

ECONOMICS & COMPUTATION

Spring 2026 | Lecture 20

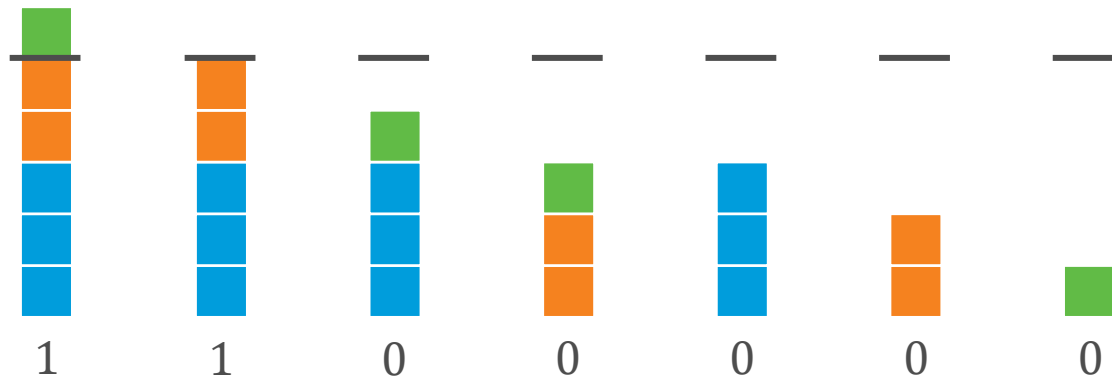
Feature Attribution

Ariel Procaccia | Harvard University

COOPERATIVE GAMES

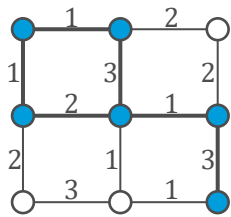
- A **cooperative game** is a pair (N, v) , where:
 - $N = \{1, \dots, n\}$ is the set of players
 - $v: 2^N \rightarrow \mathbb{R}^+$ is the **value function**, which assigns a value to each **coalition** $S \subseteq N$
 - Assume that $v(\emptyset) = 0$
- The central questions in cooperative game theory are:
 - What is the “best” coalition structure?
 - How should payoffs be divided among the players?

EXAMPLE: VOTING GAME

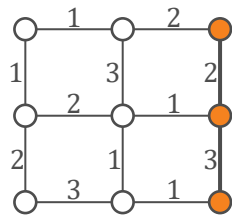


Each player i has a weight $w_i \in \mathbb{N}$ and there is a threshold $q > \frac{1}{2} \sum_i w_i$. For a coalition S , $v(S) = 1$ if $\sum_{i \in S} w_i \geq q$, otherwise $v(S) = 0$.

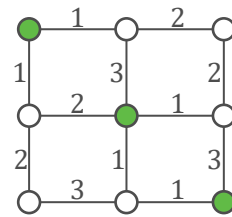
EXAMPLE: INDUCED SUBGRAPH GAME



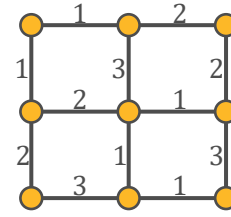
$$v(S) = 11$$



$$v(S) = 5$$



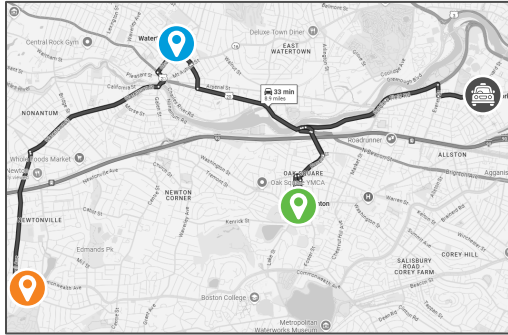
$$v(S) = 0$$



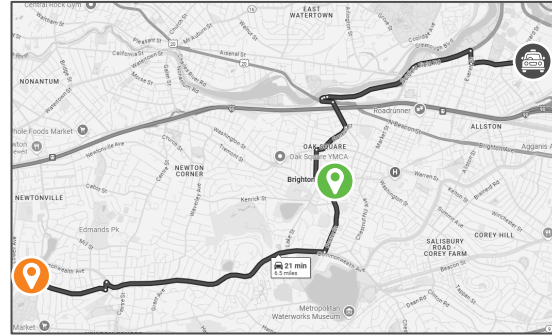
$$v(S) = 22$$

Players are nodes in an undirected, weighted graph with non-negative weights. The value of a coalition is the total weight of the edges in its induced subgraph.

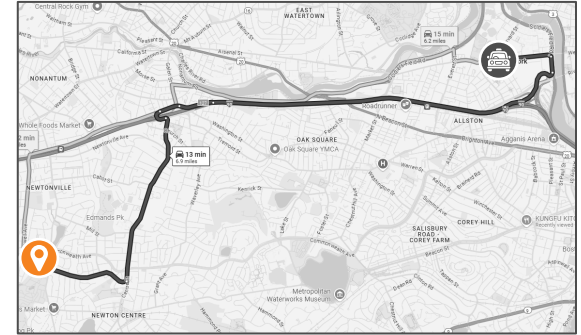
EXAMPLE: TAXI FARE GAME



33 minutes



21 minutes



13 minutes

Assume for simplicity that a taxi costs \$1 per minute. There is a common source x and a destination y_i for each player i . The value of a coalition S is $\max\{0, \sum_{i \in S} c(x, y_i) - c(x, S)\}$, where $c(x, S)$ is the shortest travel time from s to $\cup_{i \in S}\{y_i\}$.

SUPERADDITIVE GAMES

- A cooperative game is **superadditive** if for every pair of disjoint coalitions S, T ,
$$v(S \cup T) \geq v(S) + v(T)$$
- If the game is superadditive, it is rational for the **grand coalition** to form

Poll 1

Which game is **not** superadditive?

- Voting
- Taxi fare
- Induced subgraph
- All are superadditive



SUPERMODULAR GAMES

- A cooperative game is **supermodular** if for all $S \subseteq T \subseteq N$ and $i \in N \setminus T$,
$$v(S \cup \{i\}) - v(S) \leq v(T \cup \{i\}) - v(T)$$

Poll 2

Which game is supermodular?

- Voting
- Induced subgraph
- Both
- Neither one



PAYOFF DIVISIONS

- Given a cooperative game (N, v) , a **payoff division** is a vector $\mathbf{p} \in \mathbb{R}^n$, where p_i is the **payoff** of player i , such that $\sum_{i \in N} p_i = v(N)$
- This assumes that the grand coalition has formed
- We will discuss concepts that formalize the idea that a payoff division is “reasonable” or “stable”

THE SHAPLEY VALUE

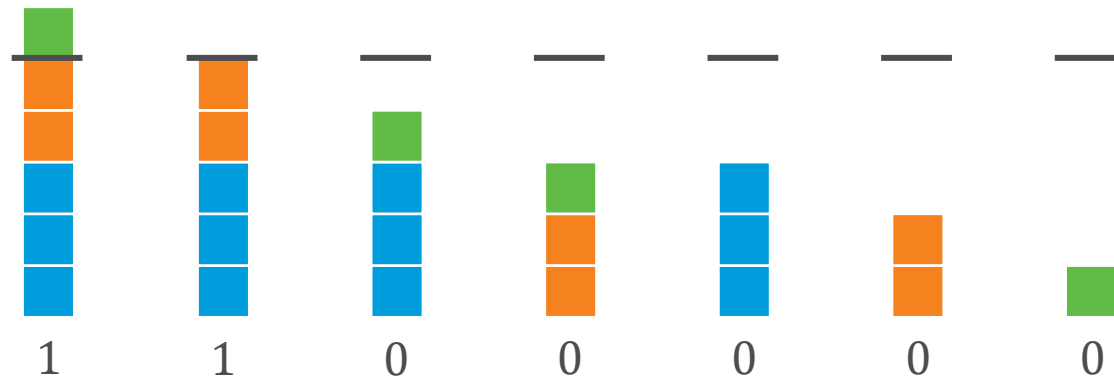
- Given a permutation π over N , let S_π^i be the coalition that consists of the prefix of π up to (and excluding) i
- The **Shapley value** of player i is

$$\sigma_i(N, v) = \frac{1}{n!} \sum_{\pi} [v(S_\pi^i \cup \{i\}) - v(S_\pi^i)]$$

- The vector of Shapley values is a valid payoff division, because

$$\begin{aligned} \sum_{i \in N} \sigma_i(N, v) &= \frac{1}{n!} \sum_{\pi} \sum_{i \in N} [v(S_\pi^i \cup \{i\}) - v(S_\pi^i)] \\ &= \frac{1}{n!} \sum_{\pi} v(N) = v(N) \end{aligned}$$

SHAPLEY VALUE: EXAMPLE



Poll 3

What is the Shapley value of orange?

- 0
- 1/6
- 2/6
- 3/6



AXIOMATIZATION

- When is a payoff division **rule** $\phi(N, v)$ “reasonable”? We take an axiomatic approach
- **Symmetry:** If $i, j \in N$ are such that for all $S \subseteq N \setminus \{i, j\}$, $v(S \cup \{i\}) = v(S \cup \{j\})$, then $\phi_i(N, v) = \phi_j(N, v)$
- **Null player:** If $i \in N$ is such that for all $S \subseteq N \setminus \{i\}$, $v(S \cup \{i\}) = v(S)$, then $\phi_i(N, v) = 0$

AXIOMATIZATION

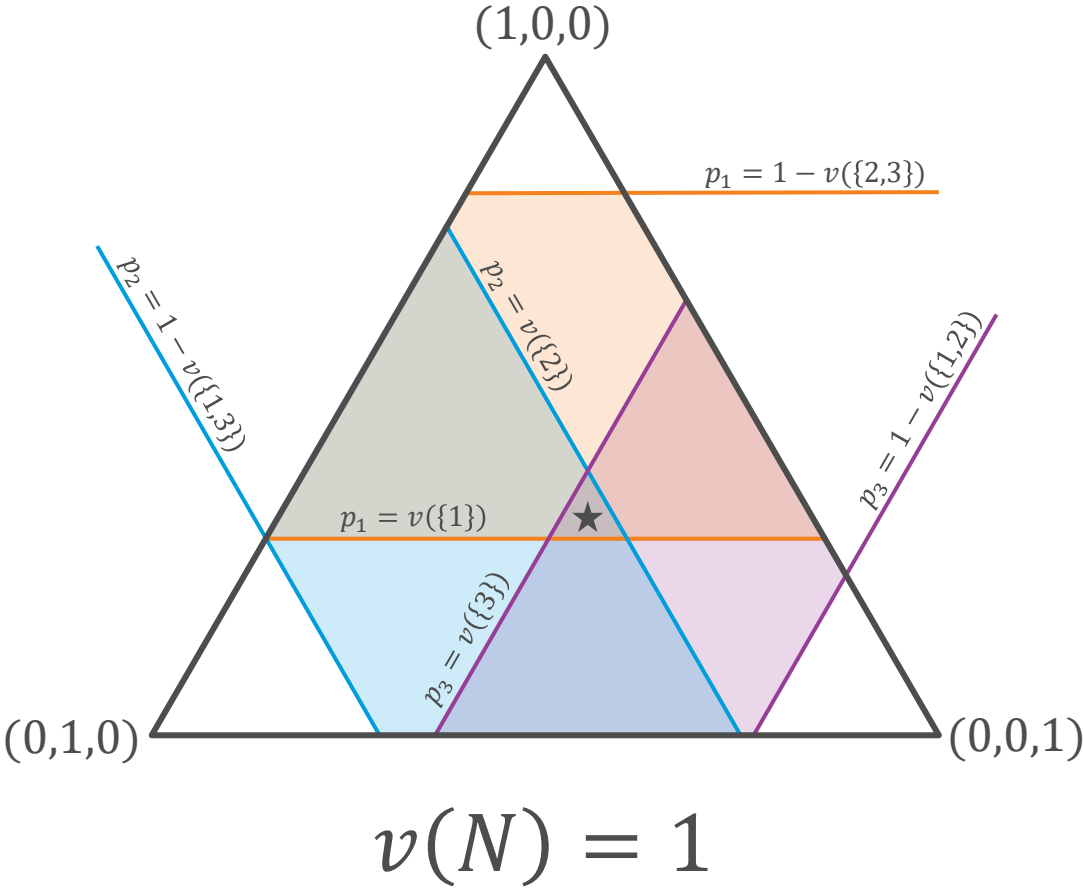
- **Additivity:** For two value functions $v_1, v_2: 2^N \rightarrow \mathbb{R}^+$, it holds that
$$\phi_i(N, v_1 + v_2) = \phi_i(N, v_1) + \phi_i(N, v_2)$$
 for all $i \in N$, where the game $(N, v_1 + v_2)$ is defined by $(v_1 + v_2)(S) = v_1(S) + v_2(S)$
- **Theorem (informal):** The Shapley value is the unique payoff division rule satisfying symmetry, null player and additivity!

THE CORE

- We would like the payoff division to be stable, in the sense that coalitions don't have an incentive to break off from the grand coalition and go it alone
- **The core** of a game (N, v) is the set of payoff divisions \mathbf{p} such that for all $S \subseteq N$,

$$\sum_{i \in S} p_i \geq v(S)$$

THE CORE: ILLUSTRATION



THE CORE

- The core is a compelling concept — but it might be empty!
- Consider a weighted voting game with three players, $w_i = 1$ for all i and $q = 2$
- If w.l.o.g. $p_1 > 0$, then
$$v(\{2,3\}) = 1 > 1 - p_1 = p_2 + p_3$$
- **Theorem:** In any supermodular game, the core is nonempty and contains the Shapley value

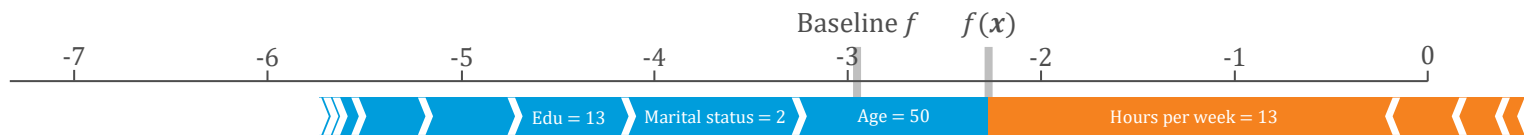
THE LEAST CORE

- The least core is a feasible relaxation of the core
- It's the set of payoff divisions \mathbf{p} arising from the (large) linear program:

$$\begin{aligned} \min \quad & \epsilon \\ \text{s.t.} \quad & \forall S \subseteq N, \sum_{i \in S} p_i \geq v(S) - \epsilon \\ & \sum_{i \in N} p_i = v(N) \\ & \forall i \in N, p_i \geq 0 \\ & \epsilon \geq 0 \end{aligned}$$

FEATURE ATTRIBUTION

Given a machine learning model $f: \mathbb{R}^d \rightarrow \mathbb{R}$ and a point $\mathbf{x} \in \mathbb{R}^d$, what is the influence of a feature over $f(\mathbf{x})$?



Shapley value explanations for a model trained to predict whether individuals have income greater than \$50k based on census data [Chen et al., 2022]

FROM ML TO COOPERATIVE GAMES

- Given a model $f: \mathbb{R}^d \rightarrow \mathbb{R}$ and a point \mathbf{x} , define a cooperative game:
 - The players are the d features
 - $v(S) = f(\mathbf{x}_S)$, where \mathbf{x}_S is \mathbf{x} with the features in $N \setminus S$ “removed”
- Now we can compute the Shapley values of the features (modulo computational challenges — see Assignment 5)
- Different notions of feature removal induce different games

REMOVING FEATURES

