ECONOMICS & COMPUTATION

Harvard CS 1360, Spring 2025
Instructor: Ariel Procaccia
Scribe: Rhea Acharya

Feature Attribution

—

Lecture 20

1 Introduction to Cooperative Games

Definition 1. A cooperative game is a pair (N, v), where:

- $N = 1, \ldots, n$ is the set of players
- $v : 2^N \to \mathbb{R}^+$ is the value function, which assigns a value to each coalition $S \subseteq N$
- We assume that $v(\emptyset) = 0$

(For most games, we also assume for all $S \subseteq T$, $v(S) \leq v(T)$, but this does not always hold.)

We can interpret a value function as: if the players work together as the group defined by the coalition and collaborate, what is the value that they can generate?

Now, we want to explore the following questions in cooperative game theory:

1. What is the "best" coalition structure? For example, one such coalition to pick from is the grand coalition where we include all the players
2. How should payoffs be divided among the players?

Cooperative games model situations where players can form binding agreements and cooperate to generate value. Unlike non-cooperative games we've studied previously, the focus here is on coalition formation and fair allocation of the resulting value, rather than strategic behavior.

2 Examples of Cooperative Games

2.1 Voting Game

In a 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 $v(S) = 0$.

2.1.1 Parliament Bill Example

Consider a scenario where we are analyzing whether a bill will pass through parliament. Suppose there are N total members in parliament. In this context, for a bill to pass, a coalition of parties must collectively represent at least a certain percentage of the total parliament—denoted by the threshold q.

Each party i is assigned a weight w_i, which corresponds to the number of members it has in parliament. Thus, a coalition's total power is the sum of its members' weights, and it can pass a bill if this sum meets or exceeds q.

For example, let's suppose we have three parties:

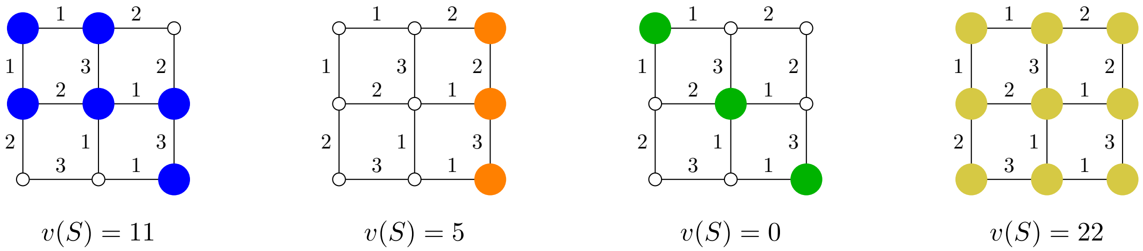
- **Blue Party:** $\frac{1}{2}$ of parliament
- **Orange Party:** $\frac{1}{3}$ of parliament
- **Green Party:** $\frac{1}{6}$ of parliament

Suppose the threshold for passing a bill is set to $q = \frac{5}{6}$
Then, we observe:

- The grand coalition (blue + orange + green) succeeds.
- The blue + orange coalition also meets the threshold.
- All other coalitions fail to reach the threshold and thus have value 0.

2.1.2 Induced Subgame Graph Example

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



$v(S) = 11$ $v(S) = 5$ $v(S) = 0$ $v(S) = 22$

Figure 1: Four different examples for induced subgame and their corresponding $v(S)$.

In these graphs, we make the following observations:

- In the blue graph on the left, the coalition consists of the nodes in blue. The edges that we include then are the thicker edges, formed between two blue nodes. The value of this coalition is $1 + 1 + 2 + 3 + 1 + 3 = 11$.
- On the orange graph, we have a value of $2 + 3 = 5$.
- For the green graph, we have a value of 0, since there are no edges between any of the green nodes.
- Finally, for the yellow graph, this is the grand coalition, so we include all the edges of total weight 22.

2.1.3 Taxi Fare Game

Let's say we have a group of students who are all leaving the SEC (in yellow) and want to go to various destinations. They would ideally like to share a taxi with at least some of the others. Assume for simplicity that a taxi costs \$1 per minute. There is a common source x and a desitnation y_i for each palyer 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 $\cup_{i \in S}\{y_i\}$. More intuitively, the first component $\sum_{i \in S} c(x, y_i)$ refers to the total time if they all took separate taxis which we subtract $c(x, S)$ from. The reason that we take the max compared to zero is because sometimes, it does not make sense for two people to share a taxi. For example, if two people are going in completely opposite directions, it would take more time for one taxi to go all the way in one direction and then go all the way back plus some in the other direction to drop the second person off, compared to both people taking their own individual taxis. In this case, we won't force them to share and will instead just have them take their own taxis.

3 Superadditive Games

Definition 2. 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, since more collaboration never hurts in superadditive games. Thus, in superadditive games, we focus simply on how to divide the payoffs and not on which coalition structures to form.

Which of the games from above are superadditive?

- **The voting game is superadditive.** If the union of 2 coalitions exceeds the threshold and has value 1, then either 0 or at most 1 of the individual coalitions can be greater than the threshold. We can't have both exceed the threshold, since the threshold is at least more than half of the weight, so we can't have two disjoint thresholds that are each more than half of the weight. In the other case, where the union has value 0, then both of the individual coalition must also have value 0.

- The induced subgraph is also superadditive. With the induced subgraph, when we combine two coalitions, we are getting all of the edges that are there before and we might also get more vertices if new edges are formed between included nodes.
- **The taxi fare game is not superadditive.** Let's say we have two people going to the east, but we have one person going to the west. The two people going to the east might have a positive value, since they are saving time overall by commuting together. However, if the person going to the west is going super far west, then if we put them all into one coalition, then this extra effort might cancel out the efficiency gains made by having the two people going to the east commuting together.

3.1 Supermodular Games

Definition 3. A cooperative game is **supermodular** if for all $S \subseteq T \subseteq N$ and $i \in N \setminus T, v(S \cup \{i\}) - v(S) \leq v(T \cup \{i\}) - v(T)$.

We can use similar arguments from above to show that voting is not supermodular but induced subgraph is.

4 Payoff Divisions

Given a cooperative game (N, v), a payoff division is a vector $p \in (\mathbb{R}^+)^n$, where p_i is the payoff of player i, such that $\sum_{i \in N} p_i = v(N)$. This assumes that the grand coalitions has formed. How can we construct formal definitions for what a "reasonable" or "stable" payoff division looks like?

4.1 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. What matters here is the coalition that is formed by taking the players in the permutation before i, not the order itself.

Definition 4. The **Shapley value** of player i is

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

Here, we are summing over all the orders of the player to understand how much value player i contributes to the coalition of S_π^i, which is the marginal contribution of player i to the prefixed coalition. We then average over all possible permutations to get an expected marginal contribution from i.

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

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

We are able to make this jump because the RHS of the top equation forms a telescopic sum, where everything cancels out, except for a subtraction between the value of grand coalition and the value of the empty set. Then, the value of the grand coalitions for each permutation is the same, as order does not matter, so the average of these identical values is just the avlue of the grand coalition itself.

4.2 Axiomatization

When is a payoff division rule $\phi(N, v)$ "reasonable"? We can define the following axioms:

- **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 $\phi_i(N, v) = \phi_j(N, v)$. If the two players induce identical values then we should get identical payoffs.

- **Null player:** If $i \in N$ is such that for all $S \subseteq N \setminus \{i\}, v(S \cup \{i\}) = v(S)$, then $\phi_i(N, v) = 0$. If you never contribute anything, then you should get zero payoffs. We can determine this by seeing if the value of the coalition every changes but adding the specified player.
- **Additivity:** For two value functions $v_1, v_2 : 2^N \to \mathbb{R}^+$, 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 $(v_1 + v_2)(S) = v_1(S) + v_2(S)$ If you take two value functions and you look at the game that is induced by summing up the value functions, then the Shapley values shold be the sum of the Shapley values in the original games. This property is not as necessary to satisfy as the first two, but is still a relatively reasonable expectation.

Theorem 1. *The Shapley value is the unique payoff division rule that satisfies symmetry, null player, and additivity.*

This result should be immediate for the null player, and can be derived more generally as well.

4.2.1 The Core

In cooperative game theory, we aim to find payoff divisions that are stable, meaning no subset of players has an incentive to break away from the grand coalition.

Definition 5. The **core** of a game (N, v) is the set of payoff vectors p such that for all $S \subseteq N$,

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

This ensures that every coalition receives at least what it could obtain on its own. Each player has one constraint that acts as a lower bound and one that acts as an upper bound. The intersection of these six constraints—corresponding to the six nontrivial coalitions—forms one region which satisfies all coalition constraints and lies in the core.

Example 1. Consider a weighted voting game with three players, where $w_i = 1$ for all i, and the threshold $q = 2$. Then,

$$v(\{i, j\}) = 1 \quad \text{for any pair } \{i, j\}$$

Suppose, without loss of generality, that $p_1 > 0$. Then the combined payoff to players 2 and 3 is $p_2 + p_3 = 1 - p_1 < 1 = v(\{2, 3\})$. So players 2 and 3 would have an incentive to form their own coalition. The same reasoning applies for any player receiving a positive payoff. Therefore, the core is empty, and there are no payoff divisions that satisfy the core constraints for this game.

However, we do have a reassuring theorem for supermodular games:

Theorem 2. *In any supermodular (also called convex) game, the core is nonempty and contains the Shapley value.*

4.2.2 The Least Core

To help avoid the empty cores, we can define the least core, a feasible relaxation of the core.

The least core is the set of payoff divisions p arising from the (large) linear program:

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

5 Feature Attribution

How does this all relate to AI? There is a hot field in AI called *interpretability*, where we want to understand how black-box models make predictions or computations.

The Shapley value was a concept originally known only within cooperative game theory, until it gained widespread popularity in the AI community following a 2017 NeurIPS paper. It has since become one of the most well-known and widely used methods for interpreting complex models.

Here is an example of how the Shapley value is useful for interpretability:

Given a machine learning model $f : \mathbb{R}^d \to \mathbb{R}$ and a point $\mathbf{x} \in \mathbb{R}^d$, what is the influence of each feature in $\mathbf{x}$ on the model's prediction $f(\mathbf{x})$?

For example, we might have a model trained to predict whether individuals have income greater than \$50k based on U.S. census data, as demonstrated by Chen et al. (2022). The input features might include age, education level, marital status, and hours worked per week. We want to quantify the contribution of each feature to the model's output.

In this setup, the model output $f(\mathbf{x})$ is a real number that we interpret in the exponent: the actual predicted probability is $e^{f(\mathbf{x})}$. This means a lower value of $f(\mathbf{x})$ corresponds to a lower predicted probability. For example, if $f(\mathbf{x}) = -2$, the predicted probability is approximately $\frac{1}{e^2} \approx 0.135$.

The Shapley value allows us to assign a numerical value to each feature's contribution, capturing how much each one ”pushes” the model's prediction up or down. These contributions are computed by considering all possible coalitions (subsets of features) and averaging marginal contributions, similar to cooperative game theory.

We can plot Shapley values along a line, while also specifying a baseline represents the model's prediction if no features are known (i.e., the expected output over all data points). Each feature then either pushes the prediction to the right (increasing the model's output) or to the left (decreasing it).

Items that fall to the left of the baseline represent features that *increase* the prediction $f(x)$, or push the prediction toward the right.

Items that fall to the right of the baseline represent features that *decrease* the prediction, or push the prediction toward the left.

The further to the right or left a feature pushes, the more influential it is in the model's final decision. Thus, this kind of visualization gives a clear and intuitive sense of which features matter most and in which direction they are influencing the output. f here could have been arbitrarily complicated, but this Shapley value representation gives us insight into the impacts of the features. The size of the bars is a function of the particular x.

5.1 From ML to Cooperative Games

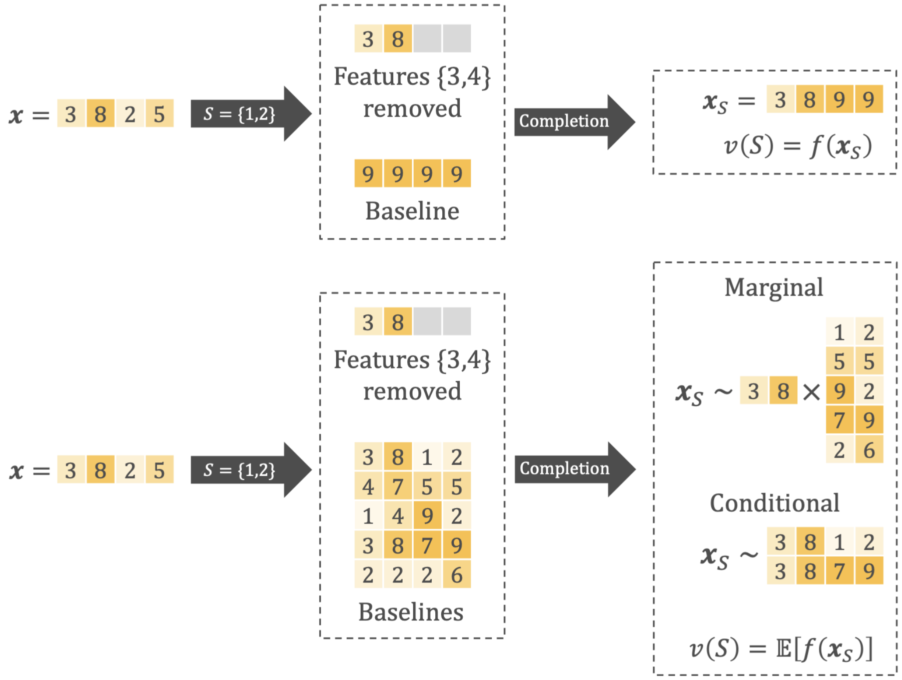
Definition 6. Given a model $f : \mathbb{R}^d \to \mathbb{R}$ and a point x, define a cooperative game:

- The players are the d features
- $v(S) = f(x_s)$, where x_s is x with the features in $N \setminus S$ ”removed”

Now, we can compute the Shapley value of the features.

5.1.1 Removing Features

When computing Shapley values, we need to evaluate the model on subsets of features. But if some features are removed, we still need to complete the input so that the model can make a prediction. There are different ways to do this, and each leads to a slightly different version of the game we're playing.



Different strategies for filling in missing features when evaluating $v(S)$.

In the top half of the figure, we remove features 3 and 4 and fill them in with a fixed baseline. here, the values $[9, 9]$, which represent noise or neutral defaults. This gives a completed input like $[3, 8, 9, 9]$, which the model can evaluate directly. This approach is simple, but the result may be artificial depending on the choice of baseline.

In the bottom half, we instead use a dataset of real examples (the new baselines) to fill in the missing features. There are two common approaches:

- **Marginal:** Sample values for the missing features from the dataset, ignoring the values of the preserved ones. This is easy to implement but may break dependencies in the data.
- **Conditional:** Sample only from rows where the preserved features match the ones in x, and use their corresponding values to complete the input. This better respects correlations in the data but may have limited matches.

Each of these methods defines a slightly different cooperative game and produces different Shapley values. There's ongoing discussion about which is best, but many approaches are reasonable and widely used in practice.