Chapter 6

Extensive-Form Games and Backward Induction

So far, players moved simultaneously. Most real interactions unfold over *time*—entry then retaliation, offer then counteroffer—and timing changes everything: threats must be *credible* to matter.

6.1 Game trees

Definition 6.1: Extensive-form game with perfect information

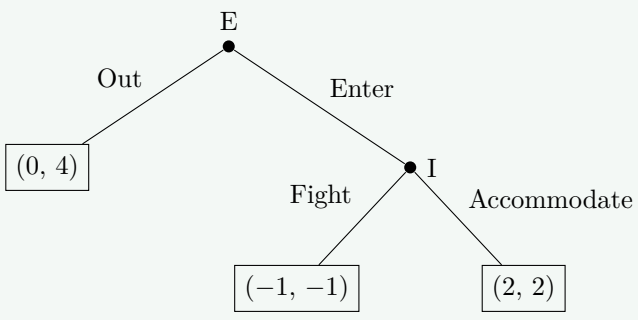
An *extensive-form game with perfect information* consists of:

- a finite *game tree*: a set of nodes with a unique root, where each non-terminal node is assigned to a player (or to *chance*, with fixed move probabilities) and each edge is an available *move*;
- *payoffs* u_i assigned to every terminal node;
- *perfect information*: each player, when moving, observes the entire history of play so far.

A *strategy* for player i is a function assigning a move to *every* node where i moves—a complete contingent plan, even for nodes that the strategy itself makes unreachable.

Example 6.1: The Entry Game

A potential entrant (E) decides whether to enter a market or stay out; if entry occurs, the incumbent (I) decides whether to fight a price war or accommodate.



Payoffs are (Entrant, Incumbent). The incumbent would love to deter entry by threatening to fight. Is the threat credible?

6.2 Backward induction

Definition 6.2: Backward induction

Solve the game from the end: at each final decision node, the mover chooses her best move; replace that node by the resulting payoff; repeat up the tree until the root. The resulting strategy profile is the *backward induction solution*.

In the Entry Game: at I's node, Accommodate (2) beats Fight (−1), so I accommodates. Anticipating this, E compares Out (0) with Enter (2) and *enters*. Backward induction outcome: (Enter, Accommodate).

Key Idea: Credibility

"Fight if you enter" is an *incredible threat*: once entry has occurred, carrying it out hurts the incumbent. Backward induction disciplines predictions by insisting that plans be optimal *at every point where they might be executed*—not just along the anticipated path. This idea, generalized as subgame perfection, is due to Selten.

Note the contrast with the normal form: if we write the Entry Game as a matrix, the profile (Out, Fight) is also a Nash equilibrium—E's "Out" is optimal given I's plan to fight, and I's plan is costless because it is never tested. Backward induction eliminates precisely such equilibria supported by empty threats.

Theorem 6.1: Zermelo's theorem (1913); Kuhn's theorem for perfect information

Every finite extensive-form game with perfect information has a pure-strategy Nash equilibrium, obtainable by backward induction. If no player is indifferent between any two terminal nodes, the backward induction solution is unique.

Corollary 6.2: Determinacy of chess

In chess (a finite two-player win/lose/draw game with perfect information), exactly one holds: White can force a win, or Black can force a win, or both can force at least a draw. (Which one is true remains computationally unknown.)

6.3 Subgames and subgame perfection

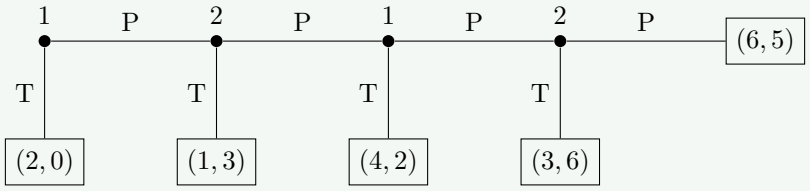
Definition 6.3: Subgame; subgame perfect equilibrium

A *subgame* is the part of the game tree starting at any single node and including all its successors. A strategy profile is a *subgame perfect equilibrium* (SPE) if it induces a Nash equilibrium *in every subgame*—including subgames off the anticipated path of play.

For perfect-information games, SPE coincides with backward induction. The concept's real power appears in games with simultaneous stages or imperfect information (Chapters 9, 11), where "backward induction" alone is not well-defined but subgame perfection still bites.

Example 6.2: The Centipede Game

Players alternate deciding whether to Take a growing pot or Pass. The pot starts at $(2, 0)$; each Pass multiplies total value but shifts advantage:



Backward induction: at the last node, 2 Takes ($6 > 5$); so at the third node, 1 Takes ($4 > 3$); so 2 Takes at the second ($3 > 2$); so 1 Takes *immediately*, yielding $(2, 0)$—even though passing throughout would give $(6, 5)$. In experiments, people pass for several rounds. The game sharpens the question of what common knowledge of rationality really implies, and whether a rational player should "believe" backward induction after observing a deviation from it.

Example 6.3: Stackelberg duopoly (first-mover advantage)

Take the Cournot setting of Example 4.2 but let firm 1 choose q_1 first, observed by firm 2. Backward induction: firm 2 best-responds, $q_2 = \frac{a-c-q_1}{2}$. Firm 1 anticipates this and maximizes

$$u_1 = q_1\Big(a - q_1 - \tfrac{a-c-q_1}{2} - c\Big) = \tfrac{q_1}{2}\,(a - c - q_1),$$

giving $q_1^* = \frac{a-c}{2} > \frac{a-c}{3}$ and $q_2^* = \frac{a-c}{4} < \frac{a-c}{3}$. Firm 1 earns $\frac{(a-c)^2}{8} > \frac{(a-c)^2}{9}$: *commitment is valuable*. Note the paradox of power: firm 2 knows more when moving (it observes q_1) yet earns less—in games, information and options can hurt you if others anticipate you have them.

Key Idea: Commitment

The ability to bind yourself—burn bridges, sign contracts, delegate to a tough agent—can be strategically valuable precisely because it removes your future flexibility and thereby changes others' behavior. Thomas Schelling received the 2005 Nobel Prize largely for developing this insight (deterrence theory in the Cold War).

6.4 The ultimatum game and the limits of backward induction

Example 6.4: Ultimatum game

A proposer offers a split $(10 - x, x)$ of \$10, $x \in \{0, 1, \ldots, 10\}$; the responder accepts (split implemented) or rejects (both get \$0). Backward induction with purely self-interested players: the responder accepts any $x \geq 1$ (and is indifferent at 0), so the proposer offers $x = 1$ (or 0). In experiments across many cultures, proposers typically offer 40–50%, and offers below 20–30% are rejected about half the time. The resolution is not that game theory is wrong, but that *payoffs are not money* (Remark in Chapter 1): responders derive negative utility from being treated unfairly, and proposers know it. With fairness-adjusted payoffs, generous offers *are* the backward induction solution.

6.5 Imperfect information: information sets

To connect the extensive and normal forms fully we need games where a mover does *not* know exactly where in the tree she is.

Definition 6.4: Information set

An *information set* for player i is a collection of decision nodes of i such that (i) i has the same available moves at each node, and (ii) when play reaches the set, i knows only that she is at *some* node of the set, not which one. A game has *perfect information* iff every information set is a singleton.

A simultaneous-move game is the special case where player 2's single information set contains all nodes following player 1's moves. Conversely, every extensive-form game has a normal-form representation (list each player's full contingent strategies and the resulting payoffs). Mixed strategies over full plans are called *mixed*; move-by-move randomizations are *behavior strategies*.

Theorem 6.3: Kuhn's theorem (1953)

In every extensive-form game with *perfect recall* (no player forgets what she once knew or did), mixed and behavior strategies are equivalent: each induces the same distributions over outcomes as some strategy of the other kind.

Subgame perfection extends verbatim: a subgame starts at any *singleton* information set whose successors don't cut across other information sets. When information sets prevent the tree from decomposing, subgame perfection loses power, and we will need sequential rationality with explicit beliefs—perfect Bayesian equilibrium (Chapter 11) and sequential equilibrium (Chapter 13).

Exercises

Exercise 6.1: Backward induction practice

Player 1 chooses L or R; after L, player 2 chooses ℓ (payoffs $(3,1)$) or r ($(1,2)$); after R, player 2 chooses ℓ' ($(2,1)$) or r' ($(0,0)$). Draw the tree, solve by backward induction, then write the 2×4 normal form (player 2 has four strategies) and find all pure Nash equilibria. Which are subgame perfect?

Exercise 6.2: Burning money / commitment

In the Battle of the Sexes (Example 3.3), suppose before play the row player can publicly burn 1 unit of her own utility (or not), after which the game is played with all past moves observed. Show by backward induction on the normal form (using iterated weak dominance) that the row player achieves her favorite equilibrium *without actually burning anything*. Comment on how commitment opportunities can matter even when unused.

Exercise 6.3: Sequential Battle of the Sexes

Let the row player move first in the Battle of the Sexes, observed by column. Solve by backward induction. Who benefits from moving first? Compare with Matching Pennies,

where moving first is fatal. What feature of the payoffs determines whether there is a first-mover or second-mover advantage?

Exercise 6.4: Pirates

Five perfectly rational pirates (seniority 1–5) divide 100 gold coins. The most senior proposes a division; all pirates (including the proposer) vote; a strict majority (or exactly half including the proposer's vote, if you prefer that variant—state your convention) passes the proposal; otherwise the proposer is thrown overboard and the next most senior proposes. Pirates value, in order: survival, gold, seeing others thrown overboard. Find the backward induction outcome with the strict-majority convention.