CS 405/6001: Game Theory and Algorithmic Mechanism Design **Jul–Nov 2026**

Lecture 8: Extensive Form Games and Subgame Perfection

Lecturer: Swaprava Nath *Scribe(s): Claude Scribe*

Disclaimer: *These notes aggregate content from several texts and have not been subjected to the usual scrutiny deserved by formal publications. If you find errors, please bring to the notice of the Instructor.*

8.1 Recap and Motivation

So far, every game we studied was in **normal form**: players move simultaneously (or as if they do), and a strategy is a single choice made once. Many real interactions are not like this. Chess, negotiations, and multi-round auctions all unfold in stages, where players observe what happened earlier and then act. The order of play, the **history** of the game, matters. Today we build the formal machinery for such **perfect information extensive form games (PIEFG)**, and ask what a sensible notion of equilibrium looks like once time and sequence enter the picture.

8.2 Perfect Information Extensive Form Games

Example 8.1 (The hackathon prize split) *Ananya and Vihaan just won a hackathon together. The prize is 2 credits. Ananya moves first and proposes a split: keep both credits (2–0), split evenly (1–1), or give both to Vihaan (0–2). Vihaan then **Accepts** or **Rejects**. If Vihaan rejects, the organizers withhold the entire prize and both get 0.*

Definition 8.2 (PIEFG) *A perfect information extensive form game is a tuple $\langle N, A, H, X, P, (u_i)_{i\in N}\rangle$, where*

- *N is the set of players;*
- *A is the set of all possible actions, across all players;*
- *H is the set of all sequences of actions (**histories**) such that: the empty history $\emptyset \in H$; every prefix of a history in H, starting at the root, is itself in H; a history $h = (a^{(0)}, \dots, a^{(T-1)})$ is **terminal** if no $a^{(T)} \in A$ extends it within H; and $Z \subseteq H$ denotes the set of all terminal histories;*
- *$X : H \setminus Z \to 2^A$ gives the set of actions available at a non-terminal history;*
- *$P : H \setminus Z \to N$ names the player who acts at a non-terminal history;*
- *$u_i : Z \to \mathbb{R}$ is player i's utility, defined only on terminal histories.*

Working this out on Example 8.1: $N = \{1, 2\}$ (Ananya, Vihaan); $A = \{\text{2-0}, \text{1-1}, \text{0-2}, A, R\}$; $H = \{\emptyset, (\text{2-0}), (\text{1-1}), (\text{0-2}), (\text{2-0}, A), (\text{2-0}, R), (\text{1-1}, A), (\text{1-1}, R), (\text{0-2}, A), (\text{0-2}, R)\}$; Z is the last six of these; $X(\emptyset) = \{\text{2-0}, \text{1-1}, \text{0-2}\}$ and $X(\text{2-0}) = X(\text{1-1}) = X(\text{0-2}) = \{A, R\}$; $P(\emptyset) = 1$ and $P(\text{2-0}) = P(\text{1-1}) = P(\text{0-2}) = 2$; finally $u_1(\text{2-0}, A) = 2$, $u_1(\text{1-1}, A) = 1$, $u_2(\text{1-1}, A) = 1$, $u_2(\text{0-2}, A) = 2$, and utilities are zero at every other terminal history.

8.2.1 Strategies and the NFG Translation

The strategy of player i is a tuple of actions, one at every history where i moves:

$$S_i = \times_{\{h \in H :\, P(h)=i\}} X(h).$$

A strategy is a **complete contingency plan**: it fixes an action even at histories that will never be reached once play actually starts, because it must be well defined no matter what happens earlier. In Example 8.1, $S_1 = \{2\text{-}0, 1\text{-}1, 0\text{-}2\}$ and $S_2 = \{A, R\}^3$, eight strategies for Vihaan since he moves at three distinct histories.

Once S_1, S_2 are known, the PIEFG can be rewritten as an NFG, a 3×8 payoff matrix listing every strategy combination. Two things stand out once we do this. First, a Nash equilibrium such as $(2\text{-}0, RRA)$ is not a sensible prediction: why would Vihaan commit to R at 1-1, a history he would never rationally want to reject if actually reached? A profile like $(2\text{-}0, RRR)$ rests on a **non-credible threat**: it deters Ananya only because Vihaan is assumed to follow through on rejecting a split he would not actually want to reject. Second, the NFG representation is enormously redundant compared to the compact tree, the extensive form is a much more succinct description of the same game.

8.3 Subgame Perfection

Example 8.3 (A feature launch race) *A startup decides to Launch Now (A) or Wait (B). If it launches, a rival can Copy (C) or Ignore (D), giving payoffs* $(3, 8)$ *or* $(8, 3)$ *respectively. If the startup waits, the rival can Launch First (E), giving* $(5, 5)$*, or Hold Back (F); if the rival holds back, the startup then chooses Launch Late (G), giving* $(2, 10)$*, or Give Up (H), giving* $(1, 0)$*.*

The startup's strategies are $\{AG, AH, BG, BH\}$ and the rival's are $\{CE, CF, DE, DF\}$. Reading this off as an NFG, the PSNEs are (AG, CF), (AH, CF), and (BH, CE). Is there a non-credible threat lurking in any of these? A better notion of rational play should require optimality not just at the start of the game, but at *every point* the game could reach.

Definition 8.4 (Subgame) *The subgame of a PIEFG G rooted at history h is the restriction of G to the descendants of h.*

The set of subgames of G is the collection of subgames rooted at every history of G (including G itself, rooted at $\emptyset$).

Definition 8.5 (Subgame perfect Nash equilibrium, SPNE) *A strategy profile $s \in S$ is a subgame perfect Nash equilibrium of a PIEFG G if, for every subgame G' of G, the restriction of s to G' is a PSNE of G'.*

That is, subgame perfection demands a best response at *every* subgame, not only along the path that is actually played. Of the three PSNEs found above for Example 8.3, which survive this stronger test, and which rely on an empty threat at some subgame the equilibrium itself never visits?

8.3.1 Computing SPNE: Backward Induction

SPNE can be computed directly by working from the leaves of the tree back to the root.

Algorithm 1: Backward Induction

Function $BACK\_IND(history\ h)$**:**
    if $h \in Z$ **then**
        return $u(h), \emptyset$
    $best\_util_{P(h)} \longleftarrow -\infty$
    foreach $a \in X(h)$ **do**
        $util\_at\_child_{P(h)}, str\_at\_child_{P(h)} \longleftarrow BACK\_IND((h,a))$
        if $util\_at\_child_{P(h)} > best\_util_{P(h)}$ **then**
            $best\_util_{P(h)} \longleftarrow util\_at\_child_{P(h)}$, $best\_str_{P(h)} \longleftarrow a + str\_at\_child_{P(h)}$
    return $best\_util_{P(h)}, best\_str_{P(h)}$

At each history, the player who moves there picks the action leading to the child with the best continuation value for them, and this optimal continuation is computed recursively, starting from the terminal histories where the utility is simply read off.

8.4 Limitations of SPNE

Advantages. SPNE is guaranteed to exist in every finite PIEFG (this requires its own proof). Since every SPNE is also a PSNE, this immediately gives a broad class of games where a PSNE is guaranteed to exist. The algorithm to compute it, backward induction, is conceptually simple.

Disadvantages and criticisms. Backward induction must, in principle, parse the entire tree, which can be computationally prohibitive: chess has roughly 10^{150} vertices, far beyond exhaustive traversal. There is also a behavioural objection: real players may not have the cognitive capacity to actually carry out full backward induction, even when it is well defined.

Example 8.6 (The centipede game) *Two players alternate moves over a shared, growing pot. At each turn the player to move can Defect (D), ending the game immediately and splitting the current pot in their favor, or continue by playing A, letting the pot grow further. With five decision points, the terminal payoffs are:* $(1,0)$ *if Player 1 defects first,* $(0,2)$ *if Player 2 defects next,* $(3,1)$*,* $(2,4)$*,* $(4,3)$ *at the following three defection points, and* $(3,5)$ *if both players choose A throughout.*

By backward induction, Player 1 should defect immediately, at the very first move, since every later stage unravels the same way once analyzed from the end. What is the unique SPNE prediction here, and what is uncomfortable about it as a description of how people actually play?

This game has been run experimentally on many populations, random participants, university students, and even grandmasters. Most subjects, apart from grandmasters, continue for several rounds rather than defecting immediately. Proposed explanations include altruism, limited computational capacity, and heterogeneous incentives among the subjects. There is also a structural criticism of SPNE itself: its defining principle asks "what is optimal if this history is reached," yet an equilibrium reasoned about from the top can simultaneously claim that history will *never* be reached, an internal tension between off-path optimality and on-path prediction. SPNE remains a powerful and often accurate tool, but the centipede game is a standing reminder of its limits; extending the analysis using **players' beliefs** is one way forward, which is a preview of what comes with imperfect information games.

8.5 Summary and Preview

We introduced PIEFG as a tuple $\langle N, A, H, X, P, (u_i)_{i \in N} \rangle$, defined a strategy in an EFG as a complete contingency plan, and showed that translating a PIEFG into an NFG can validate PSNE built on non-credible threats. Subgame perfection fixes this by requiring optimality at every subgame, computed via backward induction, and is guaranteed to exist in every finite PIEFG. The centipede game showed that this guarantee comes with a behavioural cost: the SPNE prediction can diverge sharply from how people actually play. Next time: extensive form games with *imperfect* information, where players may not observe everything that came before their move.

References

[RSV92] R. Rosenthal, "Games of perfect information, predatory pricing and the chain-store paradox," *Journal of Economic Theory*, 25(1), 1981, pp. 92–100.

[MSZ13] M. Maschler, E. Solan and S. Zamir, *Game Theory*, Chapter 3, Cambridge University Press, 2013.