } r. \text{ (Fig.9.3B)}\\
\text{and let } \mathbf{C}_r &:= \{\mathbf{x} \in \mathbf{D} \,;\, |\mathbf{x}| = r\} && \text{be the circle of radius } r. \text{ (Fig.9.3B)}\\
\text{If } r \geq 0 \text{ and } \theta \in [0, 2\pi], \text{ let } r\angle\theta &:= (r\cos(\theta), r\sin(\theta)) && \text{be the point with polar}\\
& && \text{coordinates } r \text{ and } \theta. \text{ (Fig.9.3A)}\\
\text{For each } \theta \in [0, 2\pi], \text{ let } \mathbf{R}_\theta &:= \{r\angle\theta \,;\, r \in [0,1]\} && \text{be the ray at angle } \theta. \text{ (Fig.9.3C)}\\
\text{For any } \Theta \in [0, 2\pi], \text{ let } \Delta(\Theta) &:= \{r\angle\theta \,;\, r \in [0,1],\ \theta \in [0, \Theta)\} && \\
& && \text{be the wedge between angles 0 and } \Theta. \text{ (Fig.9.3C)}
\end{aligned}$$

Beck's apportionment game assumes that the territory $\mathbf{X}$ is the unit disk $\mathbf{D}$, so that $\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_I$ are subsets of the unit circle $\mathbf{C}$ (because $\partial\mathbf{D} = \mathbf{C}$). This assumption causes no loss of generality, because Beck shows that we can accurately represent any territorial dispute using an analogous territorial dispute on the unit disk, as shown in Figure 9.4, and described by the next lemma:

Lemma 9E.2 *Let $\mathbf{X} \subset \mathbb{R}^2$ be an open, simply connected region, and let $\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_I \subset \partial\mathbf{X}$ be disjoint connected sets such that $\partial\mathbf{X} = \mathbf{B}_1 \sqcup \mathbf{B}_2 \sqcup \cdots \sqcup \mathbf{B}_I$. Let $\mu_1, \ldots, \mu_I$ be nonatomic utility measures on $\mathbf{X}$.*

There exists a homeomorphism $\phi : \mathbf{X} \longrightarrow \mathbf{D}$ such that, if $\nu_1 = \phi(\mu_1), \ldots, \nu_I = \phi(\mu_I)$, then for all $i \in [1..I]$,

$\nu_i(\mathbf{R}_\theta) = 0$, *for all* $\theta \in [0, 2\pi]$, *and* $\nu_i(\mathbf{C}_1) = 0$, *for all* $r \in [0,1]$. □

Now, if we can achieve a partition $\mathcal{Q} = \{\mathbf{Q}_1, \ldots, \mathbf{Q}_i\}$ of $\mathbf{D}$ which is connected and proportional with respect to the measures $\nu_1, \ldots, \nu_I$, and we define $\mathbf{P}_i = \phi^{-1}(\mathbf{Q}_i)$ for all $i \in [1..I]$ (see Fig.9.4), then Proposition 9E.5 of the *Appendix* says that $\mathcal{P} = \{\mathbf{P}_1, \ldots, \mathbf{P}_i\}$ is a partition of $\mathbf{X}$ which is connected and proportional with respect to the measures $\mu_1, \ldots, \mu_I$. Thus, it suffices to find a proportional, connected partition of the disk $\mathbf{D}$.

To do this, we use a procedure similar to the Dubins-Spanier Moving Knife, only now the 'knife blade' is a *circle* of varying radius.

1. Each player submits a 'bid' for the largest radius r such that the disk $\mathbf{D}_r$ has mass $\frac{1}{I}$ in that player's estimation. In other words, for all $i \in [1..I]$, we define r_i by

$$r_i = \max\left\{r \in [0,1] \,;\, \mu_i\left[\mathbf{D}(r)\right] = \frac{1}{I}\right\}.$$

Claim 1: *Such an r_i exists.*

Proof: Define $f_i : [0,1] \longrightarrow [0,1]$ by $f_i(r) = \mu_i[\mathbf{D}(r)]$. Then f_i is a *continuous* function of r, because Lemma 9E.2 says $\mu_i[\mathbf{C}_r] = 0$ for all r. Observe that $f_i(0) = 0$ and $f_i(1) = 1$. Thus, the Intermediate Value Theorem yields some r with $f_i(r) = \frac{1}{I}$. ◇ Claim 1

2. Consider the player(s) whose value(s) of r_i are minimal. There are two cases:

   Case 1: There is a *unique* player whose value of r_i is minimal (highly probable)

   Case 2: There are several players whose values of r_i are equal and all minimal (highly improbable, but still theoretically possible).

   We deal with these cases separately.

   Case 1: Assume without loss of generality that r_1 is minimal (if necessary, permute the players to achieve this). Also assume without loss of generality that $\mathbf{B}_1$ is the arc of the circle $\mathbf{C}$ between angles 0 and Θ^*:

$$\mathbf{B}_1 \quad = \quad \{1\angle\theta \; ; \; \theta \in [0, \Theta^*)\}.$$

   (if necessary, rotate the disk to achieve this). Thus, to connect the disk $\mathbf{D}(r_1)$ to her territory, Owen must define some sort of 'corridor' from $\mathbf{D}(r_1)$ to $\mathbf{B}_1$. This corridor will take the form of a 'wedge'. For each $i \in [2..I]$, let i propose some $\Theta_i \in [0, \Theta^*]$ so that

$$\mu_i \Big[\mathbf{D}(r_1) \cup \Delta(\Theta_i)\Big] \quad = \quad \frac{1}{I}. \tag{9.2}$$

   Claim 2: *Such a Θ_i exists.*

   Proof: Define $f_i : [0, 2\pi] \longrightarrow [0,1]$ by $f_i(\theta) = \mu_i \left[\mathbf{D}(r) \cup \Delta(\theta)\right]$. Then f_i is *continuous* as a function of θ, because Lemma 9E.2 says $\mu_i[\mathbf{R}_\theta] = 0$ for all θ. Now, $f_i(0) = \mu_i \left[\mathbf{D}(r)\right] < \frac{1}{I}$ (by hypothesis), while $f_i(2\pi) = 1$. Thus, the Intermediate Value Theorem yields some θ such that $f_i(\theta) = \frac{1}{I}$. ◇ Claim 2

   Now let $\Theta = \min_{i \in [2..I]} \Theta_i$. Define $\mathbf{P}_1 \; = \; \mathbf{D}(r_1) \cup \Delta(\Theta)$, as shown in Figure 9.5.

   Case 2: This case is resolved through a sequence of 'subsidiary auctions'. We'll only sketch the idea here.

   Assume without loss of generality that there is some $J \leq I$ so that players $1, \ldots, J$ all tied for the minimum bid in the first auction. Then we hold a second auction, where each of $1, \ldots, I$ submits a 'bid' for the *smallest* radius r such that the disk $\mathbf{D}(r)$ has mass $\frac{1}{I}$ in that player's estimation. In other words, for all $j \in [1..J]$, we define $r'_j <$ by

$$r'_j \quad = \quad \min \left\{ r \in [0,1] \; ; \; \mu_j \left[\mathbf{D}(r)\right] = \frac{1}{I} \right\}.$$

   Now, if there is a unique minimal bid in this auction (say, r'_1), then we move onto the 'wedge' procedure from **Case 1**, and give Owen the disk $\mathbf{D}(r'_1)$ plus a wedge-shaped corridor connecting $\mathbf{D}(r'_1)$ to her territory. If there is also a tie in the second

auction, then Beck introduces a complex procedure to resolve the tie; the details are in [Bec87].

3. Assume without loss of generality that Owen won the auction(s), and was awarded portion $\mathbf{P}_1$. Owen exits the game. Let $\mathbf{X}_1 = \mathbf{D} \setminus \mathbf{P}_1$, as shown in Figure 9.5. Then $\mathbf{X}_1$ is open and simply connected, and $\mu_i[\mathbf{X}_1] \geq \frac{I-1}{I}$ for each $i \in [2..I]$. The remaining players repeat the game to divide $\mathbf{X}_1$.

The Beck procedure provides an elegant constructive proof of Hill's 'existence' theorem. However, it is not clear that Beck's procedure could resolve a real-world territorial dispute between two or more adversarial parties, for two reasons:

- The Hill-Beck theorem specifically requires that the utility measures $\mu_1, \ldots, \mu_I$ be *non-atomic* (this is necessary for the proof of Lemma 9E.2). But in real-world territorial conflicts, nontrivial value is often concentrated at a single indivisible entity (e.g. a city, an oil well, a gold mine).

- Beck's partition *procedure* does not lend itself to a partition *game* because there is no incentive for players to bid honestly. It's true that the players will bid honestly for the 'minimal radius disk' $\mathbf{D}(r)$ in **Step 1** (for the same reason that we expect honest bidding in Banach-Knaster or in Dubins-Spanier). However, consider the construction of the 'wedge' corridor $\Delta(\Theta)$ in **Case 1** of **Step 2**. In this stage, players $2, \ldots, i$ have no incentive to provide honest bids satisfying equation (9.2); instead, they all have an incentive to make the wedge *as small as possible*. It's certainly true that, as long the wedge $\Delta(\Theta)$ has nonzero width, Owen will think that she has obtained 'more' than her fair share, because $\mu_1[\mathbf{P}_1] > \frac{1}{I}$. However, rival countries can 'shave' this wedge to be so thin that it is useless for practical purposes. If $\Delta(\Theta)$ cuts across rugged terrain, it can so thin that it becomes impassable. If $\Delta(\Theta)$ is flanked by enemies, it can be made so thin that it is indefensible.

9E.1 Appendix on Topology

Let $\mathbf{X} \subset \mathbb{R}^2$ and $y \in \mathbb{R}^2$, we say that y is **adjacent** to $\mathbf{X}$ if there is a sequence of points $\{x_i\}_{i=1}^{\infty} \subset \mathbf{X}$ such that $\lim_{i \to \infty} x_i = y$ (Figure 9.6A) . If $\mathbf{P}, \mathbf{Q} \subset \mathbb{R}^2$ are disjoint subsets, we say that $\mathbf{P}$ and $\mathbf{Q}$ are *adjacent* if either $\mathbf{P}$ contains a point adjacent to $\mathbf{Q}$ or $\mathbf{Q}$ contains a point adjacent to $\mathbf{P}$ (Figure 9.6B). A set $\mathbf{X} \subset \mathbb{R}^2$ is *connected* if, whenever we partition $\mathbf{X}$ into two disjoint subsets $\mathbf{P}$ and $\mathbf{Q}$ so that $\mathbf{X} = \mathbf{P} \sqcup \mathbf{Q}$, we find that $\mathbf{P}$ and $\mathbf{Q}$ are adjacent (Figure 9.6C and 9.6D).

If $\mathbf{X} \subset \mathbb{R}^2$, then the *complement* of $\mathbf{X}$ is the set $\mathbf{X}^{\complement} := \mathbb{R}^2 \setminus \mathbf{X}$ (Figure 9.6E). The *boundary* of $\mathbf{X}$ is the set $\partial\mathbf{X}$ of all points which are both adjacent to $\mathbf{X}$ and adjacent to $\mathbf{X}^{\complement}$ (Figure 9.6F). For example, if $\mathbf{D} := \{\mathbf{x} \in \mathbb{R}^2 \,;\, |\mathbf{x}| \leq 1\}$ is the *unit disk* (Figure 9.6G) , then $\mathbf{D}^{\complement} = \{\mathbf{x} \in \mathbb{R}^2 \,;\, |\mathbf{x}| > 1\}$. If $\mathbf{C} := \{\mathbf{x} \in \mathbb{R}^2 \,;\, |\mathbf{x}| = 1\}$ is the *unit circle* (Figure 9.6H), then $\partial\mathbf{D} = \mathbf{C}$.

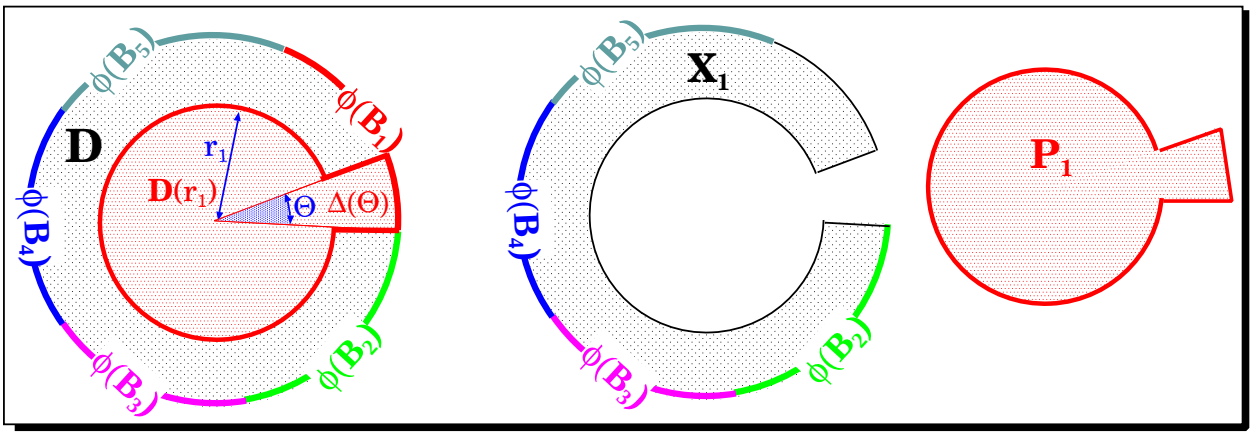


Figure 9.5: The Beck procedure.

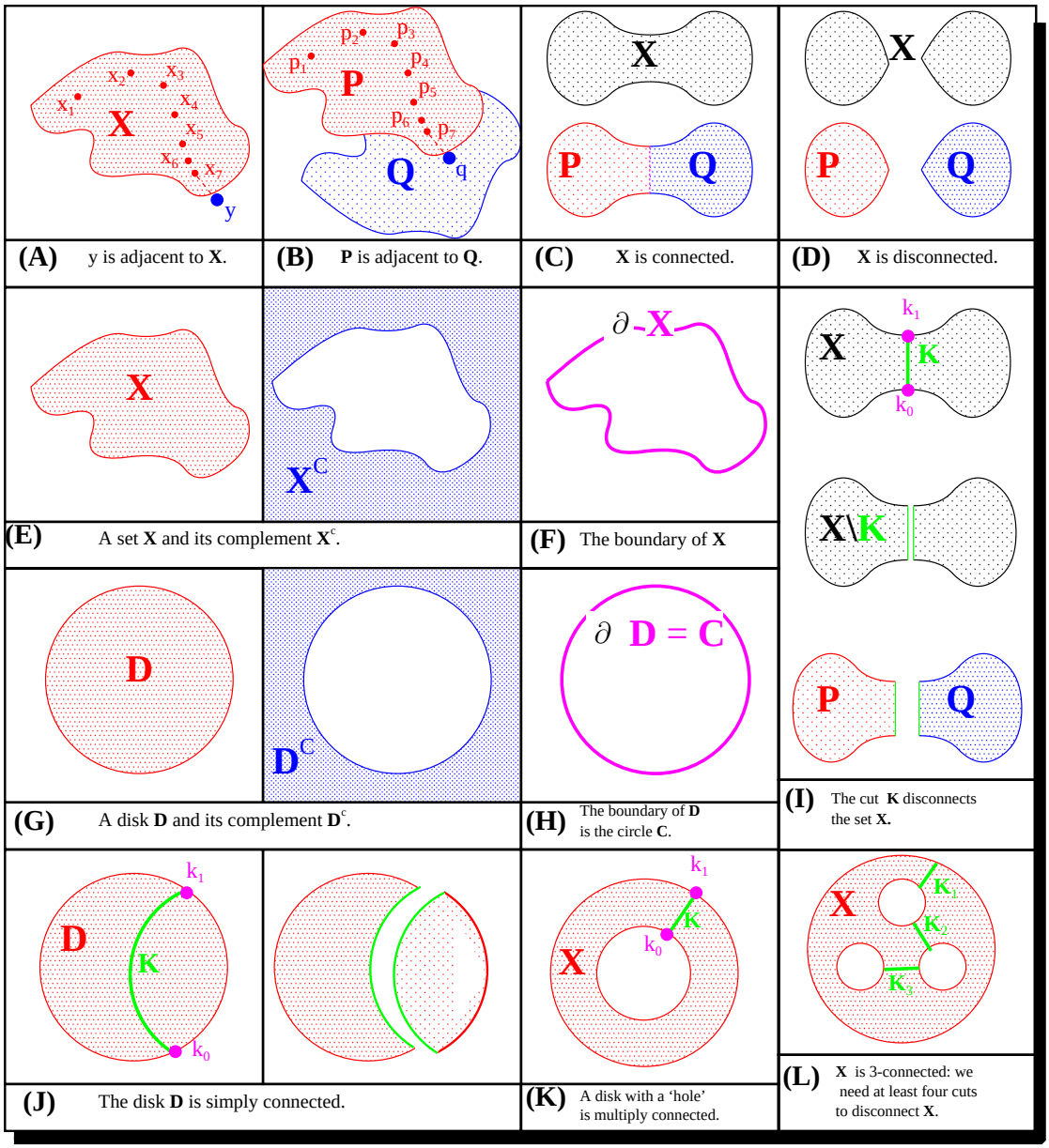


Figure 9.6: Some concepts in planar topology

A subset $\mathbf{X} \subset \mathbb{R}^2$ is *open* if no point in $\mathbf{X}$ is adjacent to $\mathbf{X}^{\complement}$. Equivalently, $\mathbf{X}$ is open if $\partial\mathbf{X} \subset \mathbf{X}^{\complement}$. For example, the disk $\mathbf{D}$ is *not* open, because $\partial\mathbf{D} = \mathbf{C} \subset \mathbf{D}$. However, the set $\mathbf{O} = \{\mathbf{x} \in \mathbb{R}^2 \,;\, |\mathbf{x}| < 1\}$ *is* open, because $\partial\mathbf{O} = \mathbf{C} \subset \mathbf{O}^{\complement}$.

Suppose $\mathbf{X} \subset \mathbb{R}^2$ is a connected domain. A *cut* is a curve $\mathbf{K} \subset \mathbf{X}$ which goes from one boundary point $k_0 \in \partial\mathbf{X}$ to another boundary point $k_1 \in \partial\mathbf{X}$, as in Figure 9.6I. We say that $\mathbf{K}$ *disconnects* $\mathbf{X}$ if $\mathbf{X} \setminus \mathbf{K}$ is disconnected, as in Figure 9.6I. In other words, 'cutting' $\mathbf{X}$ along the curve $\mathbf{K}$ splits $\mathbf{X}$ into two pieces. We say that $\mathbf{X}$ is *simply connected* if *any* cut $\mathbf{K}$ disconnects $\mathbf{X}$. For example, the unit disk $\mathbf{D}$ is simply connected (Figure 9.6J).

However, suppose $\mathbf{X}$ has a 'hole', as in Figure 9.6(K). If k_0 is a point on the 'exterior' boundary of $\mathbf{X}$, and k_1 is a point on the 'hole' boundary, then a cut from k_0 to k_1 will *not* disconnect $\mathbf{X}$. Thus, $\mathbf{X}$ is *not* simply connected. We say $\mathbf{X}$ is *multiply connected*, meaning that there is some cut $\mathbf{K}$ which does *not* disconnect $\mathbf{X}$ (i.e. $\mathbf{X} \setminus \mathbf{K}$ is still connected). More generally, we say that $\mathbf{X}$ is *I-connected* if there are I cuts $\mathbf{K}_1, \ldots, \mathbf{K}_I \subset \mathbf{X}$ such that the set $\mathbf{X} \setminus (\mathbf{K}_1 \cup \cdots \cup \mathbf{K}_I)$ is simply connected, as in Figure 9.6(L). (hence, at this point, one more cut will disconnect $\mathbf{X}$). Loosely speaking, $\mathbf{X}$ is I-connected if $\mathbf{X}$ has I distinct 'holes' in its interior.

If $\mathbf{X}, \mathbf{Y} \subset \mathbb{R}^2$, then a *homeomorphism* is a function $\phi : \mathbf{X} \longrightarrow \mathbf{Y}$ such that:

- ϕ is bijective [and thus, has a well-defined inverse function $\phi^{-1} : \mathbf{Y} \longrightarrow \mathbf{X}$].
- ϕ is *continuous* [i.e. if $\{x_1, x_2, \ldots\} \subset \mathbf{X}$, and $\lim_{i\to\infty} x_i = x$, then $\lim_{i\to\infty} \phi(x_i) = \phi(x)$].
- ϕ^{-1} is also continuous.

Heuristically, ϕ provides a method to 'represent' $\mathbf{X}$ using the region $\mathbf{Y}$; all topological phenomena (e.g. adjacency, connectivity) on $\mathbf{X}$ are transformed by ϕ into analogous phenomena on $\mathbf{Y}$, as follows:

Lemma 9E.3 *Suppose $\phi : \mathbf{X} \longrightarrow \mathbf{Y}$ is a homeomorphism.*

(a) *If $\mathbf{P}, \mathbf{Q} \subset \mathbf{X}$, then* $\Big(\mathbf{P}$ *is adjacent to* $\mathbf{Q}\Big) \iff \Big(\phi(\mathbf{P})$ *is adjacent to* $\phi(\mathbf{Q})\Big)$.

(b) $\Big(\mathbf{X}$ *is connected*$\Big) \iff \Big(\mathbf{Y}$ *is connected*$\Big)$.

(c) $\Big(\mathbf{X}$ *is simply connected*$\Big) \iff \Big(\mathbf{Y}$ *is simply connected*$\Big)$.

(d) $\Big(\mathbf{X}$ *is I-connected*$\Big) \iff \Big(\mathbf{Y}$ *is I-connected*$\Big)$.

(e) *If $\mathbf{P} \subset \mathbf{X}$, then* $\Big(\mathbf{P}$ *is connected*$\Big) \iff \Big(\phi(\mathbf{P})$ *is connected*$\Big)$.

Proof: **Exercise 9.6** □

If μ is a utility measure on $\mathbf{X}$, and $\phi : \mathbf{X} \longrightarrow \mathbf{Y}$ is a homeomorphism, then we define a new utility measure $\nu := \phi(\mu)$ on $\mathbf{Y}$ by the equation:

$$\nu[\mathbf{Q}] \quad = \quad \mu\left[\phi^{-1}(\mathbf{Q})\right], \qquad \text{for any } \mathbf{Q} \subset \mathbf{Y}.$$

[Recall: if $\mathbf{Q} \subset \mathbf{Y}$, then $\phi^{-1}(\mathbf{Q}) = \{\mathbf{x} \in \mathbf{X} \; ; \; \phi(\mathbf{x}) \in \mathbf{Q}\}$, hence $\phi^{-1}(\mathbf{Q}) \subset \mathbf{X}$.]

Lemma 9E.4 *Let $\phi : \mathbf{X} \longrightarrow \mathbf{Y}$ be a homeomorphism, and let μ be a utility measure on $\mathbf{X}$. Let $\nu = \phi(\mu)$. Then:*

(a) *ν is also a utility measure.*

(b) $\Big(\mu$ *is nonatomic*$\Big) \iff \Big(\nu$ *is nonatomic*$\Big)$.

Proof: **Exercise 9.7** □

Homeomorphisms can transform 'good' partitions of $\mathbf{X}$ into 'good' partitions of $\mathbf{Y}$ as follows:

Proposition 9E.5 *Let $\mathbf{P}_1, \ldots, \mathbf{P}_I \subset \mathbf{X}$ be some subsets of $\mathbf{X}$. Let $\phi : \mathbf{X} \longrightarrow \mathbf{Y}$ be a homeomorphism, and for all $i \in [1..I]$, define $\mathbf{Q}_i := \phi(\mathbf{P}_i)$. Let $\mathcal{P} = \{\mathbf{P}_1, \ldots, \mathbf{P}_I\}$ and let $\mathcal{Q} = \{\mathbf{Q}_1, \ldots, \mathbf{Q}_I\}$. Then:*

(a) $\Big(\mathcal{P}$ *is a partition of* $\mathbf{X}\Big) \iff \Big(\mathcal{Q}$ *is a partition of* $\mathbf{Y}\Big)$.

(b) *For all $i \in [1..I]$,* $\Big(\mathbf{P}_i$ *is connected*$\Big) \iff \Big(\mathbf{Q}_i$ *is connected*$\Big)$.

(c) *Let $\mu_1, \ldots, \mu_I$ be utility measures on $\mathbf{X}$. Let $\nu_1 = \phi(\mu_1), \ldots, \nu_I = \phi(\mu_I)$. Then*

$$\Big(\mathcal{P} \text{ is a proportional partition of } \mathbf{X}, \text{ relative to } \mu_1, \ldots, \mu_I\Big)$$
$$\iff \quad \Big(\mathcal{Q} \text{ is a proportional partition of } \mathbf{Y}, \text{ relative to } \nu_1, \ldots, \nu_I\Big).$$

Proof: **Exercise 9.8** □

Chapter 10

Pareto Optimality

10A Introduction

Prerequisites: §8 **Recommended:** §9A

Let $\mathcal{P} = \{\mathbf{P}_1, \ldots, \mathbf{P}_I\}$ and $\mathcal{Q} = \{\mathbf{Q}_1, \ldots, \mathbf{Q}_I\}$ be two partitions. We say that $\mathcal{P}$ is *Pareto-preferred* to $\mathcal{Q}$ if:

- For all $i \in [1..I]$, $\quad \mu_i[\mathbf{Q}_i] \leq \mu[\mathbf{P}_i]$.
  (i.e. every player gets *at least* as much in $\mathcal{P}$ as she does in $\mathcal{Q}$).
- For some $i \in [1..I]$, $\quad \mu_i[\mathbf{Q}_i] \ < \ \mu[\mathbf{P}_i]$
  (i.e. at least one player feels that she got strictly *more* in $\mathcal{P}$.)

The partition $\mathcal{P}$ is better (or at least, no worse) for every individual. Clearly, given a choice, we should choose partition $\mathcal{P}$ over partition $\mathcal{Q}$.

We say that $\mathcal{P}$ is *Pareto-optimal*[1] if there does not exist any other partition $\mathcal{Q}$ which is Pareto-preferred to $\mathcal{P}$. A partition procedure is *Pareto-optimal* if it always yields a Pareto-optimal outcome. Clearly, this is is desirable. After all, if the procedure produced a partition that was *not* Pareto-optimal, then by definition, we could suggest *another* partition which was at *least* as good for everyone, and strictly *better* for someone.

Example 10A.1: 'I cut, you choose' is *not* Pareto Optimal

Suppose the left half of the cake is orange cream, and the right half is tartufo. Owen only likes orange cream, and Twyla only likes tartufo. Clearly, a Pareto-optimal, proportional partition exists: cut the cake in half, and let Owen take the orange cream portion and Twyla take the tartufo portion. Both players receive a payoff of 1 (they get everything they value).

Unfortunately, this is *not* the partition which will be generated by 'I cut, you choose' (Game 8C.1). Not knowing Twyla's preferences (or not trusting her to abide by them), Owen must use

[1] Sometimes this is called *Pareto efficient*, or even just *efficient*.

his maximin strategy (see Example 8C.2), which cuts the cake into two portions, each having *half* the orange cream. Twyla will then choose the portion which has more tartufo. Twyla will end up with a payoff of at least $\frac{1}{2}$ (and possibly even a payoff of Owen, if one of the portions happens to have all the tartufo). However, Owen has *ensured* that he only gets a payoff of $\frac{1}{2}$. This is not Pareto optimal, because, as we've seen, *both* players *could* have gotten a payoff of Owen, with the right partition. ♢

Similarly, the Banach-Knaster and Dubins-Spanier procedures are not Pareto-optimal. However, this doesn't mean that proportionality and Pareto-optimality are mutually exclusive.

Lemma 10A.2 *Let $\mathcal{Q}$ be a proportional partition. If $\mathcal{P}$ is Pareto-preferred to $\mathcal{Q}$, then $\mathcal{P}$ is also proportional.*

Proof: **Exercise 10.1** □

Thus, given any proportional partition $\mathcal{Q}$ (e.g. the outcome of the Banach-Knaster procedure), Lemma 10A.2 says we can always find a proportional, Pareto-optimal partition $\mathcal{P}$ which is Pareto-preferred to $\mathcal{Q}$. We can achieve this through 'trade', as we next discuss.

10B Mutually Beneficial Trade

Prerequisites: §10A

Let $\mathcal{P} = \{\mathbf{P}_1, \mathbf{P}_2\}$ be a two-individual partition. A *mutually beneficial trade* is a pair $\mathcal{T} = (\mathbf{T}_1, \mathbf{T}_2)$, where $\mathbf{T}_1 \subset \mathbf{P}_1$ and $\mathbf{T}_2 \subset \mathbf{P}_2$, such that

$$\mu_1[\mathbf{T}_2] \quad \geq \quad \mu_1[\mathbf{T}_1], \qquad \text{and} \qquad \mu_2[\mathbf{T}_1] \quad \geq \quad \mu_2[\mathbf{T}_2],$$

and at least one of these two inequalities is strict. Thus, if player Owen gives $\mathbf{T}_1$ to Twyla in return for $\mathbf{T}_2$, then at least one of them (and possibly both) will benefit, and neither will suffer.

We say that the partition $\mathcal{Q} = \{\mathbf{Q}_1, \mathbf{Q}_2\}$ is *obtained* from $\mathcal{P}$ via the trade $\mathcal{T}$ if

$$\mathbf{Q}_1 \quad = \quad (\mathbf{P}_1 \setminus \mathbf{T}_1) \sqcup \mathbf{T}_2 \qquad \text{and} \qquad \mathbf{Q}_2 \quad = \quad (\mathbf{P}_2 \setminus \mathbf{T}_2) \sqcup \mathbf{T}_1. \qquad \text{(see Figure 10.1A)}$$

The interpretation is: Owen gives $\mathbf{T}_1$ to Twyla in return for $\mathbf{T}_2$, while Twyla gives $\mathbf{T}_2$ to Owen in return for $\mathbf{T}_1$. The *utility* of the trade $\mathcal{T}$ for Owen is defined:

$$\mu_1(\mathcal{T}) \quad := \quad \mu_1[\mathbf{T}_2] - \mu_1[\mathbf{T}_1].$$

This is the amount Owen gains by the trade. Likewise, the utility of the trade $\mathcal{T}$ for Twyla is defined:

$$\mu_2(\mathcal{T}) \quad := \quad \mu_2[\mathbf{T}_1] - \mu_2[\mathbf{T}_2].$$

The relation between trade and Pareto-optimality is the following:

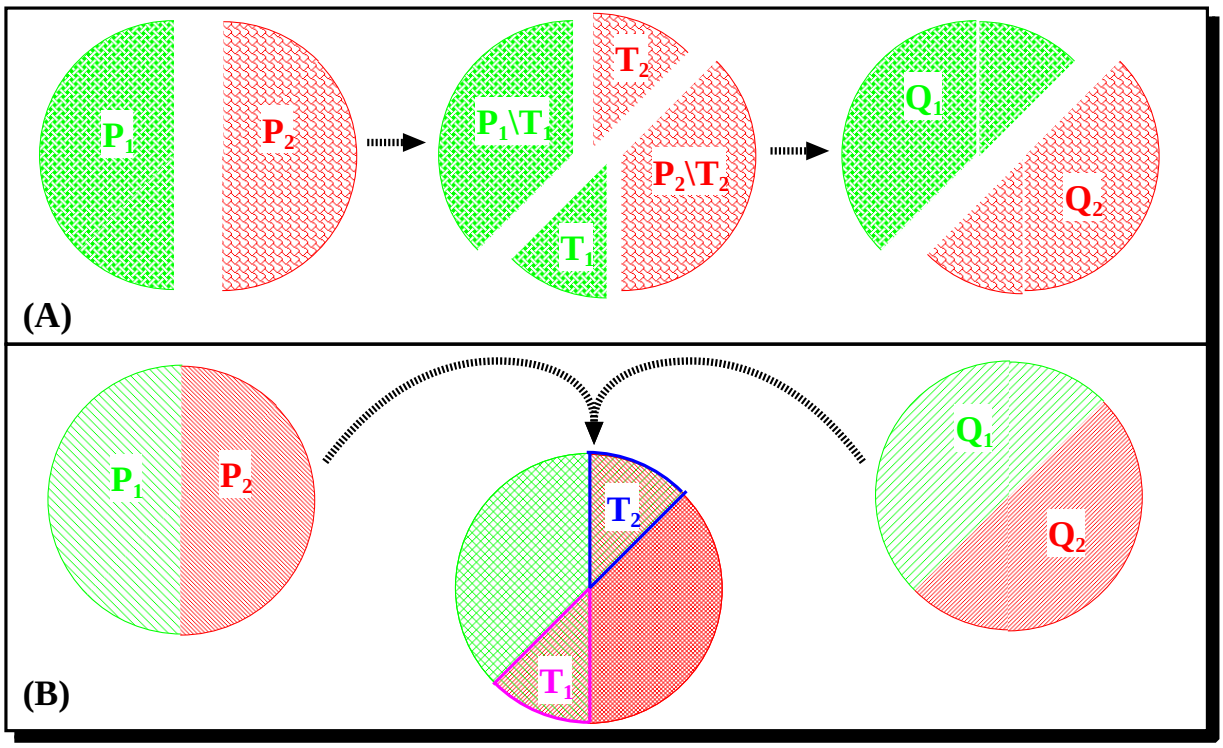


Figure 10.1: **(A)** Starting with partition $\mathcal{P} = \{\mathbf{P}_1, \mathbf{P}_2\}$, Owen gives $\mathbf{T}_1$ to Twyla in return for $\mathbf{T}_2$, to obtain partition $\mathcal{Q} = \{\mathbf{Q}_1, \mathbf{Q}_2\}$. **(B)** Given any partitions $\mathcal{P}$ and $\mathcal{Q}$, we can obtain $\mathcal{Q}$ from $\mathcal{P}$ through some kind of trade.

Proposition 10B.1 *Let $\mathcal{P} = \{\mathbf{P}_1, \mathbf{P}_2\}$ be a two-individual partition.*

(a) *If $\mathcal{Q} = \{\mathbf{Q}_1, \mathbf{Q}_2\}$ is another partition which is Pareto-preferred to $\mathcal{P}$, then $\mathcal{Q}$ can be obtained from $\mathcal{P}$ by a mutually beneficial trade.*

(b) *Thus, $\mathcal{P}$ is Pareto-optimal iff no mutually beneficial trades can be made from $\mathcal{P}$.*

Proof: **(a)** Suppose $\mathcal{Q}$ is Pareto-preferred to $\mathcal{P}$. Let $\mathbf{T}_1 = \mathbf{Q}_2 \cap \mathbf{P}_1$ and let $\mathbf{T}_2 = \mathbf{Q}_1 \cap \mathbf{P}_2$ (see Figure 10.1B). . It is left as **Exercise 10.2** to check that:

- $\mathcal{T}$ is a mutually beneficial trade.
- $\mathcal{Q}$ is obtained from $\mathcal{P}$ via the trade $\mathcal{T} = (\mathbf{T}_1, \mathbf{T}_2)$.

(b) follows immediately from **(a)**. □

Given any proportional partition $\mathcal{Q}$, we can always find a proportional, Pareto-optimal

partition $\mathcal{P}$ which is Pareto-preferred to $\mathcal{Q}$ through some kind of trade[2]. If the players are able to communicate and trade, then they can achieve a Pareto-optimal partition by 'trading' bits of their $\mathcal{Q}$-portions with each other. Each player will trade bits she considers 'low value' for bits she considers high value (i.e. in Example 10A.1, Owen would give tartufo to Twyla, in return for orange cream). Trading can only *increase* the utilities of all the traders, so the post-trade partition is still proportional, and is also Pareto-preferred to the pre-trade partition. We let this process continue until we've reached a partition $\mathcal{P}$ where further no further trades are possible (e.g. Owen runs out of tartufo, or Twyla runs out of orange cream). At this point, we have reached a Pareto optimal partition.

The concept of 'mutually beneficial trade' and its relation to Pareto-optimality can be generalized to three or more players. For instance, Barbanel [Bar99] has studied the possibility of *cyclic trades* amongst I players (e.g. Owen gives something to Twyla, who gives something to Trey, who gives something to Ford, who gives something to Owen; all four end up better off).

Unfortunately, the 'trading' procedure may destroy the *envy-freedom* of a three-player partition (see §11A). If there are only two players, then any proportional partition is automatically envy-free (Theorem 11A.3), so the 'trading procedure' preserves envy-freedom. However, if we have three players, and two of them trade to their mutual advantage, then the nontrader may end up envying one or both of the two traders.

10C Utility Ratio Threshold Partitions

Prerequisites: §10B; Elementary integral calculus[3].

To get a nontrivial example of a Pareto-optimal partition for two players, suppose $\mathbf{X} = [0, 1]$, and suppose there are *utility functions* $U_1, U_2 : [0, 1] \longrightarrow [0, \infty)$ which define the player's utility measures as follows: for any subinterval $[a, b] \subset \mathbf{X}$

$$\mu_1[a, b] \quad = \quad \int_a^b U_1(x)\, dx, \qquad \text{and} \qquad \mu_2[a, b] \quad = \quad \int_a^b U_2(x)\, dx.$$

We assume that $\int_0^1 U_1(x)\, dx \quad = \quad 1 \quad = \quad \int_0^1 U_2(x)\, dx.$

[2]Economists will recognize this as a version of *Coase's Theorem*, which states that an economy with 'well-defined property rights', 'perfect information', and 'costless transactions' will always converge to a Pareto-optimal state if the participants trade rationally. In the context of fair division theory, the existence of a specific partition $\mathcal{P}$ constitutes 'well-defined property rights'. The fact that all players know their own utility measures (Axiom **(Ψ1)** on page 169) constitutes 'perfect information'. We have also tacitly assumed from the beginning that all 'transactions' are costless (i.e. the cake is not damaged by repeatedly slicing and recombining pieces, the players never get tired or hungry, etc.)

[3]Antidifferentiation techniques are not required, but a basic understanding of the concept of integration is necessary.

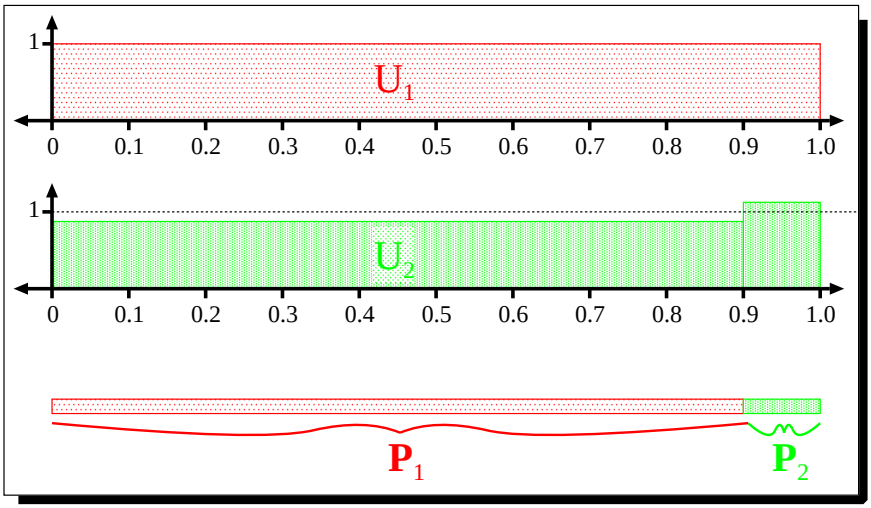


Figure 10.2: Example 10C.1: The Highest Bidder Partition favours Owen over Twyla.

The *utility ratio* is the function $R(x) = U_1(x)/U_2(x)$. Intuitively, if $R(x) > 1$, then Owen wants the point x more than Twyla does, while if $R(x) < 1$, then Twyla wants x more than Owen does. The *highest bidder* partition is the partition $\mathcal{P}^{(1)} = \{\mathbf{P}_1, \mathbf{P}_2\}$ defined

$$\mathbf{P}_1 \quad \subseteq \quad \{x \in \mathbf{X} \,;\, R(x) \geq 1\} \quad \text{and} \quad \mathbf{P}_2 \quad \subseteq \quad \{x \in \mathbf{X} \,;\, R(x) < 1\}.$$

Thus, each player gets those parts of the cake she wants more than the other player. This seems like it should be a fair partition, but it often is not, as the next example shows.

Example 10C.1: Suppose $U_1(x) = 1$, for all $x \in \mathbf{X}$, while $U_2(x) = \begin{cases} \frac{9}{10} & \text{if} \quad 0 \leq x \leq \frac{9}{10}; \\ \frac{19}{10} & \text{if} \quad \frac{9}{10} < x \leq 1. \end{cases}$

(See Figure 10.2). It can be checked that $\int_0^1 U_1(x)\, dx = 1 = \int_0^1 U_2(x)\, dx$. However,

$$\mathbf{P}_1 = \left[0, \frac{9}{10}\right] \quad \text{so that} \quad \mu_1[\mathbf{P}_1] = \int_0^{9/10} 1\, dx = 0.9$$

$$\text{while} \quad \mathbf{P}_2 = \left(\frac{9}{10}, 1\right] \quad \text{so that} \quad \mu_2[\mathbf{P}_2] = \int_{9/10}^1 \frac{19}{10}\, dx = \frac{1}{10} \cdot \frac{19}{10} = 0.19.$$

Thus, the highest bidder partition is *not* proportional. Player Owen does very well, but Twyla does very badly, because of the slight asymmetry in Twyla's preferences. ♢

The problem here is that we have partitioned the cake using a 'threshold' value of 1 for the utility ratio. This seems like a good idea, but in Example 10C.1 it yields an 'unfair' partition.

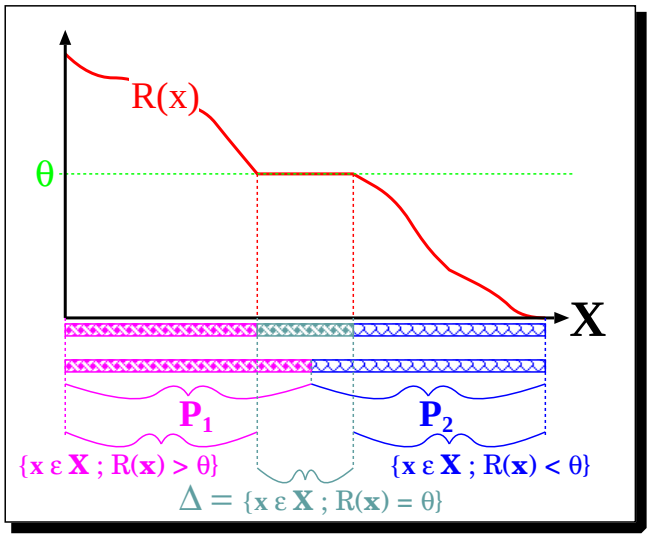


Figure 10.3: A utility ratio threshold partition.

To fix this problem, we need to use a different threshold. Given any $\theta \in [0, \infty)$, a **θ-utility ratio threshold (θ-URT)** partition is a partition $\mathcal{P}^{(\theta)} = \{\mathbf{P}_1, \mathbf{P}_2\}$ such that

$$\mathbf{P}_1 \subseteq \{x \in \mathbf{X} \,;\, R(x) \geq \theta\} \quad \text{and} \quad \mathbf{P}_2 \subseteq \{x \in \mathbf{X} \,;\, R(x) \leq \theta\} \qquad \text{(see Fig.10.3)}$$

Thus, the *highest bidder* partition is a URT partition with $\theta = 1$.

Remark: Note that the allocation of the set $\Delta = \{x \in \mathbf{X} \,;\, R(x) = \theta\}$ is ambiguous here. If the set Δ has zero measure (which is likely), then it doesn't really matter how Δ is split between $\mathbf{P}_1$ and $\mathbf{P}_2$, and there is effectively a unique θ-URT partition. However, if Δ has *nonzero* measure, then there are many ways Δ can be divided, yielding many different θ-URT partitions.

Proposition 10C.2 *For any value of θ, any URT partition $\mathcal{P}^{(\theta)}$ is Pareto-optimal.*

Proof: Suppose $\mathcal{P}^{(\theta)}$ were not Pareto-optimal, and suppose $\mathcal{Q}$ was Pareto-prefered to $\mathcal{P}^{(\theta)}$. Then Proposition 10B.1(a) says we can obtain $\mathcal{Q}$ from $\mathcal{P}^{(\theta)}$ through some mutually advantageous trade $\mathcal{T} = (\mathbf{T}_1, \mathbf{T}_2)$, where $\mathbf{T}_k \subset \mathbf{P}_k$.

Let $\Delta = \{x \in \mathbf{X} \,;\, R(x) = \theta\}$.

Claim 1: *We can assume without loss of generality that $\mathbf{T}_1 \cap \Delta = \emptyset$.*

Proof: IDEA: *Trading bits of the set Δ benefits no one. Thus, $\mathcal{T}$ can be replaced with modified trade $\mathcal{T}' = (\mathbf{T}'_1, \mathbf{T}'_2)$, which has the same utility to both players, but where $\mathbf{T}'_1 \cap \Delta = \emptyset$.*

Claim 1.1: *If $\mathbf{S} \subset \Delta$ is any subset, then $\mu_1[\mathbf{S}] = \theta \cdot \mu_2[\mathbf{S}]$.*

Proof: $\mu_1[\mathbf{S}] \;=\; \int_{\mathbf{S}} U_1(s)\,ds \;=\; \int_{\mathbf{S}} R(s)\cdot U_2(s)\,ds \;\underset{(*)}{=}\; \int_{\mathbf{S}} \theta\cdot U_2(s)\,ds$

$= \;\theta\cdot\int_{\mathbf{S}} U_2(s)\,ds \;=\; \theta\cdot\mu_2[\mathbf{S}].$

Here, $(*)$ is because $R(s)=\theta$ for all $s\in\mathbf{S}\subset\Delta$. ▽ Claim 1.1

Let $\mathbf{S}_1 := \mathbf{T}_1\cap\Delta$, and let $s_1 := \mu_2[\mathbf{S}_1]$; then Claim 1.1 says that $\mu_1[\mathbf{S}_1]=\theta\cdot s_1$.
Let $\mathbf{S}_2 := \mathbf{T}_2\cap\Delta$, and let $s_2 := \mu_2[\mathbf{S}_2]$; then Claim 1.1 says that $\mu_1[\mathbf{S}_2]=\theta\cdot s_2$.
Assume without loss of generality that $s_1\le s_2$ (otherwise switch the two players to make this the case). Let $\mathbf{S}'_2\subseteq\mathbf{S}_2$ be a subset such that $\mu_2[\mathbf{S}'_2]=s_1$; such an $\mathbf{S}'_2$ exists because $s_1\le s_2$, and because μ_2 is nonatomic. Claim 1.1 says that $\mu_1[\mathbf{S}'_2]=\theta\cdot s_1$.
Let $\mathbf{T}'_1=\mathbf{T}_1\setminus\mathbf{S}_1$ and let $\mathbf{T}'_2=\mathbf{T}_2\setminus\mathbf{S}'_2$. Let $\mathcal{T}'=(\mathbf{T}'_1,\mathbf{T}'_2)$.

Claim 1.2: *$\mu_1(\mathcal{T}')=\mu_1(\mathcal{T})$ and $\mu_2(\mathcal{T}')=\mu_2(\mathcal{T})$.*

Proof: First, note that
$$\begin{aligned}
\mu_1[\mathbf{T}'_1] &= \mu_1[\mathbf{T}_1]-\mu_1[\mathbf{S}_1] &&= \mu_1[\mathbf{T}_1]-\theta\cdot s_1,\\
\mu_2[\mathbf{T}'_1] &= \mu_2[\mathbf{T}_1]-\mu_2[\mathbf{S}_1] &&= \mu_2[\mathbf{T}_1]-s_1,\\
\mu_1[\mathbf{T}'_2] &= \mu_1[\mathbf{T}_2]-\mu_1[\mathbf{S}'_2] &&= \mu_1[\mathbf{T}_2]-\theta\cdot s_1,\\
\text{and}\quad \mu_2[\mathbf{T}'_2] &= \mu_2[\mathbf{T}_2]-\mu_2[\mathbf{S}'_2] &&= \mu_2[\mathbf{T}_2]-s_1.
\end{aligned}$$
$$\begin{aligned}
\text{Thus,}\quad \mu_1(\mathcal{T}') &= \mu_1[\mathbf{T}'_2]-\mu_1[\mathbf{T}'_1] &&= \Big(\mu_1[\mathbf{T}_2]-\theta s_1\Big)-\Big(\mu_1[\mathbf{T}_1]-\theta s_1\Big)\\
&= \mu_1[\mathbf{T}_2]-\mu_1[\mathbf{T}_1] &&= \mu_1(\mathcal{T}).\\
\text{Likewise,}\quad \mu_2(\mathcal{T}') &= \mu_2[\mathbf{T}'_1]-\mu_2[\mathbf{T}'_2] &&= \Big(\mu_2[\mathbf{T}_1]-s_1\Big)-\Big(\mu_2[\mathbf{T}_2]-s_1\Big)\\
&= \mu_2[\mathbf{T}_1]-\mu_2[\mathbf{T}_2] &&= \mu_2(\mathcal{T}).
\end{aligned}$$
▽ Claim 1.2

Thus, we can replace trade $\mathcal{T}$ with a modified trade $\mathcal{T}'$, which has exactly the same value for both players. In the modified trade $\mathcal{T}'$, notice that $\mathbf{T}_1\cap\Delta=\emptyset$. ◇ Claim 1

Claim 2: *Suppose $(\mathbf{T}_1,\mathbf{T}_2)$ is a trade, and $\mathbf{T}_1\cap\Delta=\emptyset$. Then $(\mathbf{T}_1,\mathbf{T}_2)$ cannot be a mutually advantageous trade. That is: Either $\mu_1[\mathbf{T}_2] < \mu_1[\mathbf{T}_1]$, or $\mu_2[\mathbf{T}_1] < \mu_2[\mathbf{T}_2]$.*

Proof: Suppose
$$\mu_1[\mathbf{T}_1] \;\le\; \mu_1[\mathbf{T}_2]; \tag{10.1}$$
I'll show that $\mu_2[\mathbf{T}_1] < \mu_2[\mathbf{T}_2]$. To see this, note that:
$$\mu_1[\mathbf{T}_2] \;=\; \int_{\mathbf{T}_2} U_1(t)\,dt \;\underset{(\dagger)}{=}\; \int_{\mathbf{T}_2} R(t)\cdot U_2(t)\,dt \;\underset{(*)}{\le}\; \int_{\mathbf{T}_2}\theta\cdot U_2(t)\,dt \;=\; \theta\,\mu_2[\mathbf{T}_2], \tag{10.2}$$
here, $(\dagger)$ is because $R(t) = U_1(t)/U_2(t)$, and $(*)$ is because $R(t)\le\theta$ for all $t\in\mathbf{T}_2$. Likewise,
$$\mu_2[\mathbf{T}_1] \;=\; \int_{\mathbf{T}_1} U_2(t)\,dt \;\underset{(\ddagger)}{=}\; \int_{\mathbf{T}_1}\frac{U_1(t)}{R(t)}\cdot\,dt \;\underset{(\star)}{<}\; \int_{\mathbf{T}_1}\frac{1}{\theta}\cdot U_1(t)\,dt \;=\; \frac{1}{\theta}\,\mu_1[\mathbf{T}_1], \tag{10.3}$$
where $(\ddagger)$ is because is because $R(t) = U_1(t)/U_2(t)$, and $(\star)$ is because $R(t)>\theta$ for all $t\in\mathbf{T}_1$ (because $\mathbf{T}_1\cap\Delta=\emptyset$). Thus,
$$\mu_2[\mathbf{T}_1] \;\underset{(10.3)}{<}\; \frac{1}{\theta}\,\mu_1[\mathbf{T}_1] \;\underset{(10.1)}{\le}\; \frac{1}{\theta}\,\mu_1[\mathbf{T}_2] \;\underset{(10.2)}{\le}\; \frac{\theta}{\theta}\,\mu_2[\mathbf{T}_2] \;=\; \mu_2[\mathbf{T}_2],$$

where **(10.3)** is by eqn.(10.3); **(10.1)** is by hypothesis (10.1); and **(10.2)** is by eqn.(10.2).

Hence, $\mu_2[\mathbf{T}_1] < \mu_2[\mathbf{T}_2]$, so the trade is *not* beneficial for Twyla.

Conversely, if we suppose $\mu_2[\mathbf{T}_1] \geq \mu_2[\mathbf{T}_2]$, then similar reasoning shows that $\mu_1[\mathbf{T}_2] < \mu_1[\mathbf{T}_1]$. ◇ Claim 2

Thus, a mutually beneficial trade is impossible, so Proposition 10B.1(B) says that $\mathcal{P}^{(\theta)}$ must be Pareto-optimal. □

In §11B a suitable choice of θ will yield a partition which is both Pareto-optimal and envy-free.

Remark: Throughout this section, we assumed $\mathbf{X} = [0, 1]$ only for concreteness and simplicity. Actually the definition of a URT partition and the proof of Theorem 10C.2 will work if $\mathbf{X}$ is any 'reasonable' subset of $\mathbb{R}^I$, or indeed, if $\mathbf{X}$ is any measure space, and $U_1, U_2 \in \mathbf{L}^1(\mathbf{X})$.

Exercise 10.3 Generalize the definition of *utility ratio partition* and the proof of Theorem 10C.2 to the case when $\mathbf{X} \subset \mathbb{R}^I$.

Exercise 10.4 Generalize the definition of *utility ratio partition* and the proof of Theorem 10C.2 to the case when $\mathbf{X}$ is an abstract measure space.

Further reading: Much work on Pareto-optimality in fair division has been done by Ethan Akin [Aki95] and Julius Barbanel [Bar99, Bar00].

10D Bentham Optimality & 'Highest Bidder'

Prerequisites: §10A **Recommended:** §10C

Let $\mathcal{P} = \{\mathbf{P}_1, \ldots, \mathbf{P}_I\}$ be a partition. The *total utility* of $\mathcal{P}$ is defined:

$$U(\mathcal{P}) \quad = \quad \sum_{i=1}^{I} \mu_i[\mathbf{P}_i].$$

Let $\mathcal{Q} = \{\mathbf{Q}_1, \ldots, \mathbf{Q}_I\}$ be another partition. We say that $\mathcal{P}$ is *Bentham-prefered* to $\mathcal{Q}$ if $U(\mathcal{P}) \geq U(\mathcal{Q})$. We say that $\mathcal{P}$ is *Bentham-optimal* if there does not exist any other partition $\mathcal{Q}$ which is Bentham-prefered to $\mathcal{P}$. In other words, $\mathcal{P}$ has the maximum total utility of any partition.

Lemma 10D.1 *Let $\mathcal{P}$ and $\mathcal{Q}$ be partitions.*

(a) $\Big(\mathcal{P}$ *is Pareto-prefered to* $\mathcal{Q}\Big) \Longrightarrow \Big(\mathcal{P}$ *is Bentham-prefered to* $\mathcal{Q}\Big)$.

(b) $\Big(\mathcal{P}$ *is Bentham-optimal*$\Big) \Longrightarrow \Big(\mathcal{P}$ *is Pareto-optimal*$\Big)$.

(c) *The converses of* **(a)** *and* **(b)** *are false.*

Proof: **Exercise 10.5** □

Intuitively, we can achieve Bentham optimality by giving each player the parts of $\mathbf{X}$ which she values more than any other player. Thus, every bit of cake goes to the individual who values it most, so the maximum amount of total utility has been 'extracted' from the cake.

For example, let $\mathbf{X} = [0, 1]$. Suppose $\mathcal{I} = \{1, \ldots, I\}$, and suppose there are *utility functions* $U_1, U_2, \ldots, U_i : [0, 1] \longrightarrow [0, \infty)$ which define the players' utility measures as follows: for any subinterval $[a, b] \subset \mathbf{X}$

$$\mu_i[a, b] \quad = \quad \int_a^b U_i(x)\, dx, \qquad \text{for all } i \in [1..I]. \tag{10.4}$$

We assume that $\int_{\mathbf{X}} U_i(x)\, dx \quad = \quad 1$, for all $i \in [1..I]$. The *Highest Bidder* partition $\mathcal{P} = \{\mathbf{P}_1, \ldots, \mathbf{P}_i\}$ is defined by

$$\mathbf{P}_i \quad = \quad \{x \in \mathbf{X}\ ;\ \text{for all } j \in [1..I], \text{ either } U_i(x) \ > \ U_j(x) \text{ or } U_i(x) \ = \ U_j(x) \text{ and } i < j\}. \tag{10.5}$$

Remarks: (a) Notice that we break 'bidding ties' in an arbitrary way by awarding the tied portion to the player with the lower number.

(b) Observe that, if $I = 2$, this agrees with the 'highest bidder' partition defined in §10C.

(c) As in §10C, there is nothing special about $\mathbf{X} = [0, 1]$; we could perform a similar construction if $\mathbf{X}$ was any 'reasonable' subset of $\mathbb{R}^I$, or indeed, if $\mathbf{X}$ was any measure space.

Proposition 10D.2 *If the utility measures of* $1, \ldots, i$ *are defined by utility functions as in eqn.*(10.4)*, then the Highest Bidder partition of eqn.*(10.5) *is Bentham-optimal.*

Proof: **Exercise 10.6** □

The problem with the Highest Bidder partition is that it may be far from proportional, as we saw in Example 10C.1. Indeed, with three or more players, the Highest Bidder partition may award some players an *empty* portion (**Exercise 10.7**). To remedy this, we can introduce 'side payments', whereby the losing players are compensated by the winners using some commodity *beyond* the cake (e.g. money), until all players feel they have received an equal proportion of the *value* (even if they haven't recieved an equal portion of cake).

Chapter 11

Envy and Equitability

11A Envy-freedom

Prerequisites: §8 **Recommended:** §9B

If all the players just want to get their 'fair share', then a proportional partition is all we seek, and the Banach-Knaster (§9B) or Dubins-Spanier (§9D) procedure will do the job. However, sometimes the players are jealous or hostile of one another, and each may demand not only that he gets his fair share, but that no other individual *more* than her fair share (as *he* sees it).

For example, suppose the three kingdoms *Wei*, *Wu*, and *Shu* are are quibbling over a disputed territory. Each state wants to get at least one third of the territory, but also each wants to make sure that no other state gets *more* territory, because then the other state would have a military advantage in future engagements. For example, even if the partition $\mathcal{P}$ gives Wei 40% of the territory, Wei would find $\mathcal{P}$ unacceptable if (in Wei's perception), $\mathcal{P}$ gives Shu 50% and gives Wu only 10%. The reason is not that Wei cares for the plight of Wu, but rather, that Wei fears the territorial advantage of Shu.

A partition $\mathcal{P} = \{\mathbf{P}_1, \ldots, \mathbf{P}_I\}$ is *envy-free* if, for all i and j, $\quad \mu_i[\mathbf{P}_i] \geq \mu_i[\mathbf{P}_j]$. In other words, each participant believes that she received *at least* as much as any other single participant did. A partition procedure or is *envy free* it always yields an envy-free partition.

Example 11A.1: 'I cut, you choose' is envy-free

If Owen and Twyla use 'I cut, you choose' (Procedure 8B.1), then the outcome will be envy free. To see this, recall that Owen divides the cake into two portions $\mathbf{P}$ and $\mathbf{Q}$ such that $\mu_1[\mathbf{P}] = \frac{1}{2} = \mu_1[\mathbf{Q}]$. Then Twyla chooses the portion she thinks is larger —say $\mathbf{Q}$. Thus, Twyla doesn't envy Owen, because $\mu_2[\mathbf{Q}] \geq \mu_2[\mathbf{P}]$. Conversely, Owen doesn't envy Twyla because $\mu_1[\mathbf{P}] = \mu_1[\mathbf{Q}]$. ♢

Example 11A.2: Banach-Knaster is *not* envy free

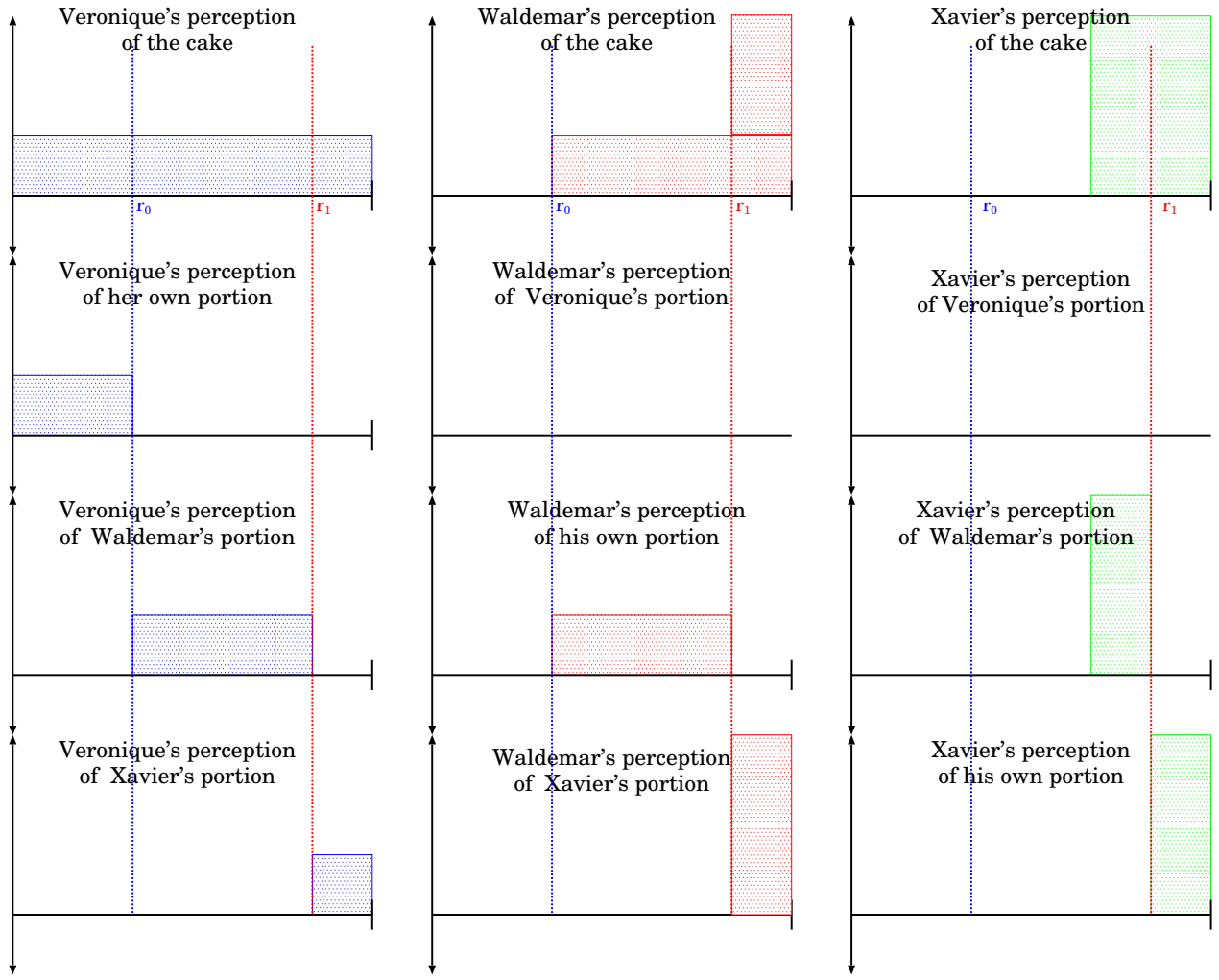


Figure 11.1: The Banach-Knaster procedure is not envy free (Example 11A.2).

Suppose $\mathbf{X} = [0, 1]$, and Owen, Twyla, and Trey are dividing $\mathbf{X}$ using 'Last Diminisher' (Game 9B.2). We'll give the three players utility measures so that the outcome of 'Last Diminisher' *cannot* be envy free. Suppose the players perceive the cake as in Figure 11.1. Thus:

- Owen values all parts of $\mathbf{X}$ equally. In other words, for any interval $[a, b] \subset [0, 1]$, $\mu_1[a, b] = (b - a)$.
- Twyla and Trey think the left-hand third of the cake is worthless. Trey also thinks the middle third is worthless. Both Twyla and Trey think that the right end of the cake is the most valuable.

Thus, Owen will choose $r_0 = \frac{1}{3}$, because $\mu_1\left[0, \frac{1}{3}\right] = \frac{1}{3}$. The other two players think this portion is worthless, so they are happy to give it to him untouched. Hence, Owen receives portion $\mathbf{P}_1 = \left[0, \frac{1}{3}\right]$, and exits the game feeling that he has a fair share.

Next, Twyla will choose $r_1 = \frac{5}{6}$, because she believes $\mu_2\left[\frac{1}{3}, \frac{5}{6}\right] = \frac{1}{2}$ —i.e. one half the remaining value of the cake (minus Owen's 'worthless' portion). Since Trey also believes $\mu_3\left[\frac{1}{3}, \frac{5}{6}\right] = \frac{1}{2}$, Trey is happy to let Twyla have this portion. So Twyla exits with $\mathbf{P}_2 = \left[\frac{1}{3}, \frac{5}{6}\right]$, and Trey is left with $\mathbf{P}_3 = \left[\frac{5}{6}, 1\right]$.

All three players believe they got a fair portion Indeed, Twyla and Trey both think they got more than a fair portion, receiving a payoff of $\frac{1}{2}$ each. However, Owen believes that Twyla's portion is bigger than his own, because $\mu_1[\mathbf{P}_1] = \frac{1}{3}$, but $\mu_1[\mathbf{P}_2] = \frac{1}{2}$. Hence he envies Twyla. ♢

'Envy freedom' is only a problem when there are *three or more* people, with *different* utility measures:

Proposition 11A.3 *Let $\mathcal{P}$ be a partition.*

(a) $\Big(\mathcal{P}$ *is envy-free*$\Big) \Longrightarrow \Big(\mathcal{P}$ *is proportional*$\Big)$.

(b) *Suppose there are only two individuals (i.e. $\mathcal{I} = \{1, 2\}$).*

$$\Big(\mathcal{P} \text{ is envy-free}\Big) \iff \Big(\mathcal{P} \text{ is proportional}\Big).$$

(c) *Suppose all individuals have the same utility measures (i.e. $\mu_1 = \mu_2 = \cdots = \mu_I$). Then*

$$\Big(\mathcal{P} \text{ is envy-free}\Big) \iff \Big(\mathcal{P} \text{ is proportional}\Big).$$

Proof: **Exercise 11.1** □

Envy-free cake division becomes a nontrivial problem when there are three or more players. A number of envy-free three-individual partition games have been devised; we will discuss one of the earliest and simplest, which was discovered independently by John L. Selfridge and John Horton Conway in the 1960s (but not published by either).

Procedure 11A.4: Selfridge-Conway 'Trimming' procedure

Suppose that $\mathcal{I} = \{1, 2, 3\}$ (say, Owen, Twyla, and Trey). Let these players have utility measures μ_1, μ_2, and μ_3. Let $\mathbf{X}$ be the 'cake'. We refer to Figure 11A.4.

(1) Let $\mathcal{Q} = \{\mathbf{Q}_1, \mathbf{Q}_2, \mathbf{Q}_3\}$ be a partition of $\mathbf{X}$ into three portions which Owen deems of equal size; i.e. $\mu_1[\mathbf{Q}_1] = \mu_1[\mathbf{Q}_3] = \mu_1[\mathbf{Q}_3] = \frac{1}{3}$.

(2) Assume without loss of generality (by reordering the portions if necessary) that Twyla ranks these portions in ascending order: $\mu_2[\mathbf{Q}_1] \leq \mu_2[\mathbf{Q}_2] \leq \mu_2[\mathbf{Q}_3]$.

Let $\mathbf{Q}'_3 \subseteq \mathbf{Q}_3$ be a subportion, such that Twyla thinks that portions $\mathbf{Q}_2$ and $\mathbf{Q}'_3$ are 'tied for largest'; i.e. $\mu_2[\mathbf{Q}_1] \leq \mu_2[\mathbf{Q}_2] = \mu_2[\mathbf{Q}'_3]$.

Let $\mathbf{L} := \mathbf{Q}_3 \setminus \mathbf{Q}'_3$ (the 'leftover' piece).

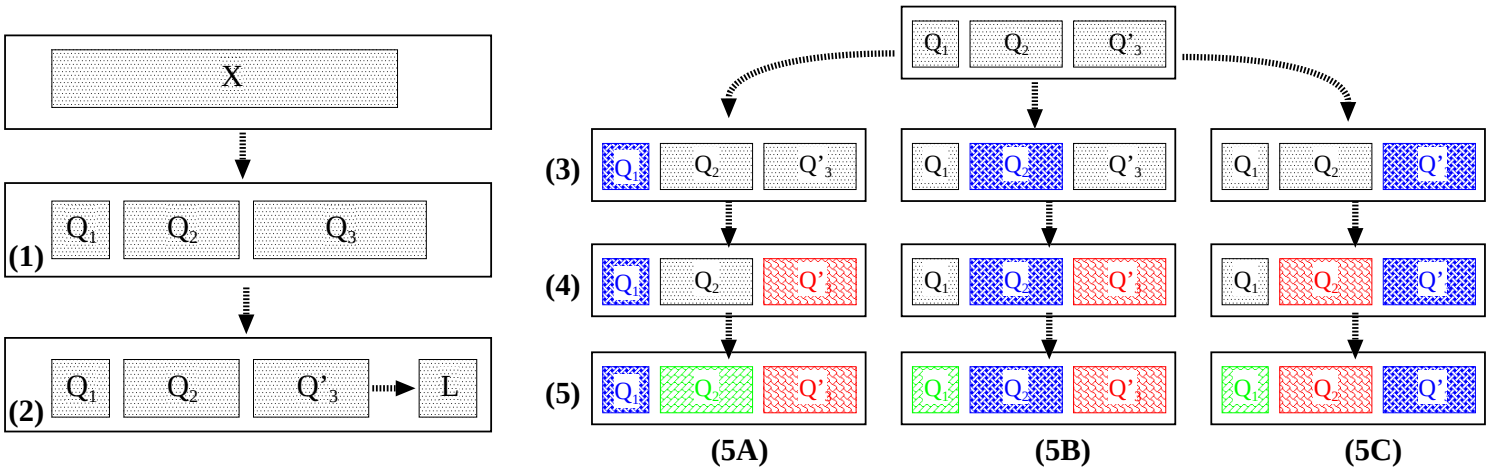


Figure 11.2: The Selfridge-Conway Trimming Procedure (Procedure 11A.4)

.

(3) Give Trey whichever of $\{\mathbf{Q}_1, \mathbf{Q}_2, \mathbf{Q}'_3\}$ he thinks is largest. Call this piece $\mathbf{P}'_3$. Observe that, no matter which piece Trey takes, at least one of the pieces $\{\mathbf{Q}_2, \mathbf{Q}'_3\}$ must remain.

(4) If only *one* of the two pieces $\{\mathbf{Q}_2, \mathbf{Q}'_3\}$ remains, then give it to Twyla.
If *both* $\mathbf{Q}_2$ and $\mathbf{Q}'_3$ remain, then give Twyla $\mathbf{Q}'_3$ (the one she trimmed).
(Twyla thinks both $\mathbf{Q}_2$ and $\mathbf{Q}'_3$ are equally large, and both are at least as big as $\mathbf{Q}_1$, so she will be happy either way).

(5) Give Owen the remaining piece, which must be either $\mathbf{Q}_1$ or $\mathbf{Q}_2$, because:

(5A) ...if Trey took $\mathbf{Q}_1$, then Twyla took $\mathbf{Q}'_3$, so $\mathbf{Q}_2$ remains.
(5B) ...if Trey took $\mathbf{Q}_2$, then Twyla took $\mathbf{Q}'_3$, so $\mathbf{Q}_1$ remains.
(5C) ...if Trey took $\mathbf{Q}'_3$, then Twyla took $\mathbf{Q}_2$, so $\mathbf{Q}_1$ remains.

Thus, Owen always gets an 'untrimmed' piece, which he thinks has size exactly $\frac{1}{3}$.

It remains to dispose of the leftover $\mathbf{L}$. Suppose that the individual who got the trimmed piece $\mathbf{Q}'_3$ has surname Short (i.e. either Twyla Short or Trey Short); and that the individual (other than Owen) who got an *untrimmed* piece has surname Taylor (i.e. either Trey Taylor or Twyla Taylor). Observe that both Short and Taylor think they got the largest piece of $\{\mathbf{Q}_1, \mathbf{Q}_2, \mathbf{Q}'_3\}$ (or at least, one of the two largest pieces), while Owen thinks he got exactly $\frac{1}{3}$ of the original cake.

Also, observe that, in partitioning $\mathbf{L}$, Owen has an *irrevocable advantage* over Short; even if Short gets *all* of $\mathbf{L}$, Owen will still think that Short got no more than $\frac{1}{3}$ of the cake, because $\mu_1[\mathbf{Q}'_3 \sqcup \mathbf{L}] = \mu_1[\mathbf{Q}_3] = \frac{1}{3}$.

(6) Taylor divides $\mathbf{L}$ into three pieces $\mathbf{L}_1, \mathbf{L}_2, \mathbf{L}_3$, with $\mu_t[\mathbf{L}_1] = \mu_t[\mathbf{L}_2] = \mu_t[\mathbf{L}_3] = \frac{1}{3}\mu[\mathbf{L}]$.

(7) Short chooses whichever of the three pieces she/he thinks is largest.

(8) Owen chooses whichever of the two remaining pieces he thinks is largest.

(9) Taylor gets the remaining piece.

At this point:

- Short thinks she/he got the largest of $\{\mathbf{Q}_1, \mathbf{Q}_2, \mathbf{Q}'_3\}$ and also the largest of $\{\mathbf{L}_1, \mathbf{L}_2, \mathbf{L}_3\}$. Hence Short envies no one.
- Owen thinks both he and Taylor got exactly $\frac{1}{3}$ of the original cake. He also thinks he got a choice of $\{\mathbf{L}_1, \mathbf{L}_2, \mathbf{L}_3\}$ which is at *least* as large as Taylor's. Hence he does not envy Taylor. Also, Owen will not envy Short because of his 'irrevocable advantage'.
- Taylor thinks she/he got the largest of $\{\mathbf{Q}_1, \mathbf{Q}_2, \mathbf{Q}'_3\}$ and also exactly $\frac{1}{3}$ of $\mathbf{L}$. Hence Taylor envies no one.

Thus, each player envies no one, so the result is an envy-free partition. ———————

A number of other envy-free cake division games exist for three players:

- Stromquist's procedure with four moving knives [Str80b].
- Levmore and Cooke's procedure with two orthogonal moving knives [LC81].
- Webb's extension [Web] to three players of Austin's two-individual moving knife procedure [Aus82].
- The Brams-Taylor-Zwicker 'rotating pie plate' procedure [BTZ97].

A summary of these and other procedures can be found in in [BTZ], or in Chapter 6 of [BT96].

The envy-free partition problem for more than three players remained unsolved for a long time. The various three-individual procedures are based on clever and elegant ideas, but it was not clear how any of them could be generalized. Recently Brams, Taylor, and Zwicker [BTZ97] found a four-individual envy-free 'moving knife' procedure. Finally, Brams and Taylor [BT95] invented a general envy-free procedure for I players, which exploits the concept of 'irrevocable advantage' developed in the Selfridge-Conway procedure. The Brams-Taylor procedure is too complex to describe here; we refer the reader to [BT95] or Chapter 7.4 of [BT96].

11B Equitable Partitions & 'Adjusted Winner'

Prerequisites: §8, §9A, §10C. **Recommended:** §11A

In §9D, we discussed how the Dubins-Spanier procedure is 'fairer' than the Banach-Knaster procedure because Dubins-Spanier is *symmetric*, meaning that it doesn't favour the 'first' player over the 'second' player or vice versa. However, Dubins-Spanier may still inadvertentely discriminate against a certain player because of the structure of his utility measure (regardless

of whether he is 'first', 'second', or whatever). We seek a procedure which treats all players equally, regardless of the nature of their preferences.

A partition $\mathcal{P} = \{\mathbf{P}_1, \ldots, \mathbf{P}_I\}$ is *equitable* if $\mu_1[\mathbf{P}_1] = \mu_2[\mathbf{P}_2] = \cdots = \mu_I[\mathbf{P}_I]$. In other words, each player's assessment of her own portion is the same as every *other* player's assesment of *his own* portion; no one has somehow been 'favoured' by the partition. A partition procedure (or game) is *equitable* if it always produces an equitable partition.

Example 11B.1: 'Moving Knife' is not equitable

Recall the Dubins-Spanier 'Moving Knife' Game (Game 9D.3). Suppose Owen is the first individual to say 'cut'; then Owen gets a portion $\mathbf{P}_1$ such that $\mu_1[\mathbf{P}_1] = \frac{1}{I}$.

But if $\mathbf{X}_1 = \mathbf{X} \setminus \mathbf{P}_1$, then everyone *else* thinks $\mathbf{X}_1$ is worth *more* than $\frac{I-1}{I}$ of the total value of $\mathbf{X}$. Thus, if Twyla is the *second* individual to say 'cut', then Twyla gets a portion $\mathbf{P}_2$ so that

$$\mu_2[\mathbf{P}_2] \quad = \quad \left(\frac{1}{I-1}\right) \mu_2[\mathbf{X}_1] \quad > \quad \left(\frac{1}{I-1}\right) \cdot \left(\frac{I-1}{I}\right) \quad = \quad \frac{1}{I} \quad = \quad \mu_1[\mathbf{P}_1].$$

Thus, Twyla is favoured over Owen.

Loosely speaking, the individual who says 'cut' first (i.e. Owen) is the 'least greedy' (at least, measuring things from the left end of the cake), whereas people who wait longer to say 'cut' are 'more greedy'. Thus, the 'Moving-Knife' procedure favours 'greedy' people. ♢

Clearly, equitability implies symmetry, but Example 11B.1 shows that symmetry does not imply equitability. Equitable partition procedures have two advantages:

1. No player is favoured or harmed by the intrinsic structure of his preferences (e.g. 'greedy' players are not favoured over 'nongreedy' players).

2. Equitability combined with Pareto-optimality yields envy-freedom, by the next lemma.

Lemma 11B.2 *Suppose $\mathcal{I} = \{1, 2\}$. If $\mathcal{P}$ is an equitable, Pareto-optimal partition, then $\mathcal{P}$ is also envy-free (and thus, proportional).*

Proof: **Exercise 11.2** □

The Brams-Taylor Adjusted Winner partition: Recall the definition of a *utility ratio threshold* (URT) partition $\mathcal{P}^{(\theta)}$ from §10C. We saw that the 'Highest Bidder' partition $\mathcal{P}^{(1)}$ seemed like a good idea, but often produced highly unbalanced outcomes (Example 10C.1). However, 'Highest Bidder' is a good starting point from which to build a fair partition.

Proposition 11B.3 *There exists a $\theta_0 \in [0, \infty)$ and a θ_0-utility ratio threshold partition $\mathcal{Q}$ which is equitable and Pareto-optimal, and thus, envy-free.*

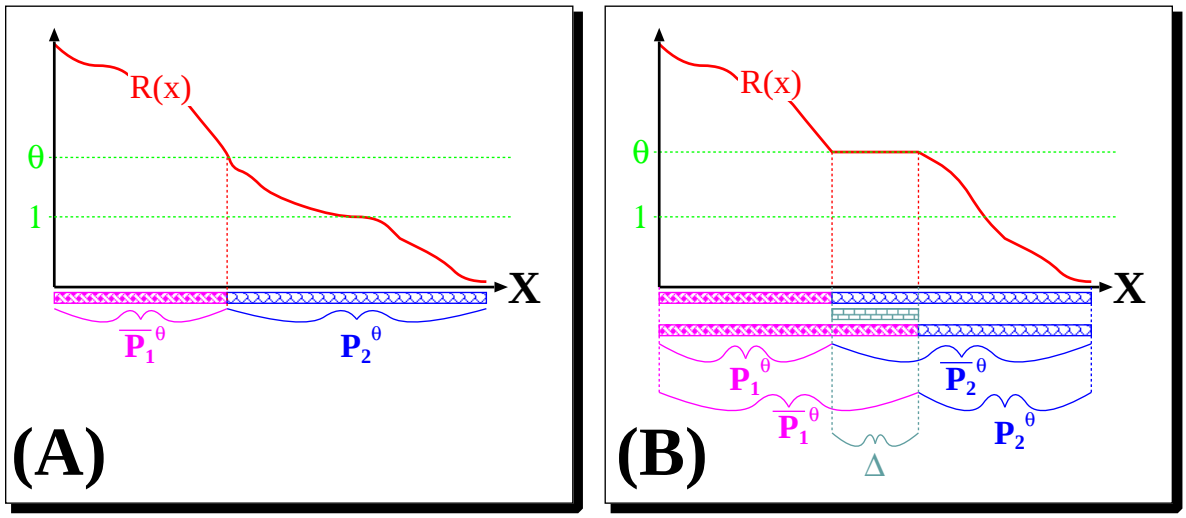


Figure 11.3: The Adjusted Winner Procedure

Proof: Let $R(x) = U_1(x)/U_2(x)$ for all $x \in \mathbf{X}$. For any $\theta \geq 1$, define the θ-URT partition $\mathcal{P}^{(\theta)} = \{\overline{\mathbf{P}}_1^{(\theta)}, \mathbf{P}_2^{(\theta)}\}$ by

$$\overline{\mathbf{P}}_1^{(\theta)} = \{x \in \mathbf{X}\,;\, R(x) \geq \theta\} \quad \text{and} \quad \mathbf{P}_2^{(\theta)} = \{x \in \mathbf{X}\,;\, R(x) < \theta\}. \qquad \text{(Fig.11.3A)}$$

Thus, if $\theta = 1$, then $\mathcal{P}^{(1)} = \{\overline{\mathbf{P}}_1^{(1)}, \mathbf{P}_2^{(1)}\}$ is the Highest Bidder partition (see §10C).

Proposition 10C.2 says $\mathcal{P}^{(1)}$ is Pareto-optimal. If $\mu_1[\overline{\mathbf{P}}_1^{(1)}] = \mu_2[\mathbf{P}_2^{(1)}]$ then $\mathcal{P}^{(1)}$ is also equitable, so set $\mathcal{Q} := \mathcal{P}^{(1)}$ and we're done.

Otherwise, assume without loss of generality that $\mu_1[\overline{\mathbf{P}}_1^{(1)}] > \mu_2[\mathbf{P}_2^{(1)}]$ like in Example 10C.1 (if not, then switch the players). As we increase θ, observe that $\mu_1[\overline{\mathbf{P}}_1^{(\theta)}]$ decreases, while $\mu_2[\mathbf{P}_2^{(\theta)}]$ increases. In the limit as θ goes to ∞, we have $\overline{\mathbf{P}}_1^{(\infty)} = \emptyset$ and $\mathbf{P}_2^{(\infty)} = \mathbf{X}$ (i.e. Owen gets nothing, and Twyla gets everything).

Let $\Theta \in [1, \infty)$ be the largest value of θ such that $\mu_1[\overline{\mathbf{P}}_1^{(\theta)}] \geq \mu_2[\mathbf{P}_2^{(\theta)}]$.

CASE 1: If $\mu_1[\overline{\mathbf{P}}_1^{(\Theta)}] = \mu_2[\mathbf{P}_2^{(\Theta)}]$, then $\mathcal{P}^{(\Theta)}$ is equitable, and Proposition 10C.2 already says $\mathcal{P}^{(\Theta)}$ is Pareto-optimal, so set $\mathcal{Q} := \mathcal{P}^{(\Theta)}$, and we're done.

CASE 2: If $\mu_1[\overline{\mathbf{P}}_1^{(\Theta)}] > \mu_2[\mathbf{P}_2^{(\Theta)}]$, then define a new Θ-URT partition $\widetilde{\mathcal{P}}^{(\Theta)} = \{\mathbf{P}_1^{(\Theta)}, \overline{\mathbf{P}}_2^{(\Theta)}\}$ by

$$\mathbf{P}_1^{(\Theta)} = \{x \in \mathbf{X}\,;\, R(x) > \Theta\} \quad \text{and} \quad \overline{\mathbf{P}}_2^{(\Theta)} = \{x \in \mathbf{X}\,;\, R(x) \leq \Theta\}. \qquad \text{(Fig.11.3B)}$$

Note: $\mathbf{P}_1^{(\Theta)} \subset \overline{\mathbf{P}}_1^{(\Theta)}$ and $\overline{\mathbf{P}}_2^{(\Theta)} \supset \mathbf{P}_2^{(\Theta)}$. Thus, $\mu_1[\mathbf{P}_1^{(\Theta)}] \leq \mu_1[\overline{\mathbf{P}}_1^{(\Theta)}]$ and $\mu_2[\overline{\mathbf{P}}_2^{(\Theta)}] \geq \mu_2[\mathbf{P}_2^{(\Theta)}]$.

Claim 1: $\mu_1[\mathbf{P}_1^{(\Theta)}] \leq \mu_2[\overline{\mathbf{P}}_2^{(\Theta)}]$.

Proof: **Exercise 11.3** Hint: First, note that, for any $\theta > \Theta$, $\overline{\mathbf{P}}_1^{(\theta)} \subset \mathbf{P}_1^{(\Theta)}$ and $\mathbf{P}_2^{(\theta)} \supset \overline{\mathbf{P}}_2^{(\Theta)}$. Thus, $\mu_1[\overline{\mathbf{P}}_1^{(\theta)}] \leq \mu_1[\mathbf{P}_1^{(\Theta)}]$ and $\mu_2[\mathbf{P}_2^{(\theta)}] \geq \mu_2[\overline{\mathbf{P}}_2^{(\Theta)}]$.

Next, prove that $\lim_{\theta \searrow \Theta} \mu_1[\overline{\mathbf{P}}_1^{(\theta)}] = \mu_1[\mathbf{P}_1^{(\Theta)}]$ and that $\lim_{\theta \searrow \Theta} \mu_2[\mathbf{P}_2^{(\theta)}] = \mu_2[\overline{\mathbf{P}}_2^{(\Theta)}]$.

Now prove the claim by using the fact that $\Theta \in [0, \infty)$ is the *largest* value of θ such that $\mu_1[\overline{\mathbf{P}}_1^{(\theta)}] \geq \mu_2[\mathbf{P}_2^{(\theta)}]$. ◇ Claim 1

Now there are two subcases

Case 2.1: If $\mu_1[\mathbf{P}_1^{(\Theta)}] = \mu_2[\overline{\mathbf{P}}_2^{(\Theta)}]$, then $\widetilde{\mathcal{P}}^{(\Theta)}$ is equitable, and Proposition 10C.2 already says $\widetilde{\mathcal{P}}^{(\Theta)}$ is Pareto-optimal, so set $\mathcal{Q} := \widetilde{\mathcal{P}}^{(\Theta)}$, and we're done.

Case 2.2: Suppose $\mu_1[\mathbf{P}_1^{(\Theta)}] < \mu_2[\overline{\mathbf{P}}_2^{(\Theta)}]$. Then loosely speaking, $\mathbf{P}_1^{(\Theta)}$ is too small, while $\overline{\mathbf{P}}_1^{(\Theta)}$ is too big. Likewise, $\overline{\mathbf{P}}_2^{(\Theta)}$ is too *big*, while $\mathbf{P}_2^{(\Theta)}$ is too *small*. Let

$$\Delta := \{x \in \mathbf{X} \,;\, R(x) = \Theta\} = \overline{\mathbf{P}}_1^{(\Theta)} \setminus \mathbf{P}_1^{(\Theta)} = \overline{\mathbf{P}}_2^{(\Theta)} \setminus \mathbf{P}_2^{(\Theta)} \qquad \text{(Fig.11.3B)}$$

Claim 2: *There exists a family of subsets* $\{\Delta_r \subseteq \Delta \,;\, r \in [0,1]\}$ *such that:*

(a) $\Delta_0 = \emptyset$ *and* $\Delta_1 = \Delta$.

(b) *If* $s < r$, *then* $\Delta_s \subset \Delta_r$.

(c) *Define* $f_1(r) := \mu_1[\Delta_r]$ *and* $f_2(r) := \mu_2[\Delta_r]$ *for all* $r \in [0,1]$. *Then* $f_1 : [0,1] \longrightarrow \mathbb{R}$ *and* $f_2 : [0,1] \longrightarrow \mathbb{R}$ *are continuous nondecreasing functions.*

Proof: **Exercise 11.4** ◇ Claim 2

Now, for each $r \in [0,1]$, define $\mathbf{Q}_1^{(r)} := \mathbf{P}_1^{(\Theta)} \sqcup \Delta_r$ and $\mathbf{Q}_2^{(r)} := \overline{\mathbf{P}}_2^{(\Theta)} \setminus \Delta_r$. Let $F(r) := \mu_1[\mathbf{Q}_1^{(r)}] - \mu_2[\mathbf{Q}_2^{(r)}]$.

Claim 3: $F(0) < 0 < F(1)$.

Proof: Observe that $\mathbf{Q}_1^{(0)} = \mathbf{P}_1^{(\Theta)}$ and $\mathbf{Q}_2^{(0)} = \overline{\mathbf{P}}_2^{(\Theta)}$. Hence $F(0) = \mu_1[\mathbf{P}_1^{(\Theta)}] - \mu_2[\overline{\mathbf{P}}_2^{(\Theta)}] < 0$ by hypothesis.

Likewise, $\mathbf{Q}_1^{(1)} = \overline{\mathbf{P}}_1^{(\Theta)}$ and $\mathbf{Q}_2^{(1)} = \mathbf{P}_2^{(\Theta)}$. Hence $F(1) = \mu_1[\overline{\mathbf{P}}_1^{(\Theta)}] - \mu_2[\mathbf{P}_2^{(\Theta)}] > 0$ by hypothesis. ◇ Claim 3

Claim 4: $F : [0,1] \longrightarrow \mathbb{R}$ *is continuous*

Proof: Observe that $\mu_1[\mathbf{Q}_1^{(r)}] = \mu_1[\mathbf{P}_1^{(\Theta)}] + f_1(r)$ and $\mu_2[\mathbf{Q}_2^{(r)}] = \mu_2[\overline{\mathbf{P}}_2^{(\Theta)}] - f_2(r)$.

$$\begin{aligned} \text{Thus,} \quad F(r) &= \mu_1[\mathbf{Q}_1^{(r)}] - \mu_2[\mathbf{Q}_1^{(r)}] = \Big(\mu_1[\mathbf{P}_1^{(\Theta)}] + f_1(r)\Big) - \Big(\mu_2[\overline{\mathbf{P}}_2^{(\Theta)}] - f_2(r)\Big) \\ &= \Big(\mu_1[\mathbf{P}_1^{(\Theta)}] - \mu_2[\overline{\mathbf{P}}_2^{(\Theta)}]\Big) + f_1(r) + f_2(r), \end{aligned}$$

and f_1 and f_2 are continuous by Claim 2(c). ◇ Claim 4

Thus, Claims 3 and 4 and the Intermediate Value Theorem together imply that there is some $r \in [0,1]$ such that $F(0) = 0$, which means that $\mu_1[\mathbf{Q}_1^{(r)}] = \mu_2[\mathbf{Q}_2^{(r)}]$ Thus, $\mathcal{Q} := \{\mathbf{Q}_1^{(r)}, \mathbf{Q}_2^{(r)}\}$ is equitable, and Proposition 10C.2 already says $\mathcal{Q}$ is Pareto-optimal, so we're done. □

The Brams-Taylor *Adjusted Winner* partition (AWP) is the URT $\mathcal{Q}$ of Proposition 11B.3. There is no practical procedure to *exactly* compute the AWP in real situations, because doing so would require complete information about the functions U_1 and U_2, which is potentially an infinite amount of information. In practical applications, we assume that we can divide $\mathbf{X}$ into some fine partition $\mathcal{R} = \{\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_M\}$ so that U_1 and U_2 are *constant* on each set $\mathbf{R}_m$. We can then ask the players to express their preferences by 'bidding' on each set $\mathbf{R}_m$. We do this as follows:

1. Each player is given some finite collection of 'points' (say 1000).
2. The players can then 'spend' her points to place 'bids' on each of the sets $\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_M$, with the understanding that the highest bidder will be (initially) awarded each set. ('Fractional' bids are allowed).
3. Each player thus has an incentive to bid more points on those subsets she values most, and not to squander points on subsets she values less. (In other words, her minimax strategy is to bid 'honestly'.) Thus, we expect that the distribution of i's bidding points will be a good approximation of her utility function U_i.
4. We then compute the Highest Bidder partition $\mathcal{P}^{(1)}$. If $\mathcal{P}^{(1)}$ equitable, then we're done. If $\mathcal{P}^{(1)}$ is not equitable, we slowly slide the threshold θ up or down (as appropriate), until we reach a partition $\mathcal{P}^{(\theta)}$ which is equitable. Then we stop.

The outcome is that each player gets those parts of the cake she desires 'most'. The balance point between the desires of Owen and the desires of Twyla is chosen to make the partition equitable.

Brams and Taylor propose this procedure as a way to divide a collection of goods (e.g. in a divorce settlement). We can imagine that each of the subsets $\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_M$ represents some physical item; the 'bidding' procedure is how the players express their preferences for different items. Brams and Taylor point out one advantage of the AWP: with possibly a single exception, every one of the subsets $\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_M$ will either fall entirely into $\mathbf{P}_1$ or entirely into $\mathbf{P}_2$. Thus, if the items are difficult or impossible to 'share', then we are reduced to the problem of 'sharing' at most *one* item. (If the players cannot come to an agreement on how to share this item, then the item can be sold and the profits split).

Brams and Taylor caution that the Adjusted Winner procedure is manipulable (see §11C.4); one player can seek to obtain an advantage by lying about her preferences (i.e. bidding dishonestly). A subtle fallacy is to think: 'since the outcome is equitable by definition, any dishonesty on the part of one player will help (or hurt) *both* players equally'. The problem here is that the outcome is only equitable according to the players' *stated* preferences, not their *true* preferences. *Manipulation* occurs exactly when one player lies about her preferences. She thereby engineers an outcome which *appears* equitable when in fact it is not.

Further reading: Brams and Taylor introduce the Adjusted Winner procedure in Chapter 4 of [BT96]. They also propose a second procedure for producing equitable, envy-free two-player partitions, called *Proportional Allocation* (PA). PA does *not* yield Pareto-optimal partitions, but Brams and Taylor claim PA is probably less manipulable than AWP (although PA is still manipulable).

In Chapter 5 of [BT96], Brams and Taylor sketch the application of AWP to various disputes (some real, some hypotetical). In a more recent book [BT00], they have fleshed out the practical applications of AWP in much greater detail, promoting it as a broadly applicable conflict-resolution procedure.

11C Other issues

In this section we'll briefly look at four other issues: *entitlements*, *indivisible value*, *chores*, and *manipulation*.

11C.1 Entitlements

All bad precedents began as justifiable measures. —Gaius Julius Caesar

So far we've only considered partition problems where all participants get an 'equal' share. However, in some partition problems, there may be good reasons (either moral, legal, or political) for giving some parties a *larger* share than others. For example:

- In settling an inheritance, the spouse of the deceased may be entitled to $\frac{1}{2}$ of the estate, and each of the three children entitled to $\frac{1}{6}$th of the estate (as each sees it).

- In settling an international territorial dispute, the peace treaty might stipulate that each country gets a portion of territory proportional to its population (or military strength, or ancient historical claim, or whatever). If country A has twice as many people as country B, then country A is entitled to twice as large/valuable a portion (where different countries may value different things).

- In a coalition government, the political parties making up the ruling coalition must divide important cabinet positions amongst themselves. In theory, each political party is entitled to a portion of 'government power' proportional to its popular support (or proportional to the number of seats it obtained in the legislature, which is not the same thing). Different parties will have different estimates of the 'power' of various government ministries (one party may think that the Defense Minister is the more powerful than the Finance Minister; another may think the opposite).

- In a multinational body (e.g. the UN, NATO, WTO, etc.), the member states send delegates to a governing Council. The problem is how to allocate power within the Council.

  One solution is that each state should receive numerical representation within the Council proportional to its population (or wealth, or military strength, or whatever). Thus, if

state A has three times the population of state B, then state A should have three times as many votes in the Council. However, in §1D, we saw the 'voting power' of a country is *not* simply the number of votes it has; some countries may have one or more votes, but actually have *no* power, while other countries get too much. So a better solution is that each state should get *voting power* on the Council proportional to its population (or wealth, strength, etc.). However, there are several different 'voting power indices', which give different measures of voting power, and it's not clear which index is correct. Different states may measure their power using different indices, and thus, have different opinions about what 'fair' representation means.

The states must also divide important government positions amongst themselves. For example, who chairs the Council? Who gets to be on important subcommittees? The problem is similar to that of a coalition government, only now with entire states instead of political parties.

If I parties are trying to partition a set $\mathbf{X}$, then an *entitlement vector* is a vector $\mathbf{E} = (e_1, \ldots, e_I)$, where $e_i \geq 0$ for all $i \in [1..I]$, such that $e_1 + \cdots + e_I = 1$. Here, e_i represents the portion that i is 'entitled' to receive. The *equidistributed entitlement* $\mathbf{E}_0 = (\frac{1}{I}, \frac{1}{I}, \ldots, \frac{1}{I})$ represents the case where all parties are entitled to an equal share.

A partition $\mathcal{P} = \{\mathbf{P}_1, \ldots \mathbf{P}_I\}$ is called **E**-*proportional* if, for each $i \in [1..I]$, $\mu[\mathbf{P}_i] \geq e_i$. Thus, each player i thinks she got *at least* her 'entitled share' of e_i. Notice that a normal 'proportional' partition is just an $\mathbf{E}_0$-proportional partition, where $\mathbf{E}_0 = (\frac{1}{I}, \ldots, \frac{1}{I})$.

A partition $\mathcal{P} = \{\mathbf{P}_1, \ldots \mathbf{P}_I\}$ is called **E**-*envy-free* if, for each $i \in [1..I]$, and each $j \neq i$,

$$\frac{\mu_i[\mathbf{P}_i]}{e_i} \quad \geq \quad \frac{\mu_i[\mathbf{P}_j]}{e_j}$$

This means: each player i thinks she got at least her entitled share e_i. Furthermore, if she thinks any other player j was 'overpaid' more than his entitled share, then she thinks that she himself got overpaid *even more* by comparison. Hence, she will not envy his[1].

Suppose that $\mathbf{E}$ is a *rational* vector, meaning that $e_1, \ldots, e_I$ are rational numbers. Then all of the I-person proportional partition procedures of §9 generalize to yield **E**-proportional partitions. Also, the Brams-Taylor I-person envy-free procedures mentioned at the end of §11A generalizes to yield **E**-envy-free partitions. The key in both cases is to increase the size of I, as follows:

1. Suppose $\mathcal{I} = \{1, \ldots, I\}$ are I players, with rational entitlement vector $\mathbf{E} = (e_1, \ldots, e_I)$.

   Let D be the greatest common denominator of $e_1, \ldots, e_i$. Thus, we can write:

   $$e_1 \; = \; \frac{c_1}{D}; \quad e_2 \; = \; \frac{c_2}{D}; \quad \ldots \quad e_I \; = \; \frac{c_I}{D}.$$

   for some $c_1, \ldots, c_I \in \mathbb{N}$ such that $c_1 + \cdots + c_I = D$.

[1] Of course, this is assuming that all players feel that the original entitlement vector $\mathbf{E}$ is fair to begin with. If $e_j = 2e_i$, then i may 'envy' j because he has twice as large an entitlement as she does. But this is beyond the scope of our discussion.

2. Now, construct a new partition problem involving a set $\mathcal{W}$ of D players:

$$\mathcal{W} \quad = \quad \{w_{1,1}, \ldots, w_{1,c_1},\ w_{2,1}, \ldots, w_{2,c_2},\ \ldots, w_{2,I}, \ldots, w_{I,c_I}\}$$

Imagine that $\{w_{1,1}, \ldots, w_{1,c_1}\}$ are c_1 distinct 'clones' of Owen (so they all have utility measure μ_1). Likewise, $\{w_{2,1}, \ldots, w_{2,c_2}\}$ are c_2 distinct 'clones' of Twyla (with utility measure μ_2), etc.

3. Apply the partition procedure of your choice to yield a proportional/envy-free partition

$$\mathcal{Q} \quad = \quad \{\mathbf{Q}_{1,1}, \ldots, \mathbf{Q}_{1,c_1},\ \mathbf{Q}_{2,1}, \ldots, \mathbf{Q}_{2,c_2},\ \ldots, \mathbf{Q}_{2,I}, \ldots, \mathbf{Q}_{I,c_I}\}$$

amongst the players of $\mathcal{W}$.

4. Now define $\mathbf{P}_1 := \mathbf{Q}_{1,1} \sqcup \cdots \sqcup \mathbf{Q}_{1,c_1}$, $\quad \mathbf{P}_2 := \mathbf{Q}_{2,1} \sqcup \cdots \sqcup \mathbf{Q}_{2,c_2}$, etc. Then $\mathcal{P} = \{\mathbf{P}_1, \ldots, \mathbf{P}_I\}$ is a partition of $\mathbf{X}$ amongst $\{1, \ldots, I\}$.

Observe (**Exercise 11.5**) that:

$$\Big(\mathcal{Q} \text{ is proportional amongst } \mathcal{W}\Big) \implies \Big(\mathcal{P} \text{ is } \mathbf{E}\text{-proportional amongst } \mathcal{I}\Big).$$
$$\Big(\mathcal{Q} \text{ is envy-free amongst } \mathcal{W}\Big) \implies \Big(\mathcal{P} \text{ is } \mathbf{E}\text{-envy-free amongst } \mathcal{I}\Big).$$

11C.2 Indivisible Value

Most of the procedures we've considered assume that the value of the cake is, in principle, infinitely divisible. In other words, they assume that the utility measures of the player contain no *atoms*, or at least, relatively few atoms. However, in *real* fair division problems, there are many large components of indivisible value. For example:

- In an inheritance or divorce settlement, there may be large, indivisible items (a house, a car, a piano) which cannot realistically be shared amongst the disputants. If a mutually agreeable division cannot be found, a standard solution is to liquidate these assets and distribute the proceeds. This may not be satisfactory to anyone, however, because the house (for example) may have a sentimental value which far exceeds its market value. Liquidating the house and dividing the cash is a very suboptimal solution.

- In a territorial dispute, there may be particular sites (e.g. mines, cities, religious shrines), which are highly valued by one or more parties, and which cannot be shared between states.

- In dividing government posts amongst the various members of a coalition government, clearly the government posts themselves are indivisible entities. Two political parties cannot 'share' the position of President or of Finance Minister.

In the extreme case, the utility measure is *entirely atomic*, which means $\mathbf{X} = \{x_1, \ldots, x_m\} \sqcup \mathbf{Y}$, where $\mu\{x_1\} + \ldots + \mu\{x_m\} = 1$ and $\mu[\mathbf{Y}] = 0$. In this case, *none* of the methods in §9 -§11B are applicable.

One elegant procedure which can acommodate entirely atomic measures is Bronislaw Knaster's method of *Sealed Bids* [Ste48a]. We do not have time to discuss this procedure; we refer the reader to Section 3.2 of Brams and Taylor [BT96], or Section 14.9 of Luce and Raiffa [LR80]. Suffice it to say that Knaster's procedure guarantees a proportional partition of $\{x_1, \ldots, x_m\}$ amongst I players, and has the further advantage that the players need not even agree on the total value of the goods (i.e. Owen may think the whole cake is bigger than Twyla thinks it is). Knaster resolves these difficulties by introducing an infinitely divisible *numéraire* commodity (i.e. money) which the players can exchange to even out the inevitable unfairness of partitioning indivisible commodities. The disadvantage of Knaster's procedure is that it requires each player to enter the division problem with a pre-existing *bankroll* (i.e. a stash of cash) which can be used to 'pay off' other players. Thus, Knaster's procedure is inapplicable if the value of the cake exceeds the amount of money which one or more parties can realistically bring to the table.

Other fair division procedures have been proposed for indivisible commodities. For example, William F. Lucas [BT96, §3.3] has proposed a variant of the Dubins-Spanier procedure, but it is not guaranteed to work in all situations; Lucas must assume a property of 'linearity', which in practice means that there are many very small atoms and few or no large ones. Also, Brams and Taylor suggest that their Adjusted Winner procedure (§11B) is good for indivisible commodities, because the players will be forced to divide at most *one* of the atoms $\{x_1, \ldots, x_m\}$. In practical situations, this may require liquidating *one* asset, which is certainly better than liquidating *all* of them.

11C.3 Chores

Instead of partitioning a cake (i.e. a 'good' thing), suppose the players must partition a set of 'chores' (i.e. a 'bad' thing). Now each player does not want to recieve as *much* as possible, but instead, wants to get as *little* as possible.

For example, a partition $\mathcal{P} = \{\mathbf{P}_1, \ldots, \mathbf{P}_I\}$ of *chores* is *proportional* if $\mu_i[\mathbf{P}_i] \leq \frac{1}{I}$ for all $i \in [1..I]$. The partition $\mathcal{P}$ is *envy-free* if, for all $i, j \in [1..I]$, $\mu_i[\mathbf{P}_i] \leq \mu_i[\mathbf{P}_j]$. Finally, $\mathcal{P}$ is *equitable* if $\mu_1[\mathbf{P}_1] = \mu_2[\mathbf{P}_2] = \cdots = \mu_I[\mathbf{P}_I]$. Pareto-preference and Pareto-optimality are defined in the obvious way.

Most of the 'cake-cutting' procedures in §9 -§11B generalize to chore-division; we must simply modify them so that each player will systematically choose the 'smallest' portion they can (rather than the 'largest'). Here's one example:

Procedure 11C.1: I cut, you choose, for chores

Let $\mathbf{X} = [0, 1]$ be the unit interval (representing a one-dimensional set of chores). Let μ_1 and μ_2 be utility measures on $\mathbf{X}$. Assume μ_1 is at most $\frac{1}{2}$ atomic.

(1) Let $r \in [0,1]$ be such that $\mu_1[0,r) = \frac{1}{2} = \mu_1[r,1]$ (i.e. Owen divides the chores into two pieces which he perceives have equal size; this is possible because μ_1 is at most $\frac{1}{2}$ atomic)

(2a) If $\mu_2[0,r) \leq \mu_2[r,1]$, then define $\mathbf{P}_2 = [0,r)$ and $\mathbf{P}_1 = [r,1]$. (If Twyla thinks that $[0,r)$ is *smaller*, she takes this piece, and Owen takes the other one).

(2b) Otherwise, if $\mu_2[0,r) > \mu_2[r,1]$, then define $\mathbf{P}_1 = [0,r)$ and $\mathbf{P}_2 = [r,1]$. (If Twyla thinks that $[r,1]$ is *smaller*, she takes this piece, and Owen takes the other one).

Now let $\mathcal{P} = \{\mathbf{P}_1, \mathbf{P}_2\}$. (**Exercise 11.6** Show that that $\mathcal{P}$ is proportional allocation of chores.)

We refer the reader to Sections 3.2.2 and 7.6.3 of [BT96] for further discussion of chore division.

11C.4 Nonmanipulability

Recall that a partitioning game Γ *yields* a partition procedure Π if each player of Γ has a unique maximin strategy, and if Γ will produce the same partition as Π when all players play their maximin strategies. A partition game is designed so that your maximin strategy is an 'honest' expression of your preferences. Thus, if you are rational and risk-averse, your 'best' strategy is simply to be honest about what you want. If everyone chooses their 'best' strategy in this way, then everyone will be 'honest', and the outcome will be a 'fair' division.

However, your maximin strategy is only your 'best' strategy when you know *nothing* about the preferences and strategies of the other players. Your maximin strategy simply optimizes your 'worst-case scenario', based on total ignorance of what everyone else is doing. If you *know* (or at least, suspect) what the other players will do, you can exploit this information by picking a strategy which is *not* your maximin, but which will yield a superior outcome if the other people play like you expect them to. This is called *manipulating* the partition game.

Example 11C.2: The Divider's Advantage in 'I cut, you choose'

Suppose the cake is half tartufo, half orange cream, and Owen knows that Twyla only likes tartufo. He likes tartufo and orange cream equally. Clearly, the 'fairest' partition is into two pieces **A** and **B** such that:

A contains *all* the orange cream (50% of the whole cake).

B contains *all* the tartufo (the other 50% of the whole cake).

Twyla will take **B**, and Owen can take **A**. Both will be happy. Twyla got everything she wanted, and Owen herself is indifferent between the two pieces.

However, Owen can also *exploit* Twyla as by cutting the cake into two pieces as follows:

A contains *all* the orange cream and 48% of the tartufo (a total of 74% of the whole cake).

B contains the remaining 52% of the tartufo (only 26% of the whole cake).

Because **B** has (slightly) more tartufo, Owen can expect Twyla to choose **B**, leaving him with the disproportionately large piece **A** of the cake.

Owen's opportunity to manipulate 'I cut, you choose' is sometimes called the *Divider's Advantage.* ♢

A partition game is *nonmanipulable* if the maximin strategy for each player is also a *dominant* strategy for that player. In other words, no matter what the other players plan to do, it is *always* rational for you to play your maximin (i.e. 'honest') strategy.

At present, no nonmanipulable partition games are known (indeed, there may be an 'impossibility theorem' lurking here). However, Brams and Taylor suggest an interesting strategy [BT96, §4.5]. Although we cannot find a *single* game where a player's 'honest' maximin strategy is dominant, perhaps we can find a *sequence* of games $\Gamma_1, \Gamma_2, \ldots, \Gamma_M$ (all using the same strategy sets for each player) such that, for each player $i \in \mathcal{I}$,

- i has the same maximin strategy in all of $\Gamma_1, \Gamma_2, \ldots, \Gamma_M$.
- For any I-tuple of strategies $\mathbf{s} = (s_1, \ldots, s_I)$ where i does *not* use her maximin strategy, there is at least one game Γ_m where $\mathbf{s}$ yields an outcome for i which is *strictly worse* than her maximin outcome on Γ_m.

The sequence $\Gamma_1, \ldots, \Gamma_M$ now defines a 'supergame' with the following rules:

1. Everyone plays Γ_1. If everyone is satisfied with the outcome, then the game ends here.
2. If even *one* player is dissatisfied, and feels that Γ_1 has been 'manipulated' by someone else, then the dissatisfied player can unilaterally nullify the results of Γ_1.

   The players move on and play Γ_2, with the proviso that *everyone must use the same strategies they used in* Γ_1.
3. If someone is dissatisfied with the results of Γ_2, then the players move onto Γ_3, and so forth.

Thus, if Owen suspects Twyla of a manipulating the outcome of Γ_1 by *not* using her maximin strategy, then Owen can force Twyla to play the same 'dishonest' strategy in Γ_2, Γ_3, etc. In at least *one* of these games, Twyla's strategy (if it really wasn't her maximin strategy) will produce a strictly inferior outcome for Twyla. The *threat* of this possibility makes it rational for Twyla to be honest in Γ_1.

Further Reading

The most extensive resource on fair division procedures and games is by Brams and Taylor [BT96]; we strongly recommend this to the interested reader. For a completely different approach to the problem of fair division (almost disjoint from that presented here), please see

[Mou03]. Brams and Taylor (and the past few chapters of this book) are concerned with games which 'implement' certain notions of fairness. In contrast, Moulin is concerned with the normative question of what the word 'fair' itself means. He contrasts a variety of definitions, each of which seems perfectly correct in some contexts and yet is inappropriate in others. Moulin's book is filled with a multitude of concrete examples, ranging from resource divison to taxation regimes to the location of public facilities. These examples highlight the strengths and weaknesses of various 'fairness' notions, and also demonstrate the relevance of the theory of fair division to problems far beyond cake cutting.

Bibliography

[Abe02] Francine F. Abeles, editor. *The Political Pamphlets and Letters of Charles Lutwidge Dodgson and Related Pieces*. Lewis Carroll Society of North America, 2002.

[Aki95] Ethan Akin. Vilfredo Pareto cuts the cake. *Journal of Mathematical Economics*, 24:23–44, 1995.

[All77] Glen O. Allen. Beyond the Voter's Paradox. *Ethics*, 88:50–61, October 1977.

[All82] Glen O. Allen. Formal decision theory and majority rule. *Ethics*, 92:199–206, January 1982.

[Ans76] G.E.M. Anscombe. On the frustration of the majority by fulfilment of the maority's will. *Analysis*, 36:161–168, 1976.

[Arr63] Kenneth J. Arrow. *Individual Values and Social Choice*. John Wiley & Sons, New York, 2nd edition, 1963.

[Aus82] A.K. Austin. Sharing a cake. *Mathematical Gazette*, 66(437):212–215, 1982.

[Axe85] Robert Axelrod. *The Evolution of Cooperation*. Basic Books, reprint edition, 1985.

[Axe97] Robert Axelrod. *The Complexity of Cooperation*. Princeton UP, 1997.

[Bar99] Julius Barbanel. Partition ratios, Pareto optimal cake division, and related notions. *Journal of Mathematical Economics*, 32:401–428, 1999.

[Bar00] Julius Barbanel. On the structure of Pareto optimal cake partitions. *Journal of Mathematical Economics*, 33:401–424, 2000.

[Bec87] Anatole Beck. Constructing a fair border. *American Mathematical Monthly*, 94(2):157–162, February 1987.

[BF78] Steven J. Brams and Peter C. Fishburn. Approval voting. *American Political Science Review*, 72:831–847, 1978.

[Bin87] Ken Binmore. Nash bargaining theory ii. In Ken Binmore and Partha Dasgupta, editors, *The Economics of Bargaining*, pages 61–76. Blackwell, Oxford, 1987.

[Bin91] Ken Binmore. *Fun and Games*. D.C. Heath, Cambridge, Mass., 1991.

[Bin93] Ken Binmore. *Game theory and the social contract I: Playing Fair*. MIT Press, Cambridge, Mass., 1993.

[Bin98] Ken Binmore. *Game theory and the social contract II: Just Playing*. MIT Press, Cambridge, Mass., 1998.

[BKZ93] Steven J. Brams, D. Marc Kilgour, and William S. Zwicker. A new paradox of vote aggregation. In *Annual Meeting*. American Political Science Association, 1993.

[Bla58] Duncan S. Black. *Theory of committees and elections*. Cambridge UP, Cambridge, UK, 1958.

[Bor81] Jean-Charles Borda. Memoire sur les elections au Scrutin. *Histoire de l'Academie Royale des Sciences*, 1781.

[Bra55] Richard B. Braithwaite. *Theory of games as a tool for the moral philosopher*. Cambridge University Press, Cambridge, UK, 1955.

[Bra90] Steven J. Brams. *Negotiating Games*. Routledge, New York, revised edition, 1990.

[BRW86] Ken Binmore, Ariel Rubinstein, and Asher Wolinsky. The Nash bargaining solution in economic modelling. *Rand J. Econom.*, 17(2):176–188, 1986.

[BT95] Stephen J. Brams and Alan D. Taylor. An envy-free cake division protocol. *American Mathematical Monthly*, 102(1):9–18, January 1995.

[BT96] Stephen J. Brams and Alan D. Taylor. *Fair Division: from cake-cutting to dispute resolution*. Cambridge UP, Cambridge, 1996.

[BT00] Stephen J. Brams and Alan D. Taylor. *The Win-Win Solution: Guaranteeing Fair Shares to Everybody*. Norton, 2000.

[BTZ] Stephen J. Brams, Alan D. Taylor, and William S. Zwicker. Old and new moving-knife schemes. *Mathematical Intelligencer*.

[BTZ97] Stephen J. Brams, Alan D. Taylor, and William S. Zwicker. A moving knife solution to the four-person envy-free cake-division problem. *Proceedings of the AMS*, 125(2):547–554, February 1997.

[Cao82] X. Cao. Preference functions and bargaining solutions. In *Proceeedings of the 21st IEEE Conference on Decision and Control*, volume 1, pages 164–171, 1982.

[Cla71] Edwin Clarke. Multipart pricing of public goods. *Public Choice*, 11:17–33, Fall 1971.

[ClMdC85] Jean-Antoine-Nicolas Caritat (le Marquis de Condorcet). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*, 1785.

[Col70] James S. Coleman. Political money. *American Political Science Review*, 64:1074–1087, 1970.

[dG77] Claude d'Aspremont and Louis Gevers. Equity and the informational basis of collective choice. *Review of Economic Studies*, 44:199–209, 1977.

[Dhi98] Amrita Dhillon. Extended Pareto rules and relative utilitarianism. *Soc. Choice Welf.*, 15(4):521–542, 1998.

[DM99] Amrita Dhillon and Jean-François Mertens. Relative utilitarianism. *Econometrica*, 67(3):471–498, 1999.

[Dod73] Charles Lutwidge Dodgson. A discussion of the various methods of procedure in conducting elections, 1873. reprinted in Black (1958) or Abeles (2002).

[DS61] Lester E. Dubins and Edwin H. Spanier. How to cut a cake fairly. *American Mathematical Monthly*, 68:1–17, 1961.

[Fin64] A.M. Fink. A note on the fair division problem. *Mathematics Magazine*, 37(5):341–342, November 1964.

[Fis73] Peter C. Fishburn. *The Theory of Social Choice*. Princeton Univ. Press, Princeton, NJ, 1973.

[FM98] Dan S. Felsenthal and Moshé Machover. *The measurement of voting power*. Edward Elgar Publishing Limited, Cheltenham, 1998. Theory and practice, problems and paradoxes.

[FR82] Peter C. Fishburn and Ariel Rubinstein. Time preference. *Internat. Econom. Rev.*, 23(3):677–694, 1982.

[Gär77] Peter Gärdenfors. A concise proof of a theorem on manipulation of socal choice functions. *Public Choice*, 32:137–140, Winter 1977.

[Gau86] David Gauthier. *Morals by Agreement*. Clarendon Press, Oxford, 1986.

[Gea01] John Geanokoplos. Three brief proofs of Arrow's Impossibility Theorem. Technical report, Cowles Foundation for Research in Economics; Yale University, June 2001. `http://cowles.econ.yale.edu`.

[Gib73] Allan Gibbard. Manipulation of voting schemes: a general result. *Econometrica*, 41:587–601, July 1973.

[GL79] Jerry R Green and Jean-Jacques Laffont. *Incentives in Public Decision Making*, volume 1 of *Studies in Public Economics*. North-Holland, Amsterdam, 1979.

[GM52] L.A. Goodman and Harry Markowitz. Social welfare functions based on individual rankings. *American Journal of Sociology*, 58:257–262, 1952.

[Goo77] I.J. Good. Justice in voting by demand-revalation. *Public Choice*, 11(2):65–70, 1977.

[Gro73] Theodore Groves. Incentives in teams. *Econometrica*, 41:617–631, 1973.

[Har53] John Harsanyi. Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy*, 61(434-435), 1953.

[Har55a] John Harsanyi. Cardinal welfare, individualistic ethics and interpersonal comparisons of utility. *J. Political Economy*, 63:309–321, 1955.

[Har55b] John Harsanyi. Cardinal welfare, individualstic ethics and interpersonal comparisons of utility. *Journal of Political Economy*, 63:309–321, 1955.

[Har56] John Harsanyi. Approaches to the bargaining problem before and after the theory of games: a critical discussion of Zeuthen's, Hicks, and Nash's theories. *Econometrica*, 24:144–157, 1956.

[Har82] Russell Hardin. Comment on formal decision theory and majority rule. *Ethics*, 92:207–210, January 1982.

[Hil53] Clifford Hildreth. Alternative conditions for social orderings. *Econometrica*, 21:81–94, 1953.

[Hil83] Theodore P. Hill. Determining a fair border. *American Mathematical Monthly*, 90(7):438–442, August 1983.

[HK05] Jonathan K. Hodge and Richard E. Klima. *The mathematics of voting and elections: a hands-on approach*, volume 22 of *Mathematical World*. American Mathematical Society, Providence, RI, 2005.

[HZ79] Aanund Hylland and Richard Zeckhauser. A mechanism for selecting public goods when preferences must be elicited. Technical Report KSG Discussion Paper 70D, Harvard University, August 1979.

[Int73] Michael D. Intriligator. A probabilistic model of social choice. *Review of Economic Studies*, 40(4):553–560, October 1973.

[Kal77] Ehud Kalai. Proportional solutions to bargaining situations: interpersonal utility comparisons. *Econometrica*, 45(7):1623–1630, 1977.

[Kar98] Edi Karni. Impartiality: definition and representation. *Econometrica*, 66(6):1405–1415, 1998.

[Kil83] Marc D. Kilgour. A formal analysis of the amending formula of Canada's Constitution Act. *Canadian Journal of Political Science*, 16:771–777, 1983.

[Kna46] Bronislaw Knaster. Sur le problem du partage de H. Steinhaus. *Annales de la Societé Polonaise de Mathematique*, 19:228–230, 1946.

[KR80] Ki Hang Kim and Fred W. Roush. *Introduction to Mathematical Consensus Theory.* Marcel Dekker, New York, 1980.

[KS75] Ehud Kalai and Meir Smorodinsky. Other solutions to Nash's bargaining problem. *Econometrica*, 43:513–518, 1975.

[Kuh] Harold W. Kuhn. On games of fair division. In Martin Shubik, editor, *Essays in Mathematical Economics in Honour of Oskar Morgenstern*, pages 29–37. Princeton UP, Princeton, NJ.

[Lai77] Charles R. Laine. Strategy in point voting: A note. *Quarterly Journal of Economics*, 91(3):505–507, August 1977.

[LC81] Saul X. Levmore and Elizabeth Early Cook. *Super Strategies for Puzzles and Games.* Doubleday, Garden City, NY, 1981.

[Lew69] David. Lewis. *Convention: A Philosophical Study.* Harvard UP, Cambridge, Mass., 1969.

[LR80] R. Duncan Luce and Howard Raiffa. *Games and Decisions.* Dover, New York, 1957,1980.

[LV05] Annick Laruelle and Federico Valenciano. Assessing success and decisiveness in voting situations. *Soc. Choice Welf.*, 24(1):171–197, 2005.

[Mas78] Eric Maskin. A theorem on utilitarianism. *Rev. Econom. Stud.*, 45(1):93–96, 1978.

[May52] Kenneth May. A set of independent, necessary, and sufficient conditions for simple majority decision. *Econometrica*, 20:680–684, 1952.

[MCWG95] Andreu Mas-Colell, Michael D. Whinston, and Jerry R. Green. *Microeconomic Theory.* Oxford UP, Oxford, UK, 1995.

[Mou84] Hervé Moulin. Implementing the Kalai-Smorodinsky bargaining solution. *J. Econom. Theory*, 33(1):32–45, 1984.

[Mou86] Hervé Moulin. Characterizations of the pivotal mechanism. *Journal of Public Economics*, 31(1):53–78, October 1986.

[Mou88] Hervé Moulin. *Axioms of cooperative decision making.* Cambridge University Press, Cambridge, UK, 1988.

[Mou03] Hervé Moulin. MIT Press, Cambridge, MA, 2003.

[MPV72] Dennis C. Mueller, Geoffrey C. Philpotts, and Jaroslav Vanek. The social gains from exchanging votes: A simulation approach. *Public Choice*, 13:55–79, Fall 1972.

[Mue67] Dennis C. Mueller. The possibility of a social welfare function: comment. *American Economic Review*, 57:1304–1311, December 1967.

[Mue71] Dennis C. Mueller. Fiscal federalism in a constitutional democracy. *Public Policy*, 19(4):567–593, Fall 1971.

[Mue73] Dennis C. Mueller. Constitutional democracy and social welfare. *Quarterly Journal of Economics*, 87(1):60–80, February 1973.

[Mue77] Dennis C. Mueller. Strategy in point voting: Comment. *Quarterly Journal of Economics*, 91(3):509, August 1977.

[Mue03] Dennis C. Mueller. *Public Choice III.* Cambridge UP, Cambridge, 2003.

[Mus59] Richard A. Musgrave. *The theory of public finance.* McGraw-Hill, New York, 1959.

[Mut99] Abhinay Muthoo. *Bargaining theory with applications*. Cambridge UP, Cambridge, UK, 1999.

[Mye77] Roger B. Myerson. Two-person bargaining problems and comparable utility. *Econometrica*, 45(7):1631–1637, 1977.

[Mye81] Roger B. Myerson. Utilitarianism, egalitarianism, and the timing effect in social choice problems. *Econometrica*, 49(4):883–897, 1981.

[Mye91] Roger B. Myerson. *Game theory: Analysis of conflict*. Harvard University Press, Cambridge, MA, 1991.

[Nap02] Stefan Napel. *Bilateral bargaining*, volume 518 of *Lecture Notes in Economics and Mathematical Systems*. Springer-Verlag, Berlin, 2002. Theory and applications.

[Nas50] John Nash. The bargaining problem. *Econometrica*, 18:155–162, 1950.

[Ng75] Yew-Kwang Ng. Bentham or Bergson? Finite sensibility, utility functions, and social welfare functions. *Review of Economic Studies*, 42:545–569, October 1975.

[Ng85] Yew-Kwang Ng. The utilitarian criterion, finite sensibility, and the weak majority preference principle. A response. *Soc. Choice Welf.*, 2(1):37–38, 1985.

[Ng00] Yew-Kwang Ng. From separability to unweighted sum: a case for utilitarianism. *Theory and Decision*, 49(4):299–312, 2000.

[Nit75] Shmuel Nitzan. Social preference ordering in a probabilistic voting model. *Public Choice*, 24(24):93–100, Winter 1975.

[Nit85] Shmuel Nitzan. The vulnerability of point-voting schemes to preference variation and strategic manipulation. *Public Choice*, 47(2):349–370, 1985.

[NPL80] Shmuel Nitzan, Jacob Paroush, and Shlomo I. Lampert. Preference expression and misrepresentation in points voting schemes. *Public Choice*, 35(4):421–436, 1980.

[Nur98] Hannu Nurmi. Voting paradoxes and referenda. *Soc. Choice Welf.*, 15(3):333–350, 1998.

[Nur99] Hannu Nurmi. *Voting paradoxes and how to deal with them*. Springer-Verlag, Berlin, 1999.

[OR94] Martin J. Osborne and Ariel Rubinstein. *A course in game theory*. MIT Press, Cambridge, MA, 1994.

[Ost03] Moise Ostrogorski. *La démocratie et l'organisation des partis politiques*. Calmann-Levy, Paris, 1903.

[Phi71] Geoffrey C. Philpotts. *Vote-trading in a competitive model*. PhD thesis, Cornell University, 1971.

[Phi72] Geoffrey C. Philpotts. Vote trading, welfare, and uncertainty. *Canadian Journal of Economics*, 5(3):358–372, August 1972.

[Rai53] Howard Raiffa. Arbitration schemes for generalized two-person games. In H.W. Kuhn and A.W. Tucker, editors, *Contributions to the Theory of Games* **II**, volume 28 of *Annals of Mathematical Studies*, pages 361–387. Princeton University Press, Princeton, NJ, 1953.

[Raw71] John Rawls. *A Theory of Justice*. Belknap Press, Cambridge, Mass., 1971.

[RD76] D. Rae and H. Daudt. The Ostrogorski paradox: a peculiarity of compound majority decision. *European Journal of Political Research*, 4:391–398, 1976.

[Rid98] Matt Ridley. *The Origins of Virtue : Human Instincts and the Evolution of Cooperation*. Penguin, 1998.

[Rik82] William H. Riker. *Liberalism against populism*. Waveland Press, Prospect Heights, Illinois, 1982.

[Roe98] John E. Roemer. *Theories of Distributive Justice*. Harvard UP, Cambridge, MA, 1998.

[Rub82] Ariel Rubinstein. Perfect equilibrium in a bargaining model. *Econometrica*, 50(1):97–109, 1982.

[Saa88] Donald G. Saari. Symmetry, voting, and social choice. *Math. Intelligencer*, 10(3):32–42, 1988.

[Saa90] Donald G. Saari. Consistency of decision processes. *Ann. Oper. Res.*, 23(1-4):103–137, 1990.

[Saa95] Donald G. Saari. *Basic Geometry of Voting*. Springer-Verlag, New York, 1995.

[Saa97] Donald G. Saari. Are individual rights possible? *Mathematics Magazine*, 70(2):83–92, April 1997.

[Sat75] Mark Satterthwaite. Strategy proofness and Arrow's conditions. *Journal of Economic Theory*, 10:187–217, October 1975.

[Seg00] Uzi Segal. Let's agree that all dictatorships are equally bad. *Journal of Political Economy*, 108(3):569–589, 2000.

[Sel65] Reinhard Selten. Spieltheoretische behandlung eines oligopmodells mit nachfrageträgheit. *Zeitschrift für die gesamte Staatswisenschaft*, 121(301-324 and 667-689), 1965.

[Sen70a] Amartya K. Sen. *Collective Choice and Social Welfare*. Holden-Day, San Francisco, CA, 1970.

[Sen70b] Amartya K. Sen. The impossibility of a Paretian liberal. *Journal of Political Economy*, 78:152–157, 1970.

[Sjo91] Tomas Sjostrom. A new characterization of the Groves-Clarke mechanism. *Economics Letters*, 36(3):263–267, July 1991.

[Sky96] Bryan Skyrms. *Evolution of the Social Contract*. Cambridge UP, 1996.

[Sky03] Bryan Skyrms. *The Stag Hunt and the Evolution of Social Structure*. Cambridge UP, 2003.

[Sob01] Joel Sobel. Manipulation of preferences and relative utilitarianism. *Games Econom. Behav.*, 37(1):196–215, 2001.

[SS54] Lloyd Shapley and Martin Shubik. A method for evaluating the distribution of power in a committee system. *American Political Science Review*, 48:787–792, 1954.

[Stå72] Ingolf Ståhl. *Bargaining Theory*. Economics Research Institute at the Stockholm School of Economics, Stockholm, 1972.

[Ste48a] Hugo Steinhaus. The problem of fair division. *Econometrica*, 16:101–104, 1948.

[Ste48b] Hugo Steinhaus. Sur la division pragmatique. *Econometrica* supplement, 17:315–319, 1948.

[Str80a] Philip Strafin. *Topics in the Theory of Voting*. Birkhauser, Boston, 1980.

[Str80b] Walter Stromquist. How to cut a cake fairly. *American Mathematical Monthly*, 87(8):640–644, October 1980.

[Str93] Philip Strafin. *Game Theory and Strategy*. Mathematical Association of America, 1993.

[Tay95] Alan D. Taylor. *Mathematics and Politics: Strategy, Voting, Power and Proof*. Springer-Verlag, New York, 1995.

[Tid77] T. Nicolaus Tideman, editor. *Public Choice*, volume 29(2). Center for the Study of Public Choice, Virginia Polytechnic Institute, Blacksburg, Virginia, Special Supplement to Spring 1977. (Special issue on the Groves-Clarke demand-revealing process).

[Tid97] T. Nicolaus Tideman. Voting and the revelation of preferences for public activities. In Dennis C. Mueller, editor, *Perspectives on Public Choice: A Handbook*, chapter 11, pages 226–244. Cambridge UP, New York, 1997.

[TT76] T. Nicolaus Tideman and Gordon Tullock. A new and superior method for making social choices. *Journal of Political Economy*, 84(6):1145–1159, December 1976.

[TZ92] Alan D. Taylor and William S. Zwicker. A characterization of weighted voting. *Proceedings of the AMS*, 115:1089–1094, 1992.

[TZ93] Alan D. Taylor and William S. Zwicker. Weighted voting, multicameral representation, and power. *Games and Economic Behaviour*, 5:170–181, 1993.

[vDSW90] Eric van Damme, Reinhard Selten, and Eyal Winter. Alternating bid bargaining with a smallest money unit. *Games Econom. Behav.*, 2(2):188–201, 1990.

[Vic61] W. Vickrey. Counterspeculation, auctions and competitive sealed tenders. *Journal of Finance*, 16:8–37, 1961.

[vM47] John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behaviour*. Princeton University Press, 2nd edition, 1947.

[Wag84] Carl Wagner. Avoiding Anscombe's paradox. *Theory and Decision*, 16(3):233–238, 1984.

[Web] W.A. Webb. An algorithm for a stronger fair division problem.

[Woo86] Douglas Woodall. A note on the fair division problem. *Journal of Combinatorial Theory* **A**, 42:300–301, July 1986.

[Zeu30] Frederik Zeuthen. *Problems of monopoly and economic warfare*. G. Routledge & Sons, London, 1930.